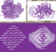
ACS SYMPOSIUM SERIES 1052

Computational Modeling in Lignocellulosic Biofuel Production



EDITED BY

Mark R. Nimlos and Michael F. Crowley

Computational Modeling in Lignocellulosic Biofuel Production

ACS SYMPOSIUM SERIES **1052**

Computational Modeling in Lignocellulosic Biofuel Production

Mark R. Nimlos, Editor
National Renewable Energy Laboratory

Michael F. Crowley, Editor
National Renewable Energy Laboratory

Sponsored by the
ACS Division of Cellulose and Renewable Material

American Chemical Society, Washington, DC

Library of Congress Cataloging-in-Publication Data

Computational modeling in lignocellulosic biofuel production / [edited by] Mark R. Nimlos, Michael F. Crowley ; sponsored by the ACS Division of Cellulose and Renewable Material.
p. cm. -- (ACS symposium series ; 1052)
Includes bibliographical references and index.
ISBN 978-0-8412-2571-8 (alk. paper)
1. Ethanol as fuel. 2. Biomass energy--Mathematical models. 3. Lignocellulose--Biotechnology. I. Nimlos, Mark R. II. Crowley, Michael F. III. American Chemical Society. Cellulose and Renewable Materials Division.
TP339.C66 2010
662'.6692--dc22
2010039839

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48n1984.

Distributed by Oxford University Press

The citation of trade names and/or names of manufacturers in this publication is not to be construed as an endorsement or as approval by ACS of the commercial products or services referenced herein; nor should the mere reference herein to any drawing, specification, chemical process, or other data be regarded as a license or as a conveyance of any right or permission to the holder, reader, or any other person or corporation, to manufacture, reproduce, use, or sell any patented invention or copyrighted work that may in any way be related thereto. Registered names, trademarks, etc., used in this publication, even without specific indication thereof, are not to be considered unprotected by law.

PRINTED IN THE UNITED STATES OF AMERICA

Foreword

The ACS Symposium Series was first published in 1974 to provide a mechanism for publishing symposia quickly in book form. The purpose of the series is to publish timely, comprehensive books developed from the ACS sponsored symposia based on current scientific research. Occasionally, books are developed from symposia sponsored by other organizations when the topic is of keen interest to the chemistry audience.

Before agreeing to publish a book, the proposed table of contents is reviewed for appropriate and comprehensive coverage and for interest to the audience. Some papers may be excluded to better focus the book; others may be added to provide comprehensiveness. When appropriate, overview or introductory chapters are added. Drafts of chapters are peer-reviewed prior to final acceptance or rejection, and manuscripts are prepared in camera-ready format.

As a rule, only original research papers and original review papers are included in the volumes. Verbatim reproductions of previous published papers are not accepted.

ACS Books Department

Chapter 1

Ab Initio Molecular Dynamics Investigation of Xylan Hydrolysis

Haitao Dong and Xianghong Qian[*]

**Department of Mechanical Engineering, Colorado State University,
Fort Collins, Colorado 80523, USA
[*]xhqian@goku.engr.colostate.edu**

Ab initio molecular dynamics (CPMD) coupled with metadynamics (MTD) simulations were used to investigate the free-energy surfaces of acid-catalyzed hydrolysis reactions of xylobiose disaccharide in the gas phase and in aqueous solution. Water and water structures were found to play a critical role in the hydrolysis reaction barrier. Proton partial desolvation associated with its migration to the ether linkage site, the protonation of the ether bond, and the subsequent breaking of the C-O bond were found to be the rate-limiting steps. The significant contribution to the reaction barrier caused by partial proton desolvation and migration could partially explain the biphasic phenomenon in xylan hydrolysis and highlight the importance of mass transport during biomass pretreatment.

Introduction

Cellulosic biomass represents an abundant renewable resource for producing bio-based products and biofuels. Cellulosic biomass is mainly composed of hemicelluloses (~15%–32%), cellulose (~30%–50%) and lignin (~15%–25%). Hemicelluloses (mostly xylan) are natural polymers of β-D-xylose and other minor sugars, whereas cellulose is made of β-D-glucose. Lignin is a polymer composed of non-fermentable phenyl-propene monomer units. Both xylose and glucose sugar monomers are connected via the β-1,4 ether linkage to form xylan and cellulose, respectively. The typical biochemical platform for converting biomass to biofuels such as ethanol includes a thermochemical pretreatment step followed by enzymatic hydrolysis and fermentation.

Before enzymatic hydrolysis and fermentation, cellulosic biomass must be pretreated to hydrolyze hemicelluloses, increase the material porosity, and render the biomass substrates more susceptible to enzymatic digestion (*1*). Thermochemical pretreatment opens up the biomass structure and has long been recognized as a critical step in producing cellulose with acceptable enzymatic digestibility (*2*). Not only is pretreatment the most costly step, it also has a significant impact on the cost of both prior (e.g., size reduction) and subsequent (enzymatic hydrolysis and fermentation) operations (*3*, *4*). Various technologies, including dilute acid (*5*, *6*), alkaline (*7*, *8*), hot water or steam (*9*, *10*), ammonia fiber explosion (AFEX) (*11*, *12*), and lime (*13*, *14*) pretreatment methods have been developed to accomplish this goal (*15*). Dilute sulfuric acid (~0.5%–3.0% sulfuric acid by weight) is one of the most common and cost-effective agents used in pretreatment to hydrolyze hemicelluloses and relocate lignin (*7*, *16–25*). Typically, dilute-acid pretreatment is carried out at an elevated temperature of 430-500K.

During dilute-sulfuric-acid pretreatment, hemicelluloses (mostly xylan) are hydrolyzed to monomer sugars, the majority of which are β-D-xylose. During this process, a small amount of β-D-glucoses are also released from hemicellulose xyloglucan and possibly from cellulose. Depending on the severity (temperature, acidity, and processing time) of the acid pretreatment, some xylose and glucose molecules undergo an undesirable degradation process that lowers the biomass conversion efficiency. 2-Furaldehyde (Furfural) (*26–29*) and 5-(hydroxymethyl)-2-furalde (HMF) (*27*, *28*, *30–33*) are major degradation products from xylose and glucose, respectively, in an acidic environment. Besides these two major products, there are several other degradation products (*48*, *50*, *53–57*). The xylose and glucose molecules could also react with each other in an acidic environment to form various disaccharides or even oligomers, particularly at higher sugar concentrations. Sugar yields decrease as temperature and acidity increase because of acid-catalyzed sugar degradation. However, at lower temperature and acidity, the processing time is much longer due to the presence of both fast and slow biphasic xylan de-polymerization reactions (*17*, *19*). So far our understanding of the biphasic phenomenon of xylan hydrolysis in the complex biomass matrix is very limited. However, laboratory evidence (*34*) supports the theory that xylan hydrolysis without the presence of other biomass components (mainly cellulose and lignin) is fast and does not exhibit biphasic kinetics. It is postulated that mass transport plays an important role in xylan hydrolysis. Here, we attempt to understand the reaction-free energy and barrier for xylan hydrolysis and associated crucial rate-limiting step(s). Because xylan hydrolysis and sugar degradation/ condensation reactions are both catalyzed by proton during dilute-acid pretreatment, the knowledge of their relatively reaction-free energies and reaction barriers is tremendously valuable for optimizing pretreatment conditions. In this chapter, we focus on the xylan hydrolysis reaction using β-1,4-linked xylobiose hydrolysis as an example.

The reaction free energy ΔG and the reaction barrier ΔE_a are extremely useful parameters for quantifying a chemical reaction. They determine the thermodynamic equilibrium constant K, the kinetic reaction rate constant k, both of which are needed to quantify the xylan hydrolysis, sugar degradation and condensation products. For a reversible chemical reaction $A + B \rightarrow C + D$, where A and B are reactants and C and D are products, the relations between the equilibrium constant K and free energy ΔG, reaction rate constant k, and activation barrier ΔE_a are shown in Equations 1 and 2, respectively. Here, R is the gas constant and T is the absolute temperature in Kelvin. A is a prefactor depending on collision frequency.

$$K = \frac{[C][D]}{[A][B]} = -RT \ln \Delta G \qquad (1)$$

$$k = Ae^{-\Delta E_a / RT} \Rightarrow \ln(k) = \frac{-E_a}{R}\frac{1}{T} + \ln(A) \qquad (2).$$

Equation 2 is the Arrhenius equation. It is an empirical relationship. It is generally assumed that prefactor A and ΔE_a are either not dependent or only weakly dependent on temperature. The prefactor A can be determined statistically as well as experimentally by plotting the natural logarithm of measured k with respect to 1/T.

Chemical reactions are complex dynamical processes involving the breaking and forming of chemical bonds and the transfer of electrons. Therefore, only electronic methods based on first-principles quantum calculations are generally able to describe these processes. The common dynamical methods, such as classical molecular dynamics (MD) simulations based on solving Newton's equation of motion, are unable to describe these chemical and electron transfer processes. Due to its dynamic nature, the reacting system changes state dramatically over a relatively short period of time, making static quantum mechanical computational methods inadequate. *Ab initio* MD simulation methods such as CPMD (*35*) are the leading techniques for investigating chemical reactions and processes. CPMD is a predictive technique that requires no empirical parameter and is one of the most accurate available. CPMD unifying molecular dynamics and density functional theory (*36*) have been successfully and extensively applied to investigate water structure, proton transfer processes, and several chemical reactions (*37–51*), many of which have been extensively tested and validated by available experimental data. While many chemical reactions and processes occur on the time scale of femtoseconds (fs) (10^{-15} s) to picoseconds (ps) (10^{-9} s), a significant number occur on nanoseconds (ns) (10^{-9} s) or even much longer time scales. Chemical reactions occur when the system migrates from one local equilibrium minimum to another, overcoming the usually large energy barriers that separate reagents from products (*52*). The probability of such an event occurring spontaneously is inversely related to the exponential of the reaction energy barrier. Depending on the reaction energy barrier, this process could easily exceed 50 ps CPU time, which is the limit our current computing technology can afford.

The typical approach of quantum chemistry to overcome this problem is to determine the local minima and saddle points on the potential energy surface to find the possible equilibrium structures and reaction pathways as used in Gaussian (*53*). These calculations are computationally very demanding, require much insight, and are generally very difficult. During the past few years, a new MTD method was developed by Parrinello and coworkers (*52*, *54*), based on the ideas of extended Lagrangian (*54–57*) and coarse-grained, non-Markovian dynamics (*54*), which allow very efficient exploration of the reactive system's free-energy surface (FES). It is suitable for implementation in *ab initio* MD simulation codes and has been incorporated into CPMD. This MTD method assumes that several collective coordinates that distinguish reactants from products are able to characterize the reaction process. These collective coordinates (e.g., distances between atoms and coordination numbers) must include the relevant modes that cannot be sampled within the typical time scale of the *ab initio* MD simulation (~50 ps). This method is a significant leap forward in simulating chemical reactions and has been successfully applied to several chemical and biological systems (*58–70*). In this work, CPMD-MTD will be used to explore the free-energy surfaces of xylobiose hydrolysis reactions. The reaction pathways, barriers, and rate constants can also be determined.

Our earlier work (*27–29*, *71*) demonstrated the unique capability of CPMD (*35*, *55*) for studying sugar reactions both in the absence and presence of explicit surrounding water molecules. Our calculations show that water and water structure play an important role in sugar reaction pathways. Water molecules can compete with the hydroxyl groups on the sugar ring for a proton. Moreover, water molecules can extract a proton from the carbocation intermediates to terminate the reaction. These results suggest that solvent molecules play a crucial role for both sugar reactions and xylan hydrolysis. Our results show that the size of the water cluster surrounding the sugar molecule has a significant effect on the reaction barrier.

Method

Metadynamics is an extended Lagrangian method designed to accelerate the energy barrier-crossing progress, which has been a major drawback for MD simulations that are limited in a sub-microsecond time scale (*52*, *54*). The basic assumption of this method is that the FES depends on n (n << 3N, with N being the total degrees of freedom of the system) and collective variables (CVs). Here the i^{th} CV is denoted as S_i. For each CV, an auxiliary particle with position s_i is coupled to S_i, and the modified Lagrangian of the system is:

$$L = L_0 + \sum_i m_i \dot{s}_i{}^2 - \sum_i \frac{1}{2} k_i \left(S_i - s_i\right)^2 - V(s) \tag{3}$$

where the first term on the right side is the original Lagrangian. In our study, it is governed by the DFT electronic structure calculations in CPMD. The second term is the kinetic energy of the fictitious particles; the third term is the harmonic coupling potential between S and s; the last term is a bias potential, which is dependent on the dynamics of s (will be considered later); and m and k are fictitious mass and coupling constants for s. If these quantities are carefully chosen so that the motions of s are much slower than the motions of S, the average force placed on *s* by *S* through the harmonic coupling is an estimate of the derivative of free energy *A* with respect to *s*,

$$\frac{\partial A}{\partial s_i} = \left\langle k_i \left(S_i - s_i \right) \right\rangle_S \qquad (4).$$

The angled bracket is the average over all configuration of *S*.

This formulation of the extended Lagrangian does not guarantee a faster barrier-crossing process. A repulsive potential *V*(*s*) needs to be introduced into the FES to drive the system in the desired direction of the reaction. Equation 4 provides an estimate of the FES along coordinates of the chosen CVs. Alternatively, the repulsive potential can conveniently take a Gaussian form

$$V(s) = \sum_i W_i \exp\left(-(s_i - s_i^0)^2 / 2\Delta s_i^{\,2} \right) \qquad (5).$$

Equation 5 is not meant to be an exact counter to the FES but merely a bias to force the system to leave the current locations s^0 (the Gaussian height W_i is usually a few percent of the barrier height). The width is controlled by Δs_i, which is the average amplitude of the fluctuation of s_i. As successive repulsive Gaussian potentials are added to the FES, the reactant well will be flattened; and the system will have a much greater chance to cross the energy barrier with its thermal energy. Equation 5 can be further modified with a second Gaussian form to reduce overlapping between successive additions, but does not lead to a significant improvement (*72*).

When both the reactant and product wells on the FES are filled up by the Gaussian potentials, the system can move freely over the configurations along the reaction pathway. After some fine tuning of the accumulated Gaussian potentials, the modified FES will be completely flattened. Therefore, the summation of all the Gaussians added with respect to the CVs is the reverse of the FES of the underlying chemical or biological process, which determines the free-energy change and the barrier height for the process. An advantage of this method is that a good knowledge of the FES does not have to be known *a priori*, as is required in umbrella sampling. The MTD progress will self-guide the system to explore the FES.

The efficiency and accuracy of MTD simulations depend on an optimal combination of parameters that control the dynamics of CVs, the shape of the Gaussian potentials, and the speed of the energy-well-filling progress. We performed a series of test simulations to determine these parameters by following the guidelines suggested by Ensing *et. al* (*73*). The errors of free-energy estimation using the MTD method are expected to be 1-2 kcal/mol if the parameters are selected properly (*74*).

Results

Protonation of the Ether Linkage in β-1,4-Linked Xylobiose in the Gas Phase and Comparison with β-1,2-Linked Xylose Disaccharides

The reaction free energies for the protonation processes of the ether linkages in β-1,4-linked as well as β-1,2-linked xylose disaccharides were initially investigated in the gas phase. The gas-phase simulations helped us understand the effects of water and water structure on the hydrolysis reactions of these disaccharides in solution. Two different linkages were investigated to study structural and conformational effects on acid-catalyzed hydrolysis.

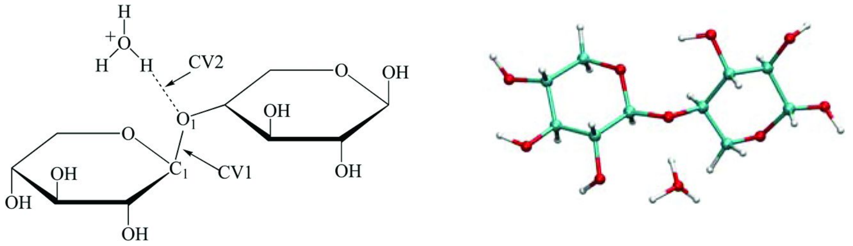
Scheme 1 shows the two CVs used in the gas-phase calculations. The first CV (CV1) is the coordination number (CN) of C1 with respect to O1. The equation of CN is given by:

$$CN = \frac{1-\left(d_{ij}/d_0\right)^p}{1-\left(d_{ij}/d_0\right)^q} \qquad (6)$$

where d_{ij} is the distance between atoms i and j, d_0 is the cutoff distance, and the high powers (p and q) distinguish between the coordinated and non-coordinated states. The second CV (CV2) is the CN difference of O1 on xylobiose and the O atom in H_3O^+ with respect to proton H, respectively. The CV dynamics are controlled by the force constant k and mass m (see Eq. 3). We used k = 8.0 a.u. and m = 600 a.m.u. for CV1 and k = 3.0 a.u. and m = 100 a.m.u. for CV2. The height and width of the Gaussian bias potentials were chosen as H = 0.002 a.u. and W = 0.1 a.u. When the first barrier crossing was observed, the value of H was reduced to 0.001 a.u. and was fixed for the rest of the simulations. The bias potentials were added whenever the CV displacements were larger than 1.5 times the width, but no shorter than 100 MD steps.

All MD calculations were carried out using the CPMD software package (*35*). In these CPMD runs, the Becke, Lee, Yang, and Parr functional was used to describe the chemically active valence electrons (*75*, *76*). The interactions between these electrons and the "frozen-cores" were described by the Goedecker pseudopotentials (*77*). We used an energy cutoff of 70 Ry for the plane-wave basis sets, which was shown to be sufficient from our earlier results (*28*, *71*, *78*, *79*). To effectively separate the electron motions from those of slow-moving nuclei, we used a fictitious mass of 800 a.m.u. and a time step of 0.125 fs. The MD simulations were carried out under NVT at 300 K with a Nosé-Hoover chain thermostat (*80*). In the gas phase, we decoupled the simulation boxes from their images using Hockney's method with an extra 4 Å added to each dimension of the simulation boxes (*81*).

Figure 1 shows the two CV trajectories of the ether linkage protonation on β-1,4-linked xylobiose during the MTD simulations. It shows that the CV1 trajectory starts approximately from 0.8 and the CV2 from -0.8, indicating the initial reaction state. The proton quickly moves towards the linkage oxygen and in the first 200 MTD step, the value of CV2 fluctuates between 0.1 and 0.8. Then the proton transfer is complete, and the C-O bond starts to break. The value of CV1 quickly



Scheme 1. Protonation of the ether linkage in xylobiose and the selection of two CVs. (see color insert)

decreased to as low as 0.1. This indicates that the ether linkage is broken, and that a cyclic carbonium-oxonium ion and a xylose molecule have been formed. This barrier crossing happens after adding 200 bias potentials. After the initial barrier crossing, the system stays in the product well on the FES for the next 760 MTD steps. The second barrier crossing then occurs as CV1 increases back to 0.8 and CV2 drops to -0.8. In the subsequent MTD steps, this barrier crossing occurs a few more times, allowing a sufficient sampling.

Figure 2 shows the FES estimated by the CPMD-MTD simulations. There are two minima located on the FES. The first one is located at CV1 = 0.84 and CV2 = -0.59, corresponding to the reactant state. The second peak is located at CV1 = 0.14 and CV2 = 0.69, representing the product state of the reactions where the protonation has been completed and the C-O bond in the ether linkage is broken. The free-energy difference between these locations is -24.1 kcal /mol, which favors the products. Along the reaction coordinates, the transition state is located at CV1 = 0.84 and CV2 = -0.07, with free energy increasing by 4.2 kcal/mol over the reactant state. In this transition state, the C-O bond stays linked; and the proton is approximately at the middle position between the linkage oxygen and the oxygen in the H_3O^+. This protonated xylobiose is located on FES at CV1 = 0.81 and CV2 = 0.69. The free-energy increase with respect to the reactant state is 2.2 kcal/mol. The C-O bond breaking occurs without an energy barrier after the proton transfer.

The two CVs for calculating the FES of the ether linkage protonation in β-1,2-linked xylose disaccharide are the same as those used for β-1,4-linked xylobiose. The same values of *k* and *m* are used as before. The trajectories of the two CVs and the constructed FES from CPMD-MTD simulations are shown in Figures 3 and 4, respectively. The CV trajectories show a slightly different progress in exploring the reaction pathway. In the first 680 MTD steps, the proton travels back and forth between the H_3O^+ oxygen and the linkage oxygen, and the C-O bond remains bonded. Then, with enough bias potentials added in the reactant well, the system crosses the energy barrier to enter the product well and stays there for the next 3180 MTD steps. The system stays much longer in the product well than in the earlier case, suggesting that the FES features may be different between the two cases. During this period, the proton occasionally goes closely back to the O atom in H_3O^+ and the C-O bond reforms, but it never has a sustained stay in those positions. After 3860 MTD steps, the energy barrier is re-crossed and the values of CV1 can be seen going back to around 0.8. In the mean time, the value of CV2 starts to decrease

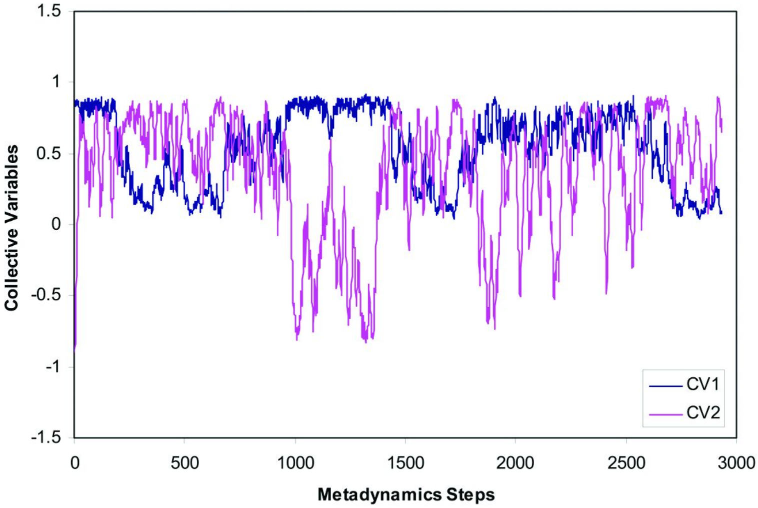


Figure 1. CV trajectory during the MTD simulations for the protonation of the ether linkage in β-1,4-linked xylobiose in the gas phase at 300 K. (see color insert)

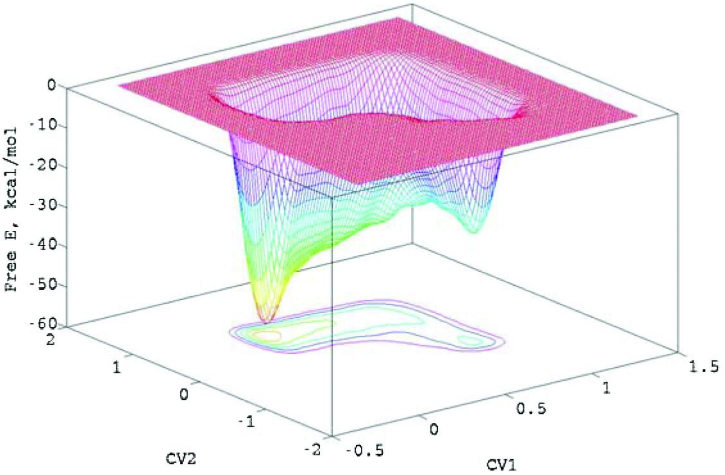


Figure 2. The free-energy surface for the protonation of the ether linkage in β-1,4-linked xylobiose in the gas phase at 300 K. (see color insert)

and fluctuates between -0.8 to 0.8 for the next 1600 MTD steps. Thereafter, CV2 stays in the negative region until it goes up to 0.8 again at MTD step 6560.

Figure 4 shows the FES from MTD simulations for protonation of β-1,2-linked disaccharide. This surface again shows two minima. The first minimum is located at CV1 = 0.90 and CV2 = -0.51, corresponding to the reactant state. The second is located at CV1 = 0.14 and CV2 = 0.74, corresponding to the product state. The free-energy difference between these locations is -27.4 kcal/mol compared to -24.1 kcal/mol in the previous case. Along the reaction coordinates, the transition state

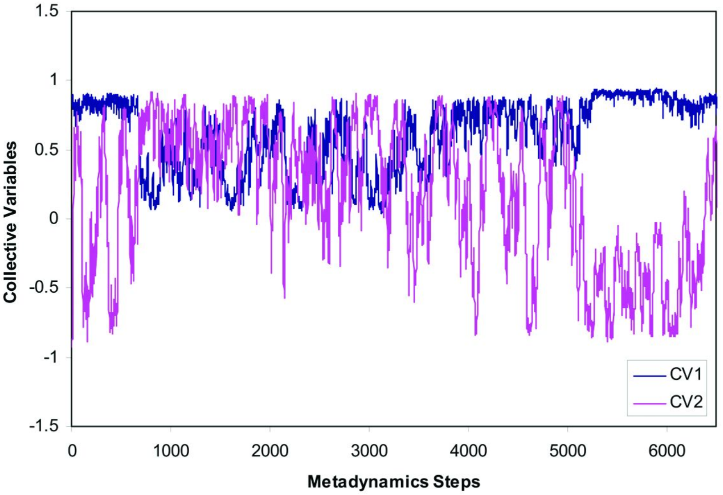


Figure 3. CV trajectories during the metadynamics simulation for the protonation of the ether linkage in β-1,2-linked xylose disaccharide in the gas phase at 300 K. (see color insert)

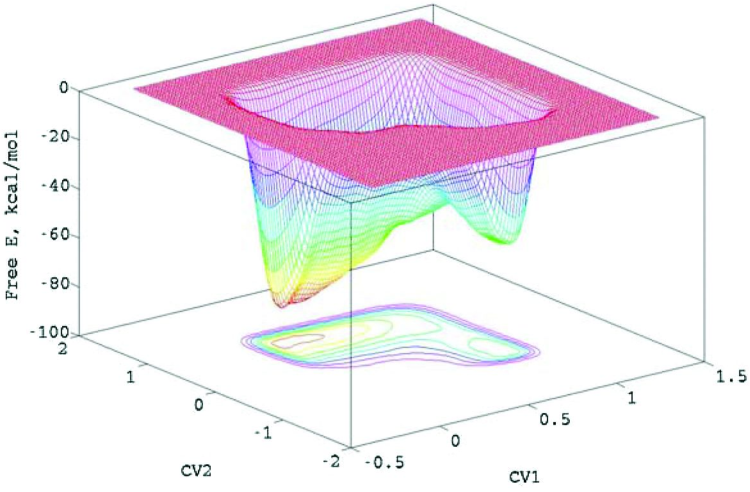


Figure 4. The free-energy surface for the protonation of the ether linkage in β-1,2-linked xylose disaccharide in the gas phase at 300 K. (see color insert)

is located at CV1 = 0.86 and CV2 = 0.00 showing a free-energy increase of 11.0 kcal/mol with respect to the reactant state. This activation energy is much larger than the previous case. On the RC, the state of protonated 1,2- xylobiose is located at CV1 = 0.80 and CV2 = 0.74. The free-energy difference to the reactant state is 6.5 kcal /mol. Again, the C-O bond break occurs without an energy barrier after completing the proton transfer.

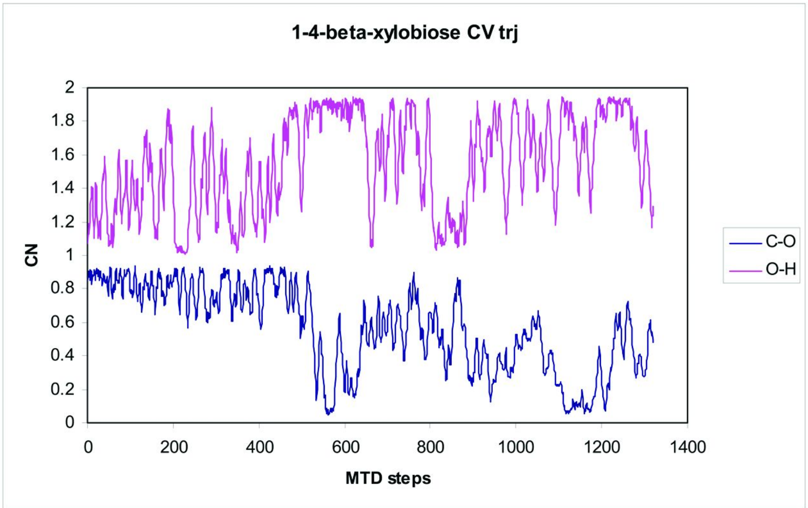


Figure 5. The two CV trajectories for protonation of the ether linkage in xylobiose and the subsequent breaking of the C-O bond in water at 300 K during CPMD-MTD simulations. (see color insert)

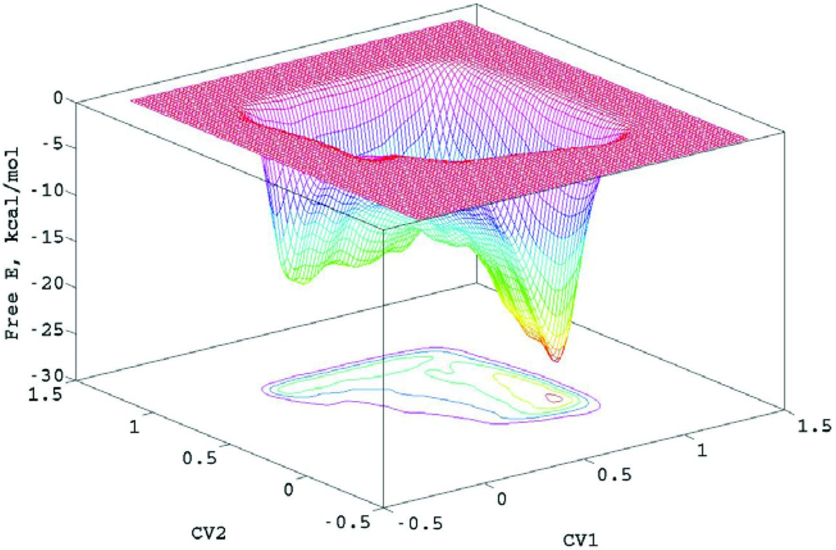


Figure 6. Free-energy surface for protonation of the ether linkage in xylobiose and the subsequent breaking of the C-O bond in water at 300 K from CPMD-MTD simulations. (see color insert)

Acid-Catalyzed Hydrolysis of β-1,4-Linked Xylobiose in Bulk Water

Acid-catalyzed hydrolysis of xylobiose in water was investigated by studying the ether linkage protonation and breakage of the C1-O bond in xylobiose. Sixty-eight water molecules were included in the xylobiose simulation box with a corresponding water density of 0.93 g/cm^3. In addition, one H_3O^+ ion and one counter ion Cl^- were introduced into the system to mimic the acidic environment and to neutralize the charge. Periodic boundary conditions were applied. The simulations were carried out at 300 K. For MTD simulations, CV1 is chosen to be the same as that in the gas phase. However, CV2 is slightly different. Here, CV2 is the coordination number of the O atom on the ether linkage with respect to one proton on H_3O^+, as shown in Scheme 1.

Figures 5 and 6 show the trajectories of the reaction coordinates during CPMD-MTD simulations and the constructed FES, respectively, for protonation of the ether linkage on β-1,4-xylobiose in aqueous solution at 300 K. Figure 5 shows that it takes more than 500 MTD steps to complete the sampling of the protonation process. At around 520 MTD steps, the C-O bond starts to break. The FES shown in Figure 6 exhibits two energy minima. The first one is located at CV1 = 0.9 and CV2 = 0.1, corresponding to the reactant state where the proton remains close to the H_2O molecule. The second minimum is located at CV1 = 0.1 and CV2 = 0.8, corresponding to the product well where the proton has been transferred to the ether linkage and the C1–O1 bond is broken. The overall free-energy change is 7 kcal/mol. The transition state is located at CV1 = 0.8 and CV2 = 0.8, with a free-energy barrier of 10 kcal/mol over the reactant state.

The reaction barrier of 10 kcal/mol obtained from the current CPMD-MTD simulations for protonation of the ether linkage and breakage of the C1-O1 bond is much smaller compared to the experimental value of 30 kcal/mol (*82*, *83*). The discrepancy between the experimental and calculated reaction barriers can be explained by taking into account the reaction barrier for partial desolvation of the hydronium ion as it moves closer to the xylobiose molecule in solution. In this case, the calculated partial desolvation free energy for a proton is approximately 15-20 kcal/mol (*79*). The total activation energy for protonation of the ether linkage and the breaking of the C-O bond during xylobiose hydrolysis is about 25-30 kcal/mol, which is in reasonable agreement with experiments.

Conclusions

The effects of water and water structure on acid-catalyzed xylan hydrolysis reaction can be inferred by investigating the xylobiose hydrolysis reaction both in the gas phase and in solution. Water and water structure both play a critical role in determining the reaction barriers in solution. Proton partial desolvation and its migration to the ether linkage site, protonation of the ether bond and the subsequent breaking of the C-O bond is the rate-limiting step. The calculated reaction barrier is in reasonable agreement with the corresponding experimental value. Because of the large contribution of proton partial desolvation to the overall reaction barrier, acid concentration and proton transport will play critical roles in xylan hydrolysis.

Acknowledgments

This work was supported by the Department of Energy, Office of the Biomass Program via a subcontract from the National Renewable Energy Laboratory (ZCO-

7-77386-01). The authors would like to thank Mark Nimlos, David Johnson, Michael Himmel, Robert N. Goldberg, Bernd Ensing, and Nisanth Nair for helpful discussions. Calculations were carried out using computing resources at Colorado State University, National Renewable Energy Laboratory, and Teragrid.

References

1. Sun, Y.; Cheng, J. Y. *Bioresour. Technol.* **2002**, *83*, 1–11.
2. Mosier, N.; Wyman, C.; Dale, B.; Elander, R.; Lee, Y. Y.; Holtzapple, M.; Ladisch, M. *Bioresour. Technol.* **2005**, *96*, 673–686.
3. Wooley, R.; Ruth, M.; Glassner, D.; Sheehan, J. *Biotechnol. Prog.* **1999**, *15*, 794–803.
4. Lynd, L. R.; Elander, R. T.; Wyman, C. E. *Appl. Biochem. Biotechnol.* **1996**, *57−8*, 741–761.
5. Knappert, D.; Grethlein, H.; Converse, A. *Biotechnol. Bioeng.* **1980**, *22*, 1449–1463.
6. Knappert, D. R. Ph.D. Thesis, Dartmouth College, Hannover, NH, 1981.
7. Hsu, T. A. In *Handbook on Bioenthanol Production and Ultlization*; Wyman, C., Ed.; Taylor & Francis: Washington, D.C., 1996.
8. Clarke, M. A.; Edye, L. A.; Eggleston, G. In *Advances in Carbohydrate Chemistry and Biochemistry*; 1997; Vol. 52.
9. Heitz, M.; Capekmenard, E.; Koeberle, P. G.; Gagne, J.; Chornet, E.; Overend, R. P.; Taylor, J. D.; Yu, E. *Bioresour. Technol.* **1991**, *35*, 23–32.
10. Saddler, J. N.; Ramos, L. P.; Breuil, C. In *Bioconversion of Forest and Agricultural Plant Residues*; Saddler, J. N., Ed.; CAB International: Oxford, 1993.
11. Dale, B. E.; Moreira, M. J. *Biotechnol. Bioeng.* **1982**, 31–43.
12. Dale, B. E.; Leong, C. K.; Pham, T. K.; Esquivel, V. M.; Rios, I.; Latimer, V. M. *Bioresour. Technol.* **1996**, *56*, 111–116.
13. Chang, V. S.; Burr, B.; Holtzapple, M. T. *Appl. Biochem. Biotechnol.* **1997**, *63-5*, 3–19.
14. Chang, V. S.; Nagwani, M.; Holtzapple, M. T. *Appl. Biochem. Biotechnol.* **1998**, *74*, 135–159.
15. Wyman, C. E.; Dale, B. E.; Elander, R. T.; Holtzapple, M.; Ladisch, M. R.; Lee, Y. Y. *Bioresour. Technol.* **2005**, *96*, 1959–1966.
16. Himmel, M. E.; Ding, S. Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D. *Science* **2007**, *315*, 804–807.
17. Esteghlalian, A. R.; Hashimoto, A. G.; Fenske, J. J.; Penner, M. H. *Bioresour. Technol.* **1997**, *59*, 129–136.
18. Shiang, M.; Linden, J. C.; Mohagheghi, A.; Grohmann, K.; Himmel, M. E. *Biotechnol. Prog.* **1991**, *7*, 315–322.
19. Kim, S. B.; Lee, Y. Y. *Biotechnol. Bioeng. Symp.* **1987**, *17*, 71.
20. Maloney, M. T.; Chapman, T. W.; Baker, A. J. *Biotechnol. Bioeng.* **1985**, *27*, 355–361.
21. Mayans, O.; Scott, M.; Connerton, I.; Gravesen, T.; Benen, J.; Visser, J.; Pickersgill, R.; Jenkins, J. *Structure* **1997**, *5*, 677–689.

22. Chen, R.; Lee, Y. Y.; Torget, R. *Appl. Biochem. Biotechnol.* **1996**, *57/58*, 133–146.
23. Liu, C. G.; Wyman, C. E. *Ind. Eng. Chem. Res.* **2004**, *43*, 2781–2788.
24. Keller, F. A.; Hamilton, J. E.; Nguyen, Q. A. *Appl. Biochem. Biotechnol.* **2003**, *105−108*, 27–41.
25. Torget, R.; Walter, P.; Himmel, M.; Grohmann, K. *Appl. Biochem. Biotechnol.* **1991**, *28-29*, 75–86.
26. Antal, M. J.; Leesomboon, T.; Mok, W. S.; Richards, G. N. *Carbohydr. Res.* **1991**, *217*, 71–85.
27. Qian, X. H.; Nimlos, M. R.; Johnson, D. K.; Himmel, M. E. *Appl. Biochem. Biotechnol.* **2005**, *121*, 989–997.
28. Qian, X. H.; Nimlos, M. R.; Davis, M.; Johnson, D. K.; Himmel, M. E. *Carbohydr. Res.* **2005**, *340*, 2319–2327.
29. Nimlos, M. R.; Qian, X. H.; Davis, M.; Himmel, M. E.; Johnson, D. K. *J. Phys. Chem. A* **2006**, *110*, 11824–11838.
30. Antal, M. J.; Mok, W. S. L.; Richards, G. N. *Carbohydr. Res.* **1990**, *199*, 91–109.
31. Torget, R. W.; Kim, J. S.; Lee, Y. Y. *Ind. Eng. Chem. Res.* **2000**, *39*, 2817–2825.
32. Kuster, B. F. M. *Starch* **1990**, *42*, 341–21.
33. Chen, S. F.; Mowery, R. A.; Castleberry, V. A.; van Walsum, G. P.; Chambliss, C. K. *J. Chromatogr., A* **2006**, *1104*, 54–61.
34. Nimlos, M. R. Private Communication.
35. *CPMD 3.11.1*; copyrighted jointly by IBM Corp and by Max-Planck Institute: Stuttgart.
36. Kohn, W. *Rev. Mod. Phys.* **1999**, *71*, 1253–1266.
37. Kuo, I. F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I.; VandeVondele, J.; Sprik, M.; Hutter, J.; Chen, B.; Klein, M. L.; Mohamed, F.; Krack, M.; Parrinello, M. *J. Phys. Chem. B* **2004**, *108*, 12990–12998.
38. VandeVondele, J.; Mohamed, F.; Krack, M.; Hutter, J.; Sprik, M.; Parrinello, M. *Journal of Chemical Physics* **2005**, *122*, 6.
39. Martonak, R.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2005**, *122*, 10.
40. Stirling, A.; Iannuzzi, M.; Parrinello, M.; Molnar, F.; Bernhart, V.; Luinstra, G. A. *Organometallics* **2005**, *24*, 2533–2537.
41. Martonak, R.; Laio, A.; Bernasconi, M.; Ceriani, C.; Raiteri, P.; Zipoli, F.; Parrinello, M. *Z. Kristallogr.* **2005**, *220*, 489–498.
42. Raiteri, P.; Martonak, R.; Parrinello, M. *Angew. Chem., Int. Ed.* **2005**, *44*, 3769–3773.
43. Oganov, A. R.; Martonak, R.; Laio, A.; Raiteri, P.; Parrinello, M. *Nature* **2005**, *438*, 1142–1144.
44. Barducci, A.; Chelli, R.; Procacci, P.; Schettino, V.; Gervasio, F. L.; Parrinello, M. *J. Am. Chem. Soc.* **2006**, *128*, 2705–2710.
45. Rodriguez-Fortea, A.; Iannuzzi, M.; Parrinello, M. *J. Phys. Chem. B* **2006**, *110*, 3477–3484.
46. Dyer, P. J.; Cummings, P. T. *J. Chem. Phys.* **2006**, *125*, 6.

47. Kuo, I. F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I. *J. Chem. Theory Comput.* **2006**, *2*, 1274–1281.
48. Williams, R. W.; Malhotra, D. *Chem. Phys.* **2006**, *327*, 54–62.
49. Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 9.
50. Chen, B.; Ivanov, I.; Park, J. M.; Parrinello, M.; Klein, M. L. *J. Phys. Chem. B* **2002**, *106*, 12006–12016.
51. Mundy, C. J.; Colvin, M. E.; Quong, A. A. *J. Phys. Chem. A* **2002**, *106*, 10063–10071.
52. Iannuzzi, M.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *90*, 238302 .
53. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2001.
54. Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.
55. Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
56. Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384–2393.
57. Nose, S. *Mol. Phys.* **1984**, *52*, 255–268.
58. Andreoni, W.; Marx, D.; Sprik, M. *ChemPhysChem* **2005**, *6*, 1671–1676.
59. Tateyama, Y.; Blumberger, J.; Sprik, M.; Tavernelli, I. *J. Chem. Phys.* **2005**, *122*, 234505 .
60. Cucinotta, C. S.; Ruini, A.; Catellani, A.; Stirling, A. *J. Phys. Chem. A* **2006**, *110*, 14013–14017.
61. Bulo, R. E.; Donadio, D.; Laio, A.; Molnar, F.; Rieger, J.; Parrinello, M. *Macromolecules* **2007**, *40*, 3437–3442.
62. Piana, S.; Laio, A. *J. Phys. Chem. B* **2007**, *111*, 4553–4559.
63. Amat, M. A.; Kevrekidis, I. G.; Maroudas, D. *Appl. Phys. Lett.* **2007**, *90*, 171910 .
64. Lelievre, T.; Rousset, M.; Stoltz, G. *J. Chem. Phys.* **2007**, *126*, 134111 .
65. Spiwok, V.; Lipovova, P.; Kralova, B. *J. Phys. Chem. B* **2007**, *111*, 3073–3076.
66. Rodriguez-Fortea, A.; Iannuzzi, M.; Parrinello, M. *J. Phys. Chem. C* **2007**, *111*, 2251–2258.
67. Bulo, R. E.; Siggel, L.; Molnar, F.; Weiss, H. *Macromol. Biosci.* **2007**, *7*, 234–240.

68. Michel, C.; Laio, A.; Mohamed, F.; Krack, M.; Parrinello, M.; Milet, A. *Organometallics* **2007**, *26*, 1241–1249.

69. Branduardi, D.; Gervasio, F. L.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 054103.

70. Boero, M.; Ikeda, T.; Ito, E.; Terakura, K. *J. Am. Chem. Soc.* **2006**, *128*, 16798–16807.

71. Qian, X.; Nimlos, M. R. In *Biomass Recalcitrance*; Himmel, M., Ed.; Blackwell Publishing Ltd: Oxford, 2007.

72. Babin, V.; Roland, C. *J. Chem. Phys.* **2006**, *125*, 204909.

73. Ensing, B.; Laio, A.; Parrinello, M.; Klein, M. L. *J. Phys. Chem. B* **2005**, *109*, 6676–6687.

74. Laio, A.; Rodriguez-Fortea, A.; Gervasio, F. L.; Ceccarelli, M.; Parrinello, M. *J. Phys. Chem. B* **2005**, *109*, 6714–6721.

75. Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

76. Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

77. Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703–1710.

78. Qian, X. H. *Mol. Sim.* **2008**, 183–191.

79. Dong, H.; Nimlos, M. R.; Himmel, M. E.; Johnson, D. K.; Qian, X. *J. Phys. Chem.* **2009**, *113*, 8577–8585.

80. Nose, S. *J. Chem. Phys.* **1984**, *81*, 511.

81. Hockney, R. W. *Methods Comput. Phys.* **1970**, *9*, 136.

82. Capon, B. *Chem. Rev.* **1969**, *69*, 407–498.

83. Tewari, Y. B.; Goldberg, R. N. *J. Biol. Chem.* **1989**, *264*, 3966–3971.

Chapter 2

Simulations of the Structure of Cellulose

James F. Matthews,[1],* Michael E. Himmel,[1] and John W. Brady[2]

[1]Biosciences Center, National Renewable Energy Laboratory, Golden, CO 80401
[2]Department of Food Science, Cornell University, Ithaca, NY 14853
***james.matthews@nrel.gov**

Cellulose is the homopolymer of (1→4)-β-D-glucose. The chemical composition of this polymer is simple, but understanding the conformation and packing of cellulose molecules is challenging. This chapter describes the structure of cellulose from the perspective of molecular mechanics simulations, including conformational analysis of cellobiose and simulations of hydrated cellulose Iβ with CSFF and GLYCAM06, two sets of force field parameters developed specifically for carbohydrates. Many important features observed in these simulations are sensitive to differences in force field parameters, giving rise to dramatically different structures. The structures and properties of non-naturally occurring cellulose allomorphs (II, III, and IV) are also discussed.

Introduction

This chapter will describe molecular mechanics simulations of cellulose structure, starting with an overview of the origin and uses of different cellulose crystal forms including the chain-packing features that define them. A detailed examination of cellobiose conformation follows as an introduction to carbohydrate nomenclature and molecular mechanics simulations. The conformational analysis of cellobiose illustrates how the behavior observed in simulations depends on the differences between the force field parameters used, which is also seen in the cellulose simulations. Last are presented simulations of cellulose Iβ fibrils in water, which are analysed in terms of hydrogen bonding, conformation of the exocyclic hydroxymethyl groups, conformation of the glycosidic linkages, and overall shape of the fibrils. A methods section describing the computational approaches used to conduct this work concludes the chapter.

Uses of Cellulose and Background

As the single largest component of dry biomass, cellulose is one of the most important molecules on Earth. It is the principal structural polymer in the cell walls of plants and algae. It can be hydrolyzed into simple sugars to provide an energy source for bioreactors, primarily fermenters. Cellulose is important not only because it is abundant, but also because it is renewable, making it an attractive alternative to petrochemicals and other fossil fuels (*1*). The enzymatic hydrolysis of crystalline cellulose is a slow process, primarily because the polymer is insoluble and difficult to decrystallize (*2*); but there is also a significant amount of non-crystalline cellulose in plants. It is not immediately obvious why the polymer of (1→4)-β-D-glucose is so insoluble, considering that the monomer unit is one of the most soluble organic compounds known.

Whereas cellulose has a simple chemical composition, determining how the polymer chains pack together has been a complicated and controversial subject. Depending on the synthesis conditions and treatment history, seven different types of crystal packing have been proposed (*3*). This number does not include alternative structures for each allomorph, and does not include chemical modifications or structures containing intercalated small molecules or ions (*4*). Some forms are more readily hydrolyzed than others (*5*, *6*). Therefore, understanding the differences in the structure and surface properties of these allomorphs may lead to improved enzymatic hydrolysis rates.

Cellulose does not generally occur as a single chain, but rather is synthesized in close proximity to many other polymer chains, which organize into fibrils as the fundamental structural unit (*7–9*). Cellulose oligomers longer than cellohexaose are almost completely insoluble (*2*). The glucan chain length (degree of polymerization, DP) varies from about 2000 to more than 15,000 glucose residues (*10*). Cellulose can vary from the so-called elementary fibrils in plants, which contain approximately 36 cellodextrin chains, to the large microfibrils and macrofibrils of cellulosic algae, which contain more than 1200 chains (*11–13*). The shape of a cellulose fibril is thought to be determined by the geometry of the cellulose synthase complex and by the local environment (*14*). The seven allomorphs of crystalline cellulose can be separated into two groups based on the polarity of chain arrangement within the crystal lattice (see Figure 1). Allomorphs with parallel chains are grouped into the cellulose I family, and allomorphs with neighboring chains arranged anti-parallel are grouped into the cellulose II family (*15*).

Natural Cellulose Allomorphs

The cellulose I family includes natural cellulose, which has been shown to be made up of two different crystal phases: a triclinic form with one chain per unit cell, designated as Iα, and a lower energy, more stable monoclinic form with two non-equivalent chains per unit cell, designated as Iβ (*16*, *17*). The Iα form is

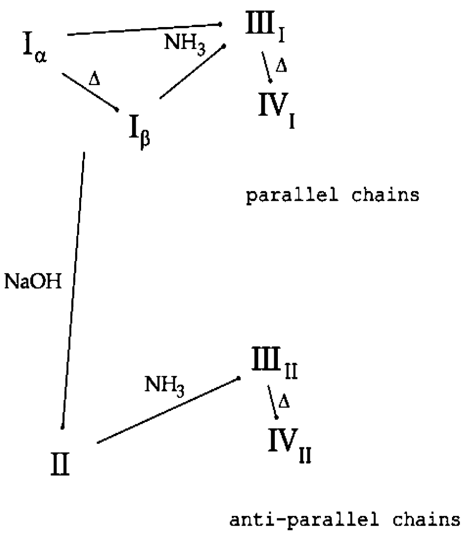


Figure 1. Relationship between chain polarity in cellulose allomorphs. The cellulose I family on top has parallel chains, while the cellulose II family on bottom has anti-parallel chains.

more susceptible to hydrolysis. Because all of the hydroxyl groups in cellulose are equatorial, all of the axial positions are occupied by non-polar (and non-hydrogen-bonding) aliphatic protons. The sides of the cellulose chain are polar and can form hydrogen bonds, while the “tops” and “bottoms” are mostly hydrophobic. The chains can stack in layers defined by hydrogen bonding, with hydrophobic chain faces meeting between layers. These two crystal forms have similar molecular conformations and lateral packing (Figure 2), but alternating sheets differ (*18*, *19*).

Whereas the dominant phase in higher plants is the Iβ form and algae contain a higher proportion of Iα, both phases can coexist along and across the same fibril (*20*). Cellulose occurs in the plant cell wall embedded in a matrix of hemicelluloses and lignin, which, in the biomass conversion process, is disrupted prior to enzymatic treatment. The newly exposed cellulose fibrils are not well characterized as to the exact number of chains on each surface, but most of the fibril surface exposes the hydrophilic sides of the glucose monomers to solution. These surfaces meet at a corner where the more hydrophobic faces of the glucose monomers are exposed. Other surfaces may be exposed by mechanical rounding of corners, or by selective hydrolysis (*21*). Although these surfaces are not likely to be a significant portion of the fibril surface area in plants, they may play a critical role in the activity of fungal cellulases. It has been shown that fungal-type cellulose binding modules bind specifically to the hydrophobic surfaces of large Iα crystals (*22*). Another important structural feature of cellulose from plant cell walls is the overall fibril shape, which take on a right-handed twist with a period on the order of hundreds of nanometers (*23–25*). This twist eliminates the possibility of a true crystallographic unit cell. Not all native cellulose fibrils are twisted; wide fibrils such as those from *Valonia* or *Halocynthia* are straight (*26*), and the relationship between diameter and twist will be explored below.

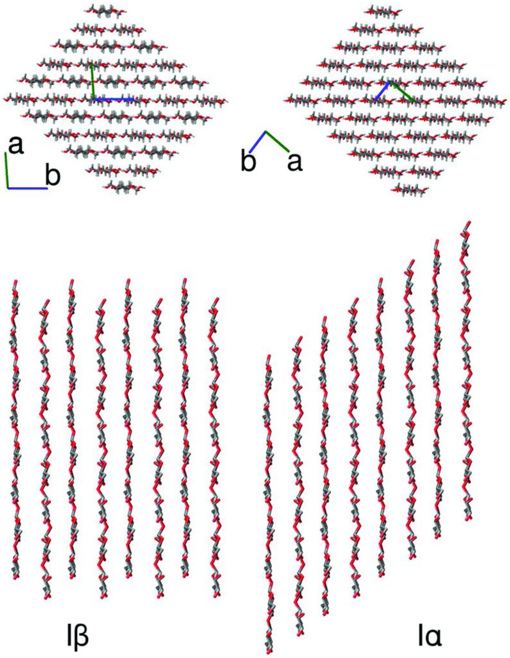


Figure 2. Comparison of cellulose Iα and Iβ crystal packing. The packing is nearly identical in cross section, but the displacement of the hydrogen-bonded sheets of Iβ are staggered alternately up and down by half a glucose monomer length, while Iα hydrogen-bonded sheets align on a constant inclined axis. This relationship allows these crystal forms to be easily inter-converted by thermal annealing. (see color insert)

Man-Made or Uncommon Cellulose Allomorphs

Cellulose II can be formed by several processes. Treating cellulose I with alkali while the fibers are under tension is known as mercerization, which is an important process in the textile industry (*1–4*). Regenerated cellulose is also important in the textile industry for producing rayon. Regeneration requires preparing a solution of cellulose, either with special solvents or by chemical modification (*5*). During subsequent precipitation, the chains recrystallize in an anti-parallel arrangement. Cellulose II is more energetically stable than cellulose I, and the transformation to anti-parallel chains is irreversible. Cellulose II can be directly synthesized by some bacteria in confined spaces or at low temperature, forming a highly corrugated structure with the individual chains folding back upon themselves, but this is not commonly found in nature (*6*). While regenerating anti-parallel chains from solution presents no special conceptual problems and the folded structure of bacterial cellulose II can be observed, it is not immediately obvious how conversion from parallel to anti-parallel chains is accomplished without fiber dissolution during the mercerization process.

Cellulose III can be formed by treating either cellulose I or II with small nitrogen-containing compounds such as ammonia or ethylenediamine (*7–9*). The

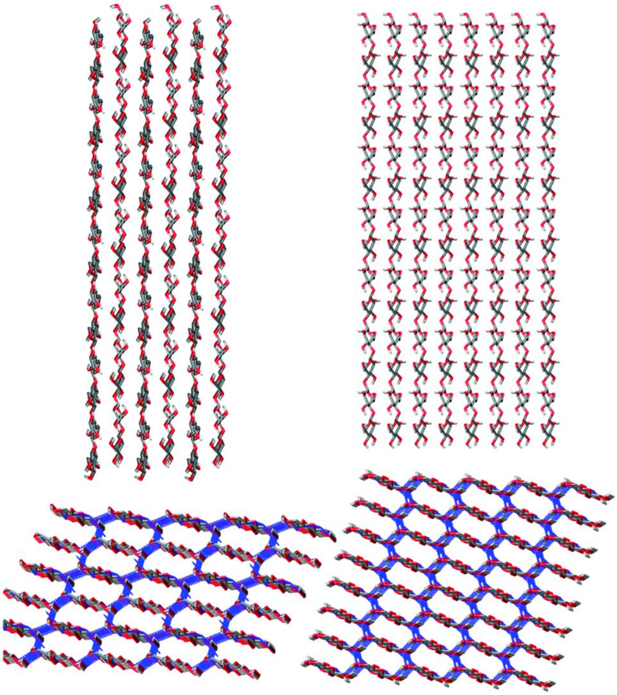


Figure 3. Cellulose II (left) and cellulose III$_I$ (right) shown from the edge and from the end, with hydrogen bonds shown in blue. (see color insert)

resulting structures are either III$_I$ or III$_{II}$ depending on the starting material and cannot be interconverted. Cellulose III has a morphology that is distinctly different from natural cellulose – with the fibril structure largely disrupted – which imparts a softer texture to textiles. Cellulose II and III$_I$ are shown in Figure 3.

Cellulose IV can be prepared by heating cellulose III and, likewise, is either cellulose IV$_I$ or IV$_{II}$ depending on the starting material (*10–12*). Cellulose IV$_I$ can also be formed by regenerating short chains at elevated temperatures (*13*). It has recently been suggested that cellulose IV$_I$ is not a separate allomorph, but rather is Iβ with a large amount of lateral disorder (*14*).

The overarching goal of our research is to improve the efficiency of cellulase enzymes acting on crystalline celluose, and understanding the structure and behavior of crystalline cellulose is the first step towards understanding enzyme/substrate interactions on crystal surfaces. While the simulations presented in this chapter are small compared to the size of plant cell walls, the behavior of the model systems presented here should give insight into the structure of cellulose, starting from the smallest cellulose repeating unit, cellobiose, and proceeding to larger systems that more closely approximate the structure of interest, cellulose fibrils in plant cell walls.

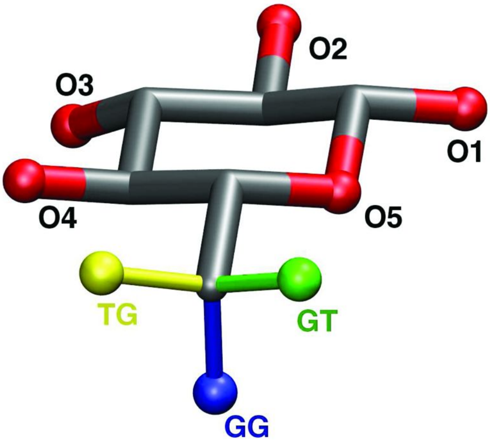


Figure 4. Hydroxymethyl rotamers of β-D-glucose. (see color insert)

Cellobiose

Conformational Energy Mapping

As the disaccharide with the same glycosidic linkage as cellulose, cellobiose has been studied extensively as a model for cellulose structure (*15–19*). Low-energy cellobiose conformations can be expected to give insight into possible low-energy cellulose conformations, but the immediate environment is an important factor influencing the shape of cellulose chains. Cellotetraose is the smallest oligomer that contains a glycosidic linkage connecting monomers that are not at a chain end and may be a better model than cellobiose for cellulose. This is an important observation because in cellotetraose the hydroxymethyl group at the non-reducing end cannot form a strong hydrogen bond between residues across the glycosidic linkage, and both chain ends are less constrained than interior residues. Nevertheless, low-energy conformations of the glycosidic linkage of cellobiose indicate which regions of φ and ψ space is allowed for cellulose chains. Conformational energy mapping of disaccharide structure has a long history; and while the parameters, methods, and computational power used to calculate energies have advanced, the basic idea of finding low-energy regions of (φ,ψ) space remains the same (*20*, *21*).

As shown in Figure 4, the exocyclic hydroxymethyl group containing C6 and O6 has three low-energy staggered conformations, which are named GT, GG, and TG. The first letter in these labels specifies the position of the O6 atom as either *trans* or *gauche* with respect to the O5 atom, and the second letter specifies its relationship to the C4 atom (see Figure 4). This dihedral angle is called ω and can be reported as a single number (i.e., ± 60, 180), but it is more precise to use the two-letter code. Remembering that the first letter always relates to the ring oxygen, GT, GG, and TG are not ambiguous names when describing the rotameric state, whereas reporting a numeric value gives no clue as to which atoms were chosen to define the dihedral angle. Glucose is special among the monosaccharides in that the sides of the ring are hydrophilic, and the faces of the ring are hydrophobic. When in the GT or TG conformation, all hydroxyl groups are in the same approximate plane, which allows the possibility for the ring faces to interact via hydrophobic stacking.

The glycosidic linkage of cellobiose can be characterized by consecutive torsion angles containing C1, the glycosidic oxygen, and C4′. These angles are φ (H1-C1-O1-C4′) and ψ (C1-O1-C4′-H4′) and are analogous to the backbone dihedral angles in proteins. Conformational analysis of combinations of φ and ψ allows us to construct energy maps for disaccharides that are similar to amino acid Ramachandran plots. These maps show regions of glycosidic linkage conformation with low potential energy. Cellobiose is synthesized so that neighboring residues are rotated by approximately 180 degrees relative to each other. In this conformation, two hydrogen bonds across the glycosidic linkage can occur: the nearly always present (HO3′-O5) hydrogen bond, and the conditional (HO6′-O2) hydrogen bond possible only when the reducing-end hydroxymethyl group is TG or GG. Hydrogen bonds are very important in cellulose structure, and these particular hydrogen bonds may help stabilize conformations that produce flat ribbons. However, these hydrogen bonds across the glycosidic linkage are not symmetrical and, therefore, are not expected to make the underlying energy landscape symmetrical about the axis of two-fold symmetry.

The mapping procedure reveals low-energy regions in (φ,ψ) space, but the lowest energy conformation at each point does not necessarily lie on a smooth transition path between neighboring points. Also, the lowest energy ring shape, hydroxyl orientation, and hydroxymethyl conformation at each (φ,ψ) point may not be the same for different parameter sets. This chapter will explore cellobiose conformations in vacuum with adiabatic potential energy maps, where the energy at each (φ,ψ) point is obtained by minimizing the potential energy of many independent starting conformations while allowing each atom to move without restraints, except on the glycosidic dihedral angles. This chapter will also explore crystalline cellobiose structures using the CSFF and GLYCAM06 force fields, two additive force fields developed for carbohydrates.

Cellulose is synthesized from monomers of UDP-glucose, but cellulose synthase enzymes have two binding sites that are assumed to be rotated relative to each other by 180° (*22–25*). This geometry of the synthase active site makes the monomers in the growing cellulose chain alternate the direction of each hydroxymethyl group point. This arrangement, with near two-fold symmetry along with the stereochemistry of the glycosidic linkage, makes a nearly flat chain shape possible.

There is no experimental way to generate a conformational energy map, but nearly all observed small-molecule cellulose analog crystal structures are observed to have glycosidic linkage angles that fall near low-energy regions in the calculated maps (*26*). To a close approximation for cellobiose, the line connecting points with φ + ψ = 0 is the line of two-fold helical symmetry. Changes in ring shape allow small deviations away from this line to also have two-fold symmetry. There is a small low-energy region through which this line passes, but for most calculated maps this line does not coincide with a minimum on the energy surface. Previous studies of cellobiose conformations have described two low-energy regions in the center of the map that are just to either side of the two-fold helical axis and two additional wells near the fringes of the map (*17*). In the absence of water, the global minimum energy is near φ = 180 and ψ = 0. This unexpected global low-energy region is caused by a concerted intramolecular hydrogen bond pattern with all hydroxyl groups in a single, unbroken head-to-tail arrangement. It is unlikely that this hydrogen bond pattern persists in solution, and upon adding one to four hydrating water molecules that interrupt this pattern, this region is no longer lower in energy than the central regions (*27*, *28*).

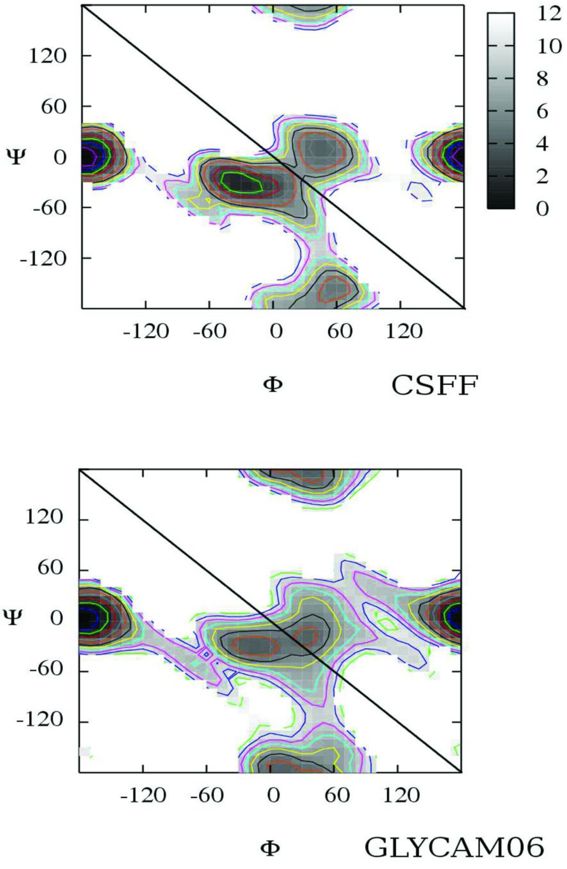


Figure 5. Adiabatic potential energy maps for the CSFF and GLYCAM06 force fields, on a 10-degree grid. Energy is in kcal/mol above global minimum for each map. Two-fold helical axis is drawn on the diagonal. (see color insert)

Figure 5 displays the adiabatic potential energy maps for cellobiose in vacuum using the CSFF and GLYCAM06 force fields. This discussion will focus on the central region of the maps, the region relevant to native cellulose structure. It should be noted that this vacuum map does not necessarily correspond to the map that would be applicable in solution or in a crystalline environment. However, these maps are useful as a first approximation to the allowed conformations of cellulose chains, and for examining preferred internal hydrogen bonding and hydroxymethyl rotamers at all conformations of the glycosidic linkage. The contour lines in the maps are spaced by 1 kcal/mol up to 12 kcal/mol above the global minimum. Both maps have the same general outline, but the locations of the minima differ in interesting ways. CSFF has minima on both sides of the two-fold axis, and GLYCAM06 has one minimum on the right-handed side, and one minimum just to the left-handed side of the two-fold axis. A survey of the glycosidic linkages of small-molecule cellulose analogs shows that most conformations have left-handed departures from the two-fold axis. Slight left-handed departures from the two-fold axis favor the formation of the HO3-O5′ hydrogen bond, and slight right-handed departures can bring O6 of the reducing end within hydrogen bonding distance of O2′ on the non-reducing end. An aqueous free-energy map for cellobiose has been calculated for the CSFF force field, and the location of the free-energy minimum in solution is on the left-handed side of the two-fold axis, the opposite handedness of the favored conformation in vacuum.

Conformations near the center (0,0) of these maps have the glucose monomers "flipped" by approximately 180° relative to each other; that is, when hydrogen atoms attached to C1 and C4′ across the glycosidic linkage are eclipsed, the hydroxymethyl groups point in opposite directions. For example, a left-handed twist of 5° away from a two-fold helix means neighboring residues are flipped by -175°, whereas a right-handed twist of the same amount means neighboring residues are flipped by +175°. The region of left-handed twist on these maps is to the right of the two-fold helical axis, and right-handed twist is to the left of the two-fold axis.

Figure 6 contains nine vacuum adiabatic potential energy maps corresponding to the nine possible combinations of hydroxymethyl rotamers for cellobiose for both the CSFF and GLYCAM06 force fields. Each column has the same rotamer at the reducing end, and each row has the same rotamer at the non-reducing end. The maps in Figure 5 contain the overall lowest energy for each (φ,ψ) point, but these maps do not give information about the hydroxymethyl conformations at these (φ,ψ) points. Breaking the maps into components by hydroxymethyl conformation shows an important difference in behavior between CSFF and GLYCAM06. The CSFF map is denominated by the conformation of hydroxymethyl groups, with the GG,GG map being nearly indistinguishable from the overall map. The GLYCAM06 maps are dramatically different depending on hydroxymethyl conformations, with low-energy regions where good intramolecular hydrogen bonds are possible. Figure 7 shows the lowest energy conformation from the central region of the nine maps for the GLYCAM06 force field. Again, this is ignoring "flipped" conformations from the edges of these vacuum maps that are not significantly populated in native cellulose. Hydrogen bonds across the glycosidic linkage are possible for certain conformations of the reducing-end hydroxymethyl group. The GG or TG conformations at the reducing end bring the O2 and O6 hydroxyl groups within hydrogen bond distance, but the GT conformation at the reducing end makes a hydrogen bond across this side of the linkage geometrically impossible.

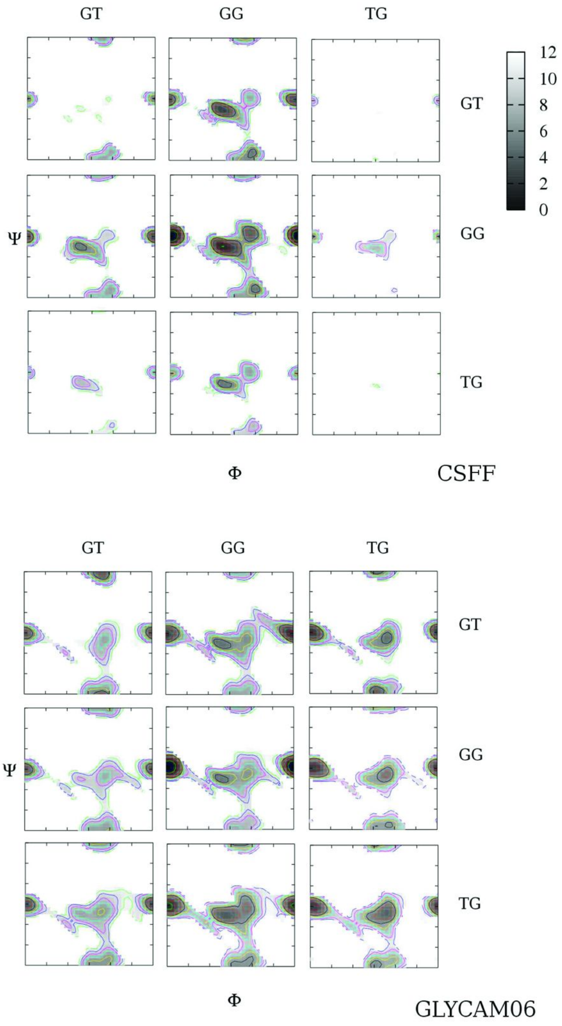


Figure 6. Maps for the nine cellobiose hydroxymethyl combinations using the CSFF and GLYCAM06 forcefields. Each map has the same range as the maps in Figure 5. Energy is in kcal/mol above the global minimum. Columns have the same hydroxymethyl rotameric state at the reducing end, rows at the non-reducing end. (see color insert)

The potential energy difference between the lowest energy hydroxymethyl conformations for each point with GLYCAM06 is on the order of 3.5 kcal/mol, whereas with CSFF the energy difference between conformers is on the order of 7 kcal/mol. This magnitude is significant, because hydrogen bond energies are on the order of 6 kcal/mol, making hydrogen bond energy contributions more important than hydroxymethyl rotamers for GLYCAM06, and less important for CSFF.

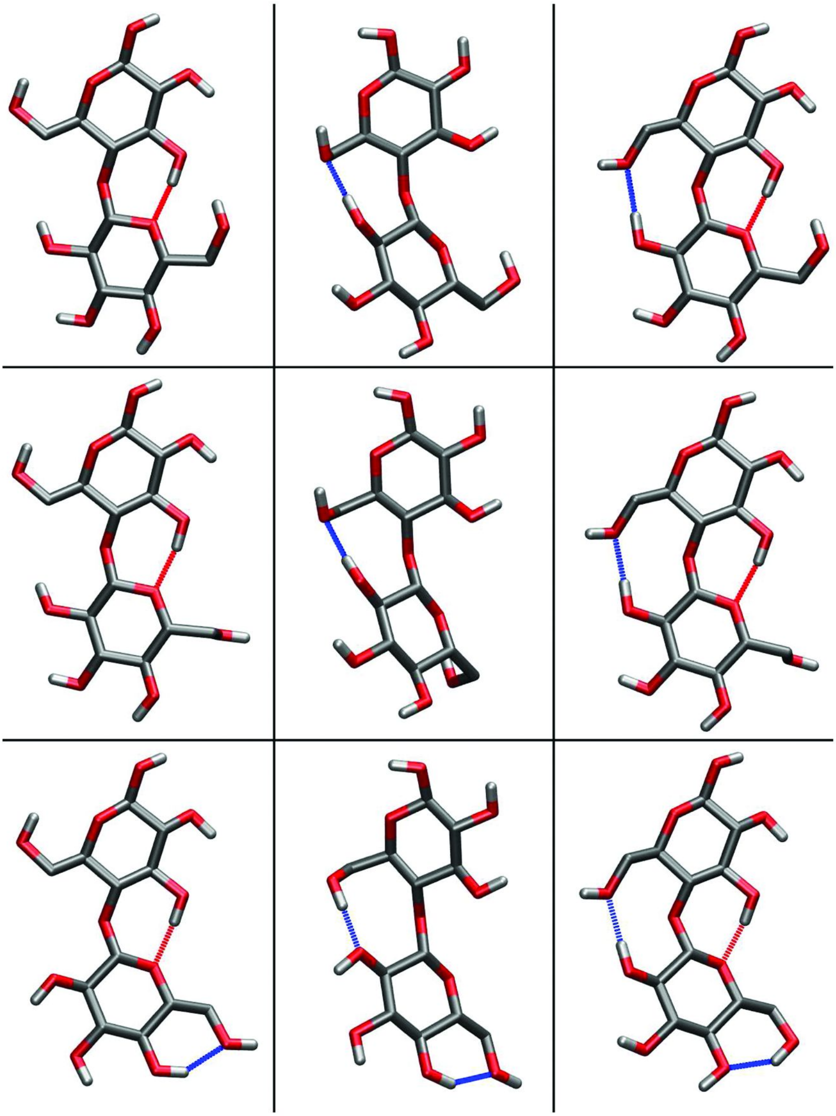


Figure 7. Lowest energy conformation in the central region of the maps in Figure 6 for the GLYCAM06 force field. (see color insert)

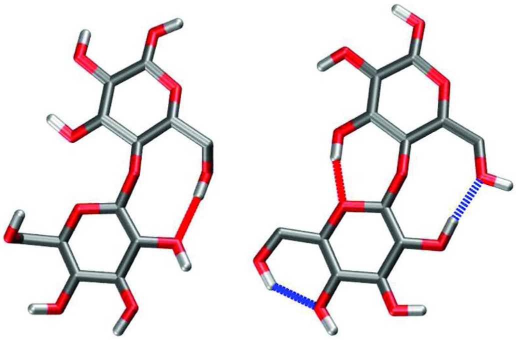


Figure 8. Lowest energy conformations for cellobiose at the φ,ψ point (30°,-30°) for the CSFF (left) and GLYCAM06 (right) forcefields. (see color insert)

Table 1. Energy in kcal/mol of cellobiose at the φ,ψ point (30°,-30°) with CSFF and GLYCAM06. Each entry in the table has been rescaled to the minimum at this φ,ψ by subtracting the amount to the right of the force field name. Columns have the same hydroxymethyl rotameric state at the reducing end, rows at the non-reducing end

CSFF			*7.73*
GT	GG	TG	
4.07	1.17	4.30	GT
1.60	0.00	2.36	GG
4.39	2.38	4.98	TG
GLYCAM06			*4.94*
GT	GG	TG	
3.18	1.97	0.91	GT
3.33	2.13	1.07	GG
1.92	1.16	0.00	TG

The (φ,ψ) points on the two-fold helical axis are a region of interest to the cellulose structure. On these maps, a point near the conformation determined for crystalline cellulose is (30°, -30°). This point is not a local energy minimum for either CSFF or GLYCAM06, which indicates that crystal structures with glycosidic linkages in this region may favor these conformations due to crystal packing interactions. Figure 8 shows the lowest energy conformers, and Table 1 contains the relative energy for the nine combinations of hydroxymethyl rotamers at the point (30°, -30°) for CSFF and GLYCAM06. The energies have been normalized for each force field relative to the lowest energy at this (φ,ψ) point, which is shown at the top of the table. It can be seen that the lowest energy conformer for GLYCAM06 has both hydroxymethyl groups TG, which is interesting because

Table 2. Unit cell parameters for cellobiose crystal

cellobiose	*Expt. crystal*	*CSFF*	*GLYCAM06*
A	10.972	10.939	10.992
B	13.048	13.478	13.469
C	5.091	5.211	4.985
gamma	90.83	90.25	92.12
volume	728.76	768.28	737.53

this is the same conformation found in cellulose I, but in relatively few other crystal structures. This structure was also found to be low in energy in the absence of water using density functional theory.

Cellobiose Crystal

As a tool to examine the correctness of force field parameters, adiabatic vacuum maps are of limited use because the information is not directly accessible to experiment. High-level *ab initio* calculations have been used to study several cellobiose conformations, but adding a small number of water molecules changes which conformations are lowest in energy; and these calculations are currently too computationally expensive to do a comprehensive mapping study. Density functional theory has been used to map a sparse subset of cellobiose (φ,ψ) space, but it is unclear if the results would be unchanged with a more complete treatment of electronic structure (*29*). Combined solid-phase crystallographic and spectroscopic data from small molecules are best suited for comparing conformational properties and are available for cellobiose (*30*, *31*).

Cellobiose crystallizes into a monoclinic unit cell with $P2_1$ symmetry that contains two molecules. A periodic crystal of cellobiose consisting of 3x3x8 primitive unit cells was constructed using the CSFF and GLYCAM06 forcefields. The average unit cell parameters after 1 ns of dynamics at 300 K are in Table 2. The unit cell volume in the solid state is most closely related to the van der Waals (VDW) parameters. These parameters are generally chosen to reproduce solution properties, and it is neither practical nor desirable to create new parameters for solid-state simulations. The VDW parameters for CSFF were chosen from analogous atom types in the CHARMM protein force field, so it is not entirely surprising that the unit cell volume of cellobiose is not ideal.

The average bond lengths and internal coordinates from the simulations are given in Table 3. Both the CSFF and GLYCAM06 force fields have average φ values that differ from the crystal structure values by ~7°. Solid state nuclear magnetic resonance (NMR) studies of crystalline cellobiose show that there are few transitions in hydroxyl and hydroxymethyl conformation at 300 K, and both CSFF and GLYCAM06 reproduce this observation. Average bond lengths for GLYCAM06 are generally 0.01-0.02 Å longer than bond lengths using CSFF, but for GLYCAM the VDW radii are smaller by a similar amount (see Chapter 7 on hydrogen bonding). These small differences in VDW and bond length

Table 3. Internal coordinates for cellobiose crystal

	Expt. crystal	*CSFF*	*GLYCAM06*
Φ	45.67	39.02	38.54
ψ	-17.05	-16.22	-18.83
glycosidic angle	116.09	116.97	116.62
sum of ring bonds	8.97	8.90	9.09
ave C-OH	1.41	1.41	1.43
chi RE	70.52	59.60	66.52
chi NR	48.70	39.85	53.55

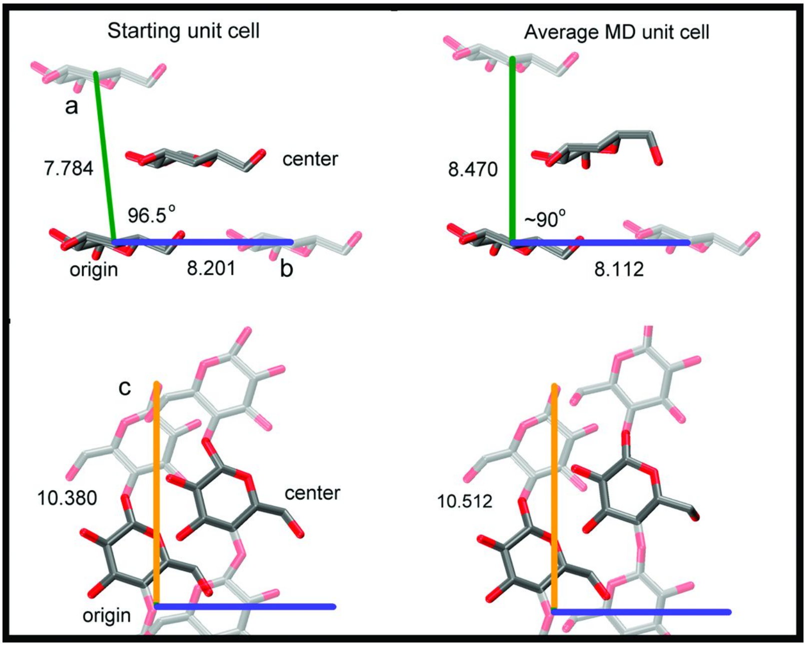


Figure 9. Left: the cellulose Iβ crystal unit cell determined by fiber diffraction; right: the trajectory-averaged unit cell for the simulation of the diagonal crystal. Hydrogen atoms are omitted for clarity and positions obtained by symmetry operations are transparent. (see color insert)

parameters, along with different partial atomic charges, angle bending constants, and rotational barriers makes isolating the cause of different conformational preferences in molecules the size of cellobiose difficult. The challenge is greater for larger molecules such as cellulose, and harder still for aggregates or crystals of polymers.

Crystal Structure Reorganization

The discussion following this paragraph describes the results of the 1-ns cellulose Iβ simulations using the CSFF force field published in 2006 (*32*). Yui and coworkers submitted results of very similar simulations of hydrated microcrystalline cellulose Iβ with GLYCAM04 at the same time as ours, but their paper was published later (*33*). The different parameters used in the two papers gave very similar results in terms of overall structure shape, but GLYCAM04 more closely reproduced the expected hydroxymethyl and hydrogen bond conformations. To enable a more thorough comparison of the structures produced with the two parameter sets, 10-ns simulations of the diagonal crystal were run with both CSFF and GLYCAM06. To explore the effect of chain length and diameter on the fibril structure, hydrated DP 40 diagonal crystals with either 36 or 16 chains were built and run for 3 ns with both CSFF and GLYCAM06. These simulations will be presented following the results of the previously published work.

One ns Simulation with CSFF

The starting structure for the cellulose crystal was built up from the hypothetical crystal conformation deduced from fiber diffraction studies. However, during the course of the simulations, several structural fluctuations and changes occurred. In simulating the diagonal crystal, the rms difference between the instantaneous structure and the starting crystal structure, averaged over the simulation, was 1.46 Å; while for the square crystal simulation, the rms difference was 1.72 Å. Over the length of the simulations, the average unit cell dimensions shifted away from those reported in the diffraction study. These dimensions varied with position in the crystal, relative to the surfaces and the chain termini. Average values were calculated for the three cellobiose units in the middle of the chains, for the three middle chains of the middle three layers of the diagonal crystal (that is, the central core of the crystal). The results, averaged over these cellobiose units, are summarized in Figure 9 and compared to the crystallographic unit cell. As can be seen, in the simulation the crystal underwent an expansion that saw the value of the lattice constant a increase from 7.784 to 8.470 Å, while the b value decreased slightly from 8.201 to 8.112 Å. The c value expanded significantly, from 10.380 to 10.512 Å. In addition, the γ angle decreased from 96.5° to almost orthogonal, γ ~90°. Reported unit cell dimensions for cellulose Iβ vary depending on the source material and expand anisotropically upon heating. However, the unit cell a-axis (corresponding to the distance between hydrogen bonded sheets) in this simulation is too different to be considered a good fit to the experimental measurements for cellulose Iβ, as is the "monoclinic" angle near 90°.

These unit cell parameters do compare favorably with those determined from the crystallographic equatorial d-spacings reported for cellulose IV_I (a = 8.02 Å, b = 8.43 Å, γ = 90°). While it was previously believed that cellulose IV_I was a unique allomorph, the experimental data currently available from diffraction, NMR, and FTIR suggests that cellulose IV_I can also be regarded as disordered cellulose Iβ.

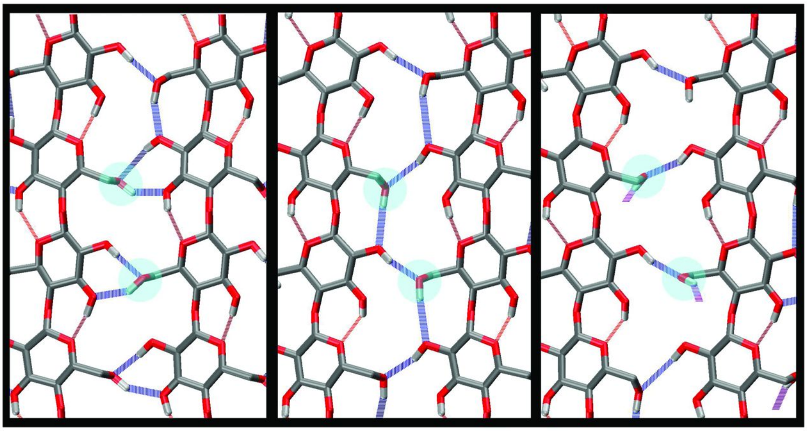


Figure 10. Single frames from the center chain layers illustrating three different hydrogen bond patterns. Left: similar to the predominant pattern from the crystal structure, but the rotation to GG makes the HO2-O6 hydrogen bond across the glycosidic linkage impossible; center: hydrogen bond pattern very similar to the less-occupied pattern from the crystal structure; right: hydrogen bonds from HO6 in a center chain to O2 in an origin layer chain, which is not shown for clarity (see Figure 12). These three patterns interchange very rapidly and are not mutually exclusive, meaning concerted motion of hydroxyl hydrogen atoms is not required to change the local hydrogen bond pattern. (see color insert)

Another extremely significant change in the crystal structure that occurred during the simulations is that many of the C6 primary alcohol groups underwent rotational transitions away from the conformation reported for the diffraction structure. In the Iβ diffraction structure, all primary alcohol groups are in the TG conformation. In this conformation, the exocyclic hydroxyl group can make hydrogen bonds along the chain or to adjacent chains in the same layer, but no hydrogen bonds between layers. For those crystal layers made up of the origin chains, there was little structure change in the molecular dynamic (MD) simulation from that of the diffraction structure; and the hydrogen bonding pattern remained the same. This result is remarkably similar to the reported experimental hydrogen bond network in origin chains, where the O2 hydroxyl group was refined to just one of the two possible hydrogen bond positions.

However, in the MD simulations, in every other layer in the interior of the crystal, made up of the center chains in the diffraction structure, this primary alcohol group rotated from the starting TG conformation to the GG position. The transitions occurred randomly, and were not unidirectional or permanent; rare transitions back or to the GT conformation also occurred. On average, these transitions brought all of the residues in the center chain layers into the GG conformation. In contrast, the layers made up of origin chains remained in their original TG conformations.

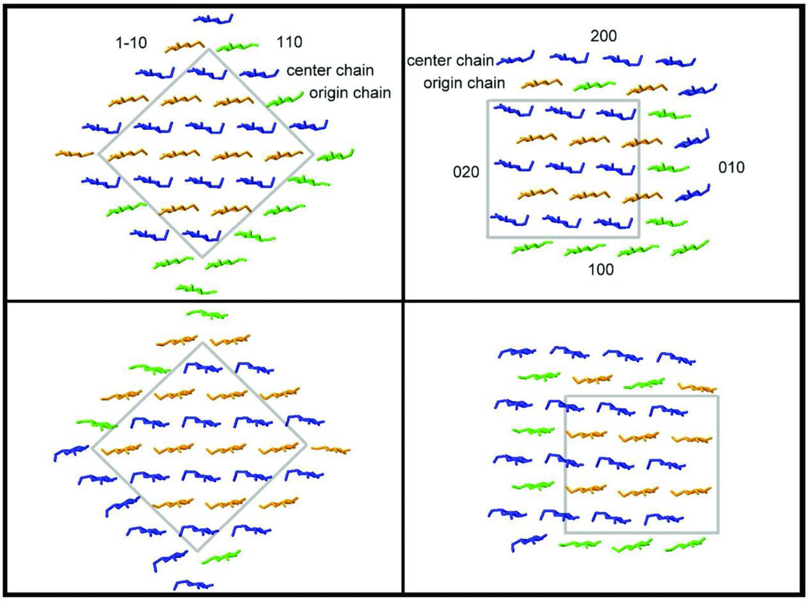


Figure 11. End views of the final structures of the two crystallites, colored to indicate the dominant primary alcohol conformations for the sugar rings in each chain. Surface molecules generally explored two rotameric states with the color here chosen to represent the more predominant conformation. The top panels show glucosyl residues with primary alcohols pointing to the right, and the bottom panels show the next glucosyl residue along the chain which is flipped by ~180°. The regions enclosed by the grey boxes indicate that these primary alcohol conformations are nearly constant, and represent the interior of the fibril. (see color insert)

In this GG conformation, three rapidly interchanging hydrogen bond patterns were possible, as shown in Figure 10. One of these patterns allowed hydrogen bonding between layers, which was not possible when the hydroxymethyl groups were in the TG conformation. On the surfaces, where the sugar monomers were in direct contact with water, the hydrogen bonds to the freely diffusing water molecules helped introduce considerable disorder into these primary alcohol conformations and promoted frequent transitions, but the interior portions of the crystal developed a clear alternating pattern of hydroxymethyl conformations between the center and origin layers. Primary alcohol groups in surface chains alternate between facing towards the interior and facing the solvent; and the conformation of these surface groups corresponds to the local environment. Figure 11 shows a trajectory-averaged picture of hydroxymethyl conformations according to location within the fibril. The chains are colored according to the key at the bottom of the figure, with the color determined by the predominant conformation. The top two panels show hydroxymethyl groups pointing to the right, while the bottom two panels show hydroxymethyl groups pointing to the left. Interior monomer units showed only rare transitions away from the GG

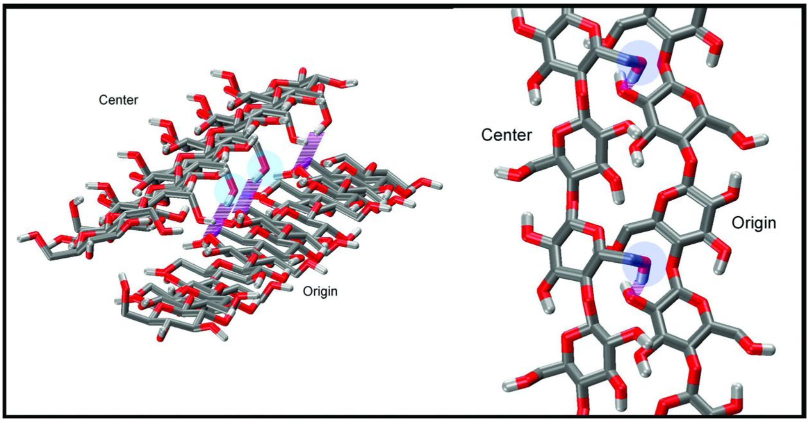


Figure 12. Hydrogen bonds between layers in the diagonal crystal. These hydrogen bonds are possible because of GG conformations in the center layer. (see color insert)

conformer in center chains and no transitions from TG in origin chains. Both inward- and outward-facing primary alcohol groups in surface chains showed much more diversity, sometimes exchanging between two conformations four or five times during the 1-ns simulation. As a result, there are some differences between the (110) and (1-10) surfaces on opposite sides of the diagonal crystal. Heiner and coworkers also found primary alcohol conformational changes in the layers adjacent to water (*34–37*). Residues in their simulations were found to rotate to the GT conformation in the surface layers as was found here. Several NMR studies have determined that the conformations of surface cellulose chains are different from the interior and, as in the present simulation, contain both GG and GT rotamers. This analysis has ignored the conformation of the four glucose units at each end of the chains, as the ends show greater disorder and swelling than does the interior. Both ends of the fibrils swelled enough to allow some water molecules to sit between layers, but water did not penetrate significantly into the interior.

In the GG conformation, the primary alcohol groups are essentially perpendicular to the average planes of the sugar rings and, as a result, are pointing up and down toward the origin chains of the layers above and below. In this conformation, the exocyclic groups can make good O6-O2 hydrogen bonds between layers, as is illustrated in Figure 12. Because under normal conditions cellulose apparently exhibits no tendency for layers to slip relative to one another experimentally, the existence of such stabilizing hydrogen bonds may not seem so implausible. However, in this conformation, steric clashes between these center-chain primary alcohol groups and the origin layers above and below force the center chains to tilt significantly with respect to the plane of their own layer (illustrated in Figure 12).

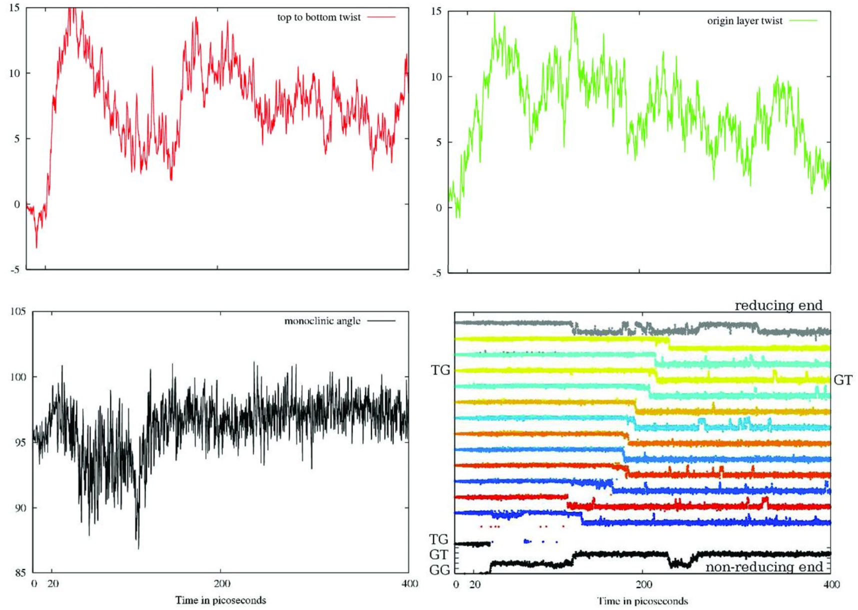


Figure 13. Time series from the heating and equilibration period of the 10-ns Iβ simulation with CSFF. Top to bottom twist (top left in red), middle plane twist (top right in green), monoclinic angle (bottom left in black), and hydroxymethyl conformation along two origin chains (bottom right). (see color insert)

Ten ns Simulations of the Diagonal Crystal and DP 40 Crystals

During the repeat of the previous CSFF diagonal simulation that was extended to 10 ns, several unexpected changes in structure occurred. The hydroxymethyl groups in the origin layers rotated to GT in the interior of the crystal, with the transitions progressing from the non-reducing end. A regular three-dimensional hydrogen bond pattern developed, involving both center and origin chains as proton donors. The hydrogen bond pattern in the center chains settled almost exclusively into the third pattern described previously, where O6 in a center chain donates a proton to O2 in an origin chain in the next layer. Origin chain O6 accepts a proton from O2 in a neighboring origin chain, and donates a proton to the glycosidic oxygen in a center chain. The non-reducing end hydroxymethyl is not part of a cooperative hydrogen bond network and is free to rotate to the lower-energy GT conformation. Each transition to GT forms a hydrogen bond between two origin chains from O6 to O2, destabilizing the previously present (TG) O6 to O2 hydrogen bond along a single chain. This destabilization causes transitions in alternating hydroxymethyl groups in neighboring chains, and it also occurs in concert with the chain tilt.

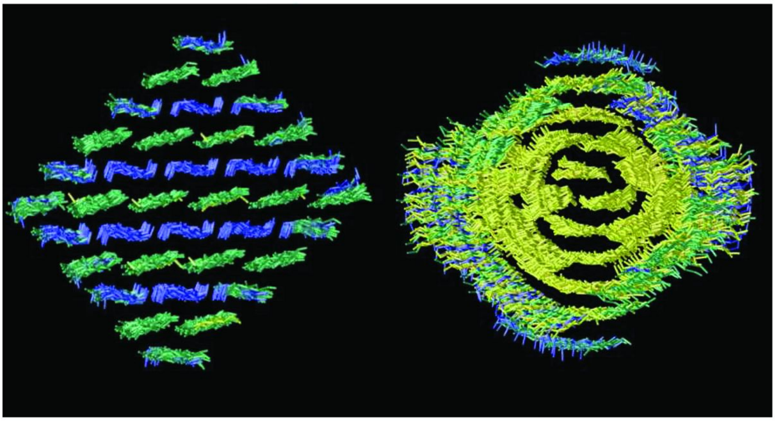


Figure 14. End view of Iβ crystals colored as in Figure 11 for CSFF and GLYCAM06 after 10 ns. (see color insert)

The average unit cell over the 10-ns, DP 14 simulation with CSFF has $P2_1$ symmetry with two chains per unit cell, each containing a single anhydroglucose monomer as the asymmetric unit. The unit cell is also monoclinic, with an angle of 96.97°. In this unit cell, the origin chains are tilted relative to each other, in a direction opposite to the tilted center chains. This packing with tilting in opposite directions in alternate layers is similar to the packing of mannan I and high-temperature chitosan. Because of this tilt, the origin chains no longer form an isolated sheet but interact with the layers above and below. Figure 13 (top left) shows a time series of the twist of a single origin layer during the heating and equilibration period. Figure 13 (top right) shows a time series of the angle between the top and bottom single chain “layers” in the crystal. Figure 13 (bottom left) shows a time series of the monoclinic angle measured at the midpoint of the crystal’s interior. The monoclinic angle starts at the crystallographic 96.5°, and after a brief deviation towards 100° during the heating period, transitions rapidly to the previously reported orthogonal unit cell, followed by a slower return to the monoclinic unit cell over the equilibration period. The time evolution of the sheet twist is similar to the twist between the top and bottom chains. Figure 13 (bottom right) shows a time series of hydroxymethyl orientation along one pair of origin chains, with the non-reducing end at the bottom and the reducing end on top. The hydroxymethyl groups involved in this transition are on two chains, distinguished by the alternating red and blue colors in this graph. The transitions to GT progress in succession from the non-reducing end to the reducing end. The transition to the GT conformer forms a hydrogen bond from O6 to O2 in a neighboring chain, disrupting the previous hydrogen bond pattern across the linkage involving a TG conformation on the reducing end. This disruption allows the TG hydroxymethyl group to rotate, forming an intermolecular hydrogen bond back to the first chain, and causing a similar disruption and propagating the transition to GT.

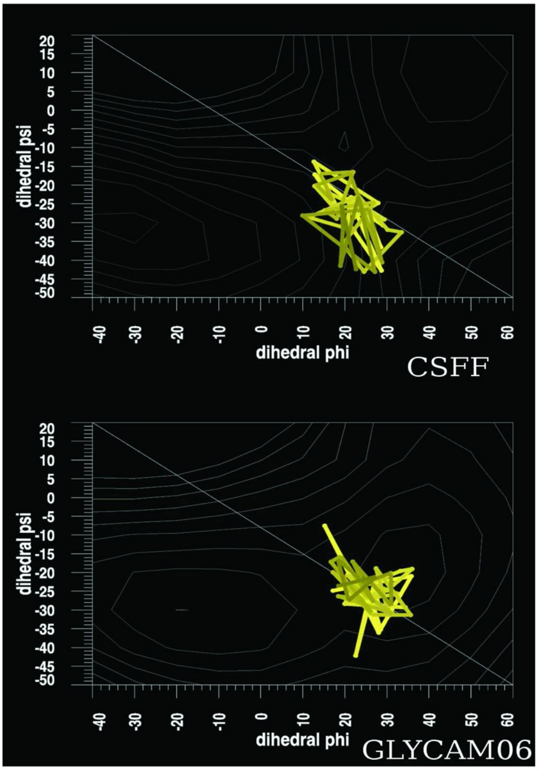


Figure 15. φ,ψ conformation of all glycosidic linkages along a single origin chain from the middle of the DP 40, 36-chain crystals. (see color insert)

It is difficult to determine whether it is the chain tilting, the primary alcohol conformation changes, or a combination of both that initially causes the untwisting to occur; but it is clear that an extensive three-dimensional hydrogen bond pattern rules out the possibility of twist. The only difference in simulation protocol between the initial 1-ns simulations with CSFF and the 10-ns simulation reported here is in the treatment of electrostatic interactions. In this longer simulation, the particle mesh Ewald method is used, which is equivalent to using an infinite cutoff, whereas the earler simulations ignored interactions more than 13 Å away. It is expected that either the change in treatment of electrostatics or the longer simulation time made these transitions more likely to happen, but with the CSFF force field this pattern is lower in energy than the previously reported pattern.

The GLYCAM06 force field reproduced the hydroxymethyl conformation and hydrogen bond pattern proposed from diffraction studies, but unit cell dimensions do not match exactly. The fiber axis (**c**) is 10.79 Å, much longer than the experimental value of 10.38 Å. This dimension includes mostly through-bond interactions, and the extra length comes from a combination of longer bonds and wider angles spread throughout the anhydrocellobiose repeat units. Figure 14 shows the DP 40, 36-chain fibrils of CSFF and GLYCAM06 with residues colored by hydroxymethyl conformation after 3 ns of dynamics. The interior of the GLYCAM06 crystal remains as a stable unchanging unit, with surface chains more free to move. CSFF shows more dynamic behavior in the interior, with occasional transitions from GT back to TG in the origin layers. The transitions to the GG and GT conformations in the CSFF simulations are not surprising because when using this force field, the TG conformation is the highest-energy conformation for the isolated glucose residue in solution, which agrees with NMR experiments. The TG conformation is high energy in single cellulose chains and in the glucose crystal, so the transitions are the result of the cellulose crystal gradually annealing to lower energy states.

Figure 15 shows the φ and ψ conformation of all glycosidic linkages along a single origin chain from the middle of the DP 40, 36-chain crystal with CSFF and GLYCAM06. The graph is superimposed on the adiabatic potential energy surface for cellobiose in vacuum for each parameter set. Lines connect neighboring linkages, with lighter colors towards the non-reducing end, and darker colors towards the reducing end. Lines parallel to the two-fold helical axis (top left to bottom right diagonal) connect linkages with different phi and psi values but have the same handedness of twist. Lines along the opposite diagonal (bottom left to top right) connect linkages that alternate left- and right-handed twist. Even for CSFF that has a saddle point in this region, most of the conformations are near the two-fold axis, with the chain ends less constrained. This is the result of hydrogen bonding between chains, which frustrates the preference of individual linkages in isolated chains to take on conformations away from the two-fold axis.

Figures 16 and 17 show similar glycosidic linkage conformation plots for each of the 36 chains in the DP 40 crystals with the CSFF and GLYCAM06 force fields, respectively. The range of each plot is identical to those in Figure 15. There is a wider range of (φ,ψ) space explored by the surface chains, and chain ends also tend to take on conformations away from the two-fold axis. It is not immediately obvious from these plots which crystal is twisted, but over the course of the simulation, each linkage alternates between left- and right-handed twist, with chains in the core of the crystal restrained by packing to remain close to a two-fold helical conformation.

The unit cell of cellulose IV_I, measured in primary cell walls or after chemical treatments, is similar to the unit cell observed with CSFF in the previously published paper (*11*, *32*, *38*). Gardiner and Sarko published coordinates for cellulose IV_I, but the data was very limited. This structure is identical to Iβ in terms of how the hydrogen-bonded sheets are stacked; but the unit cell is orthogonal and the distance between layers is larger. A crystal very similar to the diagonal crystal was constructed with these cellulose IV_I coordinates and was run with CSFF and GLYCAM06 for 5 ns. The structure for CSFF using cellulose IV_I as the starting coordinates was indistinguishable from the structure found when starting from Iβ coordinates, as may be expected because of the structural changes described previously. However, GLYCAM06, when started in the cellulose IV_I structure, does not behave as when the starting coordinates are cellulose Iβ. Remarkably, the behavior is similar to CSFF, where center chains are GG and origin chains are GT. There are still segments where the original TG conformation remains in both the center and origin layers, but the spontaneous conformation change with parameters that can reproduce the Iβ structure suggests

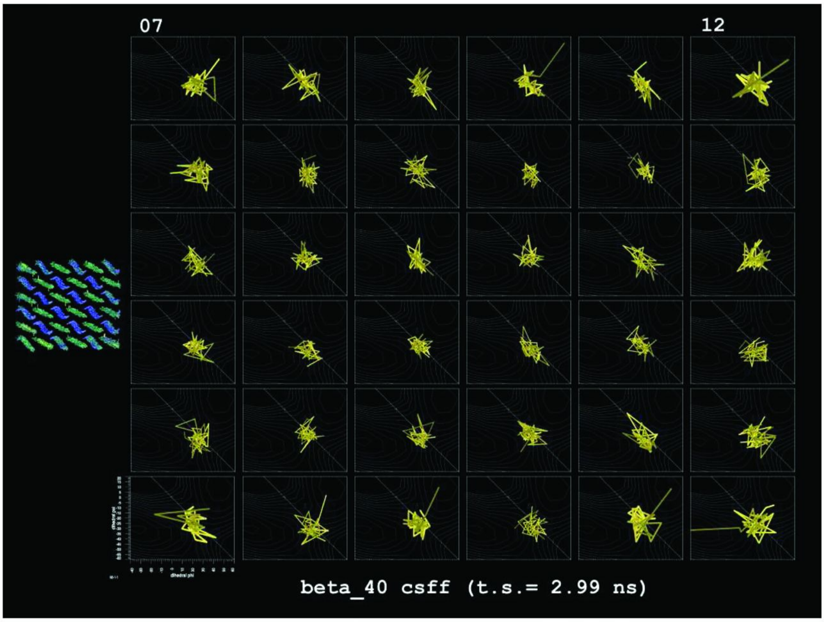


Figure 16. φ and ψ conformation of all glycosidic linkages along all chains from the DP 40, 36-chain crystal with CSFF. (see color insert)

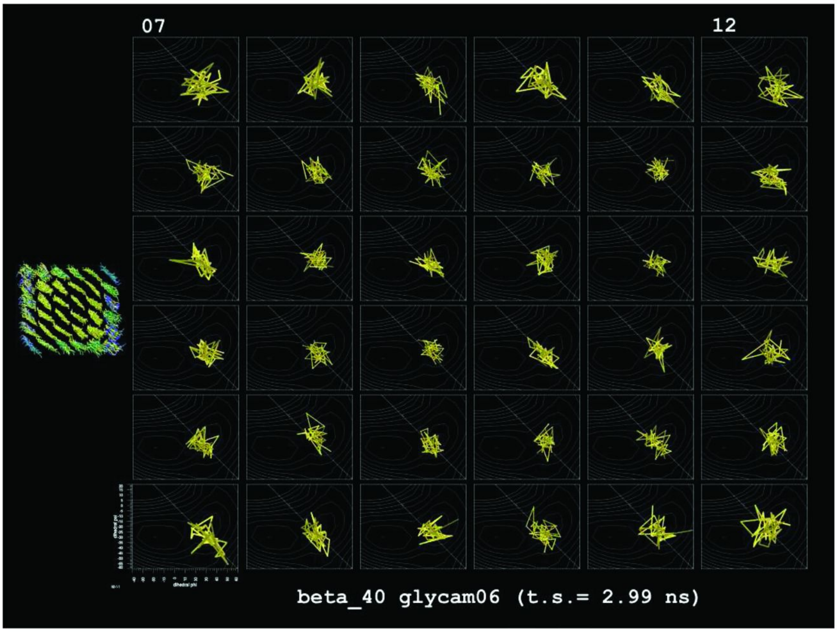


Figure 17. φ and ψ conformation of all glycosidic linkages along all chains from the DP 40, 36-chain crystal with GLYCAM06. (see color insert)

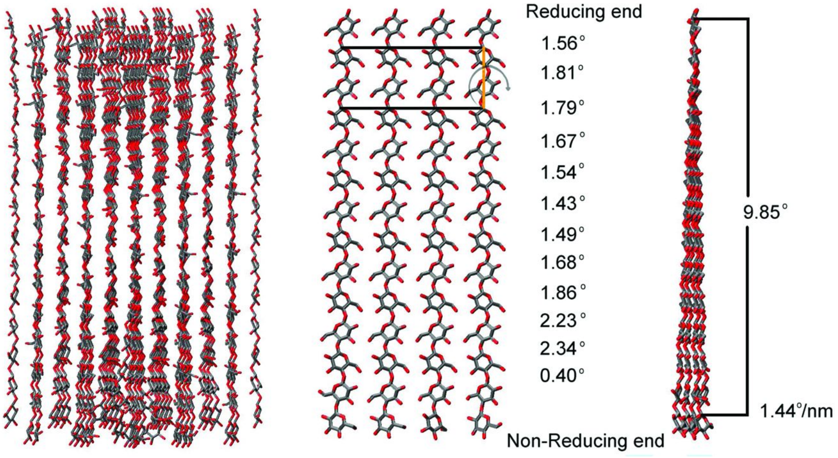


Figure 18. The trajectory average after 1 ns of the CSFF diagonal crystal, with a section of the crystal's central plane seen from above and the side, illustrating the twist that developed during the simulation. The numbers give the twist angle in degrees for each dihedral angle defined by four C1 atoms, as at the ends of the black lines shown on the left. (see color insert)

the CSFF structure is not caused by significant problems in parameters. Rather, the convergent behavior of two unrelated parameter sets, when started from the poorly resolved cellulose IV$_I$ coordinates, suggests this may be a pattern that exists under certain conditions. The energy of this structure with GLYCAM06 is higher than the energy of the structure when started with the Iβ coordinates, but it seems this is a local minimum.

The most recent cellulose IV$_I$ study (*14*) was conducted on highly crystalline materials, which enabled the collection of many spots in the diffraction patterns at each step of the transformation process from Iβ to cellulose III$_I$ to cellulose IV$_I$. Observed unit-cell parameters similar to the ones proposed here have only been collected from less-crystalline, small-diameter materials. It is most likely that cellulose IV$_I$ is a high-energy, high-temperature form of cellulose Iβ that becomes kinetically trapped in an intermediate state. The unusual cooperative hydrogen bond patterns present in Iβ may not be well reproduced with fixed-charge force fields, making the conversion to a "high temperature" structure possible without heating.

The conformation of the hydroxymethyl groups in origin layers is GT, and combined with the tilting of the chains, this structure is almost identical to a single layer of cellulose II. Raman spectroscopy indicates that cellulose IV$_I$ is composed of equal parts of cellulose I and cellulose II (*13*), which is somewhat similar to the structures produced with the CSFF force field.

Hydrogen Bonding and Twist in Cellulose Fibrils

Probably the most significant change that occurred in the cellulose during the simulations is the development of twist. Figure 18 illustrates this twist for the

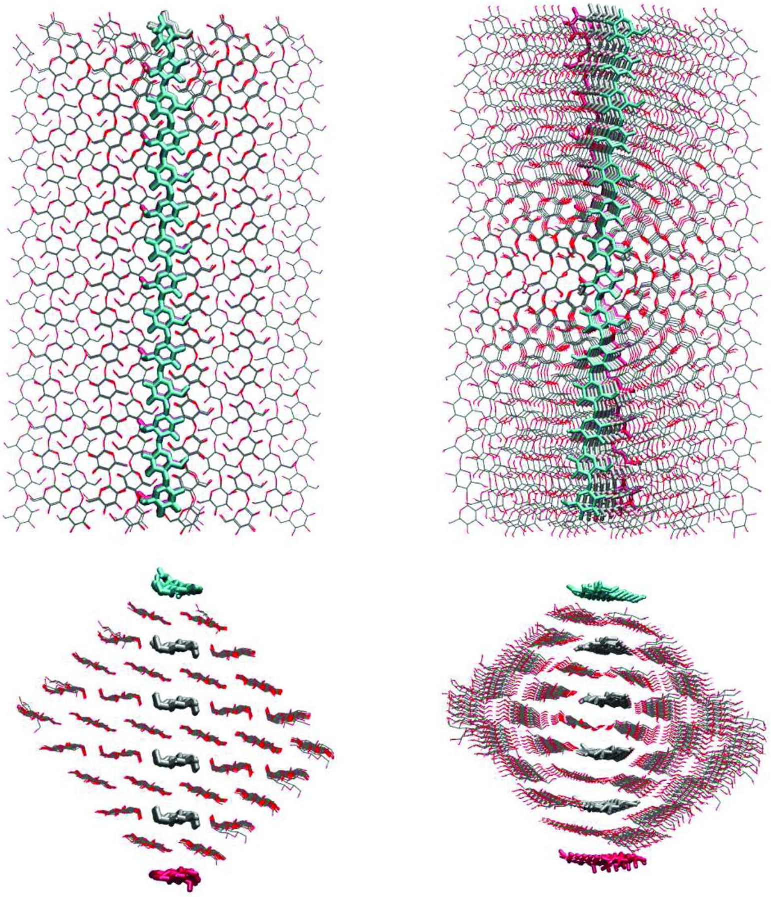


Figure 19. The trajectory average of the diagonal crystal after 10 ns with the CSFF (left) and GLYCAM06 (right) force fields. There is almost no twist in the CSFF structure, and a regular twist in the GLYCAM06 structure. (see color insert)

Iβ diagonal crystal with CSFF after 1 ns, with the middle hydrogen-bonded sheet shown in detail. In this figure, the average twist angle for each anhydrocellobiose repeat unit is shown. These angles are defined as the dihedral angle for the four C1 carbon atoms illustrated as joined by the dark lines in the figure. Although this angle varies considerably near the non-reducing end, apparently caused by edge effects, the twist in the middle of the chain is fairly constant at around 1.4-1.7° per linkage, with an overall twist for this short oligosaccharide segment of almost 9.9°, calculated from the first and last rows (which includes considerable irregularity because of the highly frayed structure of the non-reducing ends).

As mentioned, in the 10-ns simulation of the Iβ diagonal simulation with CSFF, the twist went away as the simulation progressed. This untwisting is a direct result of the development of the regular three-dimensional hydrogen bond pattern. For inelastic twisted sheets to pack as close as possible to each other,

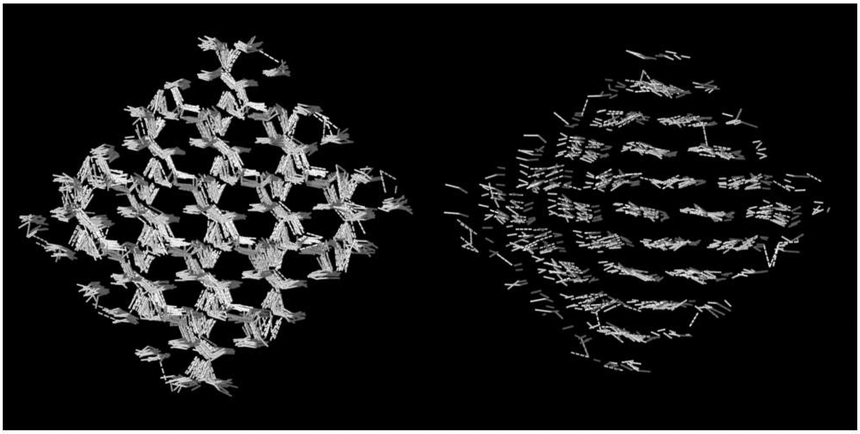


Figure 20. Hydrogen bonds in the interior of the DP 40, 36-chain crystal, CSFF on left, GLYCAM06 on right. The GLYCAM06 picture shows only the central 10 residues to make the layer structure visible, due to twist obscuring this view when all residues are shown.

the long axis of the sheets must be rotated relative to each other. That is to say, if the chains in a sheet are aligned at every point with chains in neighboring layers, the sheets cannot be twisted. This notion is similar to the rotation of the director in chiral nematic liquid crystals. Figure 19 shows the average structure of the 10-ns simulation for both CSFF and GLYCAM06 from the top and from the ends. One chain in the middle of each center layer is colored to highlight the rotation of the molecular axis between layers in the twisted structure. On the left (CSFF), the highlighted chains align almost exactly along the length of the fibril. However, in the twisted GLYCAM06 structure on the right, the chain axis in each layer is rotated relative to the layer above it. Slip between layers is possible in GLYCAM06 only because the hydrogen-bond pattern is mostly two dimensional, with very few transient hydrogen bonds between layers. Figure 20 shows just the hydrogen bonds present in a cross section of the DP 40, 36-chain Iβ crystals with CSFF and GLYCAM06 after 3 ns of dynamics. There is a drastic difference, with a regular three-dimensional hydrogen bond lattice along the entire length of the CSFF crystal and only brief transitions away from the pattern of two-dimensional, hydrogen-bonded sheets in the GLYCAM06 crystal.

Scatter plots of all hydrogen bonds present in the cellulose Iβ DP 40, 36-chain simulations are shown in Figure 21. Both plots show the same information, just exchanging which dataset is displayed on top of the other. Hydrogen bonds are usually defined as having the electronegative donor and acceptor atoms less than 3.4 Å apart and with a D-H—A angle of greater than 120°. To show the many interactions in these crystals that do not quite meet the definition of hydrogen bonds, D-H—A angles greater than 100° and distances up to 4 Å are shown. The internal structure of these crystals differ greatly, but the overall distribution of hydrogen bonds is similar. The distribution of GLYCAM06 hydrogen bonds shifts slightly towards shorter distances and higher angles.

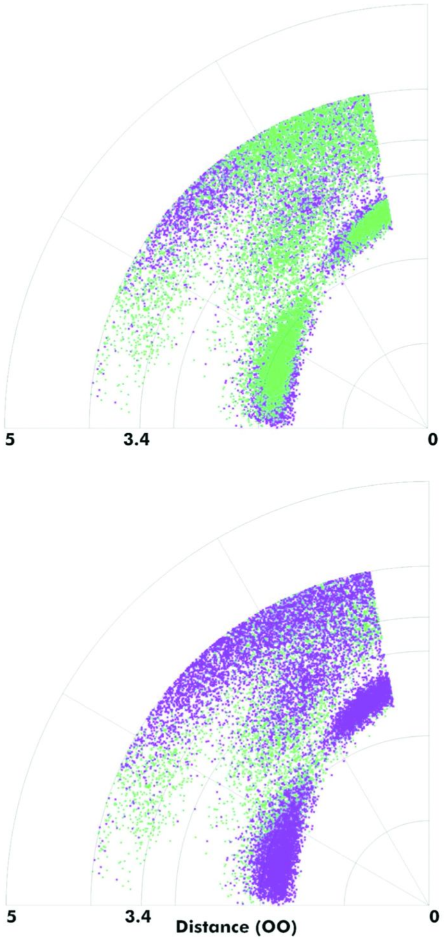


Figure 21. Scatter plot of the hydrogen bonds in Figure 20, with CSFF in green and GLYCAM06 in purple. These structures are very different, but there is no obvious difference in hydrogen-bond geometry. (see color insert)

A hydrogen bond is a short-range, angularly dependent interaction between a small electronegative donor atom (such as oxygen, nitrogen, or fluorine) that has covalently a bonded hydrogen atom and an electronegative acceptor atom. This interaction is mostly polar, but there is a partial covalent character that is strongest when the donor-hydrogen—acceptor angle is nearly linear (D-H—A = 180°). The hydrogen bonds in the proposed structures for cellulose I are exclusively in two-dimensional layers, forming hydrogen bonded sheets that do not have strong short-range interactions with neighboring layers. The attractive interaction energy of a hydrogen bond is around 6 kcal/mol, about 10 times the average energy available from thermal motions at 300 K.

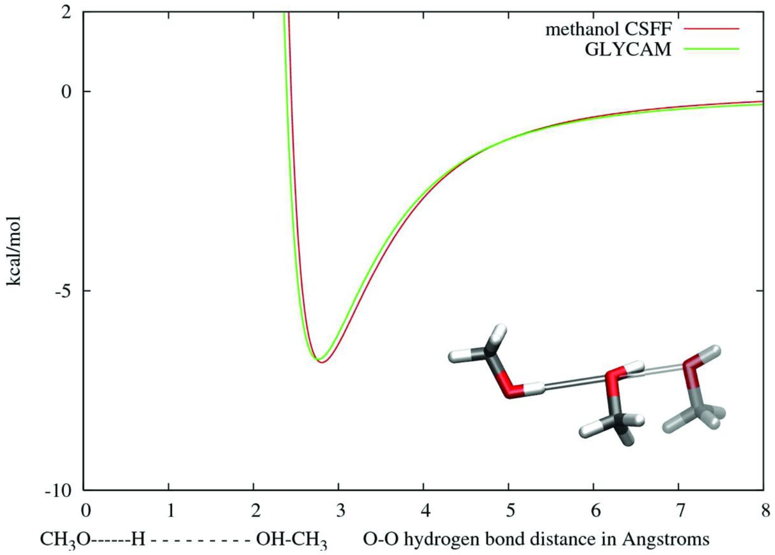


Figure 22. Interaction energy of a methanol dimer with a linear hydrogen bond as a function of oxygen-oxygen distance. (see color insert)

The hydroxyl group in methanol can be used as a simple model for hydroxyl groups in sugars. Figure 22 shows the interaction energy as a function of distance for a methanol dimer with a linear hydrogen bond, using CSFF and GLYCAM parameters. The curves are similar in the shape and magnitude of minimum energy, but CSFF is shifted to longer distances by approximately 0.02 Å. Similar potential energy surface calculations have been performed with quantum mechanical methods, but the results were reported as a function of carbon-carbon distance (*39*), as opposed to the oxygen-oxygen distance used here. A different study using *ab initio* methods found the minimum energy methanol oxygen separation distance to be 2.846 Å (*40*). The minimum energy distance for a methanol hydrogen bond pair with GLYCAM is 2.79 Å (-6.33 kcal/mol), and 2.81 Å (-6.74 kcal/mol) for CSFF.

Figure 23 shows the interaction energy for a hydrogen bonded methanol dimer as a function of O-O distance up to 4.4 Å, and D-H—A angles greater than 120°. Two contour lines in 1 kcal increments above the lowest energy are drawn on each surface. The 2 kcal/mol isoline closely follows the 3.4 Å distance. The hydrogen bonds distances and angles reported for cellulose Iα (purple) and Iβ (white) are plotted on each surface. Many hydrogen bond distances for Iα are outside the traditional definition of hydrogen bonds, either longer than 3.4 Å or slightly less than 120° D-H—A angles. There are also several hydrogen bonds that are very short and would have a repulsive interaction energy with these parameters. This may partially explain why Iα is less stable than Iβ.

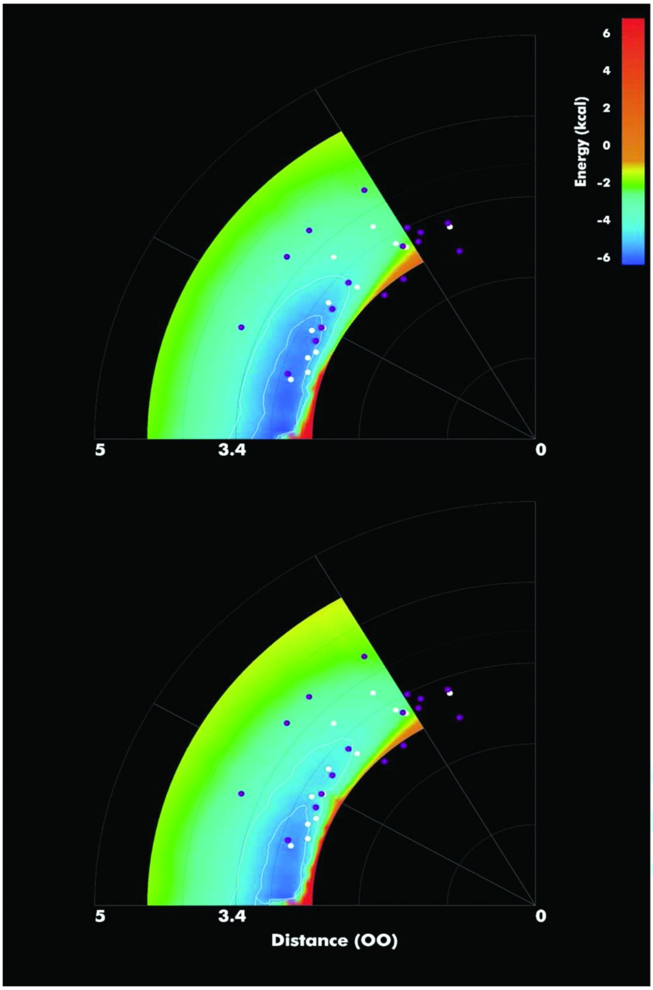


Figure 23. Interaction energy of a methanol dimer hydrogen bond as a function of oxygen-oxygen distance and angle for CSFF (top) and GLYCAM06 (bottom). Proposed hydrogen bonds from Iα and Iβ are shown in purple and white, respectively. (see color insert)

The two-dimensional hydrogen bond patterns in cellulose I give the layers a ribbon-like character. The short-range, directional nature of the hydrogen bonds prevents the chains ends from splaying or fanning out from each other and also prevents the chains from separating at elevated temperatures. The stacking of the ribbons prevents the layers from rolling into a tube shape, and the chains are not very elastic in the direction of the molecular axis. These constraints on ribbon shape allows a helicoidal twist to develop; however, the amount of twist is limited by the elasticity of the chains. Ideal helicoidal surfaces can be described by the path of an infinite line rotating and translating along an axis at a constant rate. In this case, the line corresponds to the width of the hydrogen-bonded sheet, and the

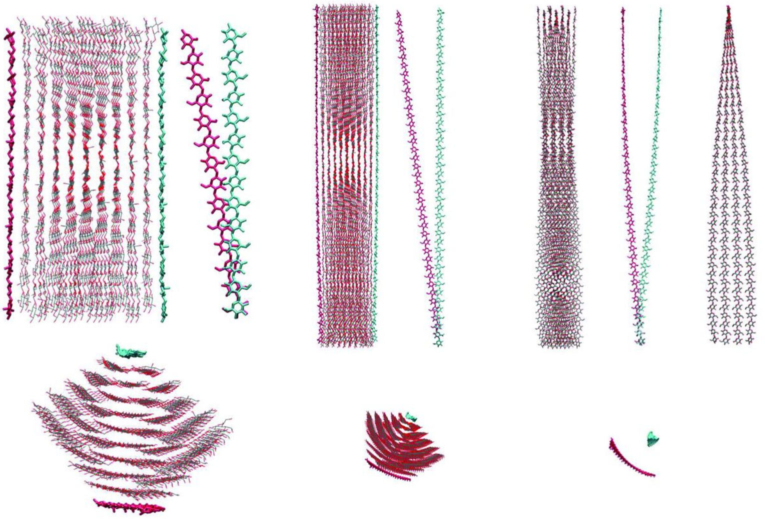


Figure 24. Twist of several cellulose fibrils from the GLYCAM06 simulations; from left to right DP 14 and 36 chains, DP 40 and 36 chains, DP 40 and 16 chains. The magnitude and character of twist depends on the width and length of the fibril. Longer and wider fibrils are less twisted than thinner or shorter fibrils. On the extreme right is the widest layer from the 16-chain fibril, showing the sheet edge at top and the full fibril width at bottom. For the 16-chain fibril, all three views are from the same direction. (see color insert)

axis of rotation and translation is parallel to the chains in the center of the fibril. Most polymer ribbons have the molecular axis perpendicular to the twist axis, but the biosynthesis of cellulose extrudes the molecules as parallel chains.

Because the cellulose chains are not able to stretch, the amount of helicoidal twist is limited. The length-to-width ratio also affects the twist of hydrogen-bonded cellulose sheets. The elastic stretching energy per unit area divided by the bending energy per unit area gives a dimensionless parameter that depends on the fourth power of the ribbon width (*41*). Three microcrystals of cellulose Iβ simulated using the GLYCAM06 force field are shown in Figure 24 from the side, from the chain ends, and with just the top and bottom chains. The effect of microfibrl length and width are apparent, where the DP 14, 36-chain fibril has a twist angle of 2.16°/nm, the DP 40, 36-chain fibril 2.09°/nm, and the DP 40, 16-chain fibril 4.51°/nm.

The longer fibril with 36 chains is less twisted per unit length than the shorter fibril. The structure of the 16-chain fibril is drastically different, with the individual top and bottom chains showing much more curvature than in the 36-chain fibril. Measuring the Gaussian curvature of these layer surfaces should be a good way to characterize these shapes. There is not such an obvious difference for CSFF, as shown in Figure 25, where the three-dimensional hydrogen bond pattern aligns the chains in the layers, preventing significant twist. In preliminary cellulose Iα simulations, the hydrogen bonding remains mostly two-dimensional, and the

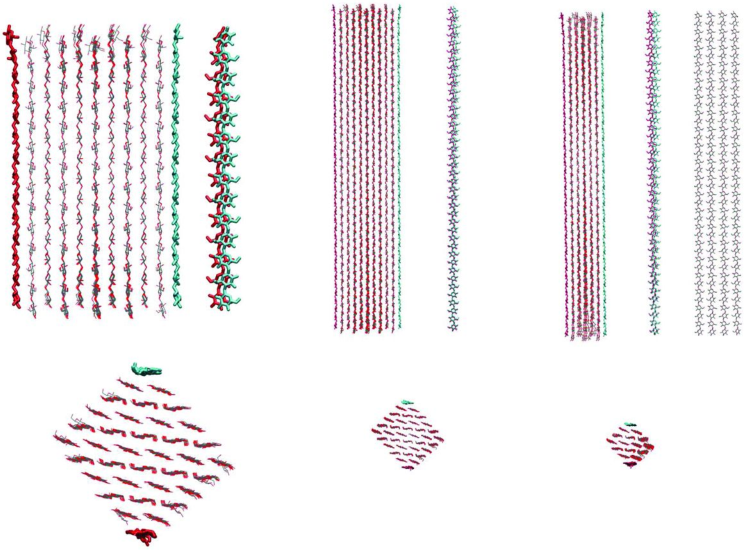


Figure 25. Three fibrils of Iβ, as shown in Figure 24, with the CSFF force field. All have extensive three-dimensional hydrogen bonding and so do not show significant twist despite the differences in length and width. (see color insert)

magnitude of the twist in Iα for both CSFF and GLYCAM06 is similar to the twist shown in Figure 24 for the GLYCAM06 Iβ simulations (data not shown).

Conclusions

Cellulose structure simulations can be very sensitive to force field parameters and treatment of long-range interactions. Differences between force field conformational preferences at the scale of a single cellobiose molecule lead to radically different macroscopic properties of cellulose fibrils. The distribution of hydrogen bond angles and distances from these dramatically different structures is similar, raising doubt as to whether Infrared or Raman spectroscopy would be able to distinguish between these possible hydrogen bond schemes on the basis of hydroxyl vibration.

Methods

All of the calculations reported here used the CHARMM molecular mechanics program (*42*). The sugar atoms were modeled using parameters specifically developed for carbohydrates, namely the CSFF (*43*) and GLYCAM (*44*) force fields. GLYCAM04L and GLYCAM06 parameters were downloaded from the GLYCAM web page and reformatted for use in CHARMM. The energy terms calculated with CHARMM for a configuration of cellobiose were validated by comparison to the energy terms produced with AMBER (*45*). Both of these carbohydrate parameter sets are all-atom force fields, but for clarity of presentation most figures in this work omit aliphatic hydrogen atoms. The water molecules were represented using the modified TIP3P force field (*46*, *47*).

Conformational analysis of cellobiose, infinite chains of cellulose, and infinite sheets of cellulose was conducted by constructing all combinations of hydroxymethyl rotamers and clockwise and counter-clockwise hydroxyl orientations (*48*, *49*). The potential energy of each starting configuration was minimized with up to 1000 steps of the conjugate gradient minimization algorithm. The energy at each point was scaled relative to the global mimimum energy for each map (*50*).

Several cellulose Iβ and Iα microcrystallites were constructed using the coordinates reported by Nishiyama *et al.*, where the structure was inferred from x-ray fiber diffraction analysis of sheets of highly crystalline cellulose (*51*, *52*). The Iβ structure was reported to be a monoclinic $P2_1$ crystal with unit cell dimension of **a** = 7.784 Å, **b** = 8.201 Å, and **c** = 10.380 Å, and **γ** = 96.5°. The Iβ unit cell consisted of two independent anhydroglucose units, with the chains containing them labeled as "center" and "origin," referring to their positions in the unit cell. The Iα structure was reported to be a triclinic P1 crystal with unit cell dimension of **a** = 6.717 Å, **b** = 5.962 Å, **c** = 10.400 Å, **α** = 118.08°, **β** = 114.80 ° and **γ** = 80.37°. The Iα unit cell contained one chain, with anhydrocellobiose as the repeating unit. Both cellulose Iα and Iβ have hydrogen bonded sheets, and the major difference between these structures is primarily in how the sheets are packed together. Figure 2 shows crystals of Iα and Iβ from the side of the hydrogen bonded sheets, emphasizing this difference in packing. Iα has sheets that are stacked along a constant inclined axis, where neighboring sheets are always shifted by +**c**/4, and Iβ has sheets that alternate between +**c**/4 and -**c**/4 packing.

Using the crystal-building facilities in CHARMM, two small crystals with different exposed crystallographic faces were fabricated. The cellooligomer chains in these crystals were either 14 or 40 monomer units in length, and hydroxyl hydrogen atoms were placed in the reported predominant hydrogen bond pattern. One of these microcrystallites contained 36 chains and emphasized hydrophilic surfaces while the other crystal was constructed for Iβ only, and contained 32 chains emphasizing the (100), (200), (010), and (020) surfaces. Different size crystals were chosen to keep the overall surface area as nearly equal as possible. For convenience, these two crystals will hereafter be referred to as the "diagonal" and "square" crystals. The square crystal has surfaces that are parallel to the unit cell axes (the 32-chain crystal), and the diagonal crystal has surfaces that cut across two unit cell axes (the 36-chain crystal). A microcrystal of the same cross-section as the diagonal crystal (36 chains total) was constructed with DP 40, as was a smaller-diameter DP 40 microcrystal with four chains per side (16 chains total).

Each of the constructed DP 14 crystallites was placed in an equilibrated rectangular box of TIP3P water molecules with dimensions 56.0 Å by 56.0 Å by 89.0 Å. The DP 40 crystallites were placed in an equilibrated box with dimensions 59.7 Å by 59.7 Å by 233 Å. All those water molecules that overlapped with the carbohydrate heavy atoms were deleted. The diagonal simulation contained 6116 water molecules and 29,040 atoms in total, and the square simulation contained

6434 water molecules and 28,806 atoms. The DP 40 crystal with 36 chains contained 19,251 water molecules and 88,101 atoms in total, and the DP 40 crystal with 16 chains contained 23,823 water molecules and 84,957 atoms in total. Simulations were performed on the Datastar and Teragrid clusters at the San Diego Supercomputer Center.

The following procedure was used for the initial publication of this work using the CSFF parameter set; slight modifications were later made to accommodate the GLYCAM parameters. Two hundred steps of steepest descent minimization, followed by 100 steps of conjugate gradient minimization were first applied to the system to relieve any serious strains resulting from the set-up procedure. MD simulations were then used to heat the system from 50 to 300 K in 50-K increments over a period of 10 ps, followed by an additional 190 ps of equilibration at 300K. After this heating and equilibration stage, the system velocities were not adjusted again, and the system was simulated in the NVE ensemble using a Verlet integrator with a step size of 1 fs. Non-bonded interactions were truncated at 15.0 Å on a neutral-group-by-neutral-group basis after being made to go smoothly to zero between 12.0 Å and 13.0 Å using ST2-type switching functions (*53*). Image non-bond interactions were also cut off at 15.0 Å. All calculations used a dielectric constant of 1. Chemical bond lengths involving hydrogen atoms were kept at fixed lengths using the constraint algorithm SHAKE (*54*). Following equilibration, trajectories were integrated for an additional 1 ns before analysis for the diagonal and square microcrystals with the CSFF parameters. For comparison, an additional simulation was conducted that was exactly like the diagonal crystal system in every respect except that dihedral angle restraining forces were used to keep the primary alcohol groups in the conformation found in the crystal. Three other simulations were carried out to test the sensitivity of the results to starting conditions. These test simulations consisted of: 1) a simulation in which the origin chains were replaced by chains in the center chain conformation; 2) a second simulation in which the center chains were replaced by chains in the origin chain conformation; and 3) a third simulation in which the center and origin chains were interchanged.

For the continuation of the previously published study, a few changes in protocol were made. With the lack of neutral charge groups in GLYCAM06, a radial cutoff for group pairs as used with CSFF would cause a significant discontinuity in electrostatic interaction energy at the cutoff boundary. To keep simulation conditions consistent, Particle Mesh Ewald electrostatics (*55*) were used for crystal simulations with both the CSFF and GLYCAM06 parameter sets. The integration step size was increased to 2 fs but the number of steps used for heating and equilibration was unchanged, for a total of 400 ps of heating and equilibration in each simulation. The trajectories were integrated for an additional 3 to 10 ns. A simulation was run using the structure obtained from the CSFF diagonal crystal conformation as the initial structure for a GLYCAM simulation. To test the assertion that the structure obtained with CSFF may be cellulose IV_I, the coordinates for cellulose IV_I published by Gardiner and Sarko (*38*) were used to construct a DP 14 microcrystal similar to the Iβ diagonal crystal and was run with CSFF and GLYCAM06.

Methanol dimer interaction energies as a function of oxygen-to-oxygen distance and hydrogen bond donor-hydrogen-acceptor (D-H--A) angle were calculated for the CSFF and GLYCAM06 force fields. Only hydrogen bond geometries corresponding to D-H--A angles of greater than 120° and oxygen pair distances up to 4.4 Å were considered, although hydrogen bonds are usually defined with the oxygen pair distance at 3.4 Å or less. A cylindrical coordinate grid with ρ spacing of 0.01 Å and φ spacing of 2° was constructed for each force field, and the minimized energies at each point were normalized relative to the energy at 50 nm separation. The relative orientation of the hydroxyl hydrogen atoms was restrained to prevent the formation of a doubly hydrogen-bonded pair at angles approaching 120°.

Acknowledgments

This work was supported by the DOE Office of the Biomass Program and Teragrid project number TG-MCA05S005

References

1. Kolpak, F. J.; Weih, M.; Blackwell, J. Mercerization of cellulose. 1. Determination of the structure of mercerized cotton. *Polymer* **1978**, *19* (2), 123–31.
2. Kolpak, F. J.; Blackwell, J. Mercerization of cellulose. 2. The morphology of mercerized cotton cellulose. *Polymer* **1978**, *19* (2), 132–5.
3. Nishimura, H.; Sarko, A. Mercerization of cellulose. III. Changes in crystallite sizes. *J. Appl. Polym. Sci.* **1987**, *33* (3), 855–66.
4. Nishimura, H.; Sarko, A. Mercerization of cellulose. IV. Mechanism of mercerization and crystallite sizes. *J. Appl. Polym. Sci.* **1987**, *33* (3), 867–74.
5. Kolpak, F. J.; Blackwell, J. The morphology of regenerated cellulose. *Text. Res. J.* **1978**, *48* (8), 458–67.
6. Shibazaki, H.; Saito, M.; Kuga, S.; Okano, T. Native cellulose II production by Acetobacter xylinum under physical constraints. *Cellulose (London)* **1998**, *5* (3), 165–73.
7. Clark, G. L.; Parker, E. A. X-ray diffraction study of the action of liquid ammonia on cellulose and its derivatives. *J. Phys. Chem.* **1937**, *41*, 777–86.
8. Legrand, C. Cellulose III regenerated from ammonia-cellulose. *J. Polym. Sci.* **1951**, *7*, 333–9.
9. Wada, M.; Heux, L.; Isogai, A.; Nishiyama, Y.; Chanzy, H.; Sugiyama, J. Improved Structural Data of Cellulose IIII Prepared in Supercritical Ammonia. *Macromolecules* **2001**, *34* (5), 1237–43.
10. Hutino, K.; Sakurada, I. A fourth modification of cellulose. *Naturwissenschaften* **1940**, *28*, 577–8.
11. Aravindanath, S.; Sreenivasan, S.; Bhama Iyer, P. Electron diffraction study of cellulose IV. *J. Polym. Sci., Part C: Polym. Lett.* **1986**, *24* (5), 207–9.
12. Nishimura, H.; Sarko, A. Mercerization of cellulose. 6. Crystal and molecular structure of Na-cellulose IV. *Macromolecules* **1991**, *24* (3), 771–8.

13. Atalla, R. H.; Dimick, B. E.; Nagel, S. C. Studies on polymorphy in cellulose. Cellulose IV and some effects of temperature. *ACS Symp. Series* **1977**, *48*, 30–41 (Cellulose Chemistry and Technology).

14. Wada, M.; Heux, L.; Sugiyama, J. Polymorphism of Cellulose I Family: Reinvestigation of Cellulose IVI. *Biomacromolecules* **2004**, *5* (4), 1385–91.

15. French, A. D.; Dowd, M. K. In *Conformational analysis of cellobiose with MM3*; 1993; pp 51−6.

16. Mendonca, S.; Johnson, G. P.; French, A. D.; Laine, R. A. Conformational Analyses of Native and Permethylated Disaccharides. *J. Phys. Chem. A* **2002**, *106* (16), 4115–24.

17. French, A. D.; Johnson, G. P. Advanced conformational energy surfaces for cellobiose. *Cellulose (Dordrecht, Netherlands)* **2004**, *11* (3/4), 449–62.

18. French, A. D.; Johnson, G. P. Quantum mechanics studies of cellobiose conformations. *Can. J. Chem.* **2006**, *84* (4), 603–12.

19. Strati Gina, L.; Willett Julious, L.; Momany Frank, A. Ab initio computational study of beta-cellobiose conformers using B3LYP/6-311++G**. *Carbohydr. Res.* **2002**, *337* (20), 1833–49.

20. Rees, D. A.; Skerrett, R. J. Conformational analysis of cellobiose, cellulose, and xylan. *Carbohydr. Res.* **1968**, *7* (3), 334–48.

21. Tvaroska, I.; Perez, S. Conformational energy calculations for oligosaccharides: a comparison of methods and a strategy of calculation. *Carbohydr. Res.* **1986**, *149* (2), 389–410.

22. Marks, D. L.; Dominguez, M.; Wu, K.; Pagano, R. E. Identification of active site residues in glucosylceramide synthase. A nucleotide-binding/catalytic motif conserved with processive b-glycosyltransferases. *J. Biol. Chem.* **2001**, *276* (28), 26492–8.

23. Han, N. S.; Robyt, J. F. The mechanism of Acetobacter xylinum cellulose biosynthesis: direction of chain elongation and the role of lipid pyrophosphate intermediates in the cell membrane. *Carbohydr. Res.* **1998**, *313* (2), 125–33.

24. Brett, C. T. Cellulose microfibrils in plants: Biosynthesis, deposition, and integration into the cell wall. *Int. Rev. Cytol.* **2000**, *199*, 161–99.

25. Haigler, C. H.; Ivanova-Datcheva, M.; Hogan, P. S.; Salnikov, V. V.; Hwang, S.; Martin, K.; Delmer, D. P. Carbon partitioning to cellulose synthesis. *Plant Mol. Biol.* **2001**, *47* (1−2), 29–51.

26. French, A. D.; Johnson, G. P. What crystals of small analogs are trying to tell us about cellulose structure. *Cellulose (Dordrecht, Netherlands)* **2004**, *11* (1), 5–22.

27. Bosma, W. B.; Appell, M.; Willett, J. L.; Momany, F. A. Stepwise hydration of cellobiose by DFT methods: 1. Conformational and structural changes brought about by the addition of one to four water molecules. *THEOCHEM* **2006**, *776* (1−3), 1–19.

28. Bosma, W. B.; Appell, M.; Willett, J. L.; Momany, F. A. Stepwise hydration of cellobiose by DFT methods: 2. Energy contributions to relative stabilities of cellobiose.bul.(H2O)1-4 complexes. *THEOCHEM* **2006**, *776* (1−3), 21–31.

29. Strati, G. L.; Willett, J. L.; Momany, F. A. Ab initio computational study of b-cellobiose conformers using B3LYP/6-311++G. *Carbohydr. Res.* **2002**, *337* (20), 1833–49.

30. Tang, H. R.; Belton, P. S. Molecular dynamics of polycrystalline cellobiose studied by solid-state NMR. *Solid State Nucl. Magn. Reson.* **2002**, *21* (3/4), 117–33.

31. Chu, S. S. C.; Jeffrey, G. A. Refinement of the crystal structures of b-D-glucose and cellobiose. *Acta Crystallogr., Sect. B* **1968**, *24* (Pt. 6), 830–8.

32. Matthews, J. F.; Skopec, C. E.; Mason, P. E.; Zuccato, P.; Torget, R. W.; Sugiyama, J.; Himmel, M. E.; Brady, J. W. Computer simulation studies of microcrystalline cellulose Ib. *Carbohydr. Res.* **2005**, *341* (1), 138–52.

33. Yui, T.; Nishimura, S.; Akiba, S.; Hayashi, S. Swelling behavior of the cellulose Ib crystal models by molecular dynamics. *Carbohydr. Res.* **2006**, *341* (15), 2521–30.

34. Heiner, A. P.; Sugiyama, J.; Teleman, O. Crystalline cellulose I.alpha. and I.beta. studied by molecular dynamics simulation. *Carbohydr. Res.* **1995**, *273* (2), 207–23.

35. Heiner, A. P.; Teleman, O. Interface between Monoclinic Crystalline Cellulose and Water: Breakdown of the Odd/Even Duplicity. [Erratum to document cited in CA126:132775]. *Langmuir* **1997**, *13* (26), 7305.

36. Heiner, A. P.; Teleman, O. Interface between Monoclinic Crystalline Cellulose and Water: Breakdown of the Odd/Even Duplicity. *Langmuir* **1997**, *13* (3), 511–8.

37. Heiner, A. P.; Kuutti, L.; Teleman, O. Comparison of the interface between water and four surfaces of native crystalline cellulose by molecular dynamics simulations. *Carbohydr. Res.* **1998**, *306* (1−2), 205–20.

38. Gardiner, E. S.; Sarko, A. Packing analysis of carbohydrates and polysaccharides. 16. The crystal structures of celluloses IVI and IVII. *Can. J. Chem.* **1985**, *63* (1), 173–80.

39. Rowley, R. L.; Tracy, C. M.; Pakkanen, T. A. Potential energy surfaces for small alcohol dimers I: Methanol and ethanol. *J. Chem. Phys.* **2006**, *125* (15), 154302/1–3.

40. Fileti, E. E.; Chaudhuri, P.; Canuto, S. Relative strength of hydrogen bond interaction in alcohol-water complexes. *Chem. Phys. Lett.* **2004**, *400* (4−6), 494–9.

41. Ghafouri, R.; Bruinsma, R. Helicoid to Spiral Ribbon Transition. *Phys. Rev. Lett.* **2005**, *94* (13), 138101/1–4.

42. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4* (2), 187–217.

43. Kuttel, M.; Brady, J. W.; Naidoo, K. J. Carbohydrate solution simulations: Producing a force field with experimentally consistent primary alcohol rotational frequencies and populations. *J. Comput. Chem.* **2002**, *23* (13), 1236–43.

44. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outeirino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: a generalizable

biomolecular force field. Carbohydrates. *J. Comput. Chem.* **2007**, *29* (4), 622–55.

45. Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–88.
46. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79* (2), 926–35.
47. Durell, S. R.; Brooks, B. R.; Ben-Naim, A. Solvent-Induced Forces between Two Hydrophilic Groups. *J. Phys. Chem.* **1994**, *98* (8), 2198–202.
48. Simon, I.; Scheraga, H. A.; Manley, R. S. J. Structure of cellulose. 1. Low-energy conformations of single chains. *Macromolecules* **1988**, *21* (4), 983–90.
49. Simon, I.; Glasser, L.; Scheraga, H. A.; Manley, R. S. J. Structure of cellulose. 2. Low-energy crystalline arrangements. *Macromolecules* **1988**, *21* (4), 990–8.
50. Tran, V. H.; Brady, J. W. Disaccharide conformational flexibility. I. An adiabatic potential energy map for sucrose. *Biopolymers* **1990**, *29* (6−7), 961–76.
51. Nishiyama, Y.; Langan, P.; Chanzy, H. Crystal structure and hydrogen-bonding system in cellulose Ib from synchrotron x-ray and neutron fiber diffraction. *J. Am. Chem. Soc.* **2002**, *124* (31), 9074–82.
52. Nishiyama, Y.; Sugiyama, J.; Chanzy, H.; Langan, P. Crystal Structure and Hydrogen Bonding System in Cellulose Ia from Synchrotron X-ray and Neutron Fiber Diffraction. *J. Am. Chem. Soc.* **2003**, *125* (47), 14300–6.
53. Stillinger, F. H.; Rahman, A. Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys.* **1974**, *60* (4), 1545–57.
54. Van Gunsteren, W. F.; Berendsen, H. J. C. Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.* **1977**, *34* (5), 1311–27.
55. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–92.

Chapter 3

Atomistic Simulation of Lignocellulosic Biomass and Associated Cellulosomal Protein Complexes

Loukas Petridis,[1,*] Jiancong Xu,[1] Michael F. Crowley,[2] Jeremy C. Smith,[1] and Xiaolin Cheng[1]

[1]University of Tennessee / ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6164, USA
[2]Chemical and Biosciences Center, National Renewable Energy Laboratory, Golden, CO 80401-3393, USA
[*]Petridisl@ornl.gov

Computer simulations have been performed to obtain an atomic-level understanding of lignocellulose structure and the assembly of its associated cellulosomal protein complexes. First, a CHARMM molecular mechanics force field for lignin is derived and validated by performing a molecular dynamics simulation of a crystal of a lignin fragment molecule and comparing simulation-derived structural features with experimental results. Together with the existing force field for polysaccharides, this work provides the basis for full simulations of lignocellulose. Second, the underlying molecular mechanism governing the assembly of various cellulosomal modules is investigated by performing a novel free-energy calculation of the cohesin-dockerin dissociation. Our calculation indicates a free-energy barrier of ~17 kcal/mol and further reveals a stepwise dissociation pathway involving both the central β-sheet interface and its adjacent solvent-exposed loop/turn regions clustered at both ends of the β-barrel structure.

Introduction

Plant cell wall structure has come under renewed interest in the context of producing bioethanol from the enzymatic hydrolysis of lignocellulosic biomass

In Computational Modeling in Lignocellulosic Biofuel Production; Nimlos, M., et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2010.

(*1–5*). The plant cell wall is made of cellulose microfibrils that are embedded in a matrix of polysaccharides (hemicelluloses and pectins), lignins, and proteins (*6*). Cellulosic ethanol production is a multi-stage process often involving, first, the pretreatment of biomass, then the hydrolysis of cellulose (and hemicelluloses) by enzymes to smaller oligosaccharides, and, finally, the fermentation of sugars to ethanol. The hydrolysis step is the bottleneck in the process because of the natural resistance, or "recalcitrance," of plant cell walls to degradation (*2*).

Given the complex and heterogeneous nature of biomass materials, a better understanding of their structure, dynamics, and degradation pathways becomes a necessary first step toward overcoming their recalcitrance to hydrolysis. Through years of extensive biochemical and biophysical studies, it has been established that although biomass recalcitrance is a very complex phenotype, with many factors contributing to it, lignin plays an important role (*7*). There is evidence of an inverse correlation between the rate of biomass hydrolysis and the lignin content (the amount of lignins present in the cell wall) (*8*). Lignin acts as a physical barrier, preventing enzymes from reaching the cellulose substrate. There is also evidence that lignin-enzyme interactions significantly contribute to the decline of rate observed during hydrolysis of lignocellulose substrates (*8*). Lignin poses an additional challenge in that, unlike hemicellulose and pectins, it is not readily removed with economically sustainable pretreatment. It has been suggested that, although lignin is initially released during pretreatment, it precipitates back on the cellulose surface at the end of the process (*9*). Another factor contributing to biomass recalcitrance is the crystallinity of cellulose. Cellulose can be found in crystalline fibrils, the compact structure of which impedes enzymatic access. In comparison, amorphous cellulose is readily digested by enzymes (*10*). Lignin content and the degree of crystallinity of cellulose had the greatest impact on biomass digestibility of Poplar wood (*11*). A more recent study of *alfalfa* lines found that the efficiency of enzymatic hydrolysis and the amount of total sugars released is proportional to the plant's lignin content (*12*).

A second promising avenue for altering biomass recalcitrance is designing more efficient enzyme systems to degrade the plant cell wall. For this, we need to more completely understand the structure, mechanism, and function of these enzyme systems. Generally, two classes of enzyme systems have been observed in microorganisms (*13–15*). One class consists of several individual endoglucanases, exoglucanases, and ancillary enzymes that can act synergistically to deconstruct plant cell walls. These enzymes are usually found in aerobic fungi and bacteria, of which the glycosyl hydrolases from *Trichoderma reesei (T. reesei)* is the best studied. The other system class, which is usually found in anaerobic microorganisms, involves the formation of a large, extracellular enzyme complex called the cellulosome, which consists of a scaffolding protein and many associated enzymes. Lignocellulosic biomass is structurally heterogeneous and includes many components in addition to cellulose, so efficient decomposition requires a variety of enzymes with a wide range of specificities and activities. To this end, the multienzyme cellulosome system seems particularly advantageous and has become a paradigm for designing more efficient enzyme complexes and biomimetics. During the past few years, an increasing number of cellulosome systems have been identified (*14*). Information is also becoming available regarding the structural principles governing the interactions among various cellulosomal domains (*16*, *17*). A cellulosome consists of a fibrillar protein (called the scaffolding protein) that contains binding sites (called cohesins) for the cellulosomal enzyme modules positioned periodically along the fibrils. In addition to their catalytic domains, all cellulosomal enzymes contain a cohesin-binding site called a dockerin. The cohesin–dockerin interaction is an important factor in cellulosome assembly. For example, the *Clostridium thermocellum* cellulosome assembles through the interaction of a type I dockerin with one of several type I cohesin modules. Although cohesins and dockerins exhibit relatively high sequence homology, the interaction between cohesins and dockerins is generally species specific (i.e., cohesins from one species do not recognize and interact with dockerins present in other species) (*16*, *18*).

Although computational studies have proven useful in providing detailed insight into diverse biochemical/biophysical processes otherwise inaccessible from experiment alone, atomistic simulation of lignocellulosic models has so far been limited. With the help of high-performance computing, the foundations for accurate simulation of these materials have been laid recently (*19*, *20*); and various simulations are starting to emerge that can be employed to derive physical properties of lignocellulosic biomass, thus serving as a reference for interpreting an array of biophysical experiments. On another front, atomic-level structural information is now being accumulated for individual cellulosomal modules (*17*, *21*), although the structure of the entire cellulosome complex is still difficult to obtain. The availability of this partially complete data from different sources, however, offers great opportunity for using computational approaches to study the structure, dynamics, and assembly process of cellulosome complexes. In this chapter, we will focus on two lines of our research as the initial efforts toward our long-term goal. One is on the parameterization of a potential energy function for simulating lignocellulosic biomass. The other is on modeling cohesin-dockerin interaction in cellulosome.

Toward More Realistic Simulation of Lignocellulosic Biomass

The chemical composition and structure of lignins are highly heterogeneous, varying significantly between different plant species and even within different parts of the same plant wall. Although the exact chemical formula of lignins is not known, abundant information is available on its composition. Lignins are composed primarily of three units: *p*-hydroxyphenyl (H), guaiacyl (G), and syringyl (S), derived by oxidation of three alcohol monolignols: *p*-coumaryl, coniferyl, and sinapyl, respectively (*22*) (Figure 1a). There are various linkages that connect the units, leading to the formation of the branched lignin biopolymer. The most common linkages are β-O-4’, 5-5’, α-O-4’, and β-5’ in guaiacyl and syringyl (Figure 1b). There is an ongoing heated debate on how monolignols couple to form the lignin macromolecule. One theory suggests that lignin monomers are oxidized and then coupled in a combinatorial fashion (*54*). The second theory suggests that lignin primary structure is controlled at the proteinaceous level (*55*). To the best of our knowledge, there are no currently published reports on the exact primary structure of lignins. For this reason, our studies are based on the assumption that this primary structure is combinatorically derived. We stress that our work does not attempt to validate either of the previously mentioned theories.

Although there is a large volume of simulation work on cellulose (*23–28*), there are relatively few computational studies of lignin. Previous computational studies (*29–32*) employed the CHARMM27 empirical force field (*33*), which was developed to model proteins rather than lignin. In recent work (*20*), we presented the first essential step towards the accurate computer simulation of lignin: the derivation of an empirical molecular mechanics (MM) force field. Together with the existing force field for polysaccharides, this force field will enable full simulations of lignocellulose.

A Molecular Mechanics Force Field for Lignin

Parameterization Strategy

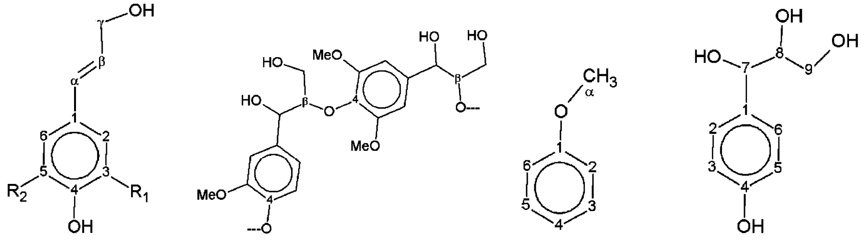
In this section, we outline the general strategy employed to obtain the lignin force field. The CHARMM potential energy function of a molecule is as follows (*33*):

$$E = \sum_{bonds} K_b (b-b_o)^2 + \sum_{angles} K_\theta (\theta - \theta_o)^2 + \sum_{U-B} K_{ub} (s-s_o)^2 + \sum_{dihedrals} K_\phi [1+\cos(n\phi - \delta)] + \sum_{impropers} K_\psi (\psi - \psi_o)^2 + \sum_{non-bonded} \left\{ \varepsilon_{ij} \left[\left(\frac{R_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\varepsilon_0 \pi r_{ij}} \right\}, \qquad (1)$$

where contributions to the energy include bonded (bond, angle, Urey-Bradley, dihedral, and improper dihedral) and non-bonded (the Lennard-Jones 6-12 potential for the van der Waals interactions and Coulomb interactions) terms. The force constants *K* and partial atomic charges *q* are molecule-dependent and must be optimized to model any specific molecule prior to performing the simulation.

This parameterization of lignin follows the main procedure of parameterization of proteins (*33*) and ethers (*34*) for the CHARMM force field. Lignin also has a linear ether bond, but it is different from those examined in (*34*) in that the oxygen is bonded to a phenyl ring and a tertiary carbon. For this reason, it was necessary to create a new atom type, OET, to represent the lignin's ether oxygen. Parameters were optimized by considering two factors. First, the target data were reproduced as closely as possible. Second, compatibility with the existing CHARMM force field was ensured by restricting optimization to the new parameters that do not already exist.

Two model compounds were used. The first model system, methoxybenzene, also known as anisole (see Figure 1c), incorporates the basic features of the β-O-4' link, an ether oxygen bonded to a tertiary and an aromatic carbon. Anisole was used to obtain all parameters involving the ether oxygen atom. The second compound (see Figure 1d) is *p*-hydroxyphenyl (PHP), the simplest lignin unit. PHP was used to obtain all lignin parameters not involving the ether oxygen.



(a) monolignols (b) G and S units with β-O-4′ link (c) anisole (d) PHP

Figure 1. (a) The tree monolignols: p-coumaryl (R_1 = R_2 = H), coniferyl (R_1 = H, R_2 = OMe), and sinapyl (R_1 = R_2 = OMe). (b) A guaiacyl unit connected with a β-O-4' linkage to a syringyl unit. (c) Model compound anisole. (d) Model compound PHP.

The optimization strategy for the new parameters is summarized below (Figure 2). Equilibrium values for bonds, angles, and dihedrals were taken from MP2/6-31G* QM-optimized geometries and were not further revised. The van der Waals parameters were taken unaltered from the CHARMM force fields (*33*), including those for the new atom type, OET. Initial values for the partial atomic charges of O_1, C_1, and Cα were deduced from a restricted fit to the quantum mechanics (QM) electrostatic potential (RESP) on selected grid points (*35*), while all other partial charges were fixed to their original CHARMM values. An iterative procedure, described in the next paragraphs, was followed until convergence was reached.

Optimization of Partial Atomic Charges

Charges were further optimized with respect to the QM interaction energies using a supramolecular approach with a model compound (anisole) interacting with one water molecule. The partial charges were adjusted to reproduce minimum distances and interaction energies between anisole and a TIP3P water molecule (*36*). Two geometries were considered in this supramolecular approach: the first d_0 with water lying on the phenyl plane, and the second d_{120} with the water hydrogen pointing at the position of the ether oxygen lone pair. A list of all final atomic charges is shown in Table 1. Only three charges (O_1, C_1, and Cα) were optimized, with the rest being kept to their CHARMM values.

To mimic the effect of electronic polarizability, which is not explicitly taken into account in additive force fields, atomic charges were purposely overestimated. This leads to an enhanced molecular dipole moment, with the QM gas-phase dipole moment being 1.42 Debyes, whereas the MM value is 1.66 Debyes. Table 2 compares the MM and QM interaction energies and distances, which were used to optimize the anisole charges. The empirical calculations successfully reproduced the scaled QM interaction energies, with the error being less than 3%. The empirical model gives distances about 0.3Å shorter than the QM values, a result of intentionally overestimating the gas-phase charges to

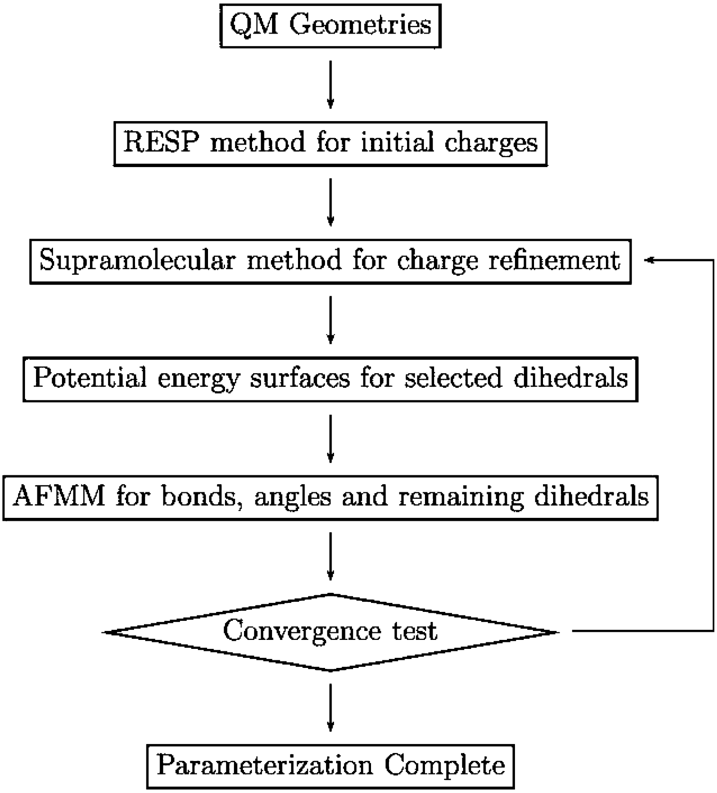

QM Geometries
RESP method for initial charges
Supramolecular method for charge refinement
Potential energy surfaces for selected dihedrals
AFMM for bonds, angles and remaining dihedrals
Convergence test
Parameterization Complete

Figure 2. Schematic representation of parameterization strategy. (Reproduced from reference (20). Copyright 2008 Wiley Periodicals, Inc.)

Table 1. A list of the anisole atoms with their respective charges[a]

Atom name	*Atom type*	*Charge*
C_α	CT3	-0.060
$H_{\alpha1}$, $H_{\alpha2}$, $H_{\alpha3}$	HA	0.090
O	OET	-0.280
C1	CA	0.070
C_2, C_3, C_4, C_5, C_6	CA	-0.115
H_2, H_3, H_4, H_5, H_6	HP	0.115

[a] Atom names refer to Figure 1c and atoms types follow the CHARMM27 force field with the new atom type labeled as OET.

obtain good condensed-phase properties. In the previous general force field for ethers, a similar behavior was observed with a 0.3Å difference between QM and MM (*34*). Finally, the electronic charge density was examined by Mulliken analysis (using the NWChem software), and the charge transfer was found to be insignificant.

After completing parameterization, we performed a further calculation to ensure that the partial atomic charges of Table 1, derived using a model compound, can be transferred to lignin. The minimum interaction energies and distances between a lignin dimer (G and S units connected with a β-O-4' linkage shown in Figure 1b) and a water molecule were obtained without further modifying the charge parameters. The excellent agreement between the QM and MM interaction energies justifies using these charges for the β-O-4' lignin linkage.

Dihedral Parameters

After completing the non-bonded terms, parameters for dihedral rotations were deduced from the QM potential energy surfaces. Six dihedral rotations were considered. The two rotations around the β-O-4' linkage (ω_1= X-C_1-O-C_α and ω_2= C_1-O-C_α-H, where X refers to any atom types) were obtained using the anisole model compound. The remaining four dihedrals that do not involve the ether oxygen (ω_3=C_2-C_1-C_7-X, ω_4=C_1-C_7-O_7-HO_7, ω_5=C_1-C_7-C_8-X, and ω_6=X-C_8-C_9-X) were deduced from the more complex rotational potential energy profiles of the second model compound, PHP. The optimization was based on reproducing the adiabatic QM energy surfaces. As an example, two plots are shown in Figure 3. In Figure 3a, the MM surface closely follows the target QM data, whereas in Figure 3b, although the agreement between the QM and MM data is not perfect, the rather complex shape is reproduced satisfactorily.

Bond and Angle Vibrations

The remaining bonded parameters (bonds and angles) were optimized to reproduce vibrational frequencies and eigenvector projections derived from QM calculations. For this, we used the Automated Frequency Matching Method (AFMM) (*37*), which optimizes the MM parameter set until the best fit with the QM reference set is obtained. AFMM requires both the eigen frequencies and eigenvectors of the MM set to match the QM data. This is an important aspect of the method, because it avoids incorrect mode matching and misleading reproduction of vibrational frequencies. The resulting plots of the vibrational frequencies obtained with QM and the MM for anisole and PHP are shown in Figures 4a and 4b, respectively. In both model compounds, the MM and QM frequencies match very well, with root mean square deviation of 51.6 cm^{-1} for anisole and 55.6 cm^{-1} for PHP, indicating that the bond and angle parameters are well-optimized.

Table 2. Minimum interaction energies (kcal/mol) and distances (Å) between water:anisole and water:lignin-dimer[a]

	Interaction energies		*Interaction distances*	
Orientations	*QM*	*MM*	*QM*	*MM*
d_0	-4.01	-3.96	2.15	1.82
d_{120}	-3.18	-3.09	2.16	1.87
dimer	-3.93	-4.02	2.10	1.81

[a] QM interaction energies were scaled by 1.16 as described in the text. Orientation geometries considered have the dihedral between the water molecule and the phenyl ring being 0, and 120 degrees, respectively and "dimer" refers to a G and S unit connected with a β-O-4' linkage.

Force Field Validation

In the final part of this work, the parameter set was tested without further adjustment against a condensed-phase experimental property of lignin that was not used during the parameterization. Because of the highly heterogeneous structure of lignin, the most appropriate experimental data to use is the crystal structure of a lignin subunit dimer, erythro-2-(2,6-dimethoxy-4-methylphenoxy)-1-(4-hydroxy-3,5-dimethoxyphenyl) propane-1,3-diol (EPD) (*38*). The chosen compound is very similar to two syringyl units connected with a β-O-4' linkage, but with a methyl group replacing the hydroxyl group of one of the phenol rings. The single crystal X-ray diffraction study revealed a triclinic $P\overline{1}$ structure whose unit cell dimensions are listed in Table 3.

To mimic as closely as possible the conditions under which the experiment was run, the MD simulation was performed for a 4×4×4 unit cell (128 dimers) using periodic boundary conditions while maintaining the temperature and pressure at their experimental values. The unit cell dimensions were allowed to vary during the simulation, and their time averages are listed in Table 4. The MD unit cell dimensions were close to the experimental values, and the system remained triclinic. The unit cell underwent a moderate expansion, with a 5% increase in volume. After aligning the MD coordinates with the experimental structure, the root mean square deviation (RMSD) between the experimental and calculated structure was 0.173±0.033 Å.

In particular, we should also note that the current force field models the β-O-4' linkage that is essential to the conformation of the lignin macromolecule very well. The time averages of the two dihedrals (d_1 and d_2) that define the β-O-4' linkage were compared with the experimental crystal values. The two dihedrals are (numbering scheme in Fig. 1b): $d_1 = C_5\text{-}C_4\text{-}O\text{-}C_8' = 77.9 \pm 6.3°$, compared to the experimental value of 80.0° and $d_2 = C_4\text{-}O\text{-}C_8'\text{-}C_7' = -148.5 \pm 5.5°$, compared to the experimental value of –152.8°. As with previous results, the simulation results agree with the experimental ones.

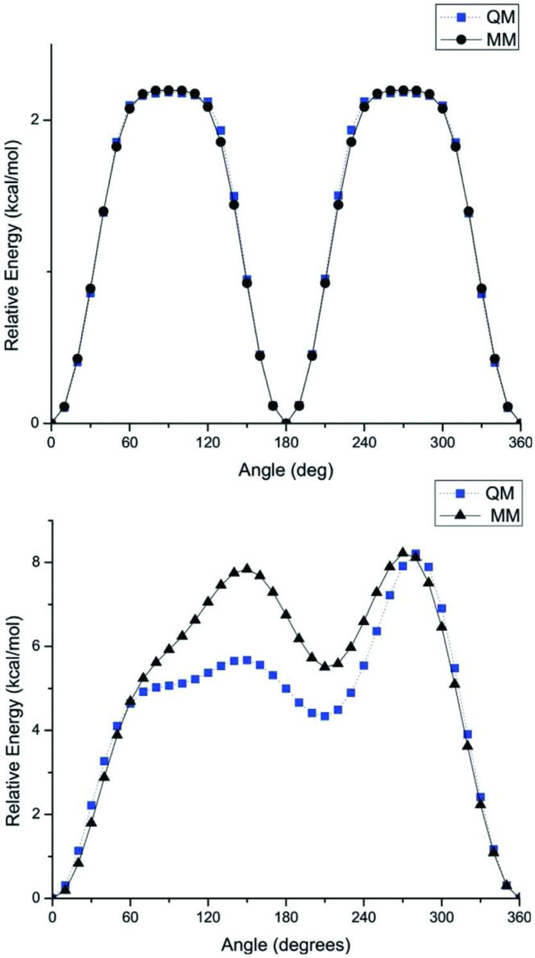


Figure 3. Potential energy profiles for rotation around the (a) ω_1 = C_2-C_1-O-C_α dihedral of anisole and (b) ω_4 = C_1-C_7-O_7-HO_7 dihedral of anisole. (Reproduced from reference (20). Copyright 2008 Wiley Periodicals, Inc.)

Building Lignocellulose Models

The accurate computer simulation of lignin presents significant challenges. Unlike many biological macromolecules that have been studied with molecular simulation, both the primary and three-dimensional structures of lignins are not known. Hence, a logical strategy is to build random multiple lignin units that have ensemble composition (and linkage) properties consistent with experimentally derived average chemical composition. In particular, emphasis was placed on ensuring that the models accurately represent the lignins found in the cell walls of softwoods. The following paragraphs describe how the atomistic lignin models were built.

We built 26 lignin molecules altogether, each with a distinct primary and tertiary structure. The initial structural models were generated by first deriving the topology of the molecules and then generating the tertiary structure. To generate molecular topologies, we used a variety of experimental data on lignin composition in softwoods. Softwoods are composed mainly of guaiacyl (G) units (*3*, *22*), so only G units are considered to be present in the model. A typical linkage composition of softwoods is: β-O-4’ 50%, 5-5’ 30%, α-O-4’ 10% and β-5’ 10%. Linkages β-O-4’, α-O-4’, and β-5’ contain chiral centers at the β and α-carbons. However, lignins are not found to be optically active (*39*). Hence, the constructed lignin molecules contain equal numbers of left- and right-hand linkages. The molecular weight of lignins is on the order of 10,000 or greater (*40*), and the models have a molecular weight of 13,000. Crosslinks are formed when one unit participates in more than one linkage. Twenty-six lignins were built with varying degrees of crosslinking, but the average crosslink density was chosen to be consistent with the experimental value of 0.052 obtained from spruce wood (*41*). With these experimental data as a guide, random lignin configurations were created using a script written in the program language Python. This method produced 26 molecules that were different to each other, but were all consistent with the experimentally determined properties of softwoods. For example, although all lignins had the same linkage composition, the order of the linkages was different. Furthermore, the number of crosslinks and their positions in individual lignin molecule were also different.

Once the topologies were derived, the 3D structures for lignin molecules were constructed using a step-wise approach. Each new unit was added to the existing structure using the appropriate linkage. As mentioned above, the geometries of all the units and linkages were obtained using high-level quantum chemical calculations. Subsequently, the entire new molecule was minimized using a molecular mechanics force field. The procedure was repeated until the maximum molecular weight of 13,000 was reached. As indicated earlier, our approach, while consistent with the average chemical properties of lignin, is limited by the lack of primary and tertiary structures of these molecules.

In contrast to lignins, the chemical structure of cellulose is known. It is a straightforward process to build cellulose microfibrils using the molecular structure of cellulose Iα (*42*) and Iβ (*43*), obtained with a combination of X-ray and neutron diffraction. In the present model, as in other studies, cellulose is in the Iβ form; and the MD simulation starts with the fibril being a perfect crystal. A preliminary model of cellulose surrounded by lignin molecules in solution is shown in Figure 5. Such initial models can probe the interactions between lignin and cellulose at the atomic level, as well as provide a way to parameterize coarse-grained mesoscale models.

Modeling Cellulosomes: Cohesin-Dockerin Interaction

Insight of Type I Cohesin-Dockerin Recognition from the Crystallographic Structure

The first 1.9-Å crystal structure of the type I cohesin-dockerin complex from the cellulosome of *C. thermocellum* has been determined (*17*) (Figure 6), providing insight into the structure and mechanism by which the cellulosome assembles. The

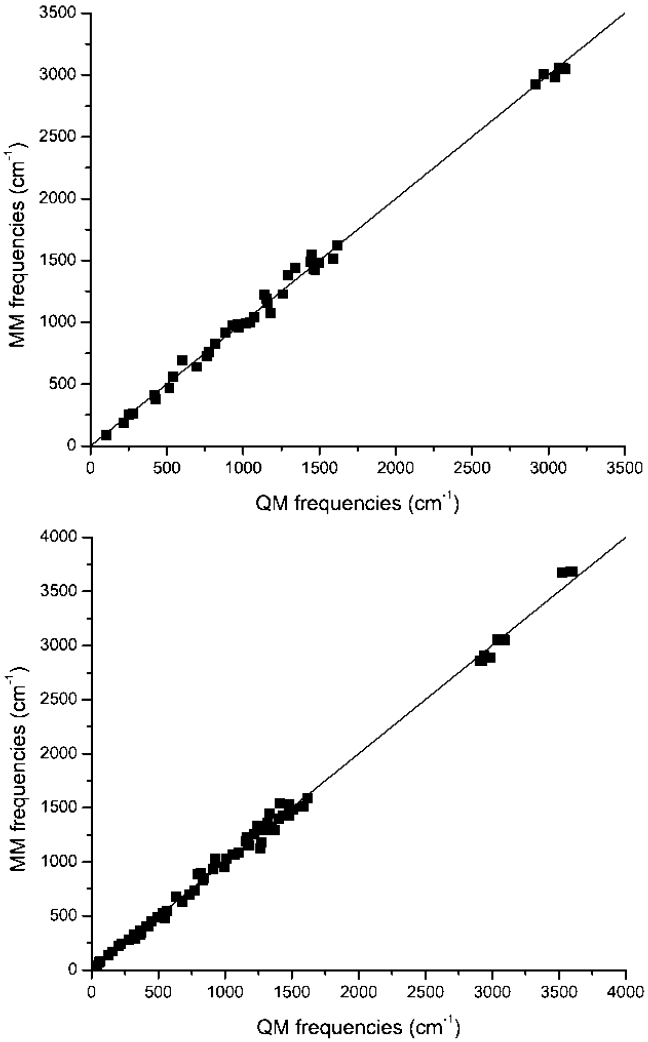


Figure 4. Vibrational frequencies of model compounds anisole (left) and PHP (right). The plotted line shows an ideal fit between QM and MM data. (Reproduced from reference (20). Copyright 2008 Wiley Periodicals, Inc.)

cohesin module forms an elongated, nine-stranded β-barrel in a classical jelly-roll topology with a tightly-packed aromatic/hydrophobic core. The two faces of the β-barrel are composed of strands 5, 6, 3, and 8 on the contact face with the dockerin, and strands 4, 7, 2, 1, and 9 on the opposite face. The dockerin partner of the cohesin-dockerin complex contains a duplicated 22-amino-acid sequence that comprises α-helix 1 and 3 in conformation, respectively. The dockerin structure is organized into two calcium-binding loop-helix motifs separated by a short linker region. Indeed, it has been found that Ca^{2+} plays a key role in maintaining the structural integrity of the cohesin-dockerin complex (*44*, *45*).

Table 3. Unit cell properties of small-molecule-dimer for experimental crystal structure and from molecular dynamics simulation

Cell dimension	*X-ray*	*MD*
A (Å)	8.69	8.73 ± 0.02
B (Å)	8.90	8.93 ± 0.01
C (Å)	13.11	13.68 ± 0.03
α (deg)	73.85	74.48 ± 0.05
β (deg)	86.15	86.30 ± 0.01
γ (deg)	83.06	83.06 ± 0.02
Cell volume (Å^3)	966	1020

Table 4. Dihedrals defining the β-O-4' linkage $d_1 = C_5$-C_4-O-C_8 and $d_2 = C_4$-O-C_8-C_7, see Figure 2d

Dihedral	*X-ray*	*MD*
d_1 (deg)	80.0	77.9 ± 6.3
d_2 (deg)	–152.8	–148.5 ± 5.5

The cohesin structure's compact nature, together with the fact that the contact surface features no obvious binding pocket or cleft, suggests that binding between cohesins and dockerins occurs through the exposed surface residues. The cohesin in the type I complex comes into contact with the entire length of α-helix 3, but is only in contact with the C-terminal end of helix 1 from the type I dockerin. The N terminus of helix 1 is diverted away from the cohesin surface. Given the orientation of the dockerin on the cohesin surface and the two-fold structural symmetry within the dockerin domain, Carvalho et al. provided evidence for a dual binding mode of dockerin modules to cohesins (*21*).

The available crystal structures suggests that the cohesin-dockerin association is maintained mainly by hydrophobic interactions, consistent with the negative heat capacity associated with the binding event (*17*, *46*, *47*). The proteins also interact through an extensive hydrogen-bonding network between one face of the cohesin and the corresponding dockerin domain. Several hydrophilic residues play an essential role in recognizing and forming the complex: Arg77, Tyr74, Asp39, Glu86, and Gly89 of the cohesin domain, and Leu22, Arg23, Ser45, Thr46, and Arg53 from α-helices 1 and 3 of the dockerin domain (Figure 6). Biochemical mutagenesis studies have provided complementary clues to the mode of cohesin-dockerin interaction. One of the striking mutations, known to cause recognition failure, is D39N. Asp39 of the cohesin, one of the most conserved residues, is located at the protein-protein interface of the complex. This residue forms direct hydrogen bonds with Ser45 of the dockerin, the most critical residue for domain recognition (*16*, *21*, *48*), and forms water-mediated hydrogen bonds with Val21 and Ile43. It has been shown that the single substitution of Asp39 by a neutrally charged Asn reduces the affinity of the interaction by more than three

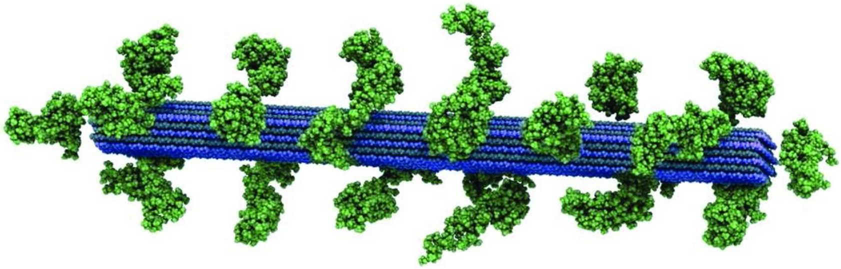


Figure 5. Atomistic model of the plant cell wall components cellulose (blue) and lignin (green).

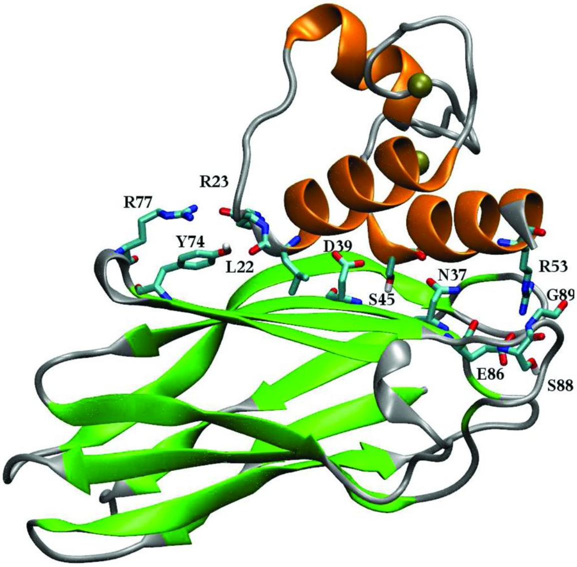


Figure 6. Crystal structure of the cohesin-dockerin complex in cartoon representation with β-sheets (cohesin) in green, α-helices (dockerin) in orange and loop regions in silver. Key residues involved in inter-domain interaction are highlighted in licorice mode, and colored by atom names.

orders of magnitude and disrupts the normal recognition of the dockerin (*49*). Thus, this residue is a hot spot for the cohesin-dockerin interaction. In addition, more recent biophysical and dockerin-mutagenesis experiments have revealed an association constant (K_a) of 8×10^7 M^{-1} for the wild-type cohesin-dockerin complex and the importance of highly conserved Ser45-Thr46 in the Ca^{2+}-binding loop for recognition of type I dockerin (*21*). It has been demonstrated that an alternative binding mode can be achieved by substituting the helix-3 Ser45/Thr46 pair with alanines; and the resultant crystal structure at 2 Å resolution shows that the dockerin module interacts with its cognate cohesin module through the helix-1, in which Ser11 and Thr12 play an equivalent role in binding.

Free Energy Landscape of Cohesin-Dockerin Dissociation

Recognizing Type I cohesins by dockerins is the essential event in assembling individual enzymatic subunits into the cellulosome complex. Even though the crystallographic structure and experimental measurements have provided essential information about the association of cohesins and dockerins, the underlying microscopic dynamic and energetic processes are not directly accessible to experiments. Consequently, aspects of the mechanism governing the assembly of cohesins and dockerins remain uncertain. It is therefore particularly informative to elucidate at the atomistic level the detailed molecular principles upon which the cohesin-dockerin interaction is based.

Understanding the underlying molecular association/dissociation mechanism in terms of structure and dynamic events is facilitated by the knowledge of the free-energy profile for the cohesin-dockerin dissociation. The effective free energy (or the potential of mean force, PMF) of cohesin-dockerin dissociation was estimated from a total of 100 ns MD simulation in bulk solution, using the adaptive biasing force (ABF) method (*50, 51*) implemented in NAMD (*52*). This method relies upon integrating the average forces acting along a reaction coordinate (ξ) that was constructed from endpoints corresponding to the average, or most probable, configurations from unconstrained MD simulations of initial and final states. The reaction coordinate for the dissociation process was defined by the separation distance between the cohesion and dockerin center-of-masses. The results are shown in Figure 7. Although this simple, low-dimensional, reaction coordinate has not been refined, if properly converged, the PMF from this reaction coordinate gives an upper bound on the barrier and, again if converged, will give a proper free-energy change between the states specified.

The free-energy profile's overall shape along the reaction coordinate shows a general uphill trend, illustrating quantitatively that the cohesin-dockerin complex exhibits a resistance against external forces, and that there is a high affinity for the two domains to remain bound. The global free-energy minimum in the profile appears at a distance separating the centers of mass equal to 22.5 Å, corresponding to the stable bound state with the key residues directly in contact. As the two domains move away from each other, the cohesin-dockerin interactions are progressively disrupted. Initially, this leads to a steep increase of the free energy before reaching the first shoulder at ~ 24 Å, at which point the hydrogen bond Asp39 (OD)-Ser45 (HG) has broken; and residues Asp39 and Ser45 at the interface of the protein complex are no longer in contact (Figure 7b). Another characteristic of the initial dissociation is the flow of water molecules into the binding area, substituting protein residues and forming new hydrogen bonds. The first dissociation step, therefore, corresponds to disrupting the hydrophobic core and overcoming the resistance imposed by the Asp39-Ser45 hydrogen bond. As the two domains move further apart, the free-energy profile reaches the second slight shoulder at ~ 26 Å. Inspection of the simulation trajectory indicates that the second shoulder corresponds to the disruption of the recognition strip interaction with the C-terminal region of α-helix 3, accompanied by the rupture of hydrogen bonds/salt bridges between Arg53 and Glu86 (Figure 7c). In contrast, at this point of the dissociation, the C-terminal of the first α-helix of the dockerin,

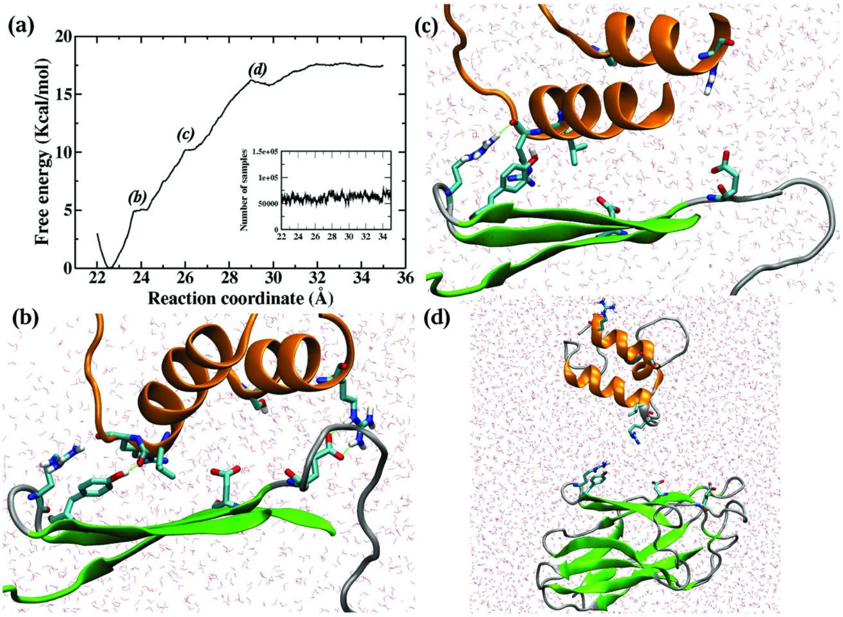


Figure 7. (a) Free-energy profile for the dissociation of cohesin and dockerin domains. The sampling distribution is included in the inset. (b) Snapshot of the cohesin-dockerin complex at ξ = 24 Å; (c) Snapshot at ξ = 27 Å; (d) Snapshot of cohesin-dockerin complex in the dissociated state, i.e., ξ > 30 Å. The two α-helices, β-strands 3, 5, 6, and loop/turn regions are represented in cartoon mode, colored orange, green, and gray, respectively. The rest of the protein structure was omitted for clarity.

and especially the backbone carbon atom of residue Arg23, is still repeatedly in contact with the side chains of the solvent-exposed Arg74 and Tyr77 in the β–strand 5/6 loop at the other end of the β-barrel, with large fluctuations of interatomic distances. The final dissociation of the two subunits corresponds to the shallow well at ~ 30 Å before the PMF eventually becomes nearly flat at > 35 Å (Figure 7d).

The experimental estimate of the overall equilibrium binding constant for the present cohesin-dockerin complex is 8×10^7, corresponding to a free-energy change of about 12 kcal/mol ($\Delta G = -RT\ln K_a$, where R is the gas constant and T is temperature). In the simulations, the overall difference in the calculated free energy between the minimum of the bound state and the barrier is ~ 17 kcal/mol. This agreement is reasonable, given that directly comparing the dissociation free energy with the experimentally determined absolute binding energy requires a knowledge of the contributions which were not considered in this study. The free energy change in the translational and rotational degrees of freedom on complexation was not included. Implementations of free energy algorithms have inherent errors. The sampling errors that may arise from the conformational flexibility of the unbound dockerin domain in solution were also not considered. Some significant extension to the present computational methodologies is needed to tackle the complex situation in the cohesin-dockerin protein complex. Furthermore, the present study focuses on a detailed view of the underlying mechanism of association and interaction in the cohesin-dockerin complex rather than calculating the absolute binding free energy.

Summary and Outlook

The accurate computer simulation of lignocellulosic biomass materials presents significant challenges. An important first step is the parameterization of a potential energy function for the system. Here, we derived an MM force field for lignin that is compatible with the CHARMM potential energy function. The parameterization was based on reproducing quantum-mechanically derived target data. Special care was taken to correctly describe the most common lignin linkage: the β-O-4' bond. The partial atomic charges of the oxygen and carbon atoms participating in the linkage were derived by examining interactions between a lignin fragment model compound and a water molecule. Dihedral parameters were obtained by reproducing QM potential energy profiles, with emphasis placed on accurately reproducing the thermally sampled low-energy regions. The remaining bond and angle parameters were derived using the AFMM method. To test the validity of the force field, we performed a simulation of a lignin-dimer crystal. The overall good agreement between the structural properties of the simulation and the experiment provide confidence that the force field can be used to simulate biomass. Furthermore, using a large body of experimental data on the average chemical composition of lignin as references, we have also constructed preliminary atomic-detail models of lignin.

Another important area of concentration is unraveling the assembly mechanism of the cellulosome complex using computer simulations. We have calculated the PMF profile for the wild-type cohesion-dockerin dissociation. The PMF reveals a high free-energy barrier and a stepwise pattern for the dissociation process. The sequential dissociation events revealed by the free-energy profile provides evidence that a set of residues lying on the flattened β-sheet surface and in the peripheral loop regions is the main obstacle to dockerin unbinding. Although examination of the crystal structure alone suggests that the formation of the cohesin-dockerin complex involves relatively large surface areas on both partners, our simulation results indicate that specific surface regions play more critical roles than others in forming and maintaining the integrity of the cellulosome complex. In turn, the insight gained from the present simulation can be used to guide protein engineering modifications to alter cohesin-dockerin binding. Efforts are underway to design engineered cellulosomal modules that can conduct more efficient biomass degradation than the corresponding wild-type protein complexes. Both atomic-detail and coarse-grained computer simulations are expected, in conjunction with appropriate biochemical and biophysical experiments (e.g., Hammel et al. 2005) (*53*), to provide a foundation for understanding the principles of domain synergy and cellulosomal activity, thus allowing the rational, structure-based design of improved cellulosomal assemblies for cellulosic ethanol production.

Acknowledgments

The authors acknowledge funding from the BioEnergy Science Center. The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Loukas Petridis is supported by DOE Field Work Proposal ERKP704. Xiaolin Cheng is also partly funded by DOE SciDAC. This research was supported in part by the National Science Foundation through TeraGrid resources provided by NICS. Additional support to Jeremy C. Smith is provided by ORNL LDRD fund and DOE INCITE award.

References

1. Lynd, L. R.; Laser, M. S.; Brandsby, D.; Dale, B. E.; Davison, B.; Hamilton, R.; Himmel, M.; Keller, M.; McMillan, J. D.; Sheehan, J.; Wyman, C. E. *Nat. Biotechnol.* **2008**, *26*, 169–172.
2. Himmel, M. E.; Ding, S. Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D. *Science* **2007**, *315*, 804–807.
3. Ragauskas, A. J.; Williams, C. K.; Davison, B. H.; Britovsek, G.; Cairney, J.; Eckert, C.A.; Frederick, W. J.; Hallett, J. P.; Leak, D. J.; Liotta, C. L.; Mielenz, J. R.; Murphy, R.; Templer, R.; Tschaplinski, T. *Science* **2006**, *311*, 484–489.
4. Gray, K. A.; Zhao, L. S.; Emptage, M. *Curr. Opin. Chem. Biol.* **2006**, *10*, 141–146.
5. Zhang, Y. H. P.; Ding, S. Y.; Mielenz, J. R.; Cui, J. B.; Elander, R. T.; Laser, M.; Himmel, M. E.; McMillan, J. R.; Lynd, L. R. *Biotechnol. Bioeng.* **2007**, *97*, 214.
6. Cosgrove, D. J. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 850–861.
7. Grabber, J. H. *Crop Sci.* **2005**, *45*, 820–831.
8. Chen, F.; Dixon, R. A. *Nat. Biotech*. **2007**, *25*, 759. Reddy, M. S. S.; Chen, F.; Shadle, G.; Jackson, L.; Aljoe, H.; Dixon, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16573.
9. Liu, C. G.; Wyman, C. E. *Ind. Eng. Chem. Res.* **2003**, *42*, 5409–5416.
10. Fan, L. T.; Lee, Y. H.; Beardmore, D. R. *Biotechnol. Bioeng.* **1981**, *23*, 419.
11. Chang, V. S.; Holtzapple, M. T. *Fundamental factors affecting biomass enzymatic reactivity*; Humana Press Inc.: Totowa, NJ, 2000.
12. Chen, F.; Dixon, R. A. *Nat. Biotechnol.* **2007**, *25*, 759–761.
13. Doi, R. H.; Kosugi, A. *Nat. Rev. Microbiol.* **2004**, *2*, 541–551.
14. Bayer, E.A.; Chanzy, H.; Lamed, R.; Shoham, Y. *Curr. Opin. Struct. Biol.* **1998**, *8*, 548–557.
15. Gilbert, H. J. *Mol. Microbiol.* **2007**, *63*, 1568–1576.
16. Mechaly, A.; Yaron, S.; Lamed, R.; Fierobe, H. P.; Belaich, A.; Belaich, J. P.; Shoham, Y.; Bayer, E. A. *Proteins* **2000**, *39*, 170–177.
17. Carvalho, A. L.; Dias, F. M.; Prates, J. A.; Nagy, T.; Gilbert, H. J.; Davies, G. J.; Ferreira, L. M.; Romao, M. J.; Fontes, C. M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13809–13814.

18. Pages, S.; Belaich, A.; Belaich, J. P.; Morag, E.; Lamed, R.; Shoham, Y.; Bayer, E. A. *Proteins* **1997**, *29*, 517–527.

19. Kuttel, M.; Brady, J. W.; Naidoo, K. J. *J. Comput. Chem.* **2002**, *23*, 1236–1243.

20. Petridis, L.; Smith, J. C. *J. Comput. Chem.* **2009**, *30*, 457–467.

21. Carvalho, A. L.; Dias, F. M.; Nagy, T.; Prates, J. A.; Proctor, M. R.; Smith, N.; Bayer, E. A.; Davies, G. J.; Ferreira, L. M.; Romao, M. J.; Fontes, C. M.; Gilbert, H. J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 3089–3094.

22. Boerjan, W.; Ralph, J.; Baucher, M. *Annu. Rev. Plant Biol.* **2003**, *54*, 519–546.

23. Matthews, J. F.; Skopec, C. E.; Mason, P. E.; Zuccato, P.; Torget, R. W.; Sugiyama, J.; Himmel, M. E.; Brady, J. W. *Carbohydr. Res.* **2006**, *341*, 138–152.

24. Nimlos, M. R.; Matthews, J. F.; Crowley, M. F.; Walker, R. C.; Chukkapalli, G.; Brady, J. V.; Adney, W. S.; Clearyl, J. M.; Zhong, L. H.; Himmel, M. E. *Protein Eng., Des. Sel.* **2007**, *20*, 179–187.

25. Yui, T.; Hayashi, S. *Biomacromolecules* **2007**, *8*, 817–824.

26. Yui, T.; Nishimura, S.; Akiba, S.; Hayashi, S. *Carbohydr. Res.* **2006**, *341*, 2521–2530.

27. Vietor, R. J.; Mazeau, K.; Lakin, M.; Perez, S. *Biopolymers* **2000**, *54*, 342–354.

28. Mazeau, K; Heux, L. *J. Phys. Chem. B* **2003**, *107*, 2394–2403.

29. Besombes, S.; Mazeau, K. *Biopolymers* **2004**, *73*, 301–315.

30. Besombes, S.; Mazeau, K. *Plant Physiology and Biochemistry* **2005**, *43*, 299–308.

31. Besombes, S.; Mazeau, K. *Plant Physiol. Biochem.* **2005**, *43*, 277–286.

32. Besombes, S.; Robert, D.; Utille, J. P.; Taravel, F. R.; Mazeau, K. *Holzforschung* **2003**, *57*, 266–274.

33. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D.T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B.* **1998**, *102*, 3586–3616.

34. Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D. *J. Chem. Theory Comput.* **2007**, *3*, 1120–1133.

35. Breneman, C. N.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.

36. Chen, I. J.; Yin, D. X.; MacKerell, A. D. *J. Comput. Chem.* **2002**, *23*, 199–213.

37. Vaiana, A. C.; Cournia, Z.; Costescu, I. B.; Smith, J. C. *Comput. Phys. Commun.* **2005**, *167*, 34–42.

38. Langer, V.; Lundquist, K.; Miksche, G. E. *Acta Crystallogr., Sect. E: Struct. Rep. Online* **2005**, *61*, O1001–O1003.

39. Ralph, J.; Peng, J. P.; Lu, F. C.; Hatfield, R. D.; Helm, R. F. *J. Agric. Food Chem.* **1999**, *47*, 2991–2996.

40. Brunow, G.; Kilpelainen, I.; Lapierre, C.; Lundquist, K.; Simola, L. K.; Lemmetyinen, J. *Phytochemistry* **1993**, *32*, 845–850.
41. Yan, J. F.; Pla, F.; Kondo, R.; Dolk, M.; McCarthy, J. L. *Macromolecules* **1984**, *17*, 2137–2142.
42. Nishiyama, Y.; Sugiyama, J.; Chanzy, H.; Langan, P. *J. Am. Chem. Soc.* **2003**, *125*, 14300–14306.
43. Nishiyama, Y.; Langan, P.; Chanzy, H. *J. Am. Chem. Soc.* **2002**, *125*, 9074–9082.
44. Chauvaux, S.; Beguin, P.; Aubert, J. P.; Bhat, K. M.; Gow, L. A.; Wood, T. M.; Bairoch, A. *Biochem. J.* **1990**, *265*, 261–5.
45. Lytle, B. L.; Volkman, B. F.; Westler, W. M.; Wu, J. H. *Arch. Biochem. Biophys.* **2000**, *379*, 237–44.
46. Lytle, B. L.; Volkman, B. F.; Westler, W. M.; Heckman, M. P.; Wu, J. H. *J. Mol. Biol.* **2001**, *307*, 745–53.
47. Spinelli, S.; Fierobe, H. P.; Belaich, A.; Belaich, J. P.; Henrissat, B.; Cambillau, C. *J. Mol. Biol.* **2000**, *304*, 189–200.
48. Schaeffer, F.; Matuschek, M.; Guglielmi, G.; Miras, I.; Alzari, P. M.; Beguin, P. *Biochemistry* **2002**, *41*, 2106–2114.
49. Handelsman, T.; Barak, Y.; Nakar, D.; Mechaly, A.; Lamed, R.; Shoham, Y.; Bayer, E. A. *FEBS Lett.* **2004**, *572*, 195–200.
50. Henin, J.; Chipot, C. *J. Chem. Phys.* **2004**, *121*, 2904–14.
51. Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *115*, 9169–83.
52. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
53. Hammel, M.; Fierober, H. P.; Czjzek, M.; Kurkal, V.; Smith, J. C.; Bayer, E. A.; Finet, S.; Receveur-Brechot, V. *J. Biol. Chem.* **2005**, *280*, 38562–38568.
54. Ralph, J.; Brunow, G.; Harris, P. J.; Dixon, R. A.; Schatz, P. F.; Boerjan, W. *Recent Adv. Polyethenol Res.* **2008**, *1*, 36–66.
55. Davin, L. B.; Lewis, N. G. *Curr. Opin. Biotechnol.* **2005**, *16*, 407–415.

Chapter 4

Modeling the Cellulosome Using Multiscale Methods

Yannick J. Bomble,* Michael F. Crowley, Qi Xu, and Michael E. Himmel

BioEnergy Science Center, Oak Ridge, TN 37831, USA, and National Renewable Energy Laboratory, Golden, CO 80401, USA
***yannick.bomble@nrel.gov**

Deriving renewable liquid fuels from biomass using microbial conversion, which utilizes free enzymes or cellulosomes for degrading cell wall material to sugars, is an attractive solution for today's energy challenges. The study of the structure and mechanism of these large macromolecular complexes is an active and ongoing research topic worldwide, with the goal of finding ways to improve biomass conversion using cellulosomes. Here, we present methods for illuminating the structure and function of systems of this size and complexity using molecular modeling. We show examples of these methods as applied to a range of sizes and time scales from atomistic models of enzymatic modules to coarse-grained models of the entire cellulosomal complex of scaffold and enzymes. Normal mode analysis, fluctuations, hydrogen-bond analysis of enzymes, as well as sampling techniques for cellulosome assembly are described and the results presented. For example, the mechanism of the immunoglobulin-like module of GH9 proteins is shown to be determined largely by hydrogen bond networks, and the exact hydrogen bonds were identified. Finally, by using coarse-grained modeling and parameter scanning techniques, the assembly of cellulosomal complexes is shown to be dominated by their size and shape and not by their mass.

Introduction

The most common processes for producing fuels from biomass require fermentation by either yeast or bacteria after fermentable sugars are produced. A new thrust in the field of cellulosic ethanol production is the study of microorganisms capable of converting biomass directly to fermentable products using a process known as Consolidated Bioprocesssing (CBP). Several organisms are good candidates for such a task, including *Clostridium thermocellum*, which produces large enzyme complexes known as cellulosomes. Cellusomes differ from free cellulases in the sense that most of the catalytic enzymes are strongly bound to a scaffolding protein.

The cellulosome concept was first introduced by Bayer and coworkers as the cellulase system of *C. thermocellum* (*1–3*). In most cases, the cellulosome is composed of two subunits – a non-catalytic scaffolding and the enzymes that attach to it by a cohesin-dockerin mechanism. A strong interaction exists between the multiple cohesin modules on the scaffoldin and the enzyme-borne dockerin modules (*4*, *5*). The primary scaffoldin of the cellulosome from *C. thermocellum*, cellulosome-integrating protein (CipA), contains a carbohydrate binding module (CBM), which binds strongly to plant cell wall polysaccharides and nine cohesins, and is thus able to accommodate nine different enzymes. The CBM modules are also present in some cellulosomal enzymes; for example, the processive endoglucanase CbhA, a family 9 glycosyl hydrolase (GH9) (*6*, *7*).

It has been recognized that different types of cohesins and dockerins exist in different microbial species and that the recognition between cohesin and dockerin is both type- and species-specific. Several research groups have used these findings to try to understand and improve the action of cellulosomes using a so-called "designer cellulosome" by assembling different types of cohesins from different microbial species. Bayer and coworkers (*1*, *8*, *9*) used this idea to probe two different questions: (1) do the enzyme patterns on the scaffoldin provide a synergistic action on crystalline cellulose, and (2) is there the potential to assemble enzymes from different species with superior activities on different substrates? The first engineered cellulosome was composed of two cohesins able to accommodate two cellulases (*10*, *11*). The resulting chimeras exhibited enhanced activity on crystalline cellulose over the same free cellulases. In 2005, Fierobe and coworkers created a new tri-functional engineered cellulosome by developing a third divergent cohesin-dockerin pair (*12*). The tri-functional engineered cellulosome was found to be superior in function when compared to the bi-functional one. When the tri-functional engineered cellulosome was decorated with one hemicellulase (GH10) and two cellulases, it performed with superior activity on both cellulose and hemicellulose in hatched straw.

Another aspect of great interest is the origin of the possible synergistic functions of the cellulosome. One of the main explanations for the cellulosome's performance is the flexibility of its quaternary structure. It has been shown that restricting enzyme mobility negatively affect cellulase activity, thus implying that flexibility is a key ingredient in the function of the cellulosome (*13*, *14*).

Molecular simulations are helpful for gaining a deeper understanding of the function and versatility of the CipA assembly. Knowing the enzymatic

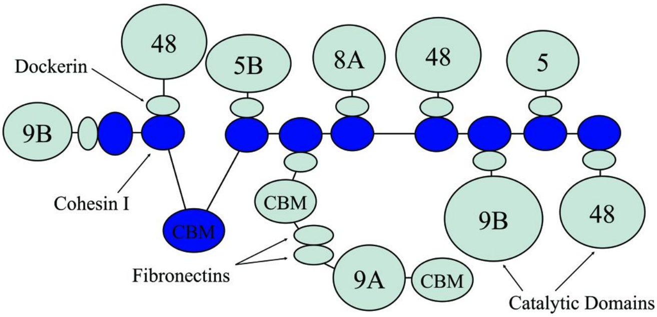


Figure 1. Concept of the first coarse-grained model for the CipA of C. thermocellum. The scaffoldin subunit (blue) contains nine cohesins and a carbohydrate binding module. The cellulolytic enzymes (grey) bind to cohesin partners with their dockerins. (see color insert)

environment necessary to attain a particular enzyme configuration on the scaffold gives insight into the way a microbial cell regulates the cellulosome population and composition near a cell wall. Probing the role of the plasticity of the cellulosome on its dynamics and self-assembly process is also an important goal. Determining the function and mode of action of the primary cellulosomal enzymes and modules may help design an improved cellulosome with improved activity. Several numerical modeling techniques can be used to answer these questions, including the more detailed all-atom molecular dynamics simulations to the less computationally expensive coarse-grained models.

Cellulosome Concept and Architecture

Cellulosomes from *C. thermocellum* can adopt different structures from the simplest three to nine-cohesin scaffoldin proteins to the more complex assemblies of multiple scaffoldin proteins organized on an additional scaffold, OlpB. In this chapter, our discussion will be solely based on the nine-cohesin stucture of CipA (Figure 1). A list of the CipA components and the cellulosomal enzymes considered in this chapter can be found in Table I.

The linkers between CipA modules vary greatly in length and are important contributors to the flexibility of the cellulosome. Cellulosomal enzymes can have simple structures, including two modules (a dockerin and a catalytic module) connected by a flexible linker, or be more complex with more than seven modules. The cellulosome is believed to bind to cellulose with the CipA-borne CBM3, but other complex enzymes whose architectures include CBMs are also believed to provide additional anchors. Moreover, many cellulosomal enzymes contain different types of carbohydrate binding modules, making them more appropriate to handle different types of substrates. Some CBMs seem to have an anchor function, whereas others have been hypothesized to be helper CBMs capable of holding a single cellulose chain and feeding it to its catalytic module partner (*15*). Several cellulosomal enzymes have protein modules with unknown function, such as immunoglobulin-like modules that are believed to stabilize

Table I. Architecture of the cellulosomal protein complexes

Protein	*Modules*	*Molecular Mass (kDa)*
CipA	2COH-CBM3a-7COH	197
Cel5B	GH5-DOC	64
Cel48A	GH48-DOC	83
CbhA	CBM4-GH9-2FN3-CBM3b-DOC	138

the catalytic modules of family 9 enzymes. Fibronectin-like modules, also known as X-domains, are another case of a protein module whose function in the cellulosome is not understood. In general, fibronectins are believed to play the role of cellulose disruptors and facilitate the digestion of cellulose.

Despite the number of different modules present in the cellulosome, its quaternary structure is stable because of the high affinity between cohesins and dockerins. As mentioned earlier, in *C. thermocellum* this affinity is non-specific, and each dockerin can equally bind to any cohesin. The type I cohesin-dockerin complex is shown in Figure 2. The recognition strip, involving two helices on the dockerin and several beta strands on the cohesin, provides an almost planar binding surface. This interaction is mediated by Ca^{2+}, which is essential for the complex to maintain structure (*4*, *16*).

The cellulosome is an amazingly complex molecular assembly that can degrade cellulose using a wide variety of enzyme combinations, which are probably adjustable as the nature of the substrate changes. Any insight into the formation and action of the cellulosome would help us understand the roles of such complex systems in the natural degradation of cellulose and cell walls by bacteria.

Function of Some Cellulosomal Modules

C. thermocellum produces a wide variety of enzyme families; among them, the family 9 enzymes are intriguing because they contain both endoglucanase and exoglucanases and can have rather complex architectures. They are divided into four groups based on their constructs (*17*), groups A through D. Group A includes enzymes containing only a catalytic module that can be linked to a dockerin. In the case of Cel9M in C. cellulyticum (*18*), group B includes an additional CBM3a located at the C-terminus (*19*). Group C includes enzymes with an immunoglobulin-like module at the N-terminus of the GH9 (*20*) catalytic module. Finally, group D includes enzymes that contain a CBM4 module and an Ig-like module at the N-terminus of the GH9 (*7*) catalytic domain.

The immunoglobulin-like module found in several of the family 9 cellulases from *C. thermocellum*, which belong to group C and D (Figure 3), is a protein module without a well-known mode of action. One of the main hypotheses for its mechanism is simply that it provides stability to the catalytic module. It has been shown that removing the Ig-like module will reduce the activity of several catalytic modules drastically (*21*). The mechanism by which the Ig provides this stability is still uncertain; and, while a possible mechanism has been proposed, there is no

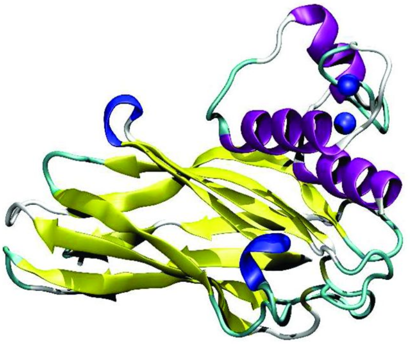


Figure 2. Structure of the cohesin-dockerin complex from the cellulosome of C. thermocellum (1OHZ) color coded by structures. (see color insert)

clear evidence supporting it. Several simulation techniques can be used to probe the hypothesized mechanism (see next section).

Several family 9 enzymes from *C. thermocellum* exhibit an immunoglobulin-like module attached to a catalytic module. Specifically, CbhA, Cel9A, and CelK have been shown to lose most of their enzymatic activity upon removing the Ig-like module. This Ig-like module consists of about 99 amino acids directly attached to the catalytic module via a interface involving close to 40 amino acids from both modules. Several studies have investigated the possible causes of such a phenomenon in CbhA. One should note that only one x-ray structure each is available for Cel9A and CbhA. Both structures exhibit the same construct, with ten hydrogen bonds at the Ig-catalytic domain interface. However, only three of the ten hydrogen bonds are conserved between CbhA and Cel9A. These three hydrogen bonds are believed to contribute to the function of the Ig-like module by stabilizing the catalytic module as well as the catalytic cleft. In both enzymes, there exist hydrogen bond networks that appear to stabilize or at least mediate catalytic residues. Both CbhA and Cel9A have a catalytic cleft with several aromatic residues able to interact with, and thus guide, a cellulose chain (Figure 4).

The hydrogen bond network described for Cel9A is shown in Figure 5. Both Thr-23 and Asp-51 form conserved hydrogen bonds with Gly-399; whereas Asp-53 forms a strong hydrogen bond with Tyr-408, which is located on a flexible loop connected to an important catalytic residue, Trp-410. Trp-410 is close to the substrate cleavage site. The experimental work of Kataeva and coworkers (in which Thr-23, Asp-51, and Asp-53 were mutated to alanyl residues) showed that several mutants could be created *in silico* to analyze the importance of each hydrogen bond on the dynamics and structure of the catalytic module. They also analyzed the configuration resulting from the removal of the Ig-like protein.

Computational Methods

Several computational methods are well suited to study these systems and span the different length scales and complexity present in cellulosomal systems.

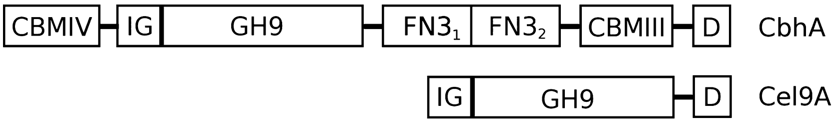


Figure 3. Constructs of CbhA and Cel9A.

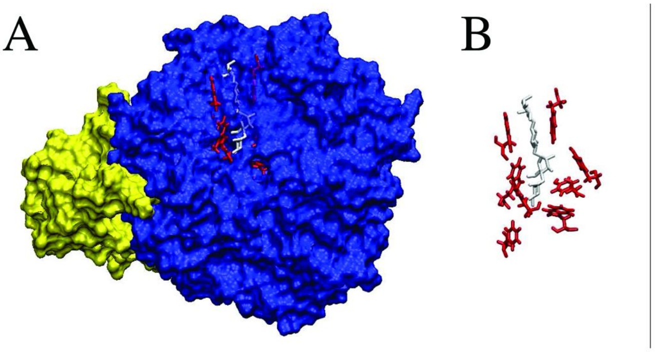


Figure 4. Surface plot of the Ig-like protein (yellow) and GH9 (Blue) with cellotetraose (white). (A) licorice representation is used to highlight the aromatic residues in the catalytic cleft. (B) Blow-up of the aromatic residues located in the catalytic cleft. (see color insert)

Here we present some strategies for using these methods to provide insight into the potential improvement of CBP microorganisms.

Coarse-Grained Modeling of the Cellulosome Assembly

Advances in computer architectures and molecular mechanics packages have allowed larger and larger simulations; systems with more than 100,000 atoms can now be routinely simulated for hundreds of nanoseconds or microseconds (*22*). Also, coarse-grained modeling has been a critical addition to the computational techniques available when simulating larger macromolecular assemblies representing millions of atoms by utilizing a reduction in the number of particles by a factor of up to 10-20 (*23–26*). While these techniques are useful, they are not well suited to the study of the formation of large macromolecular assemblies, such as cellulosomes. To understand how the cellulosome assembles close to the cell wall in a free-enzyme bath, we plan to conduct hundreds of simulations on the timescale of hundreds of nanoseconds with more than 1 million atoms. We will use the coarse-grained model proposed here to attempt to capture the most essential properties of the cellulosome and predict how these intrinsic properties will govern the enzyme configuration on the CipA scaffold. We also hope to gain insight into the dynamics of the cellulosome during and after its initial formation.

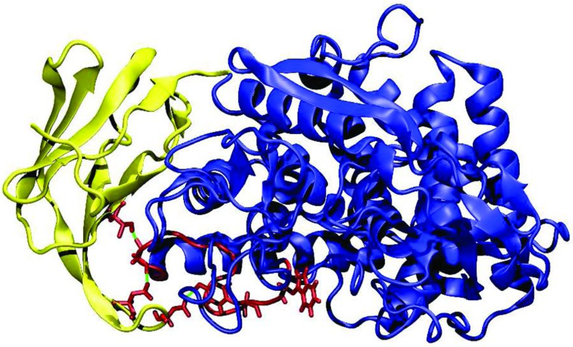


Figure 5. Structure of the catalytic (blue) and immunoglobulin-like (yellow) modules from Cel9A. The hydrogen bonding network, including an important catalytic residue, is shown in red and the three hydrogen bonds in green. (see color insert)

Functional Form and Parameters

The protein structure model consists of large spheres, called "beads," representing large regions of protein volume, up to 30 Angstroms radius, that are held together by a network of restraints to mimic the shape and flexibility of globular proteins, dockerins, cohesins, and linkers. These beads have no charge, and there is very little attractive potential between the beads. Each sphere, or bead, represents from three amino acids for linker regions to tens of amino acids in large globular protein regions. The restraints between beads are defined to be as simple as single bonds between beads in a linker, to networks of bonds between beads in globular-shaped protein modules. Special interactions are included to mimic the attraction of dockerins for cohesins. The model was developed within CHARMM (a molecular mechanics program package) (*27*). The CHARMM package offers considerable flexibility to the user for creating new pseudo atoms, has functionality for specific non-bond interactions between particular atom types, and allows additional parameters to be specified in the topology and parameters files.

Within our template, the interactions between coarse-grained beads can be expressed as a sum of traditional classical bonded and non-bonded terms as follows.

Non-Bonded Terms

The non-bonded interactions are represented by a 6-12 Lennard-Jones (LJ) potential energy function,

$$(1)\quad E_{nb} = \sum_{i,j>i} \varepsilon_{ij} \left[\left(\frac{r_{\min}}{r} \right)^{12} - 2\left(\frac{r_{\min}}{r} \right)^{6} \right]$$

where $r_{\min}$ represents the closest distance of approach between two particles, ε_{ij} is the strength of their interaction, and r is the distance between two pseudo atoms. The vdW radii are defined to accurately reproduce the radii of the module represented by the pseudo atoms, and the interaction is defined to produce a shallow LJ potential well, so as to avoid unnatural attractions between pseudo atoms. The coarse-grained beads approximate hard spheres that have limited interactions with one another.

The electrostatic effects were neglected in our model because of the limited number of pseudo atoms or beads per protein (Figures 6−11). A specific interaction was added between the pseudo atoms in the binding site of the cohesin and dockerin proteins using an additional set of non-bonded parameters between specific atom pairs. The binding energy was set to 13 kcal.mol^{-1}, a value between the experimental (12 kcal.mol^{-1}) (*5*) and theoretically determined value of 14.5 kcal.mol^{-1} (*28*).

Bonded Terms

The bonded interactions are defined by the internal energy terms,

$$(2)\quad E_b = \sum k_r (r - r_0)^2 + \sum k_\theta (\theta - \theta_0)^2 + \sum k_\varphi (1 + \cos(\varphi - \varphi_0))$$

where r, θ, and φ are the distance, angle, and torsional angles between connected coarse-grained beads; r_0, θ_0, and φ_0 are the coarse-grained bond, angle, and torsional angle equilibrium values; and k_r, k_θ and k_φ are the force constants. The force constants between beads of the same module are large, making the substructure rigid, while inter-modular linker regions have a wide range of flexibility. The distance, angles, and torsional angles were chosen to fit the original (all-atom) structure.

Scaffold Subunit

The polymeric scaffold of *C. thermocellum* CipA, includes nine cohesin proteins connected by linker peptides of 10–30 amino acids in length and an additional carbohydrate binding module, CBM3 (Figure 6 and Figure 7). To provide the flexibility of the all-atom structure, each linker bead in the coarse-grained representation represents three amino acids (Figure 7). The all-atom and the coarse-grained representations of the full-length CipA are shown in Figure 7 and Figure 6, respectively. The linker regions offer the plasticity required by the cellulosome to assume the most appropriate configuration given a particular substrate. There is a clear need for a finer grained representation of these linkers than the coarser grained representation of the other components.

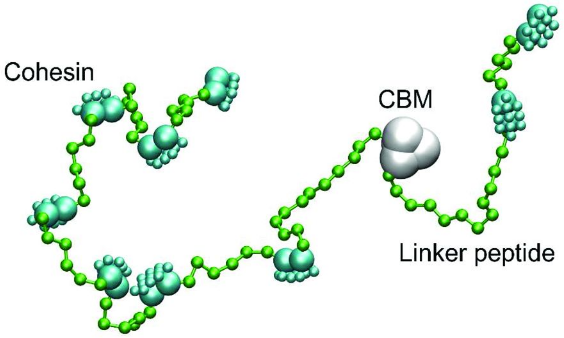


Figure 6. Coarse-grained representation of CipA from C. thermocellum. (see color insert)

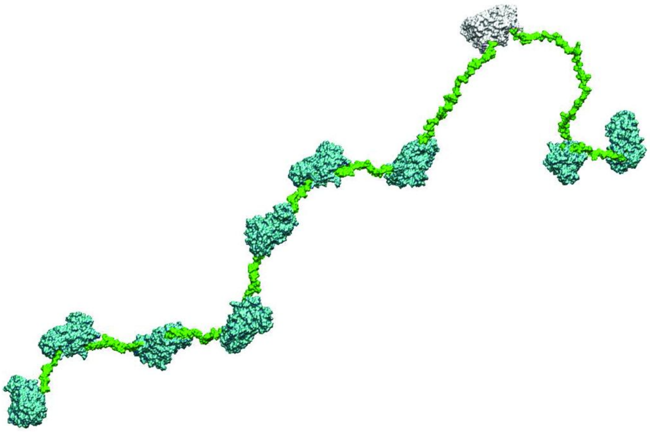


Figure 7. All-atom representation of CipA from C. thermocellum. The structure of one of the cohesins is known and reported in the literature. The other cohesins were obtained from homology modeling. (see color insert)

Cohesin and Dockerin

The cohesins have a flat binding surface able to interact with the dockerin subunits of the cellulosomal enzymes. The architecture of the coarse-grained cohesin was conceived to accurately describe the binding interaction and create a flat binding surface while conserving the overall van der Waals volume of the protein module (Figure 8). The dockerin is constructed with a mating flat surface to match the cohesin. There are three special “attractor beads” in a row across the center of the mating surface of the cohesin and dockerin that are given special attracting properties for each other. The attractor beads are surrounded on the backside of the mating surface by beads that prevent multiple bindings to the same cohesin or dockerin simply by steric hindrance.

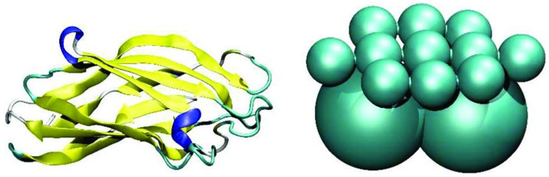


Figure 8. All-atom and coarse-grained representations of the cohesin from CipA. (see color insert)

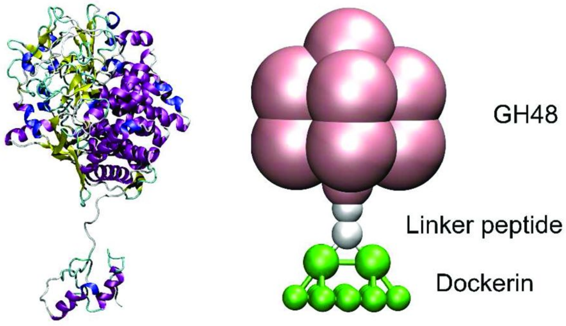


Figure 9. All-atom and coarse-grained representations of Cel48A. (see color insert)

Cellulosomal Enzymes

As mentioned earlier, *C. thermocellum* is able to produce a wide variety of enzymes with different architecture and complexity. Three of these enzymes were selected in our study: the exocellulase Cel48A, the endoglucanase Cel5B, and the processive endoglucanase CbhA. They essentially encompass the complexity of the cellulosomal enzymes found in *C. thermocellum*. The construct details for these enzymes and the scaffoldin protein can found in Table I. The linkers between modules vary greatly in length, between 3–10 amino acids. Cel5B and Cel48A have a rather simple architecture including a catalytic module, a linker, and a dockerin. CbhA is a much more complex modular protein, including modules with mostly unknown functions, such as fibronectin-like (*7*, *29*) and immunoglobulin-like modules (*7*, *21*), as well as two types of carbohydrate binding modules, CBM3b and CBM4 (*7*). All of the enzymes studied here have a dockerin protein capable of binding to any cohesin on the scaffold without specificity; and the coarse-grained model dockerins, similar to cohesins, have an engineered flat binding platform. The coarse-grained representations of these enzymes are shown in Figures 9−11 along with their all atom counterparts. Note that the shape of the enzymes is accurately reproduced; and we should be able to model some important properties in our simulations, such as volume exclusion, mass effects, and flexible linkers.

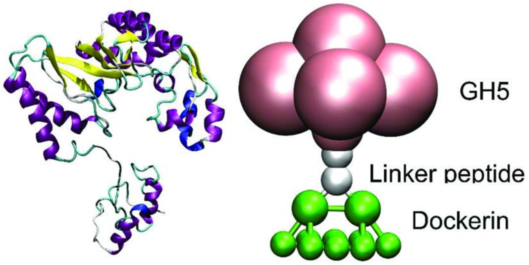


Figure 10. All-atom and coarse-grained representations of Cel5B. (see color insert)

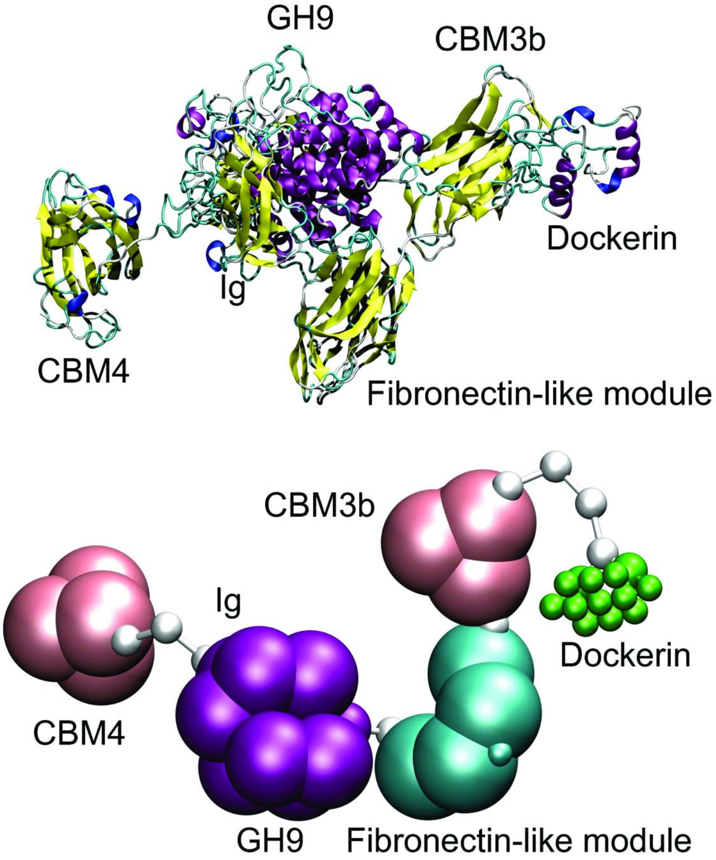


Figure 11. All-atom and coarse-grained representations of CbhA. (see color insert)

Experimental Setup

All simulations were conducted with the CHARMM package. We tried to reproduce the enzymatic environment around the scaffoldin close to the cell wall. The simulation box has a volume of $1x10^9$ Å^3 (1000Å x 1000Å x 1000Å) (Figure 12). The total enzyme concentration varies from 30–120 total enzyme molecules per scaffoldin molecule and per box. The initial configurations were always randomly generated, and different random seeds for both the initial positions and the initial velocities were used to reproduce the random nature of the enzymatic environment and to eliminate the possibility of biases in our results. Initial simulations were performed with the full-length scaffold (9 cohesin). However, for clarity, the second part of this study used a 4-cohesin scaffold. Periodic boundary conditions employing a cubic box with sides measuring 1000 Å ensured a fixed concentration of enzymes in each simulation. Non-bonded interactions were cut at 99 Å, and the individual snapshots were registered every 1000 steps. Each trajectory was equilibrated for 100,000 steps with a time step of 2 fs, and trajectories were run for 30–100 ns. In our subsequent binding studies, we performed 30 simulations of 30-ns duration for each different configuration in which total concentration, ratio of enzymes, or binding constants were varied to achieve meaningful statistical analysis.

Results and Discussion

First observations were made using a 9-cohesin scaffold in the simulation box without any enzymes in solution. The scaffold adopts compact configurations reminiscent of the TEM images by Mayer and coworkers (*30*). Starting from an extended configuration, the scaffold tends to adopt a more compact form. In this configuration, the scaffold may be more shielded from the outside, which might explain results found by Bayer and coworkers (*31*). They showed that removing enzymes docked on the scaffold was easier when the cellulosome was bound on cellulose where it would adopt a more extended configuration, but much harder when free in solution.

The second observations were made when enzymes were added to the system. When an equal ratio of each enzyme is added for a total enzyme count of 60, the scaffold is fully populated with enzymes within less than 50 ns. The behavior of CipA is greatly modified whenever CbhA binds to a cohesin, which is caused by the large mass of that enzyme; but is not as affected by the binding of smaller enzymes. CbhA seems to lock the scaffold in a given location and prevents it from freely diffusing through the box the way it did before binding occurred. This behavior contributes to the nature of sequential binding of enzymes on the remaining binding sites, because the scaffold will not be able to diffuse freely. Also, the volume excluded by the first enzymes binding to the scaffold is a contributing factor in defining the probability of other enzymes binding.

The main focus of this study was to understand the driving forces behind the different cellulosome configurations. As mentioned above, we focused on a 4-cohesin scaffold without the CBM protein. Some of the results from competitive

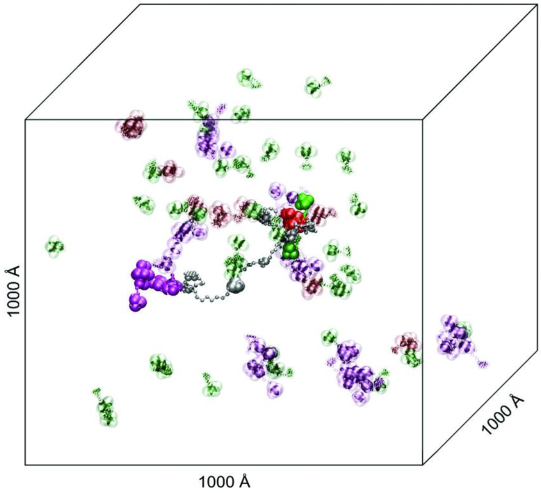


Figure 12. Simulation box with a scaffoldin molecule and 60 cellulosomal enzymes. The enzymes bound on the scaffold have solid colors. (see color insert)

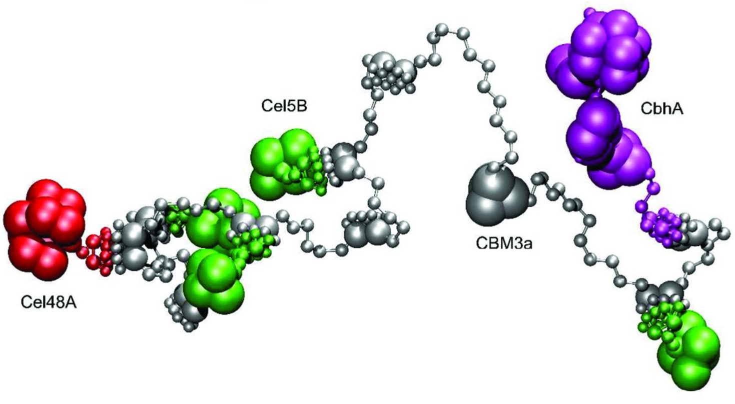


Figure 13. Coarse-grained representation of a partially populated scaffold. (see color insert)

binding studies between three cellulosomal enzymes are summarized in Table II. CbhA tends to bind to the scaffoldin protein more significantly than Cel48A and Cel5B. The size or flexibility of CbhA could be responsible for this behavior; and subsequent studies varying size, mass, and radius of gyration of a given enzyme will help to understand this phenomenon.

A detailed parameter scan of the total concentration of enzymes and enzyme ratio is being conducted and will shine more light on the binding dynamics of these enzymes that represent more than 1000 independent calculations. Response surface methodology will be used to define the environment necessary for a particular cellulosome configuration. Because of its modularity, it appears that the CbhA enzyme doesn't diffuse as quickly as the Cel5B and Cel48A because of its increased

Table II. Average cellulosome population arising from 30 replicated runs for a given ratio of enzymes in the simulation box

Enzyme in solution Cel5B/Cel48A/CbhA (percentage)	33/33/33	41/41/18	50/50/0
Enzyme on the scaffold (percentage)	20/25/55	33/36/31	45/55/0

number of internal motions and therefore has more time to “feel” a cohesin partner. However, the results shown in Table II already indicate that this model could provide great insights into the cellulosome self-assembly and how the cell might regulate its scaffold configuration. There is even the possibility that the binding behavior of CbhA could be linked to the expensive nature of its construction, and that the cell doesn’t need to secrete large amounts of this enzyme to be significantly present in cellulosomal assemblies.

Cellulosomes may attain their activity through their plasticity and special arrangements of the enzymes on the scaffold. Coarse-grained modeling proved to be an adequate tool to study these phenomena. However, more detailed simulations are needed to truly understand the interaction of the cellulosome with cellulose and the function of each individual protein involved in the hydrolysis process. These proteins include catalytic modules, carbohydrate binding modules, and modules such as the fibronectin-like or X domains. Several of these proteins, such as the fibronectins, have an unknown function, and others seem to have functions that differ from their fungal counterparts. In particular, several of the cellulosomal CBMs seem to have a unique function. In the next section, we study the family 9 enzyme of *C. thermocellum*, which contains many of these protein modules with different physical and chemical properties.

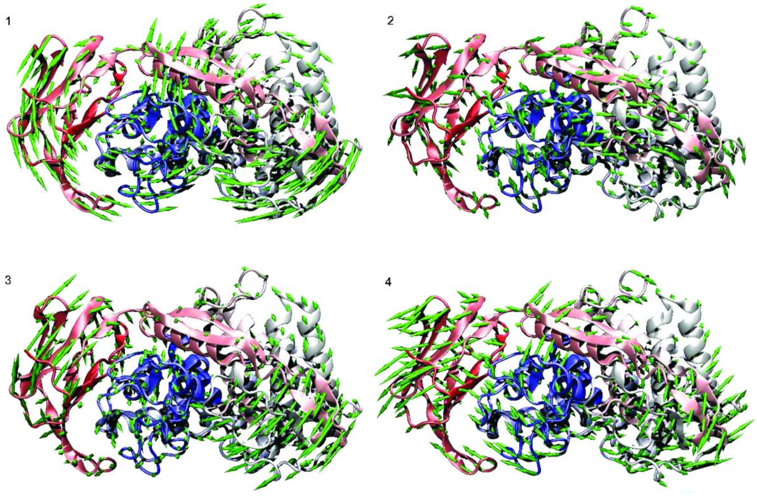


Figure 14. First four normal modes of Ig-Gh9 modules for Cel9A. The structure is color coded by amino acid sequence number. (see color insert)

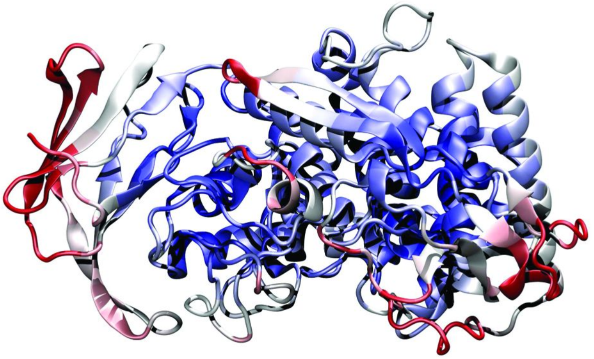


Figure 15. Stucture of Ig-Gh9 from Cel9A color coded by fluctuations (increasing from blue to red) using the first 300 normal modes. (see color insert)

Normal Mode Analysis of Cel9A

Normal mode analysis (NMA) (*32–35*) provides a computationally inexpensive way to study large-scale behaviors of molecular assemblies. NMA has several advantages over classical molecular dynamics (MD), even though it approximates the global potential by a harmonic function (*34*). First, it provides a clearer representation of the collective motions of biomolecules through a few of the lowest energy vibrational modes. Second, it makes evaluating entropy contributions and other thermodynamic properties straightforward. Finally, it is more affordable when long timescales are required for sampling times sufficient to display the low-frequency modes. While it is common practice to use elastic-network model or all-atom normal mode analysis in gas phase to approach this problem, some of the finer details may be lost in the process. Recently, NMA was extended to take advantage of the popular generalized born theory for implicit treatment of solvation effects. This new implementation (*36*, *37*) was applied to long nucleic acid duplexes and was shown to accurately describe large-scale properties of these duplexes (*37*). The same method can be used as a first approach to gather information about the possible function of the Ig-like module as well as the mechanism by which GH9 endoglucanases may acquire a cellulose chain before hydrolysis of the 1,4-beta-D-glucosidic linkage.

The normal mode analyses were carried out with the molecular mechanics program package NAB (*38*, *39*), now part of Amber10 (*40*, *41*) ambertools using the parameter set parm99SB (*42*, *43*); and we used the pairwise approach of Hawkins and coworkers for the Generalized Born (GB) model (*44*, *45*). The structures were minimized using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno Truncated Newton Conjugate minimization technique to obtain a root mean square (RMS) gradient below 1 x 10^{-8} kcal/mol-Å. This level of convergence is necessary to avoid contamination from translational and rotational modes into true internal modes. The diagonalization of the Hessian matrix was done using the ARPACK (*46*) routines in combination with a Cholesky decomposition and inversion of the Hessian matrix, therefore providing a better separation of the eigenvalues to enhance convergence. The analysis of the normal modes was performed with a modified version of the program PTRAJ with additional functionalities. The first four normal modes of the Ig-GH9 module are shown in Figure 14 using a porcupine representation. It is commonly acknowledged that the first 10-20 normal modes are enough to describe the large-scale dynamics of a given molecule. Twenty normal modes were enough to converge the root-mean square fluctuations (RMSF) shown in Figure 16, and the first five modes dominate the fluctuations. The most dominant mode (mode1) shows a hinge motion opening the catalytic cleft around the substrate chain. The dominant motion could shine some light on the possible mechanism by which the enzyme acquires a cellulose chain before catalysis. The other normal modes are more localized, but still show a lot of motion at the bottom of the cleft as well as the flexible nature of the Ig module with respect to the CD module and within itself. Also shown is another hinge motion between the Ig and CD modules, with the hinge being the linker between the two modules. Figure 15 shows the flexible regions of the Ig-CD construct for Cel9A as determined from residue fluctuations. CbhA exhibits the same basic frequency modes and overall fluctuations as Cel9A. The high flexibility regions include loops and alpha helices at the bottom of the catalytic module close to the substrate. The atomic fluctuations calculated using the normal modes for Cel9A agree with with the atomic fluctuations calculated from crystallographic temperature factors β_i using Equation 3 and are compared in Figure 16.

$$(3)\quad \left\langle (r_i)^2 \right\rangle = \left(\frac{3\beta_i}{8\pi^2} \right)$$

$$(4)\quad Corr(i,j) = \frac{\left\langle \Delta r_i \bullet \Delta r_j \right\rangle}{\sqrt{\left\langle (\Delta r_i)^2 \right\rangle \bullet \left\langle (\Delta r_j)^2 \right\rangle}}$$

$$(5)\quad \left\langle \Delta r_i \bullet \Delta r_j \right\rangle = \sum_{k=7}^{3N} \frac{k_b T}{\lambda_k} \frac{d_{ik} d_{jk}}{\sqrt{m_i m_j}}$$

$$(6)\quad \left\langle (\Delta r_i)^2 \right\rangle = \sum_{k=7}^{3N} \frac{k_b T}{\lambda_k} \frac{{d_{ik}}^2}{m_i}$$

While the amplitudes of the fluctuations are not necessarily important, in contrast the relative fluctuations are a more relevant comparison to b-factors. In this case, they describe the main features well. The relative fluctuations agree with experimental measurements of B-factors. This is reassuring and supports the accuracy of the NMA protocol used here.

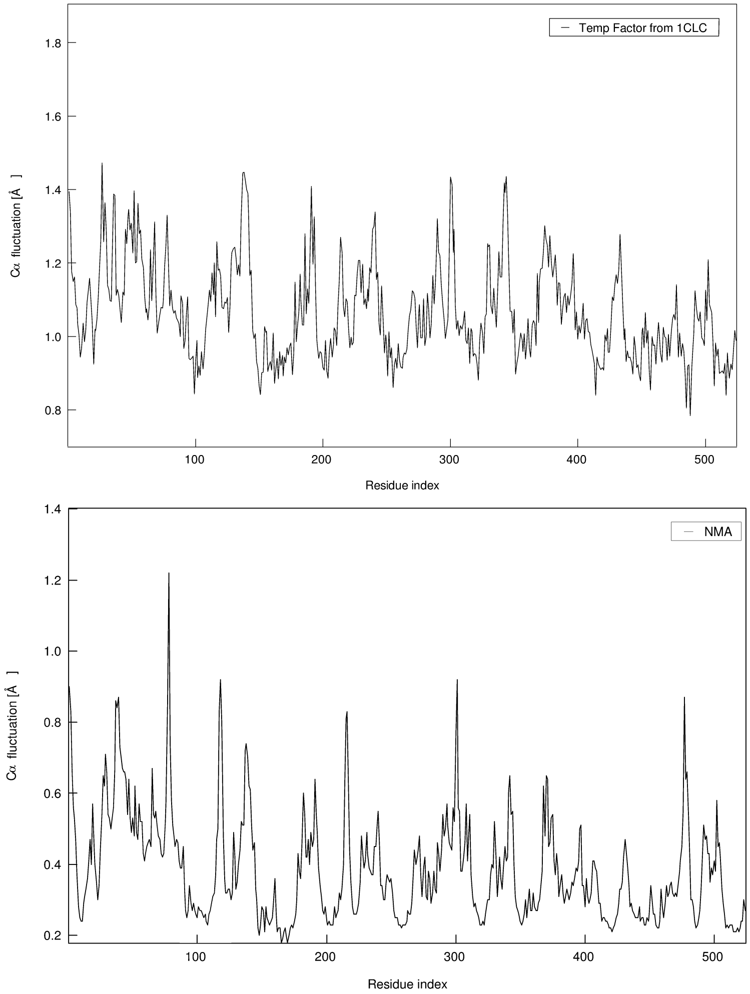


Figure 16. Atomic fluctuations for Ig-Gh9 (Cel9A-1CLC) from crystallographic temperature factor and from normal mode analysis using the first 300 normal modes. The amplitutes are in Angström.

The eigenvalues and eigenvectors can also be used to describe the correlation of motion of different protein modules. This is described by Equations 4−6 where d_{ik} and d_{jk} are the vector displacements for the kth mode and atom i or j, respectively. The cross-correlation maps of Ig-GH9 calculated for Cel9a are shown in Figure 17. The immunoglobulin-like module shows a strong correlation of motion within itself, probably due to the fact that it is composed of beta strands with strong interactions. One of the most interesting features of these maps is the fact that the Ig module, or at least several residues within the module, appear to have a strong correlation of motion with several residues of the catalytic module, including a strong positive correlation with residues 389 to 410 and also several other important loops within the vicinity of the catalytic cleft. This supports the hypothesis that these loops are closely coupled with the Ig module and that the removal of Ig or selected mutations in Ig may interfere with the dynamics of the catalytic residues, especially amino acid 410. However, a more careful investigation is required to unambiguously prove the function of the Ig module. NMA at least shows that the hypothesis mentioned earlier is relevant and deserves to be studied with a more time-consuming method such as MD simulations.

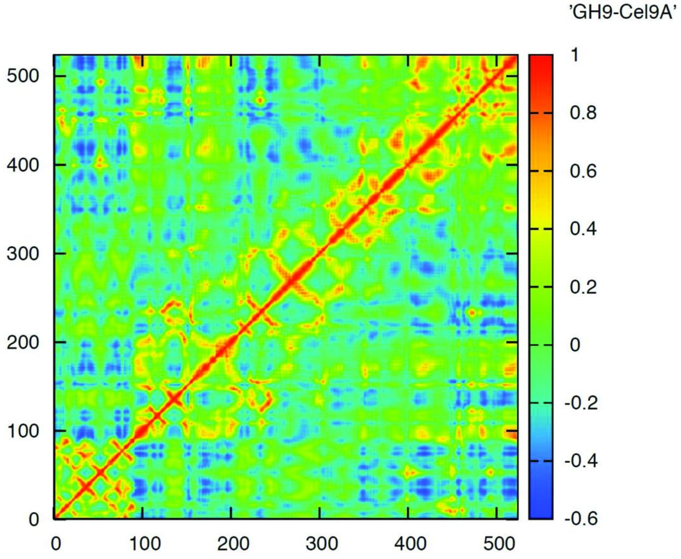


Figure 17. Residue cross-correlation map of Ig-Gh9 for Cel9A. A value of 1 shows a correlation of motion, while -1 is indicative of anti-correlation of motion, and zero represents a total lack of motion correlation. This map was calculated using the first 300 normal modes. (see color insert)

Molecular Dynamics Simulations

MD simulations were used to address the aforementioned problem – the function of the Ig module in several family 9 enzymes – in more detail using a set of analysis tools demonstrated in similar studies (*47*). All simulations in this section were carried out using the program, PMEMD, from Amber 10 and the parameter set parm99SB (*42*, *43*). The proteins were solvated in a truncated octahedral box of TIP3P water molecules extending to 12 Å from the surface of the protein. A simulation time step of 2 fs was used along with SHAKE (*48*) to constrain covalent bonds between heavy and hydrogen atoms. The particle mesh Ewald method was used along with a non-bonded cutoff of 12 Å. The calcium ions were kept in their original positions from the pdb files, and the parameters usec for the calcium ions were taken from Aqvist (*49*). After equilibration, 15 ns of unconstrained MD were performed for dynamic sampling of states. Three replicates of the same starting configuration were run with different initial velocities to check the convergence of the fluctuations and other properties extracted from the trajectory and to insure proper statistical sampling. Removing the rotations and translations from the trajectories was done by rmsd, fitting the trajectory to the backbone of the entire protein in its initial post-equilibration configuration. Using a selected area of low mobility of the protein as inferred by NMA for rmsd fitting resulted in comparable findings. However, closer inspection of the cross-correlation map in Figure 19 shows that the rms fitting procedure is of crucial importance – as Ichiye and Karplus pointed out (*50*) – where a poor choice of rms fitting parameters can result in a loss of details in such map. It was clear from the map computed in this work that even the best set of parameters does not offer as much constrast as provided by normal mode analysis.

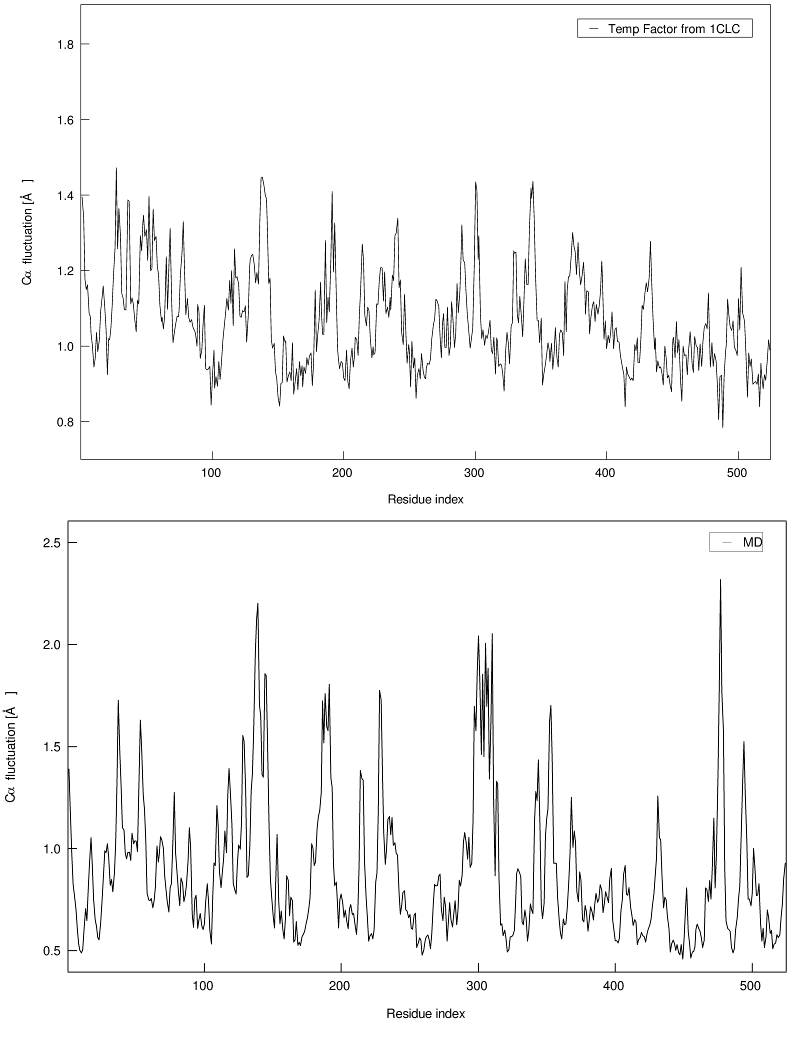


Figure 18. Atomic fluctuations for Ig-Gh9 (Cel9A-1CLC) from crystallographic temperature factor and from 15ns of molecular dynamics simulation.

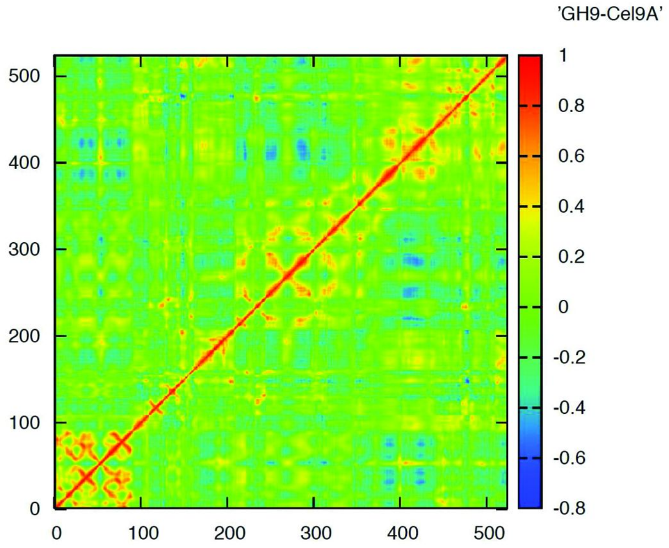


Figure 19. Residue cross-correlation maps of Ig-Gh9 Cel9A from 30 ns of molecular dynamics simulations. A value of 1 shows a correlation of motion, while -1 is indicative of an anti-correlation of motion, and zero represents a total lack of motion correlation. (see color insert)

The RMS fluctuations of the Cα atom of the wild-type Ig-GH9 are in as good agreement with those calculated from crystallographic temperature factors (Figure 18) as the fluctuations calculated from the normal mode analysis. The fluctuations from the three replicas are almost indiscernible, except for a few flexible loops where the results are not as consistent. Given the overall consistency of the results, any difference in fluctuations caused by mutation can be linked to the effect of the mutation. It is worth noticing that the fluctuations calculated from MD are overestimated, as is always the case in the literature.

Before starting experimental mutational studies, it is important to know which amino-acid residues are most likely to impact the structure or dynamics of the catalytic module. As mentioned earlier, three hydrogen bonds in Cel9A at the Ig-GH9 interface are conserved but their respective stability is unknown. The analysis program, PTRAJ, was used to follow the stability of those hydrogen bonds during 15 ns of MD simulations (Figure 20). It appears that only Asp-53 is able to create strong hydrogen bonds between the Ig and CD module in Cel9A. Thr-23 is also able to create a stable hydrogen bond in Cel9A. This analysis shows that Asp-51 is unable to strongly interact with the catalytic module as previously thought. It appears that only one or two of these conserved hydrogen bonds are good candidates for mutagenesis. A similar investigation for the remaining hydrogen bonds is being conducted; and even though these hydrogen bonds are not evolutionarily conserved, they most likely contribute to the interaction between Ig and the catalytic cleft.

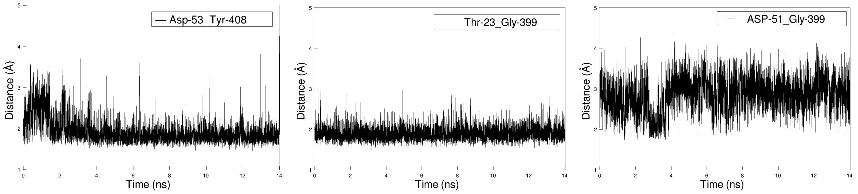


Figure 20. Distance between atoms involved in several hydrogen bonds between the Ig and catalytic modules over 15ns of simulation for Cel9A.

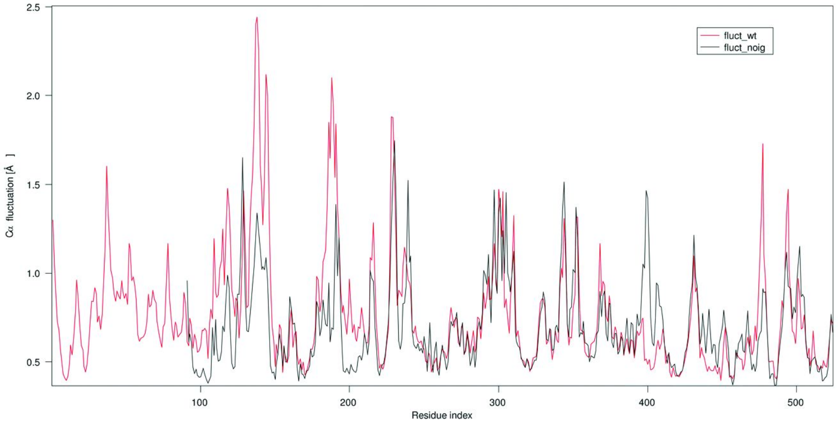


Figure 21. Atomic fluctuations for Ig-Gh9 (Cel9A) for the wild type and after ablation of the Ig module from 15ns of molecular dynamics simulation. (see color insert)

The effect of the extreme case of the Ig module's total removal is shown in Figure 21, where the fluctuations of the Cα atoms for Ig-GH9 and GH9 in Cel9A seem to present interesting differences in the vicinity of residues 390 to 425 as well as other less relevant loops. The features of the fluctuations appear to be substantially different and are not only restricted to a difference in the amplitude of a single peak. It would be encouraging to see the same behavior in some of the mutational studies for conserved or not conserved hydrogen bonds, as it would validate the hypothesis presented here. It is clear that dynamics of some of the residues inside the catalytic cleft are being perturbed, although it is not yet clear how this could affect the correct functioning of the enzyme. Substantial conformational changes have not been observed in these rather short simulations. Longer simulations with a generalized born model are being conducted as well as clustering analysis of the trajectory to better understand the difference in states visited for the wild type and mutated enzyme.

Conclusions

Whereas the results from normal mode analysis and molecular dynamics simulations to date are not enough to provide a definite answer about the function of the immunoglobulin-like module or the mode of action of the

GH9 endoglucanases, they do seem to show the close relationship between the catalytic cleft and the Ig module. These computational tools demonstrate that the hypothesis presented several years ago is viable and that more careful analysis of this problem is not only needed, but worthwhile. Understanding the function of each individual protein (modules) of the *C. thermocellum* cellulosome is essential for improving the microorganism's performance in terms of biofuels production. Such understanding would impact both the improvement of the enzymes as well as cellulosomes.

Acknowledgments

This work was supported by the DOE Office of Science, Office of Biological and Environmental Research, through the BioEnergy Science Center (BESC), a DOE Bioenergy Research Center. Computational time for this research was supported in part by the Golden Energy Computing Organization at the Colorado School of Mines using resources acquired with financial assistance from the National Science Foundation and the National Renewable Energy Laboratory. Simulations were also performed in part using the Texas Advanced Computing Center Ranger cluster under the National Science Foundation Teragrid grant number TG-MCB090159. The authors would also like to thank Professor Ed Bayer for helpful discussions.

References

1. Bayer, E. A.; Belaich, J. P.; Shoham, Y.; Lamed, R. *Annu. Rev. Microbiol.* **2004**, *58*, 521–554.
2. Demain, A. L.; Newcomb, M.; Wu, J. H. D. *Microbiol. Mol. Biol. Rev.* **2005**, *69* (1), 124+.
3. Doi, R. H.; Kosugi, A. *Nat. Rev. Microbiol.* **2004**, *2* (7), 541–551.
4. Chauvaux, S.; Beguin, P.; Aubert, J. P.; Bhat, K. M.; Gow, L. A.; Wood, T. M.; Bairoch, A. *Biochem. J* **1990**, *265* (1), 261–265.
5. Carvalho, A. L.; Dias, F. M. V.; Nagy, T.; Prates, J. A. M.; Proctor, M. R.; Smith, N.; Bayer, E. A.; Davies, G. J.; Ferreira, L. M. A.; Romao, M. J.; Fontes, C. M. G. A.; Gilbert, H. J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (9), 3089–3094.
6. Schubot, F. D.; Kataeva, I. A.; Chang, J.; Shah, A. K.; Ljungdahl, L. G.; Rose, J. P.; Wang, B. C. *Biochemistry* **2004**, *43* (5), 1163–1170.
7. Zverlov, V. V.; Velikodvorskaya, G. V.; Schwarz, W. H.; Bronnenmeier, K.; Kellermann, J.; Staudenbauer, W. L. *J. Bacteriol.* **1998**, *180* (12), 3091–3099.
8. Bayer, E. A.; Shimon, L. J. W.; Shoham, Y.; Lamed, R. *J. Struct. Biol.* **1998**, *124* (2−3), 221–234.
9. Bayer, E. A.; Lamed, R.; Himmel, M. E. *Curr. Opin. Biotechnol.* **2007**, *18* (3), 237–245.
10. Fierobe, H. P.; Mechaly, A.; Tardif, C.; Belaich, A.; Lamed, R.; Shoham, Y.; Belaich, J. P.; Bayer, E. A. *J. Biol. Chem.* **2001**, *276* (24), 21257–21261.

11. Fierobe, H. P.; Bayer, E. A.; Tardif, C.; Czjzek, M.; Mechaly, A.; Belaich, A.; Lamed, R.; Shoham, Y.; Belaich, J. P. *J. Biol. Chem.* **2002**, *277* (51), 49621–49630.
12. Ding, S. Y.; Bayer, E. A.; Steiner, D.; Shoham, Y.; Lamed, R. *J. Bacteriol.* **1999**, *181* (21), 6720–6729.
13. Gilbert, H. *J. Mol. Microbiol.* **2007**, *63* (6), 1568–1576.
14. Hammel, M.; Fierober, H. P.; Czjzek, M.; Kurkal, V.; Smith, J. C.; Bayer, E. A.; Finet, S.; Receveur-Brechot, V. *J. Biol. Chem.* **2005**, *280* (46), 38562–38568.
15. Jindou, S.; Xu, Q.; Kenig, R.; Shulman, M.; Shoham, Y.; Bayer, E. A.; Lamed, R. *FEMS Microbiol. Lett.* **2006**, *254* (2), 308–316.
16. Lytle, B. L.; Volkman, B. F.; Westler, W. M.; Wu, J. H. D. *Arch. Biochem. Biophys.* **2000**, *379* (2), 237–244.
17. Bayer, E. A.; Shoham, Y.; Lamed, R. In *The Prokaryotes, an Evolving Electronic Resource for the Microbiological Community*; 3rd ed.; Dvorkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H., Stackebrandt, E., Eds.; Springer: New York, NY, 2000; pp 1–41.
18. Belaich, A.; Parsiegla, G.; Gal, L.; Villard, C.; Haser, R.; Belaich, J. P. *J. Bacteriol.* **2002**, *184* (5), 1378–1384.
19. Sakon, J.; Irwin, D.; Wilson, D. B.; Karplus, P. A. *Nat. Struct. Biol.* **1997**, *4* (10), 810–818.
20. Chauvaux, S.; Souchon, H.; Alzari, P. M.; Chariot, P.; Beguin, P. *J. Biol. Chem.* **1995**, *270* (17), 9757–9762.
21. Kataeva, I. A.; Uversky, V. N.; Brewer, J. M.; Schubot, F.; Rose, J. P.; Wang, B. C.; Ljungdahl, L. G. *Protein Eng., Des. Sel.* **2004**, *17* (11), 759–769.
22. Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, *94* (10), L75–L77.
23. Noid, W. G.; Liu, P.; Wang, Y.; Chu, J. W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. *J. Chem. Phys.* **2008**, *128* (24), 244115.
24. Noid, W. G.; Chu, J. W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128* (24), 244114.
25. Liu, P.; Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111* (39), 11566–11575.
26. Villa, E.; Balaeff, A.; Mahadevan, L.; Schulten, K. *Multiscale Model. Simul.* **2004**, *2* (4), 527–553.
27. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4* (2), 187–217.
28. Xu, J. C.; Crowley, M. F.; Smith, J. C. *Protein Sci.* **2009**, *18* (5), 949–959.
29. Kataeva, I. A.; Seidel, R. D.; Shah, A.; West, L. T.; Li, X. L.; Ljungdahl, L. G. *Appl. Environ. Microbiol.* **2002**, *68* (9), 4292–4300.
30. Mayer, F.; Coughlan, M. P.; Mori, Y.; Ljungdahl, L. G. *Appl. Environ. Microbiol.* **1987**, *53* (12), 2785–2792.
31. Morag, E.; Yaron, S.; Lamed, R.; Kenig, R.; Shoham, Y.; Bayer, E. A. *J. Biotechnol.* **1996**, *51* (3), 235–242.
32. Tama, F. *Protein Pept. Lett.* **2003**, *10* (2), 119–132.
33. Janezic, D.; Brooks, B. R. *J. Comput. Chem.* **1995**, *16* (12), 1543–1553.

34. Case, D. A. *Curr. Opin. Struct. Biol.* **1994**, *4* (2), 285–290.

35. Brooks, B. R.; Janezic, D.; Karplus, M. *J. Comput. Chem.* **1995**, *16* (12), 1522–1542.

36. Brown, R. A.; Case, D. A. *J. Comput. Chem.* **2006**, *27* (14), 1662–1675.

37. Bomble, Y. J.; Case, D. A. *Biopolymers* **2008**, *89* (9), 722–731.

38. Macke, T. A.; Case, D. A. In *Molecular Modeling of Nucleic Acids*; Leontes, N. B., Santa Lucia, J., Jr., Eds.; American Chemical Society: Washington, DC, 1998; pp 379–393.

39. Macke, T.; Svrcek Seiler, W. A.; Brown, R. A.; Kolossvary, I.; Bomble, Y. J.; Case, D. A. *NAB*, Version 6.

40. Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; B., W.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*, University of California, San Francisco.

41. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.

42. Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21* (12), 1049–1074.

43. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65* (3), 712–725.

44. Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246* (1−2), 122–129.

45. Hawkins, C. J.; Cramer, G. D.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

46. Lehoucq, R. B.; Sorensen, D. C.; Yang, C. Presented at SIAM, Philadelphia, PA, 1999.

47. Gohlke, H.; Kuhn, L. A.; Case, D. A. *Proteins: Struct., Funct., Bioinf.* **2004**, *56* (2), 322–337.

48. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23* (3), 327–341.

49. Aqvist, J. *J. Phys. Chem.* **1990**, *94* (21), 8021–8024.

50. Ichiye, T.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1991**, *11* (3), 205–217.

Chapter 5

Meso-Scale Modeling of Polysaccharides in Plant Cell Walls: An Application to Translation of CBMs on the Cellulose Surface

Lintao Bu,[1,*] Michael E. Himmel,[2] and Mark R. Nimlos[1]

[1]National Bioenergy Center, National Renewable Energy Laboratory,
Golden, CO 80401
[2]Biosciences Center, National Renewable Energy Laboratory, Golden,
CO 80401
*Lintao.bu@nrel.gov

A coarse-grained model and force field for simulating cellulose Iβ surface (1,0,0) was derived, in which each β-D-glucose unit is represented by three beads. The coarse-grained model can reproduce a stable cellulose (1,0,0) surface with an excellent agreement with an all-atom model. When used to study the interaction of the family 1 carbohydrate-binding module (CBM1) with this cellulose surface model, the CBM "opens" as in earlier atomistic simulations. This cellulose Iβ surface model produces simulations in which the CBM translates along a broken cellodextrin chain. This processive motion of the exoglucanase cellobiohydrolase I has long been suggested by experimental studies, but has never before been observed in computer simulations.

Introduction

Utilizing biomass as a renewable energy resource typically requires degrading the plant matter into component chemicals that can be subsequently used for biochemical or thermochemical processes. However, because of the recalcitrance of biomass to deconstruction, the process of producing fuel ethanol from biomass sugars via fermentation has remained costly (*1*, *2*). Because the high cost of cellulase enzymes is a key factor in this process, a deeper understanding of the plant cell wall structure and the mechanisms of enzymatic degradation of cellulose plays a critical role in enabling a successful bioethanol industry (*3*).

At a molecular level, biomass can be divided into three main chemical components: cellulose, hemicellulose, and lignin. Understanding the chemical and ultrastructural details about the cell wall microfibrils is important for improving deconstruction. It is commonly thought that plant cell wall microfibrils are composed of a crystalline cellulose core coated with hemicellulose, a system that prevents cellulose microfibrils from self-association after their biosynthesis, and enables the formation of a strong yet flexible plant cell wall by crosslinking the microfibrils (*4–6*). Many deconstruction procedures involve stripping away the hemicellulose by hydrolyzing using acid or enzymes, leaving the crystalline cellulose core to be hydrolyzed by cellulases. Thus, understanding the chemical composition and interactions between microfibrils is essential to improving biomass conversions.

Cellulose is the primary structural polysaccharide of plant cell walls, containing glucose monomers linked by β-1,4 glycosidic bonds. Cellulose can vary from elementary fibrils in plants, containing around 36 cellodextrin chains, to the large macrofibrils of cellulosic algae, containing more than 1200 chains (*7*). The glucan chain length can vary from about 2000 to more than 15,000 glucose residues (*8–10*). The microfibril size from different plant tissues and species is estimated to be from 2 to 10 nm in diameter and can be as long as several microns. Although many important features of plant cell walls are at this meso scale, 10 nm to 1 μm, chemical analysis using existing experimental tools is not yet able to study such systems.

Molecular dynamics simulation is a very powerful tool for studying carbohydrate properties because it provides valuable structural, kinetic, and thermodynamic information (*11–17*). However, traditional computational modeling of realistic cellulose structures in water is prohibitively time-consuming because of the relatively large number of atoms per glucose residue. For example, existing atomistic molecular dynamics calculations using CHARMM (*18*, *19*) require approximately 2000 hours and 250 processors to obtain 10 ns of molecular dynamics simulation time for a system that contains one million atoms. This is a system that is approximately 10 x 10 x 50 nm, containing 32 cellulose chains, each 100 glucose residues long. To expedite the calculations, lower resolution models of cellulose need to be developed to reduce the system size dramatically.

To study large macromolecular systems at longer time scales than are accessible by atomistic simulations, coarse-grained models of these macromolecules are usually developed. In coarse-graining, a group of atoms is replaced by a single bead or particle (*20–22*). In carbohydrates, the monomer sugars are typically fairly rigid; and the interactions between sugars are well known. For instance, glucose monomers within the cellulose exist mainly in the chair conformation with equatorial hydroxyl groups. Coarse-graining the glucose by replacing each glucose unit with a few beads can significantly reduce the interactions that need to be calculated, while preserving the essential features of the glucose monomer in the cellulose. In a coarse-grained model, the beads go through some critical interactions that are more computationally efficient when compared to the atomistic interactions. Combining the efficient potential energy surface while reducing the system size results in significantly improved computational speed.

The non-catalytic carbohydrate binding modules (CBMs) are recognized as an essential component of effective cellulase action on the cellulose (*23*). CBMs are classified into 55 families based on their sequence identity (www.cazy.org) and 7 fold families based on the structural similarity (*24*). CBMs are proposed to have three primary functions: proximity effects (*23*), substrate targeting (*25–30*), and microcrystallite disruption (*31*). By binding to the cellulose surface, CBMs can promote the association of the enzyme with the substrate and increase the effective cellulase concentration (proximity effect). CBMs also have selective affinities for various soluble and non-soluble carbohydrates (targeting function). In addition, some bacterial CBMs are thought to modify the cellulose structure to render the substrate more susceptible to enzyme function (disruption function).

Family I CBMs, which are entirely fungal, are especially interesting. Of particular interest is the cellobiohydrolase (CBH) I CBM produced by *Trichoderma reesei* (*T. reesei*, also known as *Hypocrea jecorina*), the most common source of commercial cellulases today. CBH I contains a Cel7A catalytic domain and a family 1 carbohydrate-binding module separated by a highly glycosylated linker peptide. CBH I is thought to be processive, moving along a crystalline cellulose chain, pulling up that chain and feeding it into the catalytic domain tunnel where cellobiose is produced by hydrolyzing alternate β-(1,4) glycosidic linkages (*32–34*). The processivity of CBHs makes them critical for bioprocessing crystalline cellulose found in plant cell walls. However, details of the different component functions of these enzymes during processivity remain unclear. It has been postulated that the linker domain plays a role in pushing the binding domain along a cellodextrin chain on the cellulose surface as the chain being hydrolyzed advances further into the active site tunnel of the catalytic domain. It is also possible that the linker peptide serves as a hinged "spring," storing energy and pulling the catalytic domain towards the binding domain, thus propelling the cellulose chain further into the active site tunnel.

Computational modeling of CBMs holds promise for understanding molecular function, but there has been limited effort in this area. Early atomistic molecular dynamics modeling investigated the CBM from CBH I in solution and on a cellulose surface in the absence of solvent (*35*, *36*). Molecular docking calculations were used to investigate the possibility that the CBM works its way under a strand of cellulose (*37*). Recently, Nimlos and coworkers used atomistic molecular dynamics simulation to investigate the interaction of CBM1 from *T. reesei* CBH I with a model of the cellulose surface modified to display a broken chain (*38*). Their results suggested that tyrosine residues on the hydrophobic surface of the CBM, specifically Y5, Y31, and Y32, make contact with the cellulose surface; and the fourth tyrosine residue in the CBM (Y13) moves from its internal position to form hydrophobic interactions with the cellulose surface. Thus, the structure of CBM1 changed from the native "closed" state to an "opened" state during the simulation.

A shortcoming of atomistic molecular modeling of this CBM and a cellulose substrate is that long simulations are necessary to obtain biologically relevant information; and if explicit water is used, these calculations can be prohibitively time consuming. Recently, Bu and coworkers attempted to overcome this barrier by developing a coarse-grained model for the cellulose substrate (*39*). Molecular dynamics simulations of this coarse-grained cellulose with an atomistic CBM and implicit water solvent enabled long simulations and appeared to show interesting behavior of the CBM on a cellulose substrate with a broken chain. The CBM appears to process along a cellulose strand away from the reducing end. This is consistent with the hypothesized motion of the CBH I complex, but this is the first indication that the CBM may contribute to this motion.

In this chapter, we will discuss some additional observations derived from molecular modeling concerning the interactions of family 1 CBMs with coarse-grained cellulose models. This work will demonstrate how long time-scale simulations allow the investigation of protein/carbohydrate interaction events that are not accessible using atomistic simulations. The remainder of this chapter is organized as follows. In the next section, a brief review of the available coarse-grained models for carbohydrates and a detailed description of our new coarse-grained model for cellulose (1,0,0) surface are presented. (We note this new coarse-grained model was developed to study the cellulose hydrophobic surface and not suitable for other surfaces, as well as the entire crystalline cellulose. For details, see Ref (*39*).) Subsequently, the coarse-grained model of cellulose is used to study the translation of family 1 CBM1 along a broken cellulose chain on the (1,0,0) surface. For these studies, we used both atomistic models of CBM1 from two exoglucanases – CBH I and CBH II – and a coarse-grained model of CBM1 from CBH I. A short look at the future prospects of the coarse-grained modeling of plant cell walls concludes this chapter.

Method

Coarse-Grained Model of Carbohydrates

The interactions between coarse-grained beads can be expressed as a sum of bonded and non-bonded terms,

$$E = \frac{1}{2}k(r - r_0)^2 + \frac{1}{2}k_\theta(\theta - \theta_0)^2 + B(1 + \cos(\varphi - \varphi_0)) + E_{nbond}$$

where r, θ, and φ are the distance, angle, and torsional angle between connected coarse-grained beads, r_0, θ_0, and φ_0 are the coarse-grained bond, angle, and torsional angle equilibrium positions, and k, k_θ , and B are the force constants. All of these parameters must be defined before the coarse-grained force field can be used in the molecular mechanics calculations or molecular dynamics simulations. Recently, two distinct coarse-grained models for carbohydrates have been developed, largely differing in how the non-bonded parameters are derived.

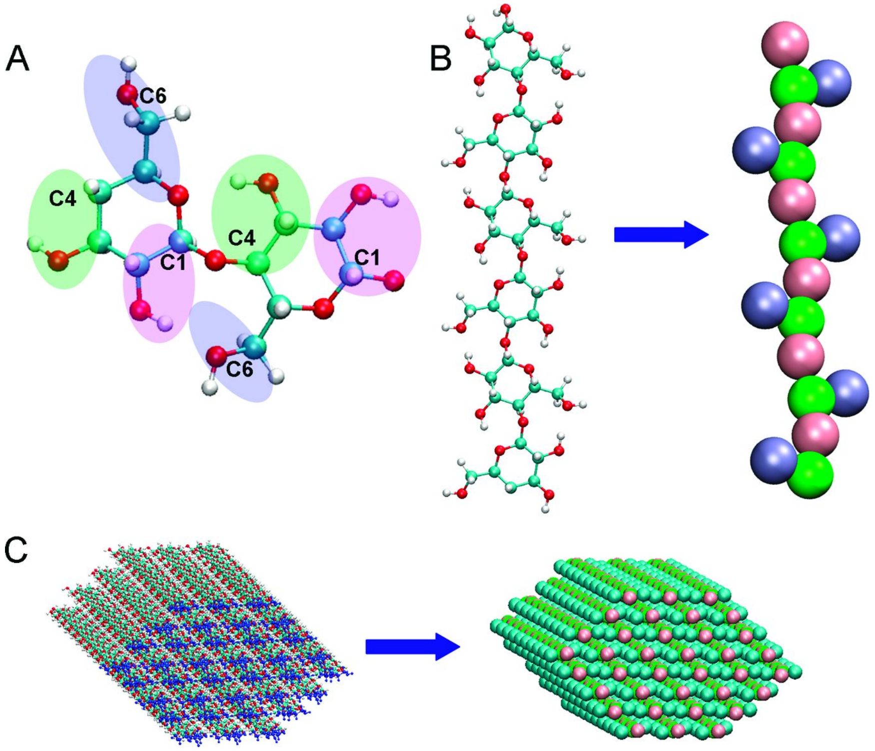


Figure 1. Parsing of the atomistic cellulose residue model among coarse-grained beads (A). The positions of the coarse-grained bead centers of mass correspond to the positions of the carbons C1, C4, and C6 in the atomistic model. The three beads in each glucose unit are connected by coarse-grained bonds. The fourth coarse-grained glycosidic bond links the glucose units in a cellulose chain. The atomistic model and coarse-grained model of a cellodextrin chain (B) and an elementary fibril (C) demonstrate the reduction of the system size. (see color insert)

Molinero and Goddard developed the first coarse-grained model, M3B, for malto-oligosaccharides, using three beads to replace each glucose unit of oligosaccharides and one bead to represent each water molecule (*40–42*). The bonded interactions were derived from Boltzmann statistical analyses of malto-oligomer atomistic trajectories in the condensed phases over a wide range of pressures. The non-bonded interactions were described with a two-body Morse potential, and the parameters were fitted from specific thermodynamics properties of glassy glucose. The M3B model was able to reproduce several properties of oligosaccharides, such as excluded volume, distribution of torsional angles, structures of left-handed and right-handed helices, and glass transition temperatures. Because neither charges nor hydrogen-bonding interactions were included in the M3B model, it was extremely efficient and resulted in an approximately 7000-fold acceleration of molecular dynamics simulations compared to an atomistic model.

Subsequently, Liu and coworkers derived a coarse-grained model for monosaccharide in aqueous solution using a systematic multiscale coarse-graining (MS-CG) algorithm (*43*). The non-bonded interactions were directly derived from the force-matching approach. Their MS-CG model was able to reproduce many structural and thermodynamic properties in the constant isothermal-isobaric ensemble (NPT). In this model, long-range interactions were effectively mapped into short-range forces with a moderate cutoff and were evaluated by table lookup, which led to molecular dynamics that was three orders of magnitude faster than the atomistic simulations. Although the model was derived at a single temperature, pressure, and concentration, it was transferable to other thermodynamics states. However, because their coarse-grained model was derived from α-D-glucose solution at one specific concentration, it was not transferable to other saccharide systems without modifications. Despite this, their MS-CG method is general and can be readily applied to other systems.

Although these coarse-grained models are very useful in simulating lengthy oligosaccharides over long times, they are not transferable to crystalline cellulose because of the significant structural difference between oligosaccharides and cellulose. To our best knowledge, no coarse-grained model has been derived for crystalline cellulose. Molecular dynamics simulation studies of crystalline cellulose have focused on using all-atom models (*11*, *44*, *45*).

Coarse-Grained Model of Cellulose

Recently, we developed a coarse-grained model and force field for the cellulose Iβ(1,0,0)surface (*39*). Similar to the M3B model (*41*) developed by Molinero and Goddard to represent oligosaccharides in solution, the coarse-grained model we proposed here also uses three neutral beads to represent each glucose unit and deposits the beads to the positions corresponding to the atoms C1, C4, and C6 in the atomistic model (shown in Figure 1). The mass of each bead is the sum of the mass of the atoms that the specific bead replaced.

The bonded parameters were derived directly from Boltzmann inversion of the bond, angle, and torsional angle distributions of the atomistic simulation of crystalline cellulose in water, which can provide the detailed dynamics information of the positions of atoms C1, C4, and C6. For example, Figure 2 plots the distribution of distances between atom C1 and atom C4 obtained from the atomistic simulations. A Gaussian function was used to fit this distance distribution to extract the equilibrium bond distance (i.e., 2.89 Å) and force constant for the virtual chemical bond between coarse-grained bead 1 and bead 4. This approach was used for all of the coarse-grain bonding interactions (bonds, angles, and dihedrals).

The non-bonded interaction between coarse-grained beads was represented by a Lennard-Jones potential. The depth of the potential energy well, D_0, was taken from the M3B model; and the distance R_0 was determined by the distance between two corresponding atoms in the atomistic model. Because the non-bonded parameters in the M3B model were derived based on the amorphous α-glucose structure, they are not transferable to the crystalline cellulose. To mimic the strong hydrogen-bonding interactions within a layer and the weak

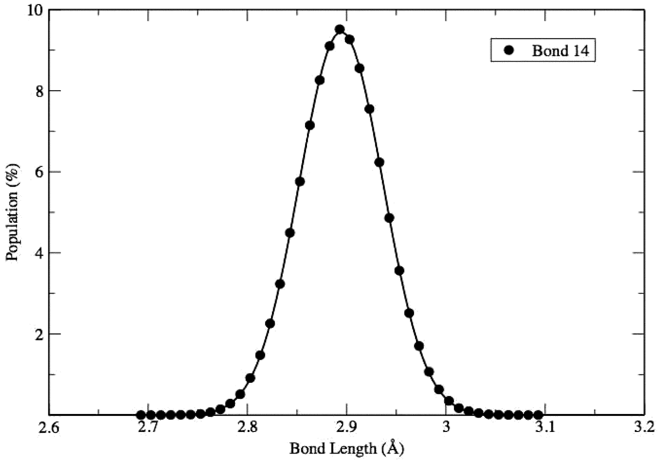


Figure 2. Bond length distribution for coarse-grained bond B1-B4 during atomistic simulation. Solid circles represent sampled values of the distance between two atoms C1 and C4 in the atomistic simulation. The solid line represents a fitted Gaussian function.

hydrophobic interactions between layers of cellulose, the interactions between different bead types were scaled by distinct factors. The attractive component between backbone beads (bead 1 and bead 4) and side chain beads (bead 6) was scaled by a factor of 1.5, while the backbone – backbone interaction was scaled by a factor of 0.1, and the side chain – side chain bead interaction was scaled by a factor of 0.2. See Ref (*39*) for details on how these rescaling factors were chosen. We confirmed that this force field was able to reproduce a stable structure of Iβ crystalline cellulose surface (*39*).

Using the cellulose bead model described here resulted in a significant decrease in the particles considered in the cellulose substrate, but further refinements were required to allow the long time-scale modeling necessary for studying protein action. Converting from an atomistic model of cellulose to the bead model resulted in a factor of 8 reduction of particles. However, the number of water molecules also needed to be reduced. In typical atomistic simulations, the number of explicit water molecule atoms is approximately two to three times the number of substrate and protein atoms. Coarse graining of water molecules to a single bead only reduced the number of particles by a factor of 3. As a result, an implicit solvent GBSW module (*46*, *47*) in CHARMM was used in all simulations discussed here. This effectively eliminated water molecule particles from the calculations and resulted in a total particle reduction from the atomistic simulation by a factor of approximately 20. Furthermore, in our simulations of the CBM with a coarse-grained cellulose, we found that using an implicit solvent had little effect upon the processive behavior of the CBM. This result was confirmed by a recent molecular dynamics simulation of cellulose atomistic models and CBM using explicit solvent, showing a similar behavior of CBM translating on the cellulose surface during a 100-ns simulation (*48*).

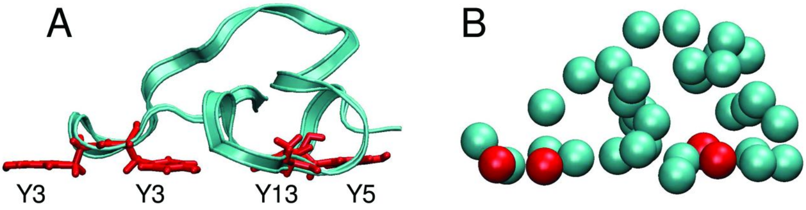


Figure 3. Comparison of an atomistic model (A) and a coarse-grained model (B) of CBM1. In the atomistic model, the CBM is shown in backbone ribbon representation with four tyrosine residues shown in stick representation. In the coarse-grained model, four tyrosine residues are shown as red beads. (see color insert)

Coarse-Grained Model of CBM

To further simplify the calculations of the CBM interacting with cellulose, in some simulations we used a coarse-grained model (Go model) of the protein. The Go model has long been used in theoretical studies of protein folding (*49*), in which each amino acid is represented by a single bead. In many early studies, the protein chain was modeled using two types of beads, one hydrophobic bead and one polar bead (*50–52*). Others also used between 3 and 20 types of beads in an effort to better capture the variability in the chemical nature encoded with protein side chains (*53*, *54*). Such coarse-grained models require identification of all possible contacts as either native (existing in the native structures) or non-native. Subsequently, a set of potentials is constructed in which native contacts are favorable, and non-native contacts are less favorable. Such a potential could significantly smooth the potential energy landscape and result in a faster protein-folding dynamic without contributing to the conformational search. Go models have been used in several studies at various resolution levels.

A Go model of the CBM was built using the Go Model Builder server (*55–57*) at MMTSB Web Service (*58*). The protein backbone was represented as a string of coarse-grained beads connected by virtual chemical bonds. Each bead represents a single residue and is located at the position of the corresponding α-carbon atom. All bond lengths were fixed, whereas bond angles were subject to a harmonic restraint, and dihedral angles were subject to potentials representing sequence-dependent flexibility and conformational preferences in the Ramachandran space. Non-bonded interactions were calculated using a Go model (i.e., only residues that were in contact in the native state interact with each other favorably). Residues not in contact in the native state interacted via a repulsive volume exclusion term.

Because the Go model potential is biased to the initial structure of a protein, the unfolding of the loop containing tyrosine 13 would not be observed if the native structure (i.e., the “closed” structure of CBM) was used to build the Go model. Therefore, we used an “opened” structure of CBM to generate the Go model. Figure 3 illustrates the comparison of an atomistic model and a Go model of the CBM. Using the Go model to investigate the interactions of CBM1 with coarse-grained cellulose is discussed in the next section.

Applications

CBM1 Translation on Cellulose Surface

Molecular dynamics simulations of an atomistic CBM from CBH I on a coarse-grained cellulose suggest that, when placed near a broken chain of cellulose, this CBM spontaneously translates along the strand away from the reducing end of the break. This motion is consistent with the processive motion of the entire CBH I enzyme complex. During procession, the reducing end of a broken strand is fed into the catalytic tunnel of the catalytic domain. The commonly accepted structure of CBH I has the CBM in front of the catalytic domain, tethered by the linker (*59*). During procession, the CBM would be in front of the catalytic domain and would be moving along the chain away from the reducing end. Thus, the simulations are consistent with the anticipated action of CBH I, but the unique suggestion here is that the CBM processes along the chain in the absence of the rest of the protein complex.

All simulations in this study were conducted using the CHARMM suite of software. A cellulose slab was generated containing four sheets, with four cellulose chains in each sheet, and 20 glucose residues in each chain. An atomistic model of the CBM was derived from a nuclear magnetic resonance (NMR) structure (*60*) and was then positioned above the hydrophobic cellulose 1β surface (1,0,0). This is believed to be the target of this CBM (*61*), and its hydrophobic face was positioned so that the three tyrosine residues (Y5, Y31, and Y32) were within 3 Å of the (1,0,0) cellulose surface. The simulations of atomistic CBM1 interacting with coarse-grained cellulose, as discussed here and subsequently, were conducted using a GBSW implicit solvent model.

During the first 200 ps of simulation on a cellulose surface with no broken chains, tyrosine Y13 moved from its internal location in the CBM and formed a contact with the cellulose surface. These simulation results support the hypotheses proposed by Nimlos and coworkers that an induced change of CBM1 near the cellulose surface plays an important role in cellulose recognition. These results also suggest that the surface of the coarse-grained cellulose model contains the same essential elements for CBM recognition as does the surface of the atomistic cellulose model. The coarse-grained simulation was continued for 40 ns, during which time no net forward movement of the CBM was observed. The motion of the CBM is due to the random diffusion instead of processive translation on an unbroken crystalline cellulose surface.

As mentioned above, CBH I recognizes the reducing end of an already broken cellodextrin chain and processively digests that chain. To simulate a cellulose surface containing a broken chain, a cellodextrin chain in the top sheet of the cellulose slab was hydrolyzed between the fourth and fifth glucans (g4 and g5 in Figure 4A) by deleting the coarse-grained bond between these two residues. In the initial conformation, the CBM was placed in front of the broken chain with tyrosine residue Y13 above the fifth glucan, which is a reducing end. The distal end of the CBM (the end that connects to the linker) faced the reducing end of the broken chain.

As shown in Figure 4A, we observed significant forward movement of the CBM in molecular dynamics simulations on a slab with a broken cellodextrin

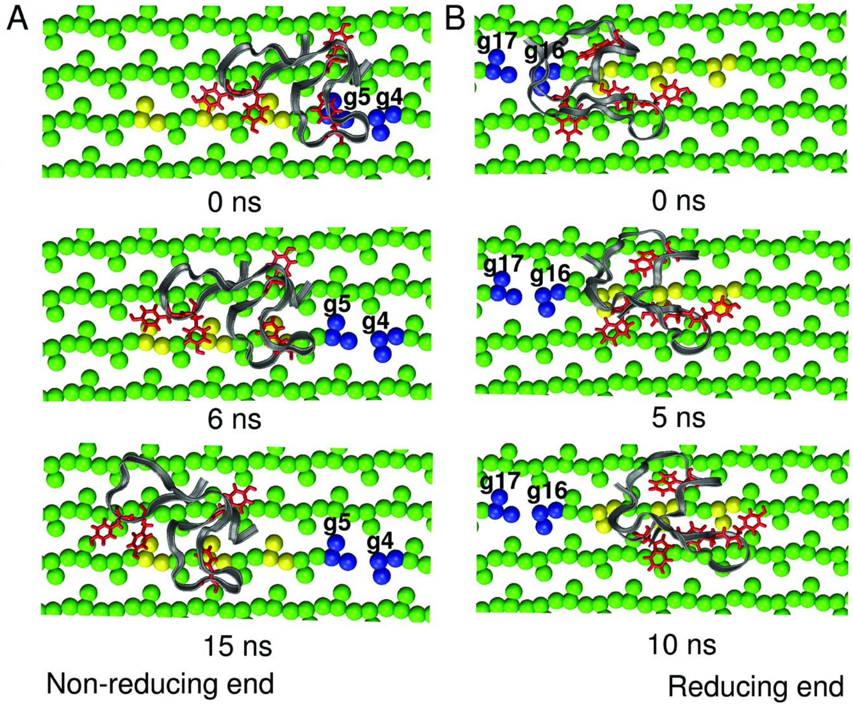


Figure 4. Plots showing details of the processivity of CBM1 during molecular dynamics simulation. The CBMs of CBH I (A) and CBH II (B) translate on the cellulose surface in the opposite directions when a chain was broken. For the purpose of clarity, only the top sheet of the cellulose slab is shown. CBM is shown in backbone ribbon representation with four tyrosine residues shown in stick representation. The gap between the two blue glucose residues (the 4th and 5th glucose residues in A, and the 16th and 17th glucose residues in B) indicates the position where a chain was broken. Yellow glucose residues represent the 7th, 9th, and 11th glucose units in A, whereas the 10th, 12th, and 14th glucose units in B. (see color insert)

chain. The CBM remained at its initial position for about 5 ns before moving to the next position, about 10 Å away from the reducing end with Y13 over the seventh glucan. In this new position, the CBM has advanced by one cellobiose unit down the chain. After staying at this new position for another 10 ns, the binding module moved another 10 Å down the chain. Interestingly, Y13, Y31, and Y32 remained roughly aligned with the cellulose chains during these jumps. The three tyrosines that are part of the hydrophobic face on the CBM (Y5, Y31, and Y32) in the closed form are spaced roughly equidistant to the spacing of cellulobiose units in cellulose. It is tempting to assume that these three tyrosines align with the cellulose chain. These simulations seem to suggest that the fourth tyrosine (Y13) along with Y30 and Y31 can also align with the cellulose. After the second jump, this simulation was then extended for another 25 ns, but no further forward movement down the chain was observed.

To determine if the movement described above was in response to a reducing end of a cellulose chain, another simulation was conducted in which CBM was placed on the cellulose surface facing in the opposite direction and on the non-reducing side of the break in the cellulose chain. In other words, the CBM was rotated by 180° along the (1,0,0) surface normal compared to the previous simulations. In the initial conformation, the CBM was in front of a broken chain with tyrosine Y13 above the sixteenth glucan (see Figure 4B). Although this figure shows the positions of CBM1 from CBH II instead of from CBH I, the CBM1 of CBH I was positioned at the same location on the cellulose surface. In this case, the distal end of the CBM faced the non-reducing end of the broken chain. During the 25-ns simulation, the CBM remained associated with the cellulose surface in a manner similar to the unbroken chain simulation. However, no net forward movement along the chain towards the reducing end was observed. This result supports the hypothesis that CBH I recognizes the reducing end of a broken chain and translates processively down that chain away from the reducing end of a cellulose chain.

Comparing the CBM1s of CBH I and CBH II

The cellobiohydrolase CBH II (Cel 6A) is also produced by the fungus *T. reseei*, but this processive enzyme hydrolyzes from the non-reducing end of a cellulose chain. The CBH II structure is very similar to that of CBH I in that they both have a binding module, a catalytic domain, and a linker. Further, the binding domain from CBH II is also from family 1 of the carbohydrate binding modules, and its sequence has a 76% homology with the CBM from CBH I. CBH I and CBH II are key components for efficient enzymatic conversion of biomass to ethanol and have been of interest for years. It is widely believed that these two enzymes move in opposite directions along the cellulose surface (i.e., the CBH I recognizes the reducing end of a broken chain and processively translates from reducing end towards non-reducing end, whereas CBH II recognizes the non-reducing end of a broken chain and translates towards the reducing end, even though their CBMs are essentially identical). Interestingly, the amino acid sequence of domains for these two enzyme families is exactly reversed (i.e., CBH I has the catalytic domain on the N-terminus and CBH II has its catalytic domain on the C-terminus).

To conduct simulations of CBM1 from CBH II, a structure of this protein was needed. Because there was no structure available in the literature, we built one by threading the backbone of this CBM onto the known structure (“closed” state) of the CBM from CBH I using the Swiss-Prot web server (http://www.expasy.ch/sprot). The high degree of homology between these proteins lends confidence in the structure obtained this way. The sequences of these binding modules are compared in Figure 5A. As can be seen, the four aromatic binding residues in CBM1 of CBH I are all tyrosines (Y5, Y13, Y31, and Y32), while two are tryptophans in CBM1 of CBH II (W5, W13, Y31, and Y32). Presumably, mutation of a tyrosine residue to a tryptophan residue increases the binding affinity of CBM to cellulose surface. However, it is also suggested that the binding affinity of CBH I has been balanced for optimal performance of the enzyme; therefore, increasing the binding affinity by a single

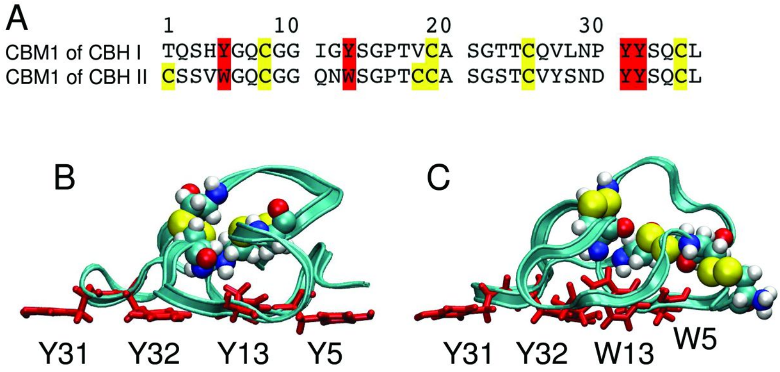

A
1
10
20
30
CBM1 of CBH I TQSHYGQCGG IGYSGPTVCA SGTTCQVLNP YYSQCL
CBM1 of CBH II CSSVWGQCGG QNWSGPTCCA SGSTCVYSND YYSQCL
B
C
Y31
Y32
Y13
Y5
Y31
Y32
W13
W5

Figure 5. Comparison of the sequences (A) and structures of CBM1 from CBH I (B) and CBH II (C). Tyrosine residues at the binding surface are shown in stick representation and Cystine residues are shown in CPK representation (cyan – Carbon, red – Oxygen, blue – Nitrogen, white – Hydrogen, yellow – Sulfur). (see color insert)

mutation was not expected to enhance the entire enzyme's ability to degrade the cellulose. Another structural difference between the two CBMs is that CBM1 of CBH I has two disulfide bonds (C8 – C25 and C19 – C35), whereas the CBM1 of CBH II has three disulfide bonds (C1 – C18, C8 – C25, and C19 – C35). The structural differences between the CBM1 of CBH I and the CBM1 from the homology model of CBH II are illustrated in Figures 5B and 5C. As a result of the three disulfide linkages, the N-terminus and C-terminus of CBH II CBM1 are closely packed together (i.e., they are connected by C18 and C19), suggesting this structure is more rigid and undergoes fewer conformational fluctuations during the dynamics.

During the first 200-ps simulation of the CBM1 from CBH II, tryptophan W13 moved from its internal location ("closed" state) in the CBM and formed a contact with the cellulose surface ("opened" state). Again, these results support the hypotheses proposed by Nimlos and coworkers that an induced change of CBM1 near the cellulose surface plays a key role in cellulose recognition. Molecular dynamics simulations of the CBM1 from CBH II showed an analogous translation to CBH I in that this CBM moved away from the non-reducing end of the broken cellulose chain. We studied the processivity of this CBM using the analogous protocol mentioned above. In the initial conformation, the CBM was in front of a broken chain with tryptophan W5 above the 15th glucan. A cellodextrin chain was broken between the 16th and 17th glucan. The distal end of the CBM faced the non-reducing end of the broken chain. As shown in Figure 4B, we observed a similar processive motion of the CBM on the cellulose surface. The CBM remained at its initial state for about 5 ns before moving to the next position. This is about 10 Å away from the non-reducing end with W5 over the 13th glucan. The CBM jumped one cellobiose unit along the chain towards the reducing end. After staying at this new position for another 10 ns, it jumped again another 5 Å along the chain.

As with the CBM1 from CBH I, another simulation was conducted in which CBM1 of CBH II was placed on the cellulose surface facing in the opposite direction. In the initial conformation, the CBM was in front of a broken chain with tyrosine W13 above the fifth glucan. A cellodextrin chain was broken between the fourth and fifth glucan. In this case, the distal end of the CBM faced the reducing end of the broken chain. During the 25-ns simulation, no net forward movement along the chain towards the non-reducing end was observed. This result supports the hypothesis that CBH II recognizes the non-reducing end of a broken chain and translates processively down that chain towards the reducing end.

Our previous studies of the potential energy landscape of the CBM1 from CBH I interacting with the cellulose surface sheds light on the mechanism of CBM1 translation when no longer near a broken cellulose chain (*39*). The potential energy landscape of CBM1 on an unbroken cellulose surface shows a uniform periodicity along the entire length of a cellulose slab, which suggests the CBM1 should undergo random diffusion on the cellulose surface instead of processive movement. However, breaking a coarse-grained chemical bond results in a potential energy barrier over the hydrolyzed cellulose chain on the order of 10 kcal/mol, which drives the CBM1 away from the broken end (*39*).

Coarse-Grained Model of CBM1

The atomistic model of CBM1 acting on the cellulose consists of 7263 atoms using an implicit solvent model. In the previous simulations (Figure 4), the cellulose was represented by a coarse-grained model and the CBM was represented by an atomistic model (1455 particles totally). To further reduce the system size, we combined the coarse-grained cellulose model and a coarse-grained CBM model (Go model) to investigate the interaction of CBM1 from CBH I with cellulose. The entire system contains only 996 particles. To compare the speed of these three models, a 100-ps molecular dynamics simulation was carried out on a Linux cluster using eight processors. The atomistic models of cellulose and CBM took 69.6 minutes, whereas the coarse-grained cellulose with atomistic CBM model took 17.1 minutes, and the coarse-grained models of cellulose and CBM took only 13.7 minutes. We should keep in mind that all these simulations used an implicit solvent model; using an explicit solvent would take longer. The advantage of using a coarse-grained model of the protein is not well demonstrated in this case, because the CBM is a small peptide containing only 36 residues. Thus, using a coarse-grained model of the CBM does not dramatically speed up the calculation. However, the method described here should be extremely useful and should significantly expedite the calculation when used to study the interactions of cellobiohydrolases (usually over 500 residues) with cellulose.

As shown in Figure 6, a similar translation of CBM1 down the cellodextrin chain was observed with the coarse-grained CBM model. The CBM remained at its initial position for about 1 ns before it moved to the next position. This is about 10 Å away from the reducing end with Y13 over the seventh glucan. The CBM jumped one cellobiose unit down the chain. After staying at this new position for another 10 ns, it jumped again another 10 Å down the chain with Y13 over the ninth glucan. These results suggest the coarse-grained models of CBM and

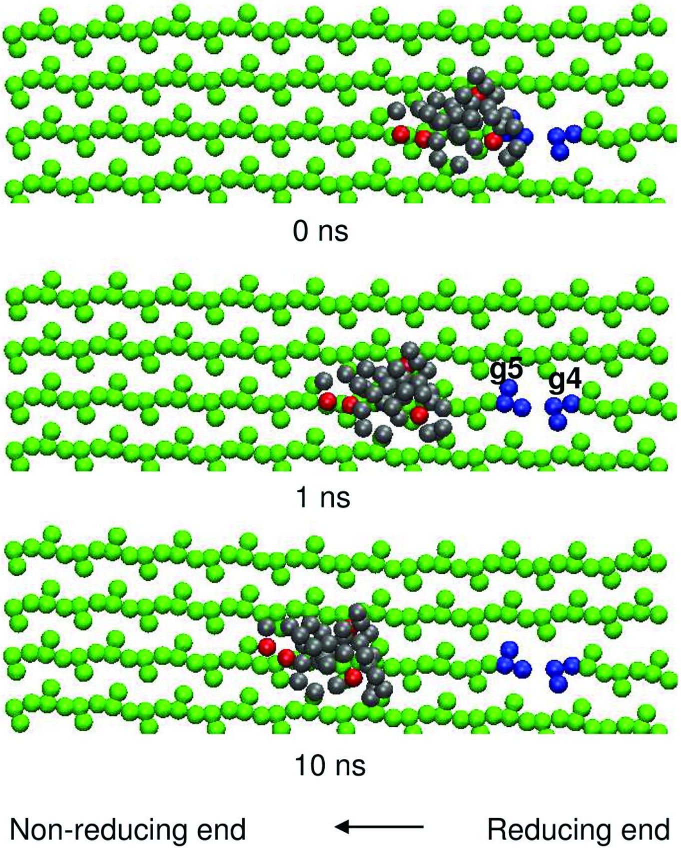


Figure 6. Plots showing details of the processivity of CBM1 during molecular dynamics simulation. The CBM1 of CBH I is shown as gray beads and four tyrosine residues are shown as red beads. The gap between the two blue glucose residues (the fourth and fifth glucose residues) indicates the position where a chain was broken. (see color insert)

cellulose contain the essential elements for CBM recognition of cellulose as the atomistic models of CBM and cellulose, indicating the coarse-grained model of enzymes could be used to study the interaction between cellulose and cellulase enzymes.

To investigate how CBM1 recognizes the reducing end of an already broken cellulose chain, another simulation was conducted in which the CBM1 was rotated by 180° along the surface and was facing the reducing end. In the initial conformation, the CBM was placed in front of a broken chain with tyrosine Y13 above the eighth glucan and tyrosine Y31 above the fifth glucan. The proximal end of the CBM faced the reducing direction of a cellodextrin chain, which was broken between the fourth and fifth glucan. In this case, the broken site is underneath the proximal end of the CBM whereas in the former simulation, the broken site is underneath the distal end of the CBM.

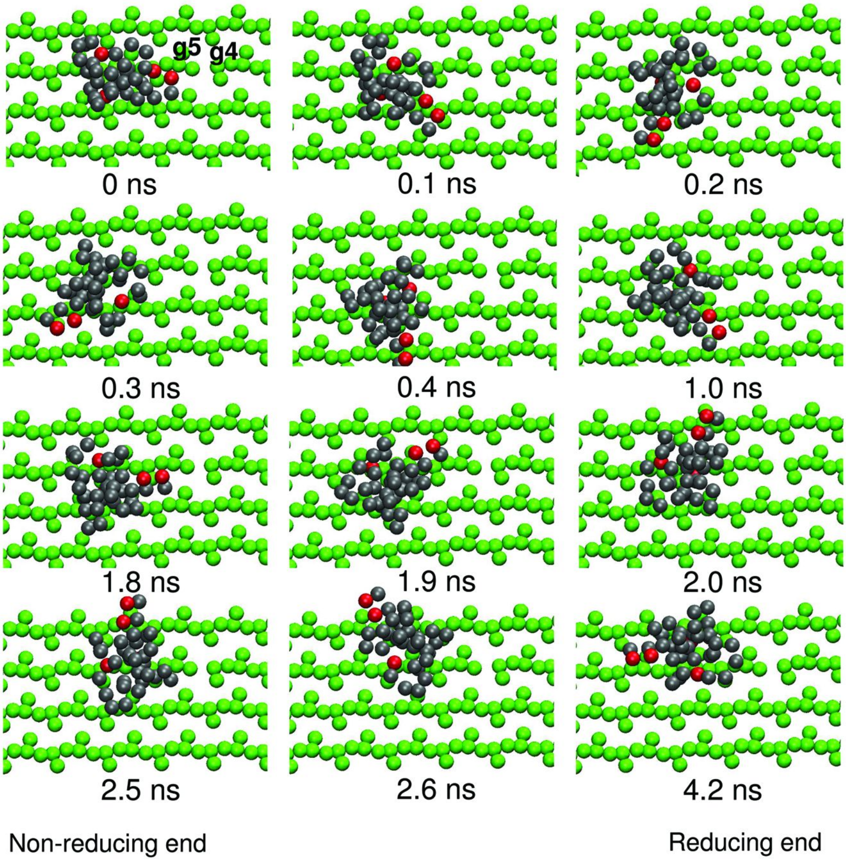


Figure 7. The CBM1 of CBH I performed a U-turn like movement on the cellulose surface when initially facing the wrong way, i.e., with the proximal end of the CBM facing the reducing end of a broken cellodextrin chain. (see color insert)

During the simulation, the CBM returned to the correct arrangement with the distal end facing the reducing end again by performing a U-turn-like movement on the cellulose surface. As shown in Figure 7, the CBM initially turned around clockwise. It rotated by ~45° in 0.1 ns and by ~135° in 0.3 ns. However, it failed to complete a 180° turn and began to rotate counter-clockwise in 0.4 ns. The CBM passed its initial position on the cellulose surface in 1.8 ns, and continued to rotate counter-clockwise. Finally, it successfully finished a U-turn on the cellulose surface in 4.2 ns and positioned itself on the correct track, with the distal end facing the reducing end of a cellodextrin chain and tyrosine residue Y13 over the sixth glucan. These results shed light on the recognition mechanism of CBM to crystalline cellulose, indicating that the initial arrangement of CBM upon binding might not be very important, since CBM could possibly adjust its initial position after binding to cellulose.

To investigate whether or not this U-turn-like movement could also happen in the atomistic model of CBM, we conducted an analogous simulation using a coarse-grained cellulose model and an atomistic CBM model. In the initial conformation, the atomistic CBM was superimposed on the coarse-grained CBM mentioned above. During a 40-ns simulation, no translational or rotational movement was observed. These simulation results indicated the coarse-grained CBM potential energy surface is smoother than the atomistic CBM potential energy surface and suggests that the coarse-grained CBM model can enhance the conformational sampling and expedite the dynamics.

These simulation results indicate that the coarse-grained models of different molecules with distinct resolutions could be combined together to expedite molecular dynamics simulations of large biomolecular systems. However, caution must be taken when using this method because, in general, the interactions between two coarse-grained models with different resolutions might need to be re-scaled to represent the real physical interactions in atomics models. For instance, to mimic the strong stacking interaction between a sugar ring and an aromatic ring, we found that the interaction between cellulose and the coarse-grained CBM beads representing aromatic residues needed to be increased by a factor of 4 compared with other coarse-grained CBM beads. A smaller scaling factor could cause the CBM to dissociate from the cellulose surface (i.e., the binding affinity is too weak). On the other hand, a larger scaling factor (i.e., a too-sticky surface) could make the CBM interact too strongly with the cellulose surface after binding and not be able to slide. To make this method useful, a detailed balance of the interaction between two coarse-grained models with distinct resolutions needs to be satisfied based on trial and error.

Conclusions

A coarse-grained model has been developed for the (1,0,0) surface of crystalline cellulose Iβ from all-atom simulations. The model decreases the number of particles by a factor of 8 and allows reliable molecular dynamics simulations to be conducted over long time scales. We are the first to show, using computer simulations, that when *T. reesei* CBM1 is applied to this new coarse-grained cellulose surface, the CBM can translate along a broken chain of cellulose without applying an artificial biasing potential. Our results demonstrate that the CBM1 of CBH I recognizes the reducing end of cellulose and migrates along the cellulose chain towards the non-reducing end, whereas the CBM1 of CBH II recognizes the non-reducing end of cellulose and migrates along the cellulose chain towards the reducing end. These observations agree with the proposed biological functions of CBH I and CBH II. We believe the coarse-grained model of cellulose presented provides a very useful tool for studying the mechanisms of other families of CBMs binding to cellulose surfaces, as well as the dynamics and functionality of the entire CBH I enzyme acting on cellulose.

Acknowledgments

The work to develop the coarse-grained cellulose model was supported by the National Renewable Energy Laboratory Discretionary Research & Development program and by the U.S. Department of Energy under contract No. DE-AC36-99GO10337 with the National Renewable Energy Laboratory. The work to model *T. reesei* CBM1 acting on this cellulose surface model was supported by the DOE Office of the Biomass Program.

References

1. Himmel, M. E.; Ding, S. Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D. *Science* **2007**, *315*, 804–807.
2. Himmel, M. E.; Ruth, M. F.; Wyman, C. E. *Curr. Opin. Biotechnol.* **1999**, *10*, 358–364.
3. Himmel, M. E.; Picattago, S. K. In *Biomass Recalcitrance: Deconstructing the Plant Cell Wall for Bioenergy*; Himmel, M. E., Ed.; Blackwell Publishing: London, U.K., 2008; pp 1−6.
4. Carpita, N. C.; Gibeaut, D. M. *Plant J.* **1993**, *3*, 1–30.
5. Hayashi, T.; Marsden, M. P. F.; Delmer, D. P. *Plant Physiol.* **1987**, *83*, 384–389.
6. Whitney, S. E. C.; Gothard, M. G. E.; Mitchell, J. T.; Gidley, M. J. *Plant Physiol.* **1999**, *121*, 657–663.
7. Sjöström, E. *Wood Chemistry*; Academic Press: San Diego, 1993.
8. Koyama, M.; Sugiyama, J.; Itoh, T. *Cellulose* **1997**, *4*, 147–160.
9. Newman, R. H. *Solid State Nucl. Magn. Reson.* **1999**, *15*, 21–29.
10. Sugiyama, J.; Harada, H.; Fujiyoshi, Y.; Uyeda, N. *Planta* **1985**, *166*, 161–168.
11. Matthews, J. F.; Skopec, C. E.; Mason, P. E.; Zuccato, P.; Torget, R. W.; Sugiyama, J.; Himmel, M. E.; Brady, J. W. *Carbohydr. Res.* **2006**, *341*, 138–52.
12. Bergenstrahle, M.; Wohlert, J.; Larsson, P. T.; Mazeau, K.; Berglund, L. A. *J. Phys. Chem. B* **2008**, *112*, 2590–2595.
13. Hanus, J.; Mazeau, K. *Biopolymers* **2006**, *82*, 59–73.
14. Mazeau, K. *J. Phys. Chem. B* **2007**, *111*, 9138–9145.
15. Mazeau, K.; Rivet, A. *Biomacromolecules* **2008**, *9*, 1352–1354.
16. Besombes, S.; Mazeau, K. *Plant Physiol. Biochem.* **2005**, *43*, 277–286.
17. Besombes, S.; Mazeau, K. *Plant Physiol. Biochem.* **2005**, *43*, 299–308.
18. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
19. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
20. Ayton, G. S.; Noid, W. G.; Voth, G. A. *MRS Bull.* **2007**, *32*, 929–934.

21. Head-Gordon, T.; Brown, S. *Curr. Opin. Struct. Biol.* **2003**, *13*, 160–167.
22. Lazaridis, T.; Karplus, M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139–145.
23. Reinikainen, T.; Ruohonen, L.; Nevanen, T.; Laaksonen, L.; Kraulis, P.; Jones, T. A.; Knowles, J. K.; Teeri, T. T. *Proteins* **1992**, *14*, 475–82.
24. Boraston, A. B.; Bolam, D. N.; Gilbert, H. J.; Davies, G. J. *Biochem. J.* **2004**, *382*, 769–781.
25. Creagh, A. L.; Ong, E.; Jervis, E.; Kilburn, D. G.; Haynes, C. A *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 1229–1234.
26. Henshaw, J. L.; Bolam, D. N.; Pires, V. M. R.; Czjzek, M.; Henrissat, B.; Ferreira, L. M. A.; Fontes, C.; Gilbert, H. J. *J. Biol. Chem.* **2004**, *279*, 21552–21559.
27. Lamed, R.; Tormo, J.; Chirino, A. J.; Morag, E.; Bayer, E. A. *J. Mol. Biol.* **1994**, *244*, 236–237.
28. Linder, M.; Teeri, T. T. *J. Biotechnol.* **1997**, *57*, 15–28.
29. Tomme, P.; Driver, D. P.; Amandoron, E. A.; Miller, R. C.; Antony, R.; Warren, J.; Kilburn, D. G. *J. Bacteriol.* **1995**, *177*, 4356–4363.
30. Tormo, J.; Lamed, R.; Chirino, A. J.; Morag, E.; Bayer, E. A.; Shoham, Y.; Steitz, T. A. *EMBO J.* **1996**, *15*, 5739–5751.
31. Din, N.; Forsythe, I. J.; Burtnick, L. D.; Gilkes, N. R.; Miller, R. C.; Warren, R. A. J.; Kilburn, D. G. *Mol. Microbiol.* **994**, *11*, 747–755.
32. Barr, B. K.; Hsieh, Y. L.; Ganem, B.; Wilson, D. B. *Biochemistry* **1996**, *35*, 586–592.
33. Teeri, T. T.; Penttila, M.; Keranen, S.; Nevalainen, H.; Knowles, J. K. *Biotechnol.* **1992**, *21*, 417–445.
34. Vršanská, M.; Biely, P. *Carbohydr. Res.* **1992**, *227*, 19–27.
35. Hoffren, A. M.; Teeri, T. T.; Teleman, O. *Protein Eng.* **1995**, *8*, 443–450.
36. Kuutti, L.; Laaksonen, L.; Teeri, T. T. *J. Chim. Phys. Chim. Biol.* **1991**, *88*, 2663–2667.
37. Mulakala, C.; Reilly, P. J. *Proteins: Struct. Funct. Bioinf.* **2005**, *60*, 598–605.
38. Nimlos, M. R.; Matthews, J. F.; Crowley, M. F.; Walker, R. C.; Chukkapalli, G.; Brady, J. W.; Adney, W. S.; Cleary, J. M.; Zhong, L.; Himmel, M. E. *Protein Eng., Des. Sel.* **2007**, *20*, 179–187.
39. Bu, L.; Beckham, G. T.; Crowley, M. F.; Chang, C. H.; Matthews, J. F.; Bomble, Y. J.; Adney, W. S.; Himmel, M. E.; Nimlos, M. R. *J. Phys. Chem. B* **2009**, *113*, 10994–11002.
40. Molinero, V.; Cagin, T.; Goddard, W. A. *J. Phys. Chem. A* **2004**, *108*, 3699–3712.
41. Molinero, V.; Goddard, W. A. *J. Phys. Chem. B* **2004**, *108*, 1414–1427.
42. Molinero, V.; Goddard, W. A. *Phys. Rev. Lett.* **2005**, *95*, 045701/1-4.
43. Liu, P.; Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 11566–11575.
44. Ford, Z. M.; Stevens, E. D.; Johnson, G. P.; French, A. D. *Carbohydr. Res.* **2005**, *340*, 827–833.
45. Nishiyama, Y.; Johnson, G.; French, A.; Forsyth, T.; Langan, P. *Biomacromolecules* **2008**, *9*, 3133–3140.
46. Im, W.; Feig, M.; Brooks, C. L. *Biophys. J.* **2003**, *85*, 2900–2918.

47. Im, W.; Lee, M. S.; Brooks, C. L., 3rd. *J. Comput. Chem.* **2003**, *24*, 1691–702.
48. Walker, R. C. Private communication.
49. Taketomi, H.; Ueda, Y.; Go, N. *Int. J. Pept. Protein Res.* **1975**, *7*, 445–459.
50. Dill, K. A.; Chan, H. S. *Nat. Struct. Biol.* **1997**, *4*, 10–19.
51. Lau, K. F.; Dill, K. A. *Macromolecules* **1989**, *22*, 3986–3997.
52. Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545–600.
53. Honeycutt, J. D.; Thirumala, D. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 3526–3529.
54. Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534–552.
55. Karanicolas, J.; Brooks, C. L. *Protein Sci.* **2002**, *11*, 2351–2361.
56. Karanicolas, J.; Brooks, C. L. *J. Mol. Biol.* **2003**, *224*, 309–325.
57. Karanicolas, J.; Brooks, C. L., 3rd. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 3954–3959.
58. Feig, M.; Karanicolas, J.; Brooks, C. L. *J. Mol. Graphics Modell.* **2004**, *22*, 377–395.
59. Zhong, L.; Matthews, J. F.; Crowley, M. F.; Rignall, T.; Talon, C.; Cleary, J. M.; Walker, R. C.; Chukkapalli, G.; McCabe, C.; Nimlos, M. R.; Brooks, C. L.; Himmel, M. E.; Brady, J. W. *Cellulose* **2008**, *15*, 261–273.
60. Kraulis, J.; Clore, G. M.; Nilges, M.; Jones, T. A.; Pettersson, G.; Knowles, J.; Gronenborn, A. M. *Biochemistry* **1989**, *28*, 7241–7257.
61. Lehtio, J.; Sugiyama, J.; Gustavsson, M.; Fransson, L.; Linder, M.; Teeri, T. T. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 484–489.

Chapter 6

Energy Storage in Cellulase Linker Peptides?

**Clare McCabe,[1,*] Xiongce Zhao,[2] William S. Adney,[3]
and Michael E. Himmel[3]**

**[1]Department of Chemical and Biomolecular Engineering and Department of
Chemistry, Vanderbilt University, Nashville, TN 37235
[2]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory,
Oak Ridge, TN 37831
[3]Chemical and Biosciences Center, National Renewable Energy Laboratory,
Golden, CO 37831, USA
[*]c.mccabe@vanderbilt.edu**

In this chapter, we discuss the use of molecular dynamics simulations and free-energy calculations to investigate the possible role the linker polypeptide, common to many cellulase enzymes, plays in the enzymatic hydrolysis of cellulose. In particular, we focus on the linker polypeptide from cellobiohydrolase I (CBH I) from *Trichoderma reesei*, which is one of the most active cellulase enzymes. CBH I is a multi-domain enzyme, consisting of a large catalytic domain containing an active site tunnel and a small cellulose binding module, which are joined together by a 27-amino-acid residue linker peptide. CBH I is believed to hydrolyze cellulose in a “processive” manner; however, the exact mechanism of the depolymerization of cellulose by CBH I is not fully understood. It has been hypothesized that the flexible interdomain linker mediates a caterpillar-like motion that enables the enzyme to move along the cellodextrin strand. Although the linker polypeptide sequence is known, the spatial conformation adopted by the linker domain and its role in the hydrolysis process, if any, has yet to be determined. The simulation results obtained to date indicate that the CBH I linker’s free energy is critically dependent on the existence of the cellulose substrate and the stretching/compression pathway adopted. In the presence of a cellulose surface, simulations suggest that

In Computational Modeling in Lignocellulosic Biofuel Production; Nimlos, M., et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2010.

the linker exhibits two stable states, which would support the hypothesis that the linker peptide has the capacity to store energy in a manner similar to a spring and facilitate a caterpillar-like motion.

Introduction

Understanding the mechanism of cellulase enzymes acting on cellulose at the molecular level helps us optimize the hydrolysis process, by allowing us to design more effective versions of nature's nanomachines and by providing insight into nature's design of nanoscale devices. The absence of a clear mechanism of action and understanding of the kinetic and thermodynamic factors important during the decrystallization and depolymerization events that release glucose molecules is a major drawback to the systematic development of improved cellulases (*1*). While atomistic molecular simulations of the complete hydrolysis process are beyond currently available computational capabilities, molecular modeling can be used to provide fundamental molecular-level insight by probing individual parts of the enzymes or the hydrolysis, or both (*2*, *3*). As described elsewhere in this volume, several computational approaches are being taken to help elucidate the mechanism of action of cellulase enzymes on cellulose that complement the ongoing experimental work; here we focus on understanding the role, if any, of the linker polypeptide commonly associated with cellulose enzymes. For completeness, we first outline the main features of cellulase enzymes, focusing on processive cellulases and CBH I from *Trichoderma reesei (T. reesei)*. We also discuss one possible mode of operation before describing efforts to use molecular dynamics modeling to probe the proposed behavior.

Cellulase Systems

The enzymatic hydrolysis of cellulose into its monomer units of glucose occurs in nature through the action of complementary cellulase enzymes. Cellulases are divided into endoglucanases, which cleave glycosidic bonds mid-chain to leave accessible ends; exoglucanases, which bind to the free reduced or non-reduced chain ends and liberate cellobiose and glucose; and β-D-glucosidases, which cleave the cellobiose units into fermentable glucose residues. Whereas endoglucanases randomly cleave glycosidic bonds, exoglucanases act in a processive manner to move along a cellulose chain and liberate cellobiose residues.

The filamentous fungus *Trichoderma reesei* has been extensively studied and is the microorganism most commonly used to study the cellulase system (*4*). The cellulytic system secreted by *T. reesei* contains two exoglucanases or cellobiohydrolases (CBH I and CBH II), at least five endoglucanases (EG1, EG2, EG3, EG4 and EG5), and several β-glucosidases (*5–7*). Of these, CBH I is the most studied and one of the most active cellulases known (*8*, *9*).

Cellobiohydrolase I

Like many of the enzymes secreted by microorganisms, CBH I is a multi-domain enzyme (*10*) consisting of a large catalytic domain containing an active site tunnel and a small cellulose binding domain (CBD) (Figure 1). With few exceptions, the CBD is connected to the catalytic module by a highly glycosylated (in fungi) flexible linker (*11*). While the binding domain serves to bring the catalytic domain into contact with the substrate (*12–14*), how the binding module and catalytic domain work together to hydrolyze a single cellodextrin strand from the crystalline cellulose surface is largely unknown (*5*). Furthermore, while much is known about the structure and composition of the binding module and catalytic domain, a complete CBH I structure has not yet been solved experimentally.

Although the sequence of the linker polypeptide has been determined, the linker's structure and its role during hydrolysis remain unclear. In cellulases, linker peptides vary in length from several to ~100 amino acids and, although linker sequences from different enzymes rarely share much sequence homology, they are generally rich in threonine, serine, proline, and glycine residues (*15–17*). Such small (glycine) and polar (threonine, serine) residues impart flexibility, yet maintain stability and conformation in solution through hydrogen bonding, and are generally common in linker domains (*18*).

As shown in Figure 2, the CBH I linker consists of 27 amino acid residues and is heavily glycosylated at the threonines and serines (*19*). Although a complete CBH I structure has not yet been solved, small angle X-ray scattering (SAXS) studies on related multi-domain cellulases have provided some insight. In particular, SAXS on cellulases from *Cellulomonas fimi* and *Humicola jecorina*, in which the catalytic and cellulose binding domains are connected by relatively extended proline-rich and serine/threonine-rich linkers, suggest that an elongated tadpole structure is adopted (*20–23*). However, subsequent additional SAXS analysis on *Humicola insolens* and *Pseudoalteromonas haloplanktis* cellulases showed that a distribution of interdomain separations is more likely exhibited, which can be attributed to the flexibility of the linker polypeptide (*24–26*).

The Role of the Linker Polypeptide?

The linker polypeptide in CBH I is thought to play several roles, including maintaining the spatial orientations of the binding module and catalytic domain, protecting the domains from proteolysis, and enabling secretion of other enzymes from *T. reesei* (*9*, *17*, *18*). Additional mutational biochemistry experiments have shown that shortening or removing the linker can result in loss or reduction in activity of CBH1 on crystalline cellulose (*16*, *20*, *27*, *28*). We therefore assume that the linker domain plays a significant role in enzyme function; however, the physical properties of the linker structure, its role in hydrolysis, and its relation to the catalytic and binding domains is largely unknown.

Given the variable nature of linker peptides, it has been suggested that their role is to provide an extended, flexible *hinge* between the catalytic and

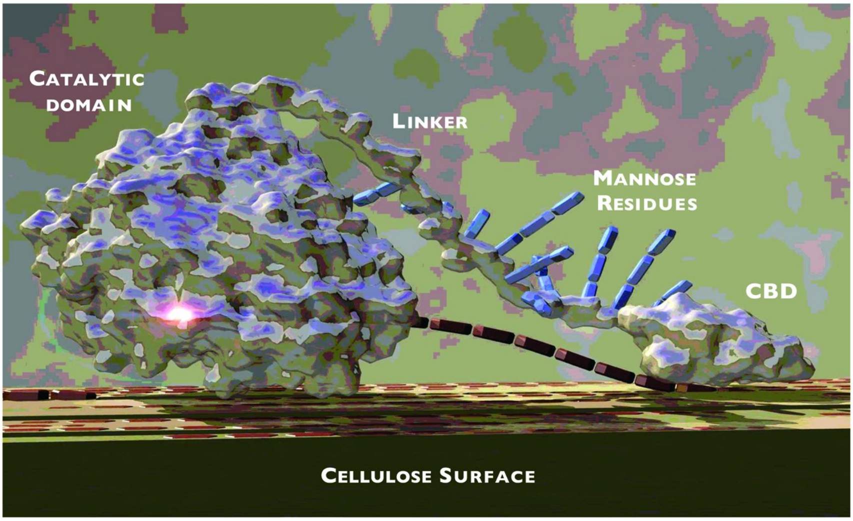


Figure 1. Artist's rendition of CBH I. The cellulose binding domain (CBD) interacts with the cellulose surface to detach cellulose molecules, which are then shepherded to the catalytic domain to undergo hydrolysis. (see color insert)

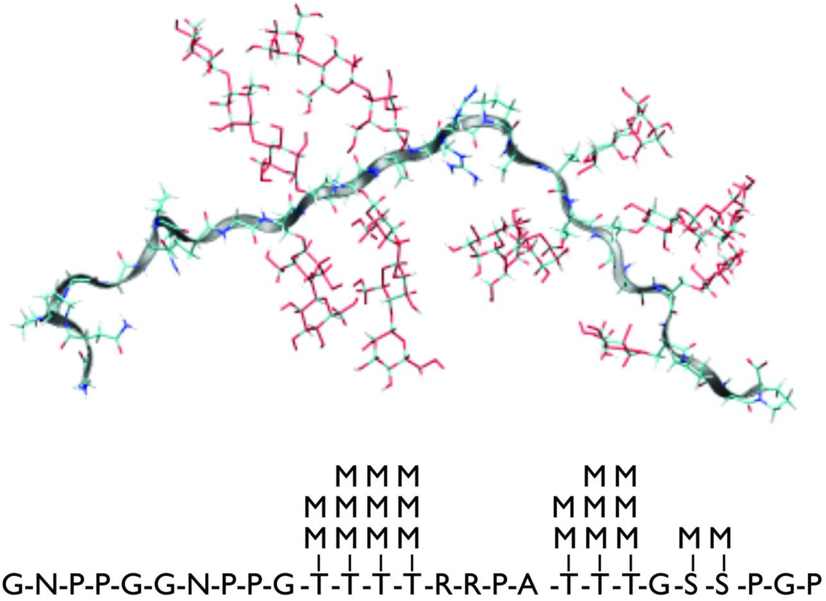


Figure 2. Molecular model of the CBH I linker peptide (top) and amino acid sequence (bottom) showing extent of glycosylation (M) at theronine (T) and serine (S) residues. The remaining residues are labeled according to the one letter amino acid shorthand scheme. (see color insert)

binding domains to facilitate the independent function of these domains and allow the cellulose binding domain to adsorb onto the cellulose surface and, subsequently, diffuse laterally along the surface leading the catalytic domain to new, enzymatically accessible sites (*16*). More recently, it has also been hypothesized that cellulases act through a caterpillar-like motion mediated by a combination of the flexible interdomain linker and diffusion of the binding domain on the cellulose surface (*24*, *25*, *29*). For the linker to work in such a caterpillar or “spring-like” motion, the linker polypeptide must be capable of storing energy. Once a short enough end-to-end distance is reached, this energy is released and allows the linker to extend. In light of the above hypothesis, it is critical to understand the mechanism of the linker peptide motion in the context of the cellulose surface to provide insight into the functioning mechanism of CBH I (*9*). Given that there are no experimental methods available to probe this complex behavior at the molecular level, molecular dynamics simulations have been used to study this system. Here, we provide an overview of the simulations performed to date.

Free Energy Calculations

Complex macro- and biomolecular systems that undergo conformational and/or structural change in response to their environment or as part of their function lie at the heart of many natural and synthetic processes. The conformational states in which macro and bio-molecular systems exist correspond to minima in free energy of the systems. Free energy is a measure of the energy of the system that takes into account entropic and thermal effects, and so differs from the internal energy (U), which is equal to the sum of the potential and kinetic energies that the atoms in a system possess by virtue of their positions and velocities, respectively. E.g., the *Helmholtz* free energy A is equal to $U - TS$ where T is temperature and S is entropy. In a system at constant temperature and density, A is minimized. Minimizing A results in a trade-off between lowering U (often achieved by having the atoms or molecules arranged in a regular structure, such as a crystal) and increasing S (achieved by increasing disorder, which generally increases U). Calculating potential and kinetic energies (and hence U) is trivial in a molecular dynamics (MD) simulation in which the positions and velocities of atoms are computed as a function of time. Free energies remain a challenge since they require higher-order information (i.e., they require entropic information about the states of the system and the probability of their occurrence) (*30*, *31*).

Many biological molecules undergo conformation changes as part of their function. Such conformation changes could be between two or more similar low-free-energy conformations or may represent switching from one minimum free-energy conformation to another, reflecting a shift in the minimum caused by changed local conditions. Understanding the relative free energies of various conformations of biological molecules, including those of the CBH I linker polypeptide, is key to understanding their structure and function. For example, the conformation into which a protein folds corresponds to the minimum free-energy conformation, thus determining its function (*32*). Developing computational methods for predicting low-free-energy conformations based on molecular composition remains an outstanding and challenging problem (*33*).

Calculating the free energy is equivalent to statistically measuring the probability of finding a system in a given state. Therefore, it depends on the extent of the phase space accessible to the molecular system, and it is not possible to calculate the absolute free energy of a system, because the entire phase space must be sampled. We can, however, conveniently calculate the free energy difference between two related states, which corresponds to estimating the relative probability of finding a system in one state to another. A good estimate of the free energy can be obtained by sampling the system along the pathway of interest. However, such sampling is usually not feasible using equilibrium molecular dynamics because high energy states are scarcely sampled and hard to overcome if two states are separated by an energy barrier. To efficiently sample all of the phase space of interest, we must use non-equilibrium molecular dynamics to facilitate the sampling. That is, an external constraint force is applied to the system to effectively flatten the energy surface so that simulations can easily access the states of interest with sufficient sampling.

To date, commonly used approaches to free-energy calculations for systems involving extensive numbers of solvent molecules include Ciccotti's constraint dynamics and the umbrella sampling methods. Both methods have been used to study the free energy of cellulase linker peptides as a function of end-to-end distance. We briefly review these methods below and provide some details about implementing these approaches for application to the current problem.

Ciccotti's Method

In Ciccotti's method (*34*), the free energy is calculated as the potential of mean force along a chosen coordinate using constraint dynamics. An external constraint is applied to the system to guarantee the sampling of phase space regions that would otherwise be highly improbable to reach in a conventional equilibrium molecular dynamics simulation. In this approach to studying the free energy of the CBH I linker polypeptide, the distance between the particles of interest, r, which corresponds to the distance between the N-terminus (N on glycine 1) and the C-terminus (the C on proline 27) on the linker backbone, is fixed by a holonomic constraint. The average force exerted on the fixed particles by the environment, which is the negative of the average force required to maintain the constraint (a quantity measured during the simulation), is then computed as a function of the distance between them.

In simulations of the linker polypeptide with a cellulose surface in aqueous solvent, the potential mean force calculation incorporates solvent effects, the intrinsic interaction between the atoms in the linker, and the interaction with the substrate. The potential of mean force at a linker end–to-end separation distance, r, corresponds to the free energy needed to stretch or compress the two termini from a distance r_0, where the potential of mean force reaches it minimum value, to the distance r. During each run at a fixed r, the force on the C ($\mathbf{F}_C$) and N ($\mathbf{F}_N$) terminus are calculated as time averages so that the mean force between the termini is given by

$$f(r) = \frac{1}{2}\left\langle \hat{r}_{CN} \cdot (\mathbf{F}_C - \mathbf{F}_N) \right\rangle, \tag{1}$$

where $\hat{r}_{CN}$ is the unit vector along the direction of the $C - N$, and $\langle \cdots \rangle$ indicates a time average over the phase-space trajectory. Integrating the $C - N$ mean force, $f(r) = -dF(r)/dr$, gives the desired potential of mean force $F(r)$,

$$F(r) = F(r_0) - \int_{r_0}^{r} f(r)dr \ . \tag{2}$$

Umbrella Sampling

Umbrella sampling (*35*) attempts to overcome the problem of sampling improbable phase space regions by modifying the potential functions so the unfavorable states are sampled sufficiently. To achieve this, a harmonic biasing potential is imposed on the linker to sample the end-to-end distance r_t of the linker,

$$u_{bias}(r) = \frac{k}{2}(r - r_t)^2 \tag{3}$$

where k is a force constant. In contrast to Ciccotti's method, umbrella sampling collects only the density of states of r along the reaction coordinate rather than the forces. The data from umbrella sampling are then processed using the weighted histogram analysis method (WHAM) (*36*) to calculate the free energy as a function of the end-to-end length of the linker.

The umbrella sampling method is computationally more efficient than Ciccotti's method, because we can readily examine the sufficiency of sampling by checking the overlap of histograms collected during the simulations. The sampling efficiency is based on the balance between the value of k, Δr, and simulation time. A small k value allows the sampling to span over a wide range of r, which requires fewer sampling windows, but each run needs a long simulation time to obtain good histogram statistics. A large k helps each sampling run to focus within a small range of r and results in histograms with better statistics in a short simulation time; however, more sampling windows are required to cover the entire phase space of interest. Sampling with small Δr and short runs is computationally more advantageous than longer runs with larger Δr (*37*). Therefore, we typically perform a series of test runs to optimize the value of Δr and the time scale of each equilibration and production run.

In the results discussed here, values of k ranged from 2.0 to 7.6 kcal/mol/A^2, which were estimated to be on the order of $k_BT/(\Delta r)^2$, where k_B is Boltzmann's constant and Δr is the sampling window interval. The histograms collected during typical samplings were analyzed for sufficient overlap and statistical smoothness. A Δr value of 0.5 Å was found to be necessary to obtain sufficient overlap between sampling windows for the given k values. Furthermore, after a series of trial simulations, we found that a production run of 1 ns was sufficient for obtaining histograms and free energy with satisfactory statistics for the system of interest.

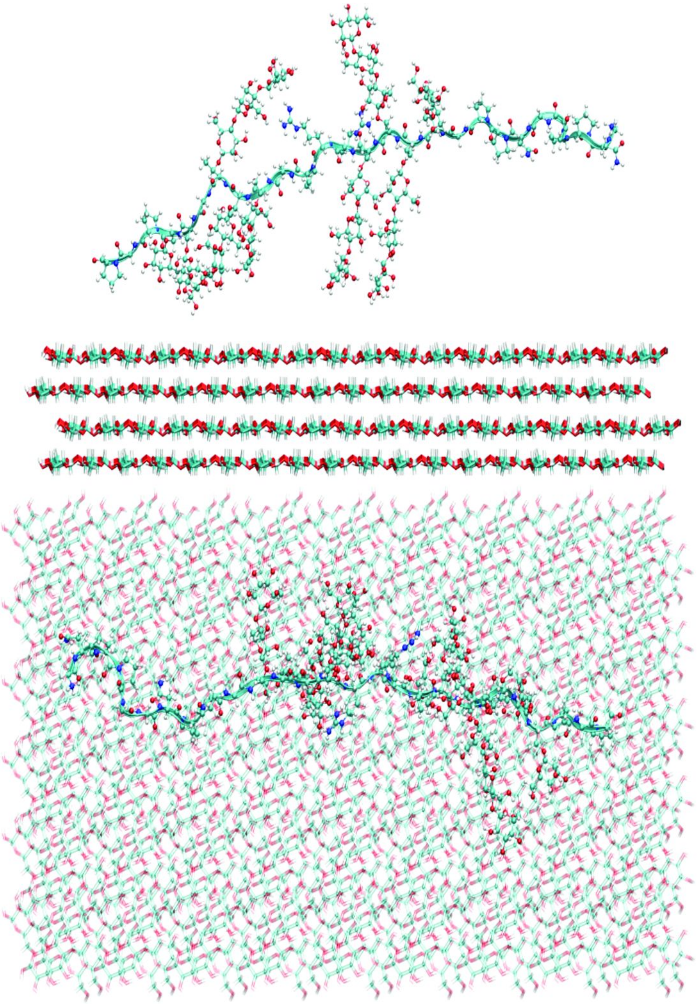


Figure 3. Snapshot showing an example initial configuration from a molecular dynamics simulation of the CBH 1 linker polypeptide above a cellulose surface in water. Top side view, bottom top down view. (see color insert)

Simulation Details

As discussed previously (and shown in Figure 2), the CBH I linker peptide from *T. reesei* consists of 27 amino acids with one to four mannose sugars glycosylated to the serine and threonine residues along the peptide backbone. The amino acids in the CBH I linker were modeled with the CHARMM27 force field (*38*), and the mannose sugar residues and cellulose were modeled by the supplemental force fields in CHARMM27, developed by Kuttel and coworkers (*39*). Water molecules were described by the TIP3P potential (*40*). For full details of the force fields used, the reader is directed to the original publications. The

NAMD (*46*) simulation package was employed in all simulations, and the VMD (*47*) tool used for visualization, snapshot extraction, and trajectory analysis.

The cellulose is modeled as a surface of the Iβ cellulose microfibril, constructed from the structure proposed by Nishiyama and coworkers (*41*). The microfibril used in the simulations reported here contains four layers of glucose sheets, with each sheet containing six glucose chains and each chain having 18 glucose units. This model gives a substrate with dimensions of approximately 9.4 nm x 4.5 nm x 1.2 nm.

The Lennard-Jones (LJ) interactions between different species were calculated using Lorentz-Bertholet combing rules, with a cutoff of 1.0 nm. An atom-based pair list with a cutoff of 1.2 nm was used and updated during the simulations. The particle-mesh Ewald summation method (*42*) with a fourth-order interpolation and direct space summation tolerance of 10^{-5} was applied to evaluate the electrostatic interactions. Periodic boundary conditions were applied in all three directions. The temperatures and pressures were kept constant where necessary using the method of Berendsen and coworkers (*43*). The SHAKE (*44*) algorithm was applied to constrain the bonds involving hydrogen atoms. An integration time step of 2 fs was used in all the runs.

In all simulations, the system was first equilibrated in the isobaric (1 bar) isothermal (300K) (*NPT*) ensemble before any production runs were performed. A typical *NPT* run included 10,000 steps of energy minimization using a conjugate gradient algorithm, followed by heating from 100 K to 300 K in 25-K increments within 10 ps. Subsequently, 150 ps of simulation with the cellulose substrate fixed (where needed) and 40 ps with all the atoms unconstrained were performed to further equilibrate the solvent and solute until the water density in the box approached a constant and stabilized value of ~1g/cm^3.

Free-energy sampling was performed under isothermal (300K) and constant volume (*NVT*) ensembles, starting from the ending configurations collected in the *NPT* simulations (Figure 3). In the free-energy simulations, the cellulose substrate was kept frozen. This assumes that the amplitude of the vibrational oscillations of the atoms in the substrate is much smaller than the typical dimension of the system at the considered temperature.

In the free-energy calculations using Ciccotti's method, the value of r_0 was chosen as the equilibrium length of the CBH I linker, r_0 = 4.95 nm, when the external force is absent and corresponds to the length at which $f(r)$ minimizes. Calculations of the mean force were performed at r ranging from 8.9 nm to 0.9 nm, with state points selected every 0.5 nm. Therefore, 160 conformational windows are involved in one full potential of mean force calculation, with each simulation covering 200 ps of phase-space trajectory. In total, we performed 10 potential of mean force calculations to obtain good statistics by averaging more than 10 potential of mean force curves.

With the umbrella sampling method, simulations have been performed for r_t in the range of 3Å to 80 Å. In each run, the system was equilibrated for 100 ps, followed by a 1-ns production run, during which time the histograms of r were collected at every step. The WHAM code developed by Grossfield (*45*) was used to analyze the histogram data and determine the free energy. We found that the convergence of the free energy calculated from the umbrella sampling

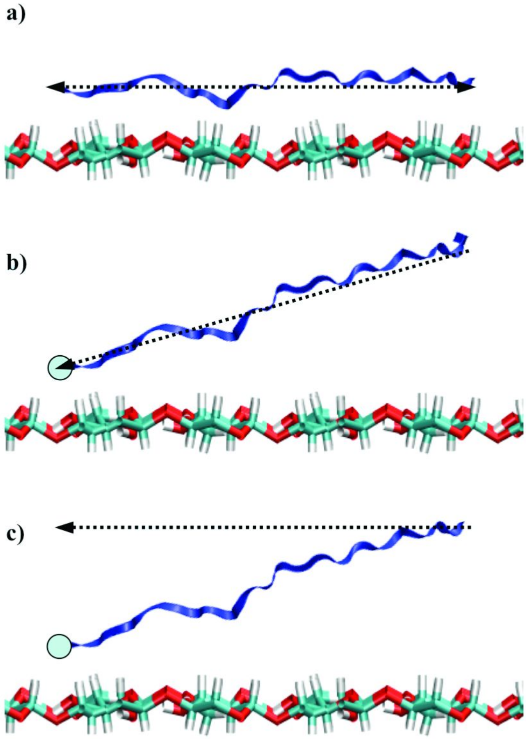


Figure 4. Schematic of the three pathways adopted in the study of the relative free energy profile of the CBH1 linker above a celluose surface. The dotted lines in each figure indicate the direction of linker motion and the solid blue circles the ends of the linker being held fixed. (see color insert)

data is sensitive to the convergence tolerance and the bin width used in WHAM processing. In this study, an optimized bin width and convergence tolerance of 0.2 Å and 10^{-5}, respectively, were used.

Results and Discussions

To probe the hypothesis that the CBH I linker acts as a spring and has the potential to store energy, simulations of the linker polypeptide were performed both with and without the cellulose surface. One important, but expected, phenomenon observed from our simulations is that the linker's free-energy profile is critically dependent on the folding pathway, or reaction coordinate, that the linker follows during the compression/stretching process. In the following paragraphs, we will discuss results from simulations in which three different pathways, shown in Figure 4, were studied.

As the first example, in Figure 5 we present the relative free-energy profile for the CBH I linker above a cellulose surface calculated along the first pathway shown in Figure 4 using the Ciccotti method (*48*). In this example, the linker is lying flat on the substrate, while its two ends are stretched or compressed by applying

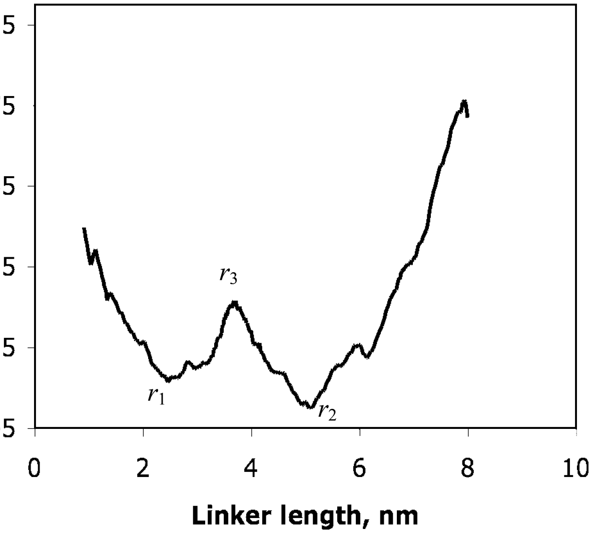


Figure 5. The potential of mean force and as a function of the linker end-end distance. Adapted from Zhao et al. Chemical Physics Letters, 460 (2008) 284–288.

constraint forces in opposite directions along the surface plane. One significant feature shown in the figure is that the free energy has two minimum state points located at r_1 = 2.5 nm and r_2 = 5.5 nm, respectively. That is, the linker is most stable at these two lengths. The free energy difference between these two linker modes, i.e., the linker at the compressed state (r_1 = 2.5 nm) and the extended state (r_2 = 5.5 nm), is about 10.5 kcal/mol, with the extended state being energetically more favorable. A local maximum is seen in the free-energy profile between these two states at r_3 = 3.7 nm, which corresponds to the energy barrier the linker must overcome as it transitions between r_1 and r_2. The energy difference between r_1 and r_3 is 17.5 kcal/mol, and the difference between r_2 and r_3 28.0 kcal/mol. When the linker is stretched or compressed to the two extreme conditions (i.e., whether it is compressed below 1.3 nm or extended above 7 nm), the free energy goes up dramatically.

For the first system shown in Figure 4, we also monitored the $C - N$ length distribution function, $P(r)$, for the linker to study the linker conformation transition between the two stable states. Figure 6 shows a schematic plot of the distribution obtained by releasing the constraints on the linker at an end-to-end separation distance of 8.9 nm and monitoring the end-to-end distribution. Two main peaks exhibited in the $P(r)$ profile indicate that the linker is predominantly distributed at the two minimum energy states. The relatively small peak corresponds to the state with free-energy minimum at r_1, and the large peak corresponds to the free-energy minimum at r_2. This indicates that the linker has the potential to switch between these two states, and that the more extended state is the preferred conformation. We note, however, that the distribution function was calculated from a relatively short simulation (20 ns) in which the linker in its initial configuration has significant stored potential energy, and that 20 ns is

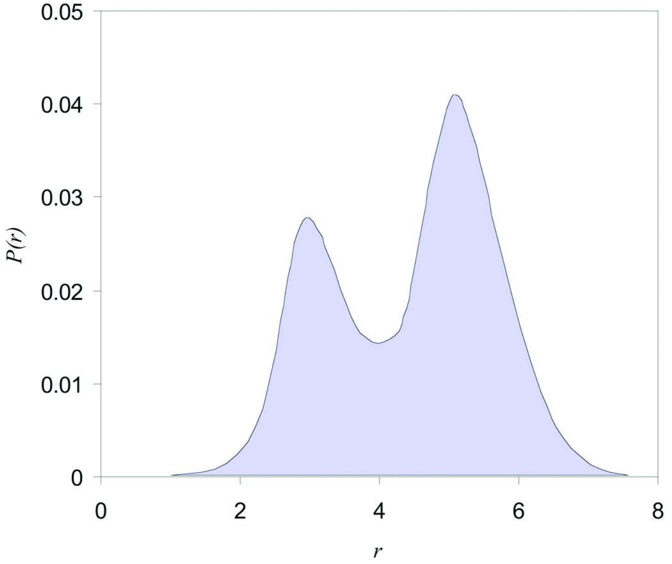


Figure 6. Schematic distribution of linker end - end distance. Adapted from Zhao et al. Chemical Physics Letters, 460 (2008) 284–288.

not long enough to obtain truly equilibrated configurations. Therefore, while the final distribution obtained is indicative of the overall distribution (i.e., shows the qualitative trends), it should not be interpreted as a reversible unconstrained sampling of the linker's energy surface from which we can infer quantitative free-energy information.

To verify the results obtained for the CBH I linker along folding pathway one, we also performed calculations using umbrella sampling. These results indicate that both methods give consistent free-energy profiles within statistical fluctuations; however, the approach using umbrella sampling plus WHAM provides better error statistics than the Ciccotti method. One possible reason could be that the WHAM technique used in umbrella sampling tends to give better statistical uncertainties in deriving the free-energy profiles than the simple integration of mean forces used in Ciccotti's constrained-dynamics method.

Another advantage of using umbrella sampling and WHAM is their practical computational efficiency. Using umbrella sampling allows us to conveniently check the sampling sufficiency by examining the overlapping extent of the umbrella histograms obtained during simulations. Once sufficient overlaps are obtained, we can stop further sampling and start simulating the next adjacent window. Therefore, we carried out most of our simulation studies on the CBH I linker polypeptide with the umbrella sampling method plus the WHAM technique to post-process the data collected and used Ciccotti's method for verification and comparison purposes only.

To further probe the behavior of the CBH I system, we also performed simulations using umbrella sampling to calculate the free energy of the linker following the same folding scheme, but without the existence of the cellulose substrate (i.e., a 'free' linker was studied) (*49*). Interestingly, in these simulations, we found that both minima observed in Figure 5 disappear; and the free-energy profile is essentially monotonic, with the free energy going up dramatically when the linker is stretched beyond 7 nm and compressed until the two ends are almost in contact. This result suggests that the substrate plays at least a partial role in determining the free energy of the linker. However, from preliminary analysis of the simulation results for the linker plus substrate system, it is not clear what the dominant force is behind the features observed in the free-energy profile. Additional analysis of the interactions between the linker and the substrate and changes in the solvent structure during the transition will be indispensable in understanding the observed free-energy features on the molecular level.

To further probe the role of the surface and effect of the folding pathway, we are also performing simulations to explore the free-energy profiles of the linker along the other pathways shown in Figure 4. In an effort to reduce the number of degrees of freedom in the system and more faithfully mimic the presence of the catalytic and binding domains, we assume that the binding domain of the enzyme is adsorbed to the cellulose surface and fixed. We also assume that the end of linker attached to the binding domain is constrained and not moving while the other end of the linker, (glycine 1) connected to the catalytic domain, moves in the direction indicated by the dashed line in the figure, which corresponds to the vector from the connecting point of the linker to the catalytic domain to the connecting point of the binding domain. Preliminary results from these calculations suggest that the linker-substrate interaction accounts, at least partially, for the distinct features observed in the free-energy profile of the linker on a cellulose surface. In particular, the free energy depends on the ability of the mannose residues on the linker backbone to interact with the cellulose surface. These results will be reported in detail in a future publication.

Conclusions

The minima observed in the free-energy profile along pathway one and the linker length distribution profile suggest that the CBH I linker polypeptide has the potential to store energy during compression/stretching and to transition between the extended and compressed states. However, we stress that it is not clear from these simulations alone whether such a mechanism is the key to understanding the normal operation of the CBH I enzyme acting on cellulose. Additional simulations are clearly needed to further probe the role of the cellulose surface, solvent, and extent of glycosylation and their effects on the observed free-energy profiles. Simulations of mutated CBH I linkers in which key residues, such as the central arginine groups around which the linker appears to hinge, also need to be performed. Furthermore, simulations of linker peptides from other enzymes could provide additional insight by enabling us to study the effect of linker length on the observed behavior. Again, the simulations reported here are focused on the linker itself; therefore, the impact of the CBH I catalytic and binding domains on the linker behavior is not included (*50*). Ideally, it would be desirable to calculate the free energy of the linker under the influence of these domains and other factors important to the hydrolysis process; however, it will be computationally very challenging to study such a large system. One possible choice is to perform such simulations using implicit solvent models or coarse-grained techniques if reliable models and force fields can be developed. Additionally, kinetic, rather than molecular modeling, approaches can be considered as in the recent study of Ting and coworkers (*51*), who developed a mechanochemical model for the dynamics of a CBH I-like cellulase as it extracts and hydrolyzes a cellulose polymer from a crystalline substrate. This work provides further support for the notion that the linker length and stiffness play a critical role in the cooperative action of the catalytic- and cellulose-binding domains.

Acknowledgments

The authors acknowledge subcontract XCO-2-32-042-01 from the National Renewable Energy Laboratory funded by the U.S. DOE Office of the Biomass Program and the Center for Nanophase Materials Sciences, which is sponsored at Oak Ridge National Laboratory by the Division of Scientific User Facilities, U.S. Department of Energy. Additionally the calculations discussed used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725, and the National Science Foundation through TeraGrid grant number MCB080117N.

References

1. Himmel, M. E.; Ding, S. Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D. *Science* **2007**, *315*, 804–807.
2. McCabe, C.; Schulthess, T.; Hirschfeld, P.; Chen, J.; McIlroy, A.; Weigand, G.; Gogotsi, Y.; Felmy, A.; Nichols, J.; Zacharia, T.; Polansky, W. M.; Strayer, M. *Scientific Impacts and Opportunities For Computing*; Office of Advanced Scientific Computing Research, Office of Science, Department of Energy, 2008.
3. Ding, S. Y.; Xu, Q.; Crowley, M.; Zeng, Y.; Nimlos, M.; Lamed, R.; Bayer, E. A.; Himmel, M. E. *Curr. Opin. Biotechnol.* **2008**, *19*, 218–227.
4. Hui, J. P. M.; White, T. C.; Thibault, P. *Glycobiology* **2002**, *12*, 837–849.
5. Zhang, Y. H. P.; Lynd, L. R. *Biotechnol. Bioeng.* **2004**, *88*, 797–824.
6. Nakazawa, H.; Okada, K.; Kobayashi, R.; Kubota, T.; Onodera, T.; Ochiai, N.; Omata, N.; Ogasawara, W.; Okada, H.; Morikawa, Y. *Appl. Microbiol. Biotechnol.* **2008**, *81*, 681–689.
7. Henrissat, B.; Driguez, H.; Viet, C.; Schulein, M. *Bio-Technology* **1985**, *3*, 722–726.
8. Cen, P. L.; Xia, L. M. *Adv. Biochem. Eng. Biotechnol.* **1999**, *65*, 69.
9. Jeoh, T.; Michener, W.; Himmel, M. E.; Decker, S. R.; Adney, W. S. *Biotechnol. Biofuels* **2008**, 1–10.

10. Gilbert, H. J.; Stalbrand, H.; Brumer, H. *Curr. Opin. Plant Biol.* **2008**, *11*, 338–348.

11. Harhangi, H. R.; Freelove, A. C. J.; Ubhayasekera, W.; van Dinther, M.; Steenbakkers, P. J. M.; Akhmanova, A.; van der Drift, C.; Jetten, M. S. M.; Mowbray, S. L.; Gilbert, H. J.; den Camp, H. *Biochim. Biophys. Acta, Gene Struct. Expression* **2003**, *1628*, 30–39.

12. Coutinho, J. B.; Gilkes, N. R.; Kilburn, D. G.; Warren, R. A. J.; Miller, R. C. *FEMS Microbiol. Lett.* **1993**, *113*, 211–218.

13. Hall, J.; Black, G. W.; Ferreira, L. M. A.; Millwardsadler, S. J.; Ali, B. R. S.; Hazlewood, G. P.; Gilbert, H. J. *Biochem. J.* **1995**, *309*, 749–756.

14. Shoseyov, O.; Shani, Z.; Levy, I. *Microbiol. Mol. Biol. Rev.* **2006**, *70*, 283–+.

15. Gilkes, N. R.; Henrissat, B.; Kilburn, D. G.; Miller, R. C.; Warren, R. A. J. *Microbiol. Rev.* **1991**, *55*, 303–315.

16. Srisodsuk, M.; Reinikainen, T.; Penttila, M.; Teeri, T. T. J. *Biol. Chem.* **1993**, *268*, 20756–20761.

17. Quentin, M.; Ebbelaar, M.; Derksen, J.; Mariani, C.; van der Valk, H. *Appl. Microbiol. Biotechnol.* **2002**, *58*, 658–662.

18. Argos, P. *J. Mol. Biol.* **1990**, *211*, 943–958.

19. Nevalainen, H.; Harrison, M.; Jardine, D.; Zachara, N. E.; Paloheimo, M.; Suominen, P.; Gooley, A. A.; Packer, N. H. *TRICEL 97 conference Carbohydrates from Trichoderma reesei and Other Microorganisms*; The Royal Society of Chemistry: Cambridge, U.K., Ghent, Belgium, 1997.

20. Shen, H.; Schmuck, M.; Pilz, I.; Gilkes, N. R.; Kilburn, D. G.; Miller, R. C.; Warren, R. A. J. *J. Biol. Chem.* **1991**, *266*, 11335–11340.

21. Abuja, P. M.; Pilz, I.; Claeyssens, M.; Tomme, P. *Biochem. .Biophys. Res. Commun.* **1988**, *156*, 180–185.

22. Abuja, P. M.; Schmuck, M.; Pilz, I.; Tomme, P.; Claeyssens, M.; Esterbauer, H. *Eur. Biophys. J. Biophys. Lett.* **1988**, *15*, 339–342.

23. Pilz, I.; Schwarz, E.; Kilburn, D. G.; Miller, R. C.; Warren, R. A. J.; Gilkes, N. R. *Biochem. J.* **1990**, *271*, 277–280.

24. Receveur, V.; Czjzek, M.; Schulein, M.; Panine, P.; Henrissat, B. *J. Biol. Chem.* **2002**, *277*, 40887–40892.

25. von Ossowski, I.; Eaton, J. T.; Czjzek, M.; Perkins, S. J.; Frandsen, T. P.; Schulein, M.; Panine, P.; Henrissat, B.; Receveur-Brechot, V. *Biophys. J.* **2005**, *88*, 2823–2832.

26. Violot, S.; Aghajari, N.; Czjzek, M.; Feller, G.; Sonan, G. K.; Gouet, P.; Gerday, C.; Haser, R.; Receveur-Brechot, V. *J. Mol. Biol.* **2005**, *348*, 1211–1224.

27. Black, G. W.; Rixon, J. E.; Clarke, J. H.; Hazlewood, G. P.; Theodorou, M. K.; Morris, P.; Gilbert, H. J. *Biochem. J.* **1996**, *319*, 515–520.

28. Black, G. W.; Rixon, J. E.; Clarke, J. H.; Hazlewood, G. P.; Ferreira, L. M. A.; Bolam, D. N.; Gilbert, H. J. *J. Biotechnol.* **1997**, *57*, 59–69.

29. Poon, D. K. Y.; Withers, S. G.; McIntosh, L. P. *J. Biol. Chem.* **2007**, *282*, 2091–2100.

30. Kofke, D. A.; Frenkel, D. In *Handbook of Molecular Modeling*; Yip, S., Ed.; Kluwer Academic Publishers: Dordrecht, 2004.

31. Kofke, D. A. *Fluid Phase Equilib.* **2005**, *228−229*, 41–48.
32. Anfinsen, C. B. *Science* **1973**, *181*, 223.
33. Wolfson, H. J.; Shatsky, M.; Schneidman-Duhovny, D.; Dror, O.; Shulman-Peleg, A.; Ma, B. Y.; Nussinov, R. *Curr. Protein Pept. Sci.* **2005**, *6*, 171–183.
34. Ciccotti, G.; Ferrario, M.; Hynes, J. T.; Kapral, R. *Chem. Phys.* **1989**, *129*, 241–251.
35. Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.
36. Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
37. Roux, B. *Comput. Phys. Commun.* **1995**, *91*, 275–282.
38. MacKerell, A. D.; Banavali, N.; Foloppe, N. *Biopolymers* **2000**, *56*, 257–265.
39. Kuttel, M.; Brady, J. W.; Naidoo, K. J. *J. Comput. Chem.* **2002**, *23*, 1236–1243.
40. Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, *103*, 335–340.
41. Nishiyama, Y.; Langan, P.; Chanzy, H. *J. Am. Chem. Soc.* **2002**, *124*, 9074–9082.
42. Darden, T.; York, D.; Pedersen, L. *J. Comput. Phys.* **1993**, *98*, 10089–10092.
43. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
44. Ryckaert, J.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
45. Grossfield, A. WHAM: the weighted histogram analysis method, version 2.0.1. http://membrane.urmc.rochester.edu/content/wham.
46. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
47. Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–&.
48. Zhao, X. C.; Rignall, T.; McCabe, C.; Adney, W. S.; Himmel, M. E. *Chem. Phys. Lett.* **2008**, *460*, 284–288.
49. Beckham, G. T.; Bomble, Y. J.; Matthews, J. F.; Resch, M. G.; Yarbrough, J. S.; Decker, S. R.; Bul, L.; Taylor, C. B.; Zhao, X. C.; McCabe, C.; Wohlert, J.; Bergenstråhle, M.; Brady, J. W.; Adney, W. S.; Himmel, M. E.; Crowley, M F. *Biophys. J.* **2010**, in press.
50. Zhong, L.; Matthews, J. F.; Crowley, M. F.; Rignall, T.; Talon, C.; Cleary, J. M.; Walker, R. C.; Chukkapalli, G.; McCabe, C.; Nimlos, M. R.; Brooks, C. L.; Himmel, M. E.; Brady, J. W. *Cellulose* **2008**, *15*, 261–273.
51. Ting, C. L.; Makarov, D. E.; Wang, Z. G. *J. Phys. Chem. B* **2009**, *113*, 4970–4977.

Chapter 7

QM/MM Analysis of Cellulase Active Sites and Actions of the Enzymes on Substrates

Moumita Saharay,[1,†] Hao-Bo Guo,[1,2,†] Jeremy C. Smith,[1,2] and Hong Guo[1,2,*]

[1]University of Tennessee / ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6164, USA
[2]Department of Biochemistry & Cellular & Molecular Biology, University of Tennessee, Knoxville, TN 37966, USA
[*]hguo1@utk.edu
[†]Co-First Author with equal contribution

Biodegradation of cellulosic biomass requires the actions of three types of secreted enzymes; endoglucanase (EC 3.2.1.4), cellobiohydrolase or exoglucanase (EC 3.2.1.91), and β-glucosidase (EC 4.2.1.21). These enzymes act synergistically to hydrolyse the β-1,4 bonds of cellulose and converts it into simple sugar. Hydrolysis of the glycosidic bond can occur either by net retention or by inversion of anomeric configuration at the anomeric center. QM/MM simulations are useful tools to study the energetics of the reactions and analyze the active-site structures at different states of the catalysis, including the formation of unstable transition states. Here, a brief description of previous work on glycoside hydrolases is first given. The results of the QM/MM potential energy and free energy simulations corresponding to glycosylation and deglycosylation processes are then provided for two retaining endoglucanases, Cel12A and Cel5A. The active-site structural features are analyzed based on the QM/MM results. The role of different residues and hydrogen bonding interactions during the catalysis and the importance of the sugar ring distortion are discussed for these two enzymes.

1. Introduction

The worldwide concern about limited storage of fossil fuel has led to the search for alternative energy sources. Biofuel generated from cellulosic biomass can be a promising substitute (*1*). Cellulose [Figure 1], the main component of plant cell walls, is a linear chain of sugar units connected by β-1,4 bonds. However, cellulose forms highly crystalline microfibrils in the plant cell wall that makes it recalcitrant to chemical and biological hydrolysis. Some microorganisms produce a battery of enzymes that work synergistically to degrade crystalline cellulose (*2–4*). Certain glycoside hydrolases (GH), such as cellulases, can play an important role in the conversion of cell wall polysaccharides into fermentable sugars through glycosidic bond cleavage (*5–7*). Non-catalytic carbohydrate binding modules (CBMs) of the enzymes attach to the polymeric surface and influence the actions of the catalytic domains (*8–11*). The enzymatic activity of cellulases in cellulose binding and hydrolysis depends on a variety of factors, including substrate composition, crystallinity, and recalcitrance (*12*, *13*).

Three types of cellulases are known to play an important role in the deconstruction of crystalline cellulose. Endoglucanases can cleave the glycosidic bonds of cellulose by producing chain ends and break down the interchain hydrogen bond in crystalline cellulose as well (*14*). Exoglucanases or cellobiohydrolases can then attach to a single fiber and break it into smaller sugars (cellobiose, cellotetraose, etc.). The simple sugars generated by these enzymes may be further hydrolysed to glucose by β-glucosidases. Molecular machines such as *T. ressei* cellobiohydrolases Cel6A, and Cel7A can act on crystalline cellulose. These enzymes may adopt two significantly different conformational states around the active site and couple the required energy for crystalline disruption to the energy released by the hydrolysis of the glycosidic bond. Activities of different cellulases are listed in Table 1. Henrissat et al. have classified the catalytic domains of these glycosyl hydrolases into over 100 families based on amino acid sequence comparison and hydrophobic cluster analysis (*15–17*) [this classification is available at http://www.cazy.org, the CAZy (Carbohydrate-Active enZymes) website].

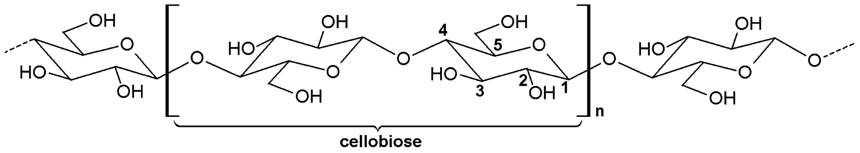


Figure 1. Cellulose, a linear polysaccharide chain with cellobiose as repeating unit. The numbering on the sugar ring shows positions of ring carbon atoms

2. Reaction Mechanism

According to the stereo chemical outcome of the hydrolysis reaction, the cellulases can be classified into two families; inverting and retaining enzymes

Table 1. Enzymes known to facilitate the hydrolysis of cellulose (*18*)

Cellulase	*Description*
Endoglucanase	Random cleavage of β-1,4 linkages of cellulose with preference for soluble and amorphous forms of the substrate. Affinity decreases with decreasing degree of polymerization with no activity on cellobiose.
Cellobiohydrolase	Release of cellobiose from the nonreducing ends with preference for crystalline forms of the substrate
β-Glucosidase	Release of β-D-glucose from the nonreducing ends of a wide variety of cellulose, cello-oligosaccharides, and a wide variety of β-1,4-glucosidases
Glucan β-1,4-glucosidase	Release of β-D-glucose from 1,4-β-D-glucans, but not cellobiose

(*19*). The inverting enzymes use a single displacement mechanism [Figure 2] resulting in inversion of the anomeric configuration. Two carboxylic/carboxylate residues at the active site play the key role in the catalytic mechanism: one acts as the general acid to cleave the glycosidic bond, and the other plays the role of the general base during the nucleophilic attack by the water molecule. Generally, these residues are Asp or Glu. On the other hand, the retention of stereochemistry at the anomeric center involves a two-step catalytic mechanism. The departure of the leaving group is first assisted by one catalytic residue with the second residue stabilizing the intermediate. The attack of a water molecule at the anomeric center then breaks the intermediate, leading to the formation of the product. The reactions in both mechanisms are believed to proceed through highly dissociative transition state structures with increasing charge formation at the anomeric center and the formation of a partial double bond between the C1 and O5 atoms leading to the oxocarbenium ion-like structure. Experimentally, the kinetic isotope effects (KIE) had been used to predict the geometry of these unstable transition states and intermediates (*20–23*). In comparison, computer simulations have been performed to provide the structures with detailed atomistic descriptions of the enzymatic reaction (*24*, *25*).

3. Structural Features of Cellulase Active Sites

The three types of active sites found in glycosyl hydrolases are i) pocket or crater, ii) cleft or groove, and iii) tunnel (*26*). The geometry of the active site depends on the endo or exo specificiity of the enzyme. Substrates bind to the open cleft in endoglucanases and xylanases which facilitates the twisting of cellulose strands along the chain to support the endo mode of action. For an example, the crystal structure of family 12 endoglucanase Cel12A (*27*) shows a concave cellulose binding cleft formed by 9 β-strands. On the other hand, the active sites of cellobiohydrolases or exoglucanases form a perfectly enclosed substrate-binding tunnel to cleave the cellulose chain from non-reducing end. The +n to n subsite nomenclature has been used for sugar-binding subsites in glycosyl hydrolases in which subsites are labelled from +n to n, with +n at the non-reducing end and

-n the reducing end (*28*). The X-ray crystal structure of cellobiohydrolase, CelS, a family 48 enzyme, showed that a ligand molecule binds to the tunnel between -7 to -2 subsites and the product (cellobiose) in the cleft region between +1 and +2. In CelS, the tunnel is intrinsically stable even without a bound substrate whereas the product binding in the cleft region is crucial to stabilize the protein conformation. A similar feature has also been observed in *C. Cellulolyticum*, CelF (*29*), supporting the suggestion that significant conformational change in the protein takes place upon the release of the product from the open cleft.

The catalytic domain of CelS folds into an $(\alpha/\alpha)_6$ barrel and forms a substrate binding tunnel at the N-terminal side of the inner α-helices [Figure 3] (*30*). This folding pattern is common among inverting glycosidases such as family 8 cellobiohydrolases and family 15 glucoamylases (*31*). It has been observed in Cel6A, another inverting enzyme, that the binding of four glucosyl units within the tunnel corresponding to the -2 to +2 sites provides the transition state stabilization (*32–34*). The intrinsically stable substrate binding tunnel and protein-carbohydrate interactions in the active site of cellobiohydrolase permits these enzymes to release the product while the remaining polysaccharide chain slides through the tunnel for the further hydrolysis. All family 48 enzymes are known to liberate cellobiose moieties by this processive mechanism when the enzyme is active and the substrate is available. A cartoon of this iterative method is shown in Figure 4.

Analyses of the active site interactions based on the crystal structure of the family 48 CelF enzyme in complex with ligands indicate that Glu55 can be the possible proton donor while Asp230 or Glu44 functions as the general base (*29*). In CelS, the equivalent residues could be Glu87 (as the general acid catalyst) and Asp255 or Glu76 (as the general base catalyst), respectively (*30*). Alzari and coworkers (*30*) suggested that Glu87 acts as proton donor in CelS due to its proximity to the active site and the favourable hydrogen bonding interaction with O4 atom of sugar unit at the +1 site. Instead, Glu76 is positioned far away from the active site and makes strong hydrogen-bonded interactions with sugar and other basic residues. Therefore, Glu76 might not participate in the reaction as a base catalyst. The sugar unit at subsite -1 was modeled to compare the role of different residues in the active site of a family 8 CelA enzyme (*30*). On the basis of their observation, Alzari *et al.* (*30*) proposed that Asp255 would possibly stabilize the ring boat conformation at -1 site rather than being a base catalyst. It was proposed that Tyr351 could participate in the reaction mechanism, although a 'direct catalytic role' might not be possible due to its high pKa value (*30*).

Another important reactant involved in the reaction is the nucleophilic water molecule in the case of inverting enzymes [Figure 2]. This water molecule may play a role in stabilizing the oxocarbenium type intermediate of the central sugar ring at -1 subsite after glycosidic bond cleavage, in addition to acting as the nucleophile. It makes a hydrogen bond with the base catalyst and donates hydroxyl ion to the anomeric carbon atom at site -1. One difficulty in experimental investigations is to predict the position of this water molecule. A recent study on family 44 endoglucanase, Cel44A, has identified the water molecule that may take part in hydrolysis based on the 3D-RISM theory (*35*, *36*). Interestingly, the atomic resolution crystal structure of CelA in complex with a single continuous cellulose chain pointed out the presence of this nucleopholic water molecule in the electron density map, even though it was difficult to identify this water molecule in the available CelS crystal structure (*30*). A reverse reaction mechanism (i.e. starting from the product state) can possibly predict the position of this water molecule in the reactant state (*37*), although the reaction path might not be the same as the forward reaction. The normal hydrolysis process then follows from this state. The T. reesei cellobiohydrolase Cel6A is found to hydrolyze α-cellobiosyl fluoride by this mechanism (*38–40*).

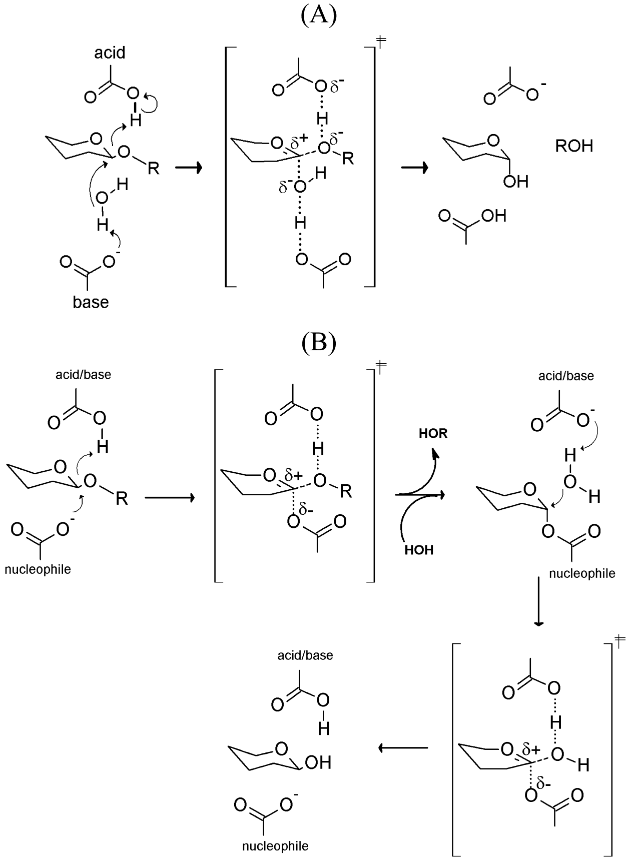


Figure 2. Schematic representations of two main enzymatic reaction mechanisms of glycosidic bond hydrolysis. (A) Inverting mechanism, (B) Retaining mechanism

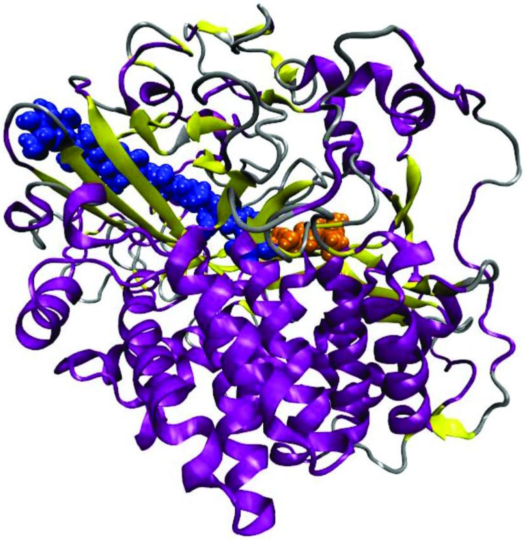


Figure 3. 3-dimensional structure of CelS (PDB ID: 1L2A) (30).

4. Sugar Ring Distortion

All accessible conformers of a sugar ring due to ring-distortion can be schematically represented by Stoddart's diagram [Figure 5].

Upon binding to the enzyme, the sugar ring at subsite -1 is believed to undergo conformational changes (*41*, *42*) from undistorted $^4C_1$ chair structure to distorted skew-boat structure [Figure 6], presumably, due in part to: (a) an increase in the charge at the anomeric carbon atom (C1), (b) an increase in the distance between C1 and O4 atoms of the leaving group, and (c) a decrease in the intra-ring O5-C1 distance. The substrate interacts with the protein mainly via hydrogen bonding and stacking interactions involving the aromatic side chains (*29*), and these interactions produce a continuous torsional strain on the substrate. Conformational changes, mainly at subsite -1, have been observed at all subsites due to these interactions (*43*). In contrast, very little structural modification was observed for the protein side chains. The most stable conformation of the protein assists the processive mechanism in which the remaining oligosaccharide chain slides through the active site after the removal of the product from the open cleft.

Further distortion in the sugar ring at subsite -1 has been observed during the hydrolysis process. Distortion in the pyranose ring at the site of potential enzymatic cleavage was first observed in the crystal structure of β-D-glycosidase (*44*, *45*). This distortion to a half-chair conformation results in a quasi-axial orientation of the glycosidic bond and leaving group to allow in-line nucleophilic attack of the anomeric carbon atom. In addition, the structural change in the pyranose ring reduces steric hindrance by hydrogen atoms [Figure 5B]. Electrophilic migration of the anomeric center along the reaction coordinate is a general feature observed in all GHs [Figure 5]. A Quantum Mechanical/Molecular Mechanical (QM/MM) study on the Michaelis complex of 1,3-1,4-β-Glucanase

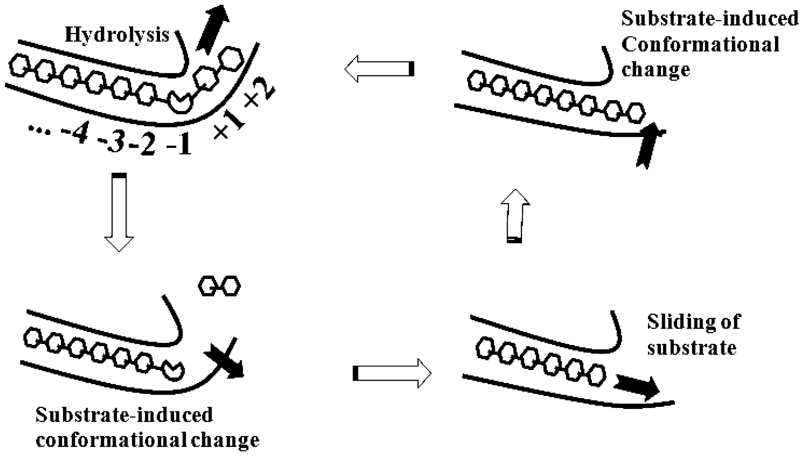

Hydrolysis
... -4 -3 -2 -1 +1 +2
Substrate-induced Conformational change
Substrate-induced conformational change
Sliding of substrate

Figure 4. Schematic description of a hypothetical processive mechanism in family 48 enzyme. Adapted from Ref. (30).

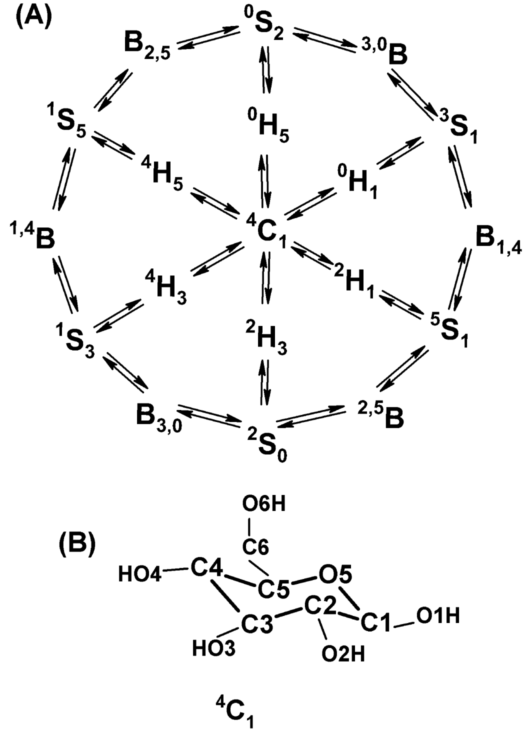


Figure 5. (A) Stoddart's diagram. (B) Chair conformation of the beta-glucose. Adapted from Ref. (50).

(*41*) indicated that there would be a further ring distortion to the ${}^4H_3$ half-chair conformation at the transition state. The extent of the distortion from the undistorted ${}^4C_1$ conformation can be quantified by a set of oligosaccharide torsion angles and ring puckering parameters (*46*). Moreover, a recent QM/MM investigation (*42*) of glycoside hydrolase family 8 (GH8) also showed that the glucosyl residue in subsite -1 in the Michaelis complex is in a distorted ring conformation, agreeing with the crystal structure.

5. Previous *ab Initio* and Molecular Dynamics Studies of GHs

Jones *et al.* performed computer simulation studies to elucidate the roles of aspartic acids D221, as the acid catalyst, and D175, as the base catalyst, in the active site of Cel6A (*25*). They modeled several reaction intermediates starting from different experimental 3D structures of the wild type and mutants. The crystal structure of unliganded enzyme shows a short hydrogen bond between D221 and D175 and a separation of 3.07 Å between the carboxyl proton of D221 and the glycosidic oxygen. Their MD simulation studies (with AMBER (*46*) all-atom force field for the protein and GLYCAM (*47*) force field for saccharides) suggested that this hydrogen bond had to be broken. An initial model structure for the active enzyme was generated by rotating the D221 side chain toward the glycosidic oxygen. The sugar ring at the -1 subsite was found to have a skew-boat conformation (${}^2S_0$) in the ground state. The water molecule for the nucleophilic attack at the anomeric center was held tightly by the hydrogen bonding interactions with S181, the carbonyl group of D401 as well as a second water molecule that connects to the base catalyst D175 via hydrogen bonds. The distance between C1 and the water molecule was 3.6 Å in the ground state, while it was 3.0 Å at the transition state. The proton donor D221 was also found to form a hydrogen bond with O4 of the glucosyl unit at the +1 site. The geometry of the high energy and unstable oxycarbenium-type transition state [Figure 7] in a boat conformation (${}^{2,5}B$) was obtained from ab inito calculations with Gaussian 94 (*25*). In the product state, the sugar ring at -1 subsite of α-cellobiose relaxed back to a chair conformation [Figure 7]. The energetic trend associated with the anomeric effect (*49*) and intramolecular hydrogen bonding was efficiently reproduced by AM1 method (*48*).

Greg and Williams (*24*) studied the free-energy pathway for the inverting reaction mechanism of human O-GlcNAcase involving substrate-assisted catalysis for the hydrolysis of N-acetyl-glucosaminides using the potential of mean force approach and the weighted histogram analysis method (see below). The reaction coordinate was the combination of the distances between anomeric carbon-glycosidic oxygen and anomeric carbon-acetamido oxygen. The free-energy profile corresponding to this reaction coordinate showed a plateau region near the transition state with the barrier height ~10.1 kcal/mol (activation energy). Ring conformational space sampled by the QM/MM trajectories along the reaction coordinate indicated that the protonated hemiacetal A (reactant state) was generally in the ${}^2S_0$ conformation whereas the distorted transition state geometry fell to the ${}^3H_4$ and ${}^4E$ regions, consistent with the existing X-ray crystallographic and KIE data (*41*). Recently, Parrinello *et al.* (*50*) performed *ab inito* metadynamics simulations on the gas-phase β-D-glucopyranose in the study of the conformational free energy landscape with respect to the ring distortion. Nine free-energy minima were observed with $^4C_1$ as the most stable conformation. The relative energies of the distorted ring conformations for $B_{3,0}$, $B_{3,0}/^2S_0$, $B_{2,5}$, $^1S_5$, $^{1,4}B/^1S_3$, $^{3,0}B$, $B_{1,4}$, and $^{2,5}B/^5S_1$ in comparison to the stable $^4C_1$ chair conformer were calculated to be 2.6, 3.0, 5.5, 5.8, 6.3, 7.2, 7.9, and 9.0 kcal/mol, respectively. The transformation of the $^4C_1$ chair conformation to any other conformer requires at least 8 kcal/mol of activation energy. Previous calculations on Michaelis complex (*41*) showed that QM/MM protocols can predict a good mimic of the transition states in GHs.

X-ray crystal structures of GHs in complex with substrate, product, and intermediate shed some lights on the different steps of the reaction pathways. However, experimental studies provide little information on the detailed conformations of high energy transition states along reaction coordinates. Recent QM/MM investigations on GH8 (*42*) and Cel5A (see below) provided important insights in this regard.

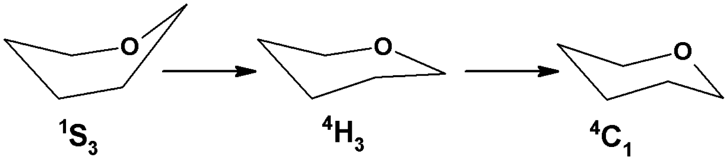


Figure 6. Conformational changes in the sugar ring at subsite -1. Adapted from Ref. (5).

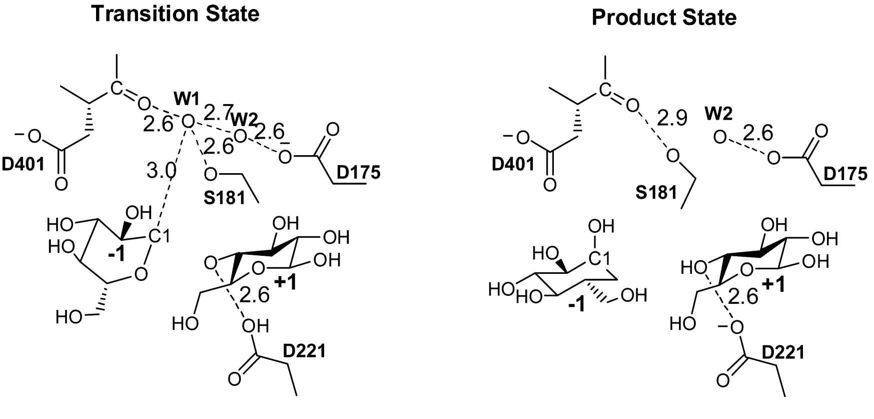


Figure 7. Geometries of transition and product states in Cel6A. Adapted from Ref. (25).

6. Model Setups and Methods for the Present Study

In the next sections, the results from our QM/MM simulations on the glycosidic bond cleavage are presented. The catalytic nucleophile and the acid/base residues in Cel12A are Glu120 (Glu228 in Cel5A) and Glu205 (Glu139 in Cel5A), respectively. Based on available crystal structures of Cel12A (*27*) and Cel5A (*51*), the two steps of the retaining mechanism, glycosylation and deglycosylation, are investigated, respectively.

The X-ray structure of Cel12A in complex with cellotetraose between the -2 to +2 sites was chosen as the initial structure for the study of glycosylation [Figure 8A] (*27*). Sugar units between the -2 to +2 subunits, Glu120 and Glu205, are included in the QM region. The deglycosylation process involving attack of a water molecule at the anomeric carbon was studied for Cel5A using the available crystal structure of trapped 2-fluoro-2-deoxy-cellotriosyl-enzyme intermediate [Figure 8B] (*51*). Free energy profiles from two independent initial models were compared. Finally, we estimated the ring distortion at subsite -1 due to the bond cleavage. Altogether, our computer simulation results provide a detailed description of the reaction mechanism.

A fast semi-empirical density-functional approach (SCC-DFTB) (*52*) implemented in the CHARMM (*53*) program was used for the QM/MM reaction calculations and free energy simulations. The efficiency of SCC-DFTB method makes it possible to sample enzyme systems with a relatively large quantum mechanical region, while high-level first-principle *ab initio* methods are not feasible to carry out such tasks even in a QM/MM framework. Moreover, recent studies indicated that the SCC-DFTB method could reliably describe the structures of hydrogen bonding systems (*54*, *55*). This was relevant in the current study as we found that the hydrogen bonding and proton transfers play important role in the retaining reaction mechanism. However, it should be pointed out that the SCC-DFTB method has systematic errors such as in the description of the hydrogen bonding energies (by about 1-2 kcal/mol) (*55*, *56*). New developments have been undertaken to improve the reliability and transferability of SCC-DFTB approach in the description of hydrogen bonding and proton affinities of biological macromolecules (*55*, *57*, *58*). The CHARMM force field was used for both enzyme and carbohydrate. A modified TIP3P water model was employed for the solvent (*59*). Stochastic boundary (SB) was used for the QM/MM simulations (*60*). The reference center for partitioning the system was chosen to be the anomeric carbon (C1) atom at subsite -1 and a sphere with radius of 22 Å was used for the reactive region of the SB boundary (*60*). The QM region included the sugar rings near the active site, nucleophile, and the acid/base residues as well as the nucleophilic water molecule in the deglycosylation process. The rest of the system was treated in MM region. The link atom approach (*61*) was used to separate the covalent bonds between the QM and MM regions.

To simulate both the glycosylation and deglycosylation processes, the reaction coordinate (RC, which is different depending on the processes. See below) was selected as the distance difference between the C-O bond that breaks and the C-O bond involved in the nucleophilic attack. For the glycosylation catalyzed by Cel12A, harmonic restraints with force constant of 500 kcal/mol/Å^2 were added to

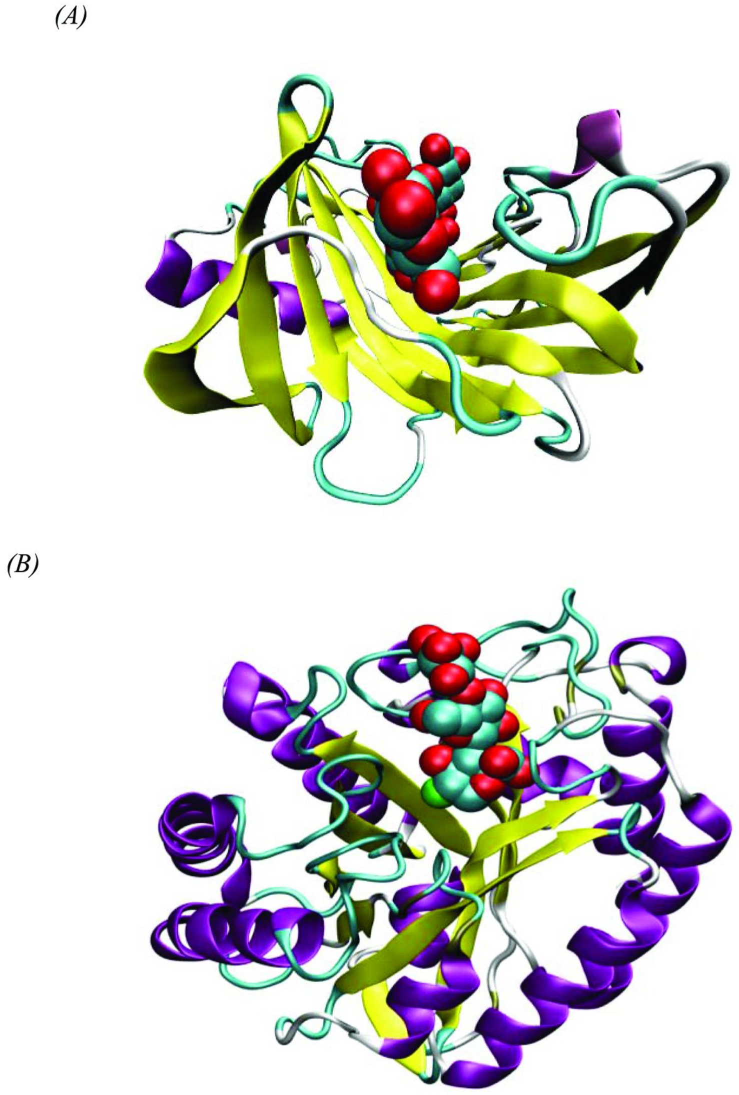


Figure 8. (A) Structure of H. grisea Cel12A (ribbons) in complex with a cellotetraose (spheres), occupying the -2 to +2 sites at the substrate binding cleft, pdb code 1W2U (27) (B) Structure of B. agaradhaerens Cel5A glycosyl-enzyme intermediate. The glycosyl occupies the -1 to -3 binding sites. The enzyme is shown in ribbons and the glycosyl in spheres, pdb code 1H11 (51).

the RC to guide the glycosylation process. A minimization protocol was adapted, in which the QM region was fixed to relax the MM region first, then the MM region was fixed to relax the QM region. This procedure was repeated until the non-restrained potential energy of the system converge to a threshold of 0.1 kcal/mol. The RC value was increased stepwise from -3.3 (the reactant complex) to 3.3 Å (the intermediate complex), with the step size of 0.1 Å; then the RC value was decreased from the product complex to the reactant complex. The forward and backward progresses were repeated in order to obtain a smooth and more reliable potential energy surface (PES) profile along the RC path.

The umbrella sampling (*62*) method was used to study the deglycosylation catalyzed by Cel5A. Harmonic potentials with a force constant of 500 kcal/mol/Å^2 were used in 23 simulating windows, each of which had 100 ps of equilibration and 50 ps of production run. The weighted histogram analysis method (WHAM) (*63*) was applied to determine the change of the free energy (potential of mean force or PMF) for the deglycosylation.

7. QM/MM Simulations on Cel12A and Cel5A

Humicola grisea Cel12A and *Bacillus agaradhaerens* Cel5A belong to the glycoside hydrolase family 12 (GH 12) and family 5 (GH 5) endoglucanases, respectively. *H. grisea* Cel12A has 224 amino acids and presents a characteristic fold of the GH 12 cellulases. A 35 Å long substrate-binding cleft on the concave surface of Cel12A is formed by a 9-strand β-sheet, see Figure 8A (*34*). *B. agaradhaerens* Cel5A (*51*) has 303 amino acids and presents a $(\alpha/\beta)_8$-barrel fold with a shallow-groove active site [see Figure 8B]. The members of the both GH 12 and GH 5 families cleave the glycosidic linkage of cellulose through the retaining mechanism. The two steps, glycosylation and deglycosylation, are involved in this mechanism [Figure 2]. Here we used the QM/MM method to study the glycosylation step catalyzed by Cel12A and the deglycosylation step catalyzed by Cel5A, respectively.

7.1. Glycosylation Catalyzed by *H. grisea* Celll2A

The first model system was built on a Michaelis complex of Cel12A and a cellulose tetramer (cellotetraose) (*64*) bound to the -2 to +2 subsites (pdb code 1W2U, resolution 1.52 Å) (*27*). Figure 9 shows the active site of the Cel12A Michaelis complex. In the glycosylation reaction, two glutamic acids, Glu120 and Glu205, act as the nucleophile and the general acid, respectively. The distance between the Oε2 atom of Glu205 and the O4 atom of glucose at the +1 site (i.e., the oxygen atom of the β-1,4-linkage) was 2.7 Å, which indicates that Glu205 may be protonated at Oε2 to function as the proton donor (general acid) to facilitate the glycosidic bond cleavage. The second glutamic acid, Glu120, is expected to function as a nucleophile to attack the anomeric carbon C1 at the subsite -1. In the study, the reaction coordinate (RC) was set as:

$$RC = r(\mathrm{C1\text{-}O4}) - r(\mathrm{C1\text{-}O\varepsilon 1}) \qquad (1)$$

The atom names are given in Figure 9. Though the reaction coordinate only included the nucleophilic attack and the corresponding bond cleavage [i.e., the bond formation involving C1(subsite -1) and Oε1(Glu120) and the bond breaking involving C1(subsite -1) and O4(subsite +1)], the proton transfer from Oε2 of Glu205 to O4 of subsite +1 happened spontaneously during the QM/MM simulations.

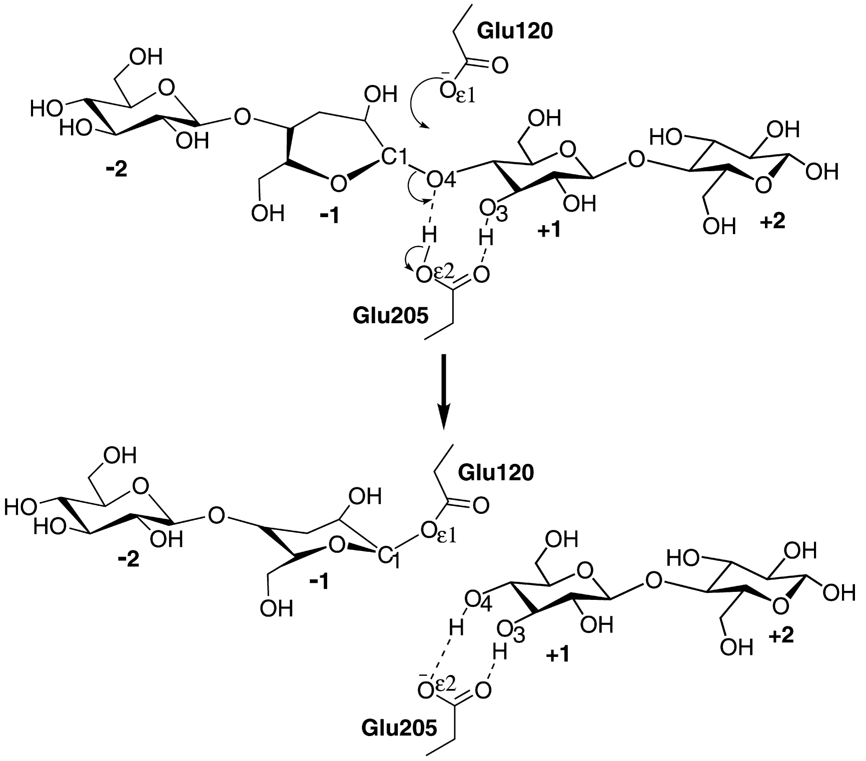


Figure 9. Active site of the H. grisea Cel12A Michaelis complex (pdb code 1W2U) (27) and the glycosylation catalyzed by Cel12A.

Figure 10A shows the potential energy profile along the RC. Three snapshots obtained from the simulations representing the Michaelis complex or reactant state (RS), the structure near transition state of glycosylation (TS1), and the glycosyl-enzyme intermediate state (IS) are shown in Figure 10B. The potential energy barrier for the glycosylation was calculated to be 20.5 kcal/mol. The glucose at subsite -1 showed the $^1S_3$ (RS) → $^4H_3$ (TS1) → $^4C_1$ (IS) comformational changes [Figure 10B]. These conformational changes are commonly observed in the GH-catalyzed reactions (see section 4 above).

It should be pointed out that the transition state (TS1) was approximated as the structure at the top of the potential energy profile [Figure 10A] without further frequency analysis. This approximation, however, should correctly reflect the distortion of glucose ring during the reaction. At TS1, *r*(C1-O4) and *r*(C1-Oε1) were 2.15 Å and 2.34 Å, respectively. At TS1, partial proton transfer from Glu205-Oe2 to O4 was observed to form a short-strong hydrogen bond with r(O4-H) = 1.18 A and r(Oe2-H) =1.27 A, respectively [Figure 10B].

Some hydrogen bonding interactions may play an important role in the glycosylation. For instance, the carboxyl sidechain of Asp103 (was treated by MM and not shown in Figure 10) formed a strong hydrogen bond with Glu120-Oe2. This hydrogen bond existed during the glycosylation, indicating that Asp103 may play a role in the catalysis by e.g., proton shuffling with Glu120.

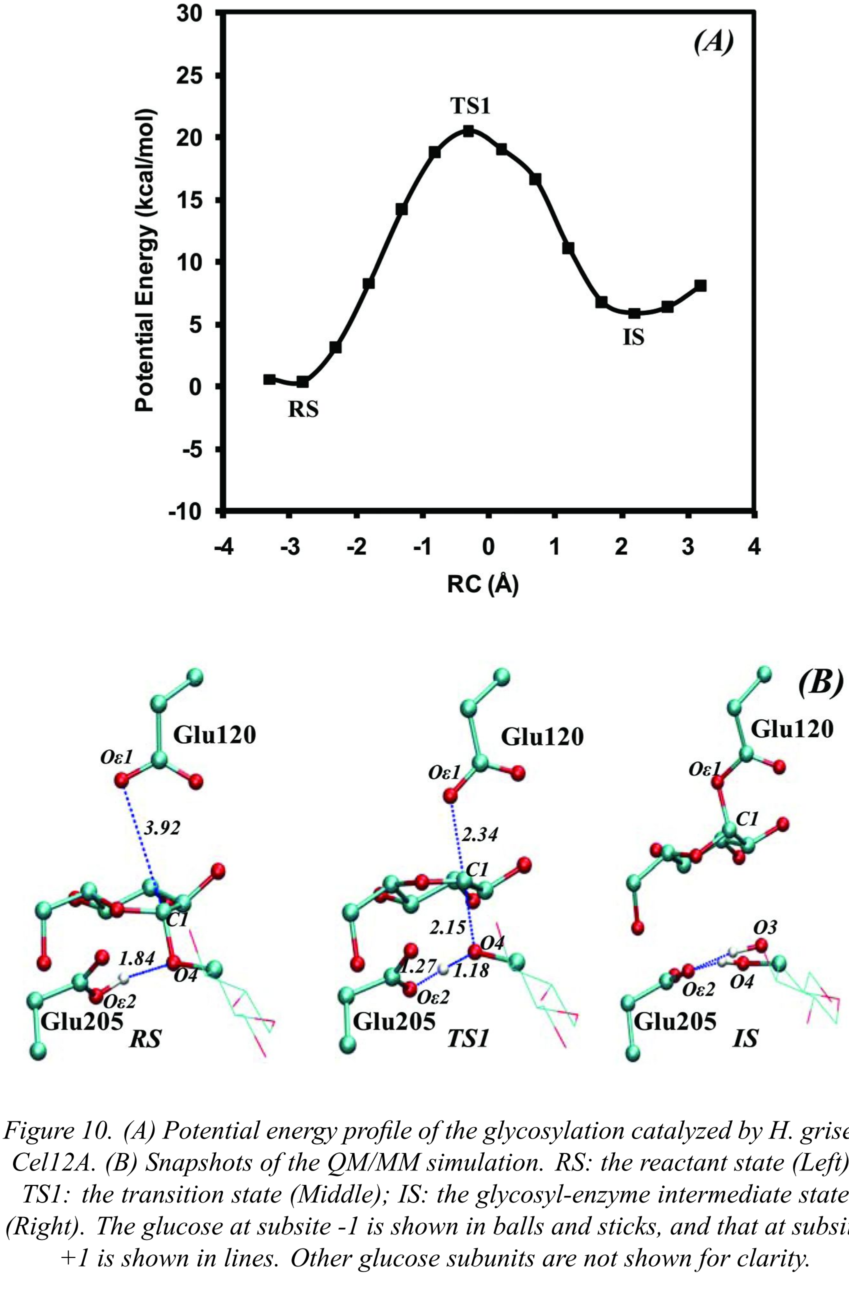


Figure 10. (A) Potential energy profile of the glycosylation catalyzed by H. grisea Cel12A. (B) Snapshots of the QM/MM simulation. RS: the reactant state (Left); TS1: the transition state (Middle); IS: the glycosyl-enzyme intermediate state (Right). The glucose at subsite -1 is shown in balls and sticks, and that at subsite +1 is shown in lines. Other glucose subunits are not shown for clarity.

Further simulations with Asp103 treated by QM may be helpful in understanding of the mechanism for Cel12A catalysis.

In the Michaelis complex, the 3-hydroxyl group at subsite +1 formed a hydrogen bond to Glu205. After the glycosylation, Glu205 acted as the hydrogen bond acceptor to the both 3-OH and 4-OH groups of the newly formed glycosyl-enzyme intermediate. Interestingly, the results were consistent with

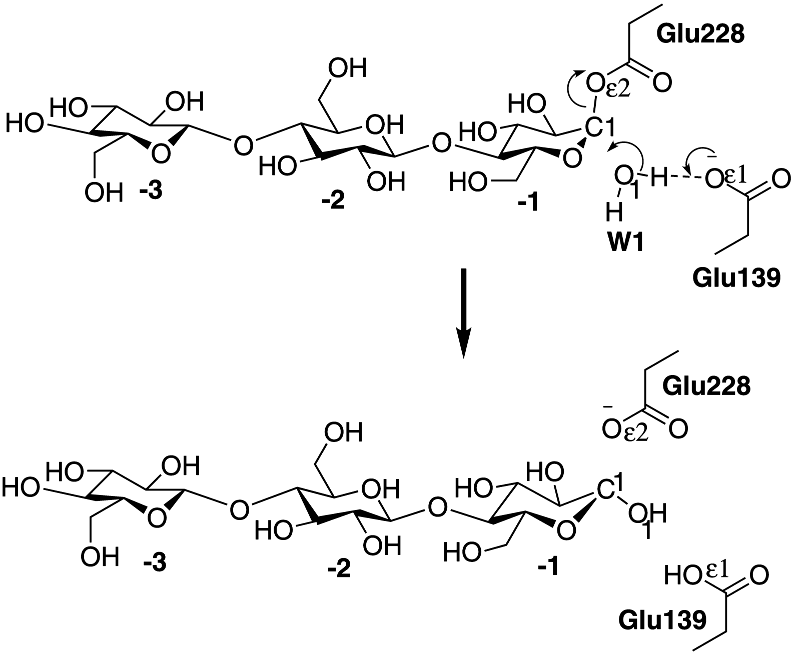


Figure 11. Active site of the B. agaradhaerens Cel5A glycosyl-enzyme intermediate (pdb code 1H11) (51), and the deglycosylation catalyzed by Cel5A.

another crystal structure of *H. grisea* Cel12A in complex with a β-D-cellobiose at subsites +1 and +2 (pdb code 1UU4) (*27*).

7.2. Deglycosylation Catalyzed by *B. agaradhaerens* Cel5A

The second model system was based on a glycosyl-enzyme intermediate of Cel5A, bound with a 2-deoxy-2-fluro-β-D-cellotrioside (pdb code 1H11, resolution 1.08 Å, Figure 11) (*51*). During the model building, the 2-fluro group of the substrate was manually changed to 2-hydroxyl group. The starting structure was therefore a glycosyl-enzyme intermediate of Cel5A with a β-D-cellotriose that occupied the -1 to -3 binding site.

The umbrella sampling (*62*) and weighted histogram analysis methods (WHAM) (*63*) were used to determine the PMF profile of the deglycosylation process in Cel5A [Figure 12]. In the glycosyl-enzyme intermediate of Cel5A, Oε2 of Glu228 was covalently bonded to C1 at subsite -1, and a water molecule, W1, hydrogen bonded to the general base residue Glu139. The QM region of this model included the subsite -1, Glu139 and Glu228 sidechains, as well as the W1 water molecule. The reaction coordinate (RC) was selected as:

$$RC = r(\text{C1-O}\varepsilon 2) - r(\text{C1-O1}) \qquad (2)$$

The atom labels are shown in Figure 11. The RC included the distance associated with the nucleophilic attack of the W1 water molecule [r(C1-O1)] and that for the bond breaking [r(C1- Oε2)]. The proton transfer from W1 to Glu139 happened spontaneously during the nucleophilic attack.

The free energy barrier for the deglycosylation was calculated to be 24.2 kcal/mol [Figure 12A]. The snapshots obtained from the QM/MM free energy simulations representing the glycosyl-enzyme intermediate state (IS), the structure near the transition state (TS2), and the product state (PS) in the deglycosylation are shown in Figure 12B. In IS, subsite -1 is in the $^4C_1$ chair conformation, similar to the glycosyl-enzyme intermediate after glycosylation catalyzed by *H. grisea* Cel12A (see above). For the structure near the transition state (TS2), subsite -1 adopted the $^4H_3$ conformation, which was converted back to an undistorted $^4C_1$ chair in the product state (PS). TS2 is the structure located at the top of free energy profile [Figure 12A] and represents the approximate transition state of deglycosylation. At TS2, *r*(C1-Oε2) and *r*(C1-O1) were 2.06 Å and 2.24 Å, respectively. The water molecule W1 did not transfer proton to Glu139 yet with *r*(Oε1-H) = 1.71 Å [Figure 12B]. The distances were averaged from 1000 frames of the corresponding window in the free energy simulations. Atom labels are shown in Figure 11 and Figure 12B.

The hydrogen bonding interactions in the deglycosylation process are also likely to play an important role. In the glycosyl-enzyme intermediate of deglycosylation in Cel5A (IS in Figure 12B), Tyr202 acts as a hydrogen bond donor to Glu139 and helps to maintain the position of Glu139 to act as the general base. With the progress of deglycosylation, this hydrogen bond was disturbed and vanished in the product state after the reaction. This may be due in part to the protonation and sidechain rotation of Glu139. Another interesting hydrogen bond donor was the 2-OH group of glucose at subsite -1. This hydrogen bond initially interacts with the Oε1 atom of the negatively charged Glu139. It was then switched to interact with the Oε2 atom of Glu228 (especially, near the product state) when the C1- Oε2 bond was broken and Glu228 was negatively charged. It seems that the orientation of the 2-OH group at subsite -1 was affected by the charge transfer from Glu139 to Glu228 during deglycosylation.

Recently, a QM/MM study (*65*) of Cel5A was published using the similar simulation approaches mentioned above (e.g., the SCC-DFTB QM method and PMF simulations), but based on a different X-ray structure. Specifically, the X-ray structure of Cel5A from *A. cellulolyticus* complexed with a cellotetrose substrate molecule (PDB ID: 1ECE) was used to generate the model for the investigation of the both glycosylation and deglycosylation processes by Liu *et al.* (*65*). One major difference between the deglycosylation process studied in Ref. (*65*) and our work discussed earlier is that the glycosyl-enzyme intermediate used by Liu *et al.* was generated from the simulations of the glycosylation step, while our study was directly based on the X-ray structure of the glycosyl-enzyme intermediate (*51*). Interestingly, our free energy barrier (24.2 kcal/mol) is somewhat lower than the one (29.7 kcal/mol) obtained by Liu *et al.* The experimental activation barrier for the hydrolysis of cellotetraose catalyzed by *Bacillus agaradherans Cel5A* was estimated to be about 19.4 kcal/mol (*65*, *66*). It is not clear at this stage as to why the free energy barriers for deglycosylation from the two simulations are different.

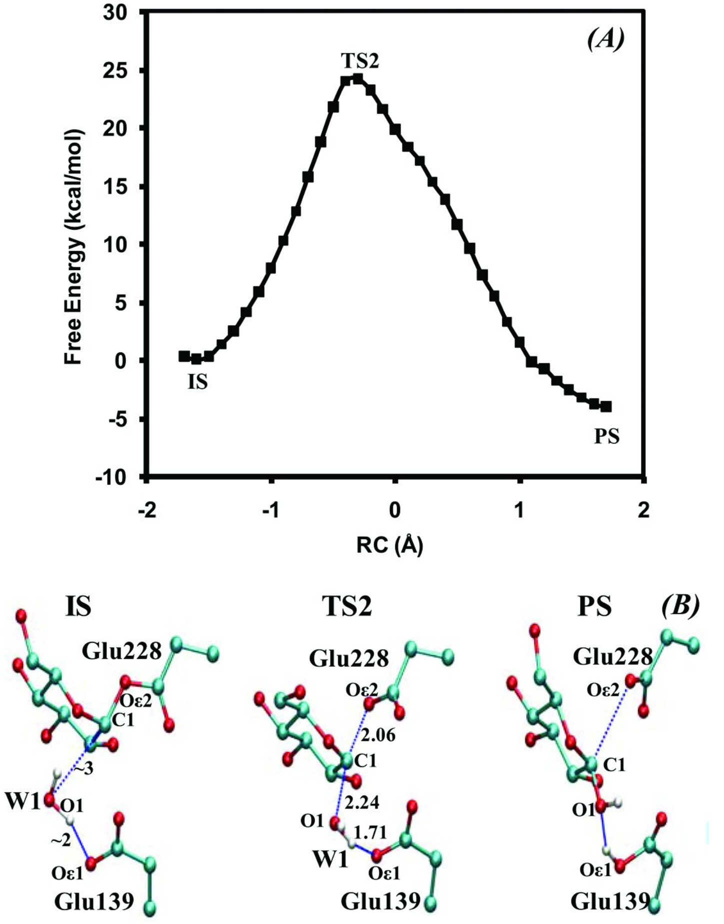


Figure 12. (A) Free energy profile of the deglycosylation catalyzed by B. agaradhaerens Cel5A. (B) Snapshots from the QM/MM free energy simulation. IS: the glycosyl-enzyme intermediate state (Left); TS2: the structure near the transition state (Middle); PS: the product state (Right). Only glucose at the subsite -1 is shown for clarity.

8. Conclusions

In summary, we selected two endoglucanases, the Michaelis complex of *H. grisea* Cel12A and the glycosyl-enzyme intermediate of *B. agaradhaerens* Cel5A, to study the glycosylation (Cel12A) and deglycosylation (Cel5A) steps, respectively, that are catalyzed through the retaining mechanism. The QM/MM methodology was used to calculate the potential energy and free energy profiles for the Cel12A and Cel5A systems, respectively. In the Cel12A Michaelis complex, the glucose at subsite -1 was in a $^{1}S_{3}$ skew boat conformation. The sugar ring distortion to a $^{4}C_{1}$ chair glycosyl-enzyme intermediate via $^{4}H_{3}$ transition state was found in the glycosylation catalyzed by Cel12A. The Cel5A glycosyl-enzyme intermediate adopts a $^{4}C_{1}$ chair conformation. Deglycosylation of Cel5A generated the cellulose product in a $^4C_1$ chair conformation through a $^4H_3$ transition state. Important hydrogen bonding interactions were observed in the enzyme active sites and shed some important light in understanding the action and mechanism of cellulase catalysis.

Acknowledgments

This work was supported in part by the National Science Foundation (Award No. 0817940 to H.G.) and the Petroleum Research Fund (to H.G.). The authors also acknowledge funding from the BioEnergy Science Center. BESC is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Additional support is provided by ORNL LDRD fund and DOE INCITE award (to J.C.S.).

References

1. Himmel, M. E.; Ding, S.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D. *Science* **2007**, *315*, 804–807.
2. Creuzet, N.; Berenger, J. F.; Frixon, C. *FEMS Microbiol. Lett.* **1983**, *20*, 347–350.
3. Henrissat, B.; Driguez, H.; Viet, C.; Schulein, M. *Biotechnology* **1985**, *3*, 722–726.
4. Irwin, D. C.; Specio, M.; Walker, L. P.; Wilson, D. B. *Biotechnol. Bioeng.* **1993**, *42*, 1002–1013.
5. Vocadlo, D. J.; Davies, G. J. *Curr. Opin. Chem. Biol.* **2008**, *12*, 539–555.
6. Bayer, E. A.; Belaich, J. P.; Shoham, Y.; Lamed, R. *Annu. Rev. Microbiol.* **2004**, *58*, 521–554.
7. Bayer, E. A.; Shimon, L. J. W.; Shoham, Y.; Lamed, R. *J. Struct. Biol.* **1998**, *124*, 221–234.
8. Bolam, D. N; Ciruela, A.; McQueen-Mason, S.; Simpson, P.; Williamson, M. P.; Rixon, J. E.; Boraston, A.; Hazlewood, G. P.; Gilbert, H. J. *Biochem. J.* **1998**, *331*, 775–781.
9. Black, G. W.; Rixon, J. E.; Clarke, J. H.; Hazlewood, G. P.; Theodorou, M. K.; Morris, P.; Gilbert, H. J. *Biochem. J.* **1996**, *319*, 515–520.
10. Din, N.; Damude, H. G.; Gilkes, N. R.; Miller, R. C.; Warren, R.; Kilburn, D. G. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 11383–11387.
11. Southall, S. M.; Simpson, P. J.; Gilbert, H. J.; Williamson, G.; Williamson, M. P. *FEBS Lett.* **1999**, *447*, 58–60.
12. Woodward, J.; Affholter, K. A.; Noles, K. K.; Troy, N. T.; Gaslightwala, S. F. *Enzyme Microb. Technol.* **1992**, *14*, 625–630.
13. Beguin, P.; Aubert, J. P. *FEMS Microbiol. Rev.* **1994**, *13*, 25–58.
14. Xiao, Z.; Gao, P.; Qu, Y.; Wang, T. *Biotechnol. Lett.* **2001**, *23*, 711–715.
15. Henrissat, B.; Bairoch, A. *Biochem. J.* **1993**, *293*, 781–788.
16. Henrissat, B.; Bairoch, A. *Biochem. J.* **1996**, *316*, 695–696.
17. Henrissat, B. *Mol. Microbiol.* **1997**, *23*, 848–849.

18. Clarke, A. J. *Biodegradation of Cellulose: Enzymology and Biotechnology*; CRC Press: Lancaster, PA, 1996.

19. Koshland, D. E. *Biol. Rev.* **1953**, *28*, 416–436.

20. Sinnott, M. L.; Souchard, I. J. *Biochem. J.* **1973**, *133*, 89–98.

21. Bause, E.; Legler, G. *Biochim. Biophys. Acta* **1980**, *626*, 459–465.

22. Uitdehaag, J. C. M.; Mosi, R.; Kalk, K. H.; van der Veen, B. A.; Dijkhuizen, L.; Withers, S. G.; Dijkstra, B. W. *Nat. Struct. Biol.* **1999**, *6*, 432–436.

23. Notenboom, V.; Birsan, C.; Nitz, M.; Rose, D. R.; Warren, R. A. J.; Withers, S. G. *Nat. Struct. Biol.* **1998**, *5*, 812–818.

24. Greig, I. R.; Williams, I. H. *Chem. Commun.* **2007**, 3747–3749.

25. Koivula, A.; Ruohonen, L.; Wohlfahrt, G.; Reinikainen, T.; Teeri, T. T.; Piens, K.; Claeyssens, M.; Weber, M.; Vasella, A.; Becker, D.; Sinnott, M. L.; Zou, J.; Kleywegt, G. J.; Szardenings, M.; Stahlberg, J.; Jones, T. A. *J. Am. Chem. Soc.* **2002**, *124*, 10015–10024.

26. Davies, G.; Henrissat, B. *Structure* **1995**, *3*, 853–859.

27. Sandgren, M.; Berglund, G. I.; Shaw, A.; Stahlberg, J.; Kenne, L.; Desmet, T.; Mitchinson, C. *J. Mol. Biol.* **2004**, *342*, 1505–1517.

28. Davies, G. J.; Wilson, K. S.; Henrissat, B. *Biochem. J.* **1997**, *321*, 557–559.

29. Parsiegla, G.; Juy, M.; Reverbel-Leroy, C.; Tardif, C.; Belaich, J. P.; Driguez, H.; Haser, R. *EMBO J.* **1998**, *17*, 5551–5562.

30. Guimaraes, B. G.; Souchon, H.; Lytle, B. L.; Wu, J. H. D.; Alzari, P. M. *J. Mol. Biol.* **2002**, *320*, 587–596.

31. (a) Guerin, D. M. A; Lascombe, M.; Costabel, M.; Souchon, H.; Lamzin, V.; Beguin, P.; Alzari, P. M. *J. Mol. Biol.* **2002**, *316*, 1061–1069. (b) Aleshin, A.; Golubev, A.; Firsov, L. M.; Honzatko, R. B. *J. Biol. Chem.* **1992**, *267*, 19291–19298.

32. Rouvinen, J.; Bergfors, T.; Teeri, T.; Knowles, J. K.; Jones, T. A. *Science* **1990**, *249*, 380–386.

33. Harjunpaa, V.; Teleman, A.; Koivula, A.; Ruohonen, L.; Teeri, T. T.; Teleman, O.; Drakenberg, T. *Eur. J. Biochem.* **1996**, *240*, 584–591.

34. Koivula, A.; Reinikainen, T.; Ruohonen, L.; Valkeajarvi, A.; Claeyssens, M.; Teleman, O.; Kleywegt, G. J.; Szardenings, M.; Rouvinen, J.; Jones, T. A.; Teeri, T. T. *Protein Eng.* **1996**, *9*, 691–699.

35. Ikuta, Y.; Karita, S.; Kitago, Y.; Watanabe, N.; Hirata, F. *Chem. Phys. Lett.* **2008**, *465*, 279–284.

36. Ikuta, Y.; Maruyama, Y.; Matsugami, M.; Hirata, F. *Chem. Phys. Lett.* **2007**, *433*, 403–408.

37. Hehre, E. J. In *Enzymatic degradation of insoluble carbohydrates*; Saddler, J. N., Penner, M. H., Eds.; ACS Symposium Series 618; American Chemical Society: Washington, DC, 1995; pp 68–78.

38. Becker, D.; Johnson, K. S.; Koivula, A.; Schulein, M.; Sinnott, M. L. *Biochem. J.* **2000**, *345*, 315–319.

39. Konstantinidis, A. K.; Marsden, I.; Sinnott, M. L. *Biochem. J.* **1993**, *291*, 883–888.

40. Damude, H. G.; Ferro, V.; Withers, S. G.; Warren, R. A. *Biochem. J.* **1996**, *315*, 467–472.

41. Biarnes, X.; Nieto, J.; Planas, A.; Rovira, C. *J. Biol. Chem.* **2006**, *281*, 1432–1441.
42. Petersen, L.; Ardevol, A.; Rovira, C.; Reilly, P. J. *J. Phys. Chem. B* **2009**, *113*, 7331–7339.
43. Sulzenbacher, G.; Driguez, H.; Henrissat, B.; Schulein, M.; Davies, G. J. *Biochemistry* **1996**, *35*, 15280–15287.
44. Tews, I.; Perrakis, A.; Oppenheim, A.; Dauter, Z.; Wilson, K.S.; Vorgias, C. E. *Nat. Struct. Biol.* **1996**, *3*, 638–648.
45. Cremer, D.; Pople, J. A. *J. Am. Chem. Soc.* **1975**, *97*, 1354–1358.
46. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
47. Woods, R. J.; Dwek, R. A.; Fraser-Reid, B. *J. Phys. Chem.* **1995**, *99*, 3832–3840.
48. Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1993**, *115*, 5745–5753.
49. Bagdassarian, C. K.; Schramm, V. L.; Schwartz, S. D. *J. Am. Chem. Soc.* **1996**, *118*, 8825–8836.
50. Biarns, X.; Ardvol, A.; Planas, A.; Rovira, C.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2007**, *129*, 10686–10693.
51. Varrot, A.; Davies, G. J. *Acta Crystallogr.* **2003**, *D59*, 447–452.
52. Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569–585.
53. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
54. Elstner, M.; Cui, Q.; Munih, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Comput. Chem.* **2003**, *24*, 565–581.
55. Riccardi, D.; Schaefer, R.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; , K.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458–6469.
56. Elstner, M.; Porezag, D.; Jungnickel, G.; Elstner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.
57. Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 5614–5621.
58. Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 10861–10873.
59. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
60. Brooks, C. L., III; Brunger, A.; Karplus, M. *Biopolymers* **1985**, *24*, 843–865.
61. Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733.
62. Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578–581.
63. Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
64. The 1W2U crystal was a complex with a soaked cellopentaose, but the fifth glucose unit did not show visible electron density. The current study used coordinates that were available from the complex with a cellotetraose.
65. Liu, J.; Wang, X.; Xu, D. *J. Phys. Chem. B* **2010**, *114*, 1462–1470.
66. Davies, G. J.; Dauter, M.; Brzozowski, A. M.; Bjornvad, M. E.; Anderson, K. V.; Schulein, M. *Biochemistry* **1998**, *37*, 1926–1932.

Chapter 8

Molecular Simulation Methods

Standard Practices and Modern Challenges

Michael Feig*

Department of Biochemistry & Molecular Biology, Department of Chemistry, Michigan State University, East Lansing, MI 48824
***feig@msu.edu**

Molecular simulations are used widely to study the structure, dynamics, and energetics of a given molecular system in atomic detail. The basic formalism underlying molecular dynamics and Monte Carlo simulations is described. Modern challenges in molecular systems are discussed. Emphasis is placed on model accuracy in current molecular force fields and the ability to reach sufficiently long time scales. Enhanced sampling methods are reviewed briefly and the chapter is concluded with an overview of emerging multi-scale techniques, in particular implicit solvent models and the construction and use of coarse-grained models.

MD: molecular dynamics; MC: Monte Carlo; QM: quantum mechanics; GB: Generalized Born; NVT: canonical ensemble; NVE: microcanonical ensemble; NPT: isothermal-isobaric ensemble; PMF: potential of mean force

The Basics

Dynamics and Conformational Sampling

The power of computer simulation methods (*1–3*) is their ability to generate single molecule dynamics and more generally conformational sampling of a given system in atomic detail. The most important information from such simulations is a description of the conformational ensemble that is visited at a given temperature.

Information about the relative probability of visiting different conformations directly translates into thermodynamic quantities that can be calculated from such data. Furthermore, it is possible to obtain kinetic information since dynamic processes that involve transitions over kinetic barriers can be observed directly. The resulting dynamic picture of a given system is highly complementary to experimental probes that provide a time- and ensemble-averaged static picture and/or do not provide full atomic resolution.

There are essentially two methods that are widely used to generate conformational sampling of a given molecular system. Molecular dynamics (MD) simulations generate deterministic trajectories in phase space, i.e. coordinates and velocities, which provide both thermodynamic and kinetic information. Monte Carlo (MC) sampling generates conformational states according to their probabilities in the canonical ensemble, but it does not result in a trajectory that is continuous in time so that kinetic information cannot be extracted directly. Both methods require a potential function to accurately describe molecular interactions and in both cases it is essential to sample all of the relevant (i.e. accessible with high probability) conformational space before otherwise anecdotal results become thermodynamically meaningful.

In the following, the basic components of molecular simulations are reviewed first before recent methods that address challenges in obtaining accurate interaction potentials and efficient conformational sampling are described.

Interaction Potential

Molecular simulations typically employ an atomistic model with a classical interaction potential as a compromise between accuracy and efficiency. The potential, often called “force field”, is commonly decomposed into bonded and non-bonded terms. The canonical form of such a force field is given in the following equation:

$$
\begin{aligned}
V(\mathbf{r}) = {} & \sum_{i=1}^{N_{bonds}} k_{bond,i}(d-d_0)^2 + \sum_{i=1}^{N_{angles}} k_{angle,i}(\theta-\theta_0)^2 \\
& + \sum_{i=1}^{N_{torsions}} k_{torsion,i}\left[1+\cos(n\phi-\phi_0)\right] + \sum_{i=1}^{N_{improper}} k_{improper,i}(\phi-\phi_0)^2 \\
& + \sum_{i=1}^{N_{atoms}-1} \sum_{j=i+1}^{N_{atoms}} \left(4\varepsilon_{LJ,ij}\left[\left(\frac{\sigma_{LJ,ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{LJ,ij}}{r_{ij}}\right)^{6}\right] + \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}} \right)
\end{aligned}
\tag{1}
$$

The bonded terms are concerned with preserving covalent bonding interactions which normally requires a quantum-mechanical treatment. However, as long as reactive processes are not considered, it is sufficient to maintain bond lengths (1-2 interactions) and angles (1-3 interactions) at equilibrium values. This is achieved with simple harmonic restraint potentials that are parameterized to match structural and vibrational data from crystallography, spectroscopy, and *ab initio* calculations (*4*).

Non-bonded terms consist of three contributions: 1) classical long-range electrostatic attraction and repulsion between partial atomic charges according to Coulomb's law; 2) short-range attraction according to van der Waals dispersion; and 3) hard-sphere like repulsion upon atomic contact formation to avoid electronic overlap. The van der Waals and contact terms are typically modeled as r^{-6} and r^{-12} terms, respectively, and combined into what is known as the Lennard-Jones potential (*5*). Non-bonded terms are applied in full strength to atoms separated by 4 or more covalent bonds (1-5 interactions and beyond) and are often scaled when applied to 1-4 interactions. 1-4 interactions are also represented by cosine series torsion potentials which are meant to explicitly reproduce the periodic energetic variation of eclipsed vs. staggered conformations. Although the non-bonded electrostatic interaction can in principle provide that effect, there is often a significant quantum mechanical component to 1-4 interactions that requires a mixing of bonded and non-bonded terms to obtain an accurate interaction potential. In practice, non-bonded terms are parameterized first with atomic partial charges and dispersion terms obtained from *ab initio* calculations while the torsion potential is parameterized last to provide the missing 1-4 interaction energies to match results from quantum mechanics.

Often, there is also a second torsion term which does not have a periodic form but instead restrains a torsion angle with a harmonic function. Such a term is typically used to restrain a set of four atoms to lie in a plane. It is called an improper torsion potential because the four atoms used in this term are not necessarily linked by consecutive covalent bonds. Such improper torsions often compensate for the lack of π-bonding interactions in the classical model that would otherwise maintain planar geometries, for example in aromatic ring systems.

Molecular Dynamics Simulations

Molecular dynamics (MD) simulations have long been known as an extremely powerful computational technique for studying the conformational dynamics of simple and complex molecular systems (*6*, *7*). In MD simulations, a dynamic trajectory is generated for each atom according to classical mechanics by following Newton's law of motion:

$$m_i \frac{d^2\mathbf{r}_i(t)}{dt^2} = \mathbf{F}_i(\mathbf{r}) = -\nabla_i V(\mathbf{r}) \tag{2}$$

where the force $\mathbf{F}_i(\mathbf{r})$ on atom i is due to interactions with all other atoms in the given system. The force in Eq. 2 is calculated from the gradient of the interatomic potential $V(\mathbf{r})$ as given in Eq. 1.

Integrator

Given a starting conformation of a molecular system with a set of velocities, Eq. 2 describes the evolution of the simulated system in phase space in a deterministic fashion. Because of the complex form of $V(\mathbf{r})$ it is not possible to find a closed-form solution. Instead, Eq. 2 is integrated in a step-wise fashion. Typically, a Verlet-type integrator (*8*) is used to maintain the energy and momentum conservation and time-reversibility that is afforded by the conservative Hamiltonian based on $V(\mathbf{r})$. In the often preferred velocity Verlet variant coordinates and velocities are advanced from the accelerations $\mathbf{a}_i = \mathbf{F}_i/m_i$ according to the following set of equations:

$$\begin{aligned} \mathbf{r}_i(t+\Delta t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{1}{2}\mathbf{a}_i(t)\Delta t^2 \\ \mathbf{v}_i(t+\Delta t) &= \mathbf{v}_i(t) + \frac{\mathbf{a}_i(t)+\mathbf{a}_i(t+\Delta t)}{2}\Delta t \end{aligned} \tag{3}$$

In this scheme, both coordinates and velocities are estimated with uncertainties of $O(\Delta t^3)$.

Choice of Time Step

The time step Δt in Eq. 3 has to be chosen so that the highest frequencies in the system are sampled sufficiently. In organic systems, the highest frequencies of about 3000 cm^{-1} are associated with C-H bond vibrations. As a rule of thumb the highest frequency should be sampled at least ten times during a full vibration which corresponds to a maximum time step of about 1 fs (*9*). A longer time step can be used if holonomic constraints are used to freeze bonds involving hydrogens (e.g. with algorithms such as SHAKE (*10*) or RATTLE (*11*)). This reduces the highest frequency in the system so that 2 fs time steps become practical in simulations of biomolecular and organic compounds. Even longer time steps usually lead to simulation instabilities due to poor energy conservation unless sophisticated multiple-time step schemes are employed (*12*, *13*).

Statistical Ensembles

MD simulations as described so far result in conformational sampling according to the microcanonical (NVE) ensemble because the total energy is conserved when integrating Newton's equation of motion. More relevant for practical applications are the canonical (NVT) and isobaric-isothermal (NPT) ensembles since environmental temperature and pressure are constant in many experiments and certainly most biological systems, at least over the time scales studied by simulation.

Simulations in the NVT ensemble can be achieved by periodically rescaling or reassigning velocities. Simple rescaling or reassignment of all velocities at fixed intervals to match a given target temperature is straightforward but only approximates an NVT ensemble. Frequent reassignment/rescaling overconstrains the temperature since even in an NVT ensemble, the instantaneous temperature fluctuates. On the other hand, infrequent reassignment/rescaling amounts to piecewise NVE simulations with occasionally altered kinetic energy rather than a true NVT simulation. A better approach is to loosely couple the system's temperature (or pressure) to a heat bath (or piston) via an appropriate thermostat. There are essentially two widely used algorithms that can correctly reproduce conformational sampling according to the NVT (or NPT) ensemble: They are Langevin dynamics (*14*, *15*) and the Nosé-Hoover thermostat (*16*, *17*).

Langevin dynamics models the effect of the atomistic interactions of a given system with the solvent in the heat bath in a physically intuitive manner. Langevin dynamics involves stochastic collisions that cause velocity reassignment of a randomly chosen particle in the system of interest as well as drag forces due to friction as a function of particle velocities in order to allow dissipation of energy back into the heat bath. Langevin dynamics employs the following modified equation of motion:

$$m_i \mathbf{a}_i(t) = -\nabla_i V(\mathbf{r}) - f_i m_i \mathbf{v}_i + \mathbf{F}_{random}(t) \qquad (4)$$

where f is the friction coefficient of the heat bath and $\mathbf{F}_{\text{random}}$ is the stochastic force simulating random collisions with solvent molecules. To maintain a physically accurate model, stochastic collisions and drag forces should only be applied to the parts of a system that would be in contact with the (fictitious) thermal bath. This poses technical challenges because typical simulations with explicit solvent involve periodic systems where the interface with the thermal bath is not clear. Instead, Langevin dynamics is often applied to all atoms in a given system thereby effectively altering the hydrodynamic characteristics of the explicitly represented solvent. Langevin dynamics is especially important for producing correct kinetics when the solvent environment is represented in an implicit fashion (see below) (*18*, *19*). In fact, Langevin dynamics by itself can be used as a primitive implicit solvent model for systems that do not interact strongly with the solvent environment (*20*).

Nosé-Hoover thermostats do not suffer from the limitations of Langevin dynamics but involve a theoretically motivated extended Lagrangian formalism that is physically less intuitive. Essentially, Nosé-Hoover dynamics introduces an extra degree of freedom to represent the thermal bath. This extra degree of freedom is propagated along with the rest of the system and can exchange kinetic energy with the rest of the system by scaling velocities of the real degrees of freedom appropriately. Nosé-Hoover dynamics involves the following modified equations of motion (*16*, *17*):

$$
\begin{aligned}
\dot{\mathbf{r}}_i &= \mathbf{p}_i / m_i \\
\dot{\mathbf{p}}_i &= -\nabla V_i(\mathbf{r}) - \zeta \mathbf{p}_i \\
\dot{\zeta} &= \frac{1}{q}\left(\sum_i \frac{\mathbf{p}_i^2}{m_i} - gkT \right)
\end{aligned}
\tag{5}
$$

where p_i is the momentum of particle i, g is the number of degrees of freedom, k is the Boltzmann constant, and T is the temperature of the thermal bath. Nosé-Hoover dynamics can be compared to Langevin dynamics by recognizing that the extended variable ζ acts like a friction constant. In contrast to Langevin dynamics, ζ is dynamically coupled to the temperature bath according to the coupling constant q and can have both positive and negative values. As a result, some amount of kinetic energy is exchange with the heat bath at every step of the simulation. This allows the kinetic energy to fluctuate appropriately and, in the end, producer correct statistics for an NVT ensemble.

Interpretation of Simulations and Convergence

The most immediate result from an MD simulation is a single molecule trajectory of a given molecular system. Although the corresponding molecular "movie" is visually impressive and can often provide important qualitative insight, the exact course of the trajectory is actually not especially meaningful. MD processes are essentially chaotic systems. As a consequence, an individual trajectory is extremely sensitive to the specific starting conditions. Changes in the starting coordinates of less than 10^{-3} Å, which is well below experimental uncertainties, typically result in an entirely different trajectory after propagation over hundreds of picoseconds. Furthermore, methodological uncertainties and the approximations inherent in the discrete integration of Newton's equation of motion result in deviations from what would be the "correct" trajectory if the same exact starting configuration could be prepared experimentally. Those deviations are amplified again because of the chaotic nature of MD. However, MD simulations *on average* generate a correctly weighted sampling of different conformations according to their relative energies. Simply stated, high energy conformations are not visited as often as low energy conformations because the gradient of the interaction potential always generates forces from high to low energies. High energy states are only reached if sufficient kinetic energy is available to overcome the opposing force due to the potential energy.

It is the averaged sampling of different conformational states (and transition paths between them) according to their relative statistical weights that is the most meaningful result from MD simulations. This requires that MD simulations can at least in principle visit all conformational states and that in reality they visit all "relevant" states. The first requirement is fulfilled if the sampling is ergodic, i.e. the time average of any thermodynamic property approaches the ensemble average in the limit of infinitely long simulations. Non-ergodic sampling is possible under certain conditions (*21*), but it is typically not an issue in common MD simulations. The second requirement goes to the heart of the issue of convergence. Simulations have to be long enough to actually visit all of the states that are within a few multiples of kT from the lowest energy conformation since those states are populated at a significant fraction and therefore contribute to any thermodynamic properties that are extracted from the simulation. While non-convergence is easy to diagnose, the achievement of full convergence can typically not be affirmed without previous characterization of a given system. Statistical sampling can be improved significantly with multiple MD simulations instead of a single long simulation and with enhanced sampling techniques (see below), but convergence is a constant concern with most MD simulations and the lack thereof has a direct impact on the reliability of quantitative predictions.

Monte Carlo Sampling

Monte Carlo (MC) simulations are conceptually much simpler than MD simulations. The basic Metropolis-MC algorithm consists of the following three steps (*22*): First, a new conformation is generated from a given initial conformation. Second, the energy of the trial conformation is calculated according to a given potential function. Third, the new conformation is either accepted or rejected as the starting conformation for the next cycle according to the probability:

$$P = \min\left(1, e^{-\frac{V(\mathbf{r}_2)-V(\mathbf{r}_1)}{kT}}\right) \tag{6}$$

The third step is commonly implemented by generating a random number between 0 and 1 that is compared against the probability P according to Eq. 6 (*22*). If the random number is less than P the new conformation is accepted, otherwise it is rejected.

MC simulations directly implement stochastic sampling according to the statistical weights in the canonical (NVT) ensemble as long as the selection of new conformations in the first step is ergodic and unbiased. In particular the detailed balance condition should be fulfilled, i.e. the probability of selecting a new conformation 2 from an initial conformation 1 should be the same as the probability of selecting 1 if 2 were instead given as the initial conformation. Otherwise, there is no restriction for how new conformations are generated during MC sampling. However, the efficiency of MC simulations depends on how often new conformations are accepted or, in other words, whether a significant fraction of the new conformations that are generated have a similar or lower energy than the initial structure. MC simulations can be potentially very powerful in cases where it is possible to “guess” new conformations on the other side of a transition barrier, thereby avoiding the kinetics of actually crossing the barrier that limits the conformational sampling in MD simulations. However, the price of such flexibility is the loss of a temporal relationship between consecutive conformations so that real-time dynamics and kinetics cannot be extracted in the same manner as from MD simulations. If kinetic rates are known, it is possible, however, to study the time evolution of a given system with the modified kinetic MC scheme (*23*).

In practice, MC simulations are used most frequently for simple systems or simple models of more complex systems where it is easy to design move sets that result in high acceptance ratios. In some cases, MC simulations are the only option, e.g. for lattice models where MD simulations that require a continuous space representation are not applicable. However, for fully solvated, condensed-phase molecular systems, MC simulations are usually only as effective as or even less effective than MD simulations because essentially all large conformational moves would result in high-energy states due to steric clashes (*24*). As a consequence, MC simulations are not widely used for such systems.

Quest for Accuracy

Force Field Accuracy

A critical factor in molecular simulations is the accuracy of the interaction potential which ultimately determines the level of realism that can be achieved. The canonical functional form of a molecular force field as given in Eq. 1 contains a large, but finite number of parameters that could in principle be optimized against a suitable set of target data to obtain an "optimal" force field. Target data may come from experimental data or *ab initio* calculations. Experimental data often involve ensemble- and time-averaged thermodynamic, structural, or dynamic quantities. In order to calculate such quantities with a given force field, time-consuming conformational sampling from long simulations is usually required. A systematic search of parameter space to optimize the agreement between the calculated and experimental data is therefore often impossible and a trial-and-error strategy is pursued instead. In the case of *ab initio* data, the calculation of target data from quantum mechanics is the limiting factor while the resulting optimized geometries, single point energies, and vibrational frequencies are readily calculated from a given force field. A systematic parameter optimization procedure base on *ab initio* data is thus feasible if enough target data is available.

To overcome some of the issues with force field parameter optimization, force fields are commonly designed in a modular fashion (*25*). In this case, each module corresponds to a certain chemical subunit that is parameterized through comparison with target data for a corresponding model compound. For example, phenol would serve as the model compound for parameterizing the amino acid side chain tyrosine and typical target data would involve crystallographic or *ab initio* geometries, heats of vaporization, free energies of solvation, and interaction energies with water molecules. For a sufficiently small chemical subunit, full parameter optimization is then possible.

A modular design also provides transferability of a given force field to a wide variety of molecules. However, it is not always straightforward to extend parameters optimized based on small compounds to larger molecules and to a variety of different environments. For example, partial charges or Lennard-Jones parameters optimized based on vacuum ab initio calculations or based on interaction energies with water molecules are not necessarily optimal for interactions in the interior of a large macromolecule or with a hydrophobic solvent. As a result, it is essentially impossible to obtain a universally optimal set of parameters (*26*). Rather, different sets of parameters may reflect different compromises between modularity, transferability, and accuracy for specific systems that is manifested in a large number of different established force fields for a given type of molecule.

There are a number of commonly used force field families (CHARMM (*25*), Amber (*27*), OPLS (*28*), and Gromos (*29*) for peptides and proteins; CHARMM (*30*) and Amber (*27*) for nucleic acids; CHARMM (*31*), Gromos (*32*), and Glycam (*33*) for lipids; CHARMM (*34*, *35*) and Glycam (*36*) for carbohydrates, TIP3P/TIP4P (*37*) or SPC/E (*38*) for water). In each family are multiple versions as a result of adjustments to better reflect target data. New target data has recently become available from experiments that focus more extensively on structure and dynamics of smaller model systems, such as NMR characterizations of short peptides. At the same time, increased computational power now allows for systematic *ab initio* calculations of a large number of different conformational states along the most important degrees of freedom of a given system (*39*). However, the main impetus for force field reparameterizations has resulted form an ability to run simulations over much longer time scales than when the force fields were originally parameterized. In many cases, such simulations have revealed, and continue to reveal, pathologies that had remained undetected previously. For example, the ability to run stable simulations of DNA over multiple nanoseconds following the introduction of the Ewald summation technique revealed that early CHARMM and Amber nucleic acid force fields strongly biased DNA conformations towards A- and B-forms, respectively (*40*). Both force fields were subsequently adjusted to shift the equilibrium towards B-DNA with the CHARMM force field and to facilitate sampling of A-DNA conformations with the Amber force field in response to A-DNA-inducing solvent environments.

A continuous focus in the development of peptide and protein force fields has been the accurate representation of the equilibrium between α, β, poly-proline II (PPII), and α_L backbone conformations because of its ramifications for the sampling of protein secondary structures. A recent innovation in this context has been the introduction of a spline-based cross-correlation map (CMAP) between overlapping ϕ and ψ backbone torsions which provides a mechanism for directly encoding a desired ϕ/ψ-map (*41*). While this approach is in principle extremely powerful, the problem remains with the choice of the target ϕ/ψ-map. An initial implementation within the CHARMM force field family based on an *ab initio* ϕ/ψ-map of alanine dipeptide has been able to resolve previous issues with overstabilized π-helices and resulted in improved stability in native protein simulations (*42*). However such a force field does not appear to provide the correct relative statistical weight between conformations in the α and β-basin (*43*). In particular, experimental data indicating that the PPII conformation within the β-basin is the preferred conformational state for short alanine-based peptides (*44*) is not reproduced well by the CHARMM force field with a CMAP potential based on alanine dipeptide *ab initio* data – or with most other force fields for that matter (*43*). One exception is the ff99sb version of the Amber force field which explicitly favors PPII conformations for short alanine peptides (*45*). However, the CHARMM/CMAP force field accurately represents the relative conformational sampling of α- and β-states in native proteins despite its apparent overstabilization of helical conformations at the peptide level (*46*). From these observations, it appears that recent force field versions might be converging when it comes to protein simulations (*47*), but it is likely that there will be further optimizations at the peptide level in future force field versions.

Polarizable Force Fields

A major approximation of commonly used force fields is the fixed nature of partial atomic charges. In reality, both solute and solvent will polarize in response to specific interactions. Polarization effects are most important when strong electrostatic fields are present or when the relative energetics of a given system in different environments are considered, for example in interactions with ions or transfer between aqueous and hydrophobic media. Polarization effects may involve conformational rearrangement and/or a redistribution of the electron densities. The conformational rearrangement is possible in fixed charge force field but electronic redistribution is not.

Over recent years, a number of efforts have attempted to incorporate electronic polarization in classical force fields.. There are essentially three main routes that have been followed (*48*): Fluctuating charges according to atomic polarizabilities (*49–52*), induced dipoles and higher-order multipoles (*53*, *54*), and Drude oscillators (*55*) that introduce a dynamically fluctuating off-center charge site. Even more so than for fixed charge force fields, parameterization is a major issue. Furthermore, polarizable force fields incur substantial additional cost, in part due to extra calculations, e.g. to calculate multipole-charge and mulipole-multipole interactions or to iteratively determine the charge response to molecular electric fields, and in part due to the need for shorter integration time steps of 0.5 to 1 fs to maintain stable trajectories. As a result, first applications of complex molecular simulations with polarizable force fields have only just began to appear (*56–58*), but it is likely that polarizable force fields will have a broader impact in the future.

QM/MM

Classical force fields, even polarizable force fields, still represent significant approximations of the quantum nature of atomic interactions. In particular, standard classical force fields do not allow bond-breaking or charge transfer and therefore exclude any type of reactive chemistry. While full quantum-mechanical sampling of larger molecular systems is not yet feasible, hybrid QM/MM (quantum mechanics/molecular mechanics) schemes are often employed to study reactive processes and other processes that are not well described by a classical force field (*59*). The idea of the QM/MM approach is to represent a small part of a given system, for example an enzyme active site, at the quantum mechanical level. The remaining part of the system and environment would be treated classically in order to save computer time.

The total energy of a QM/MM system consists of the QM energy for the quantum region, the MM energy for the classically treated surrounding parts of the system, and a coupling term that describes interactions between the MM and QM regions. While the calculation of MM and QM energies is relatively straightforward, the coupling term and the design of the QM-MM boundary present technical difficulties that are addressed in different QM/MM schemes. Although QM/MM methods have been proposed long time ago (*60*), QM/MM simulations remain very challenging and cost-intensive and are still largely at the developmental stage (*61*).

Quest for Speed

Time Scales

The need for long simulations to achieve convergence in molecular simulations essentially requires that the simulation length is on the order of the longest time scales of relevant dynamic processes. “Relevant processes” involve transitions to configurations with significant statistical weight or to configurations that are otherwise important for understanding a given system. For example, folding-unfolding transitions of proteins occur on long time scales up to seconds, but under native conditions the unfolded state of proteins is only populated to a very small extent for most proteins. Hence, simulations much shorter than folding time scales would provide meaningful quantitative results for many systems. On the other hand, rare native-state fluctuations with low statistical weight may be important for understanding dynamic intermediates or may represent conformations that are selected for in interactions with other molecular species.

Typical macromolecular time scales range from nanoseconds for minor conformational changes to hundreds of nanoseconds for loop motions and micro- to milliseconds for larger conformational rearrangements. The need for a 1-2 fs integration time step effectively limits total simulation times to microseconds, possibly tens of microseconds for small systems on current computer hardware (*62–64*). If polarizable force fields or QM/MM methods are employed, the time scales that can be reached are much shorter, on the order of tens of nanoseconds (*58*). This means that converged sampling is only possible for some dynamic molecular processes with standard all-atom MD simulations.

Many efforts have focused on bridging the discrepancy between what time scales *can* be reached with molecular simulations and what time scales *need* to be reached to study a variety of dynamic processes. There are essentially four directions that can be pursued to reach longer time scales in molecular simulations:

1) Faster computer hardware invariably results in longer simulations. Over the last three decades, improved computing power and in particular efficient use of parallel computers alone have increased simulation lengths by about 4-5 orders of magnitude. Further acceleration can come from specialized hardware such as the use of GPUs or custom-designed

hardware which has the potential to increase simulation times by another 2-3 orders of magnitude and bring millisecond simulations into the realm of possibility (*65*, *66*).

2) Algorithmic advances may allow MD simulations to be carried out faster without significant loss of accuracy. These address in particular the dominant cost of calculating non-bonded interactions. Most successful in this regard has been the Ewald summation technique which is a reformulation of the Coulomb potential for periodic systems as a sum of short- and long-range contributions in such a way that the long-range contribution can be calculated rapidly with the help of a fast Fourier transformation (*67*, *68*). Other ideas that have been met with partial success involve the use of multiple time steps for short and long-range interactions so that the long-range interactions do not have to be evaluated at every step (*12*, *13*) and the use of fast multipole schemes to approximate long-range electrostatic interactions (*69*).
3) The time scales of a molecular system are fundamentally determined by the kinetic barriers in a given energetic landscape. Simulations can be greatly accelerated if barrier crossing rates are enhanced. This may be achieved with a variety of biased and enhanced simulation techniques which increase the accessible time scales by several orders of magnitude. In such simulations, kinetic information is lost, but it is usually possible to recover unbiased equilibrium properties with appropriate reweighting schemes.
4) Model approximations decrease the system complexity so that longer simulations of larger systems can be run at reduced costs. This may be accomplished through implicit descriptions of the environment (*70*), coarse-graining (*71*, *72*), or more advanced multi-scale schemes (*73*), where lower-resolution representations and more accurate models are alternated.

In the following, enhanced sampling techniques and multi-scale simulation methodologies are described in more detail since they are widely used in modern MD simulations.

Enhanced Sampling Techniques

The general idea of all enhanced sampling techniques is to facilitate the crossing of kinetic barriers. Specific enhanced sampling techniques such as umbrella sampling methods target known barriers directly while non-specific enhanced sampling methods such as temperature replica exchange simulations accelerate the crossing of all barriers in a system and/or smooth the overall energy landscape.

Biased Sampling

The basic form of biased sampling, called umbrella sampling, involves the addition of a biasing potential to the standard interaction potential (*74*). Typically, the biasing potential has a harmonic form as given in Eq. 7:

$$U(\xi) = k\left(\xi - \xi_0\right)^2 \tag{7}$$

where ζ is a reaction coordinate along which there is a kinetic barrier at ζ_0 and k is the force constant. Such a potential biases sampling towards values of ζ near ζ_0 thereby compensating for the energetic cost associated with crossing of the barrier. The choice of a suitable reaction coordinate ζ is critical for the success of umbrella sampling methods, but not always straightforward, especially in more complex molecular systems.

It is possible to recover the unbiased probability distribution from the biased simulation. In the presence of an umbrella potential, simulations are carried out with the following effective energy function:

$$E_{biased}(r^N) = E_{unbiased}(r^N) + U_{umbrella}(\xi) \tag{8}$$

where the unbiased energy function consists of both kinetic and potential energies. The unbiased probability function as a function of the reaction coordinate ζ, $p_{unbiased}(\zeta)$, is formally given as:

$$p_{unbiased}(\xi) = \frac{\int \delta(\xi - \xi'(r^N)) e^{-\beta E_{unbiased}(r^N)} dr^N}{\int e^{-\beta E_{unbiased}(r^N)} dr^N} \tag{9}$$

Multiplication of Eq. 9 with $\int e^{-\beta E_{biased}(r_N)} dr^N / \int e^{-\beta E_{biased}(r_N)} dr^N = 1$ and substitution of $E_{unbiased}(r^N)$ according to Eq. 8 gives:

$$p_{unbiased}(\xi) = e^{\beta U_{umbrella}(\xi)} p_{biased}(\xi) e^{-\beta f} \tag{10}$$

where $e^{-\beta f}$ is a constant equal to the ratio of biased and unbiased partition functions:

$$e^{-\beta f} = \frac{\int e^{-\beta E_{biased}(r^N)} dr^N}{\int e^{-\beta E_{unbiased}(r^N)} dr^N} \tag{11}$$

The potential of mean force (PMF) or relative free energy in $w_{unbiased}(\xi)$ is then given as:

$$w_{unbiased}(\xi) = -U_{umbrella}(\xi) + w_{biased}(\xi) + f \tag{12}$$

which means that the unbiased PMF is simply the difference between the PMF from the biased simulation and the umbrella potential.

A single umbrella potential as given in Eq. 7 can compensate for the energetic cost associated with a transition barrier, but, at the same time, conformational sampling far away from the barrier is limited because of the harmonic form. In order to sample more broadly in ξ while still flattening the barrier, multiple overlapping umbrella simulations can be carried out with different values of ξ_0 (*75*). This approach also has the advantage that the exact height and location of the barrier does not need to be known *a priori*. A series of simulations each with different biasing functions $U_{umbrella,i}(\xi)$ would result in piecewise PMFs $w_i(\xi)$ that are each determined within the constant shift f in Eq. 12. If the PMFs overlap in ξ, a combined PMF $w(\xi)$ along the entire range of ξ can then be obtained by manually aligning the individual PMFs at the overlap regions or, more elegantly, with the weighted histogram analysis (WHAM) (*76*) or multi-state Bennett acceptance ratio methods (*77*).

Alternatively, it also possible to apply an umbrella potential with a different functional form, e.g. an inverted Gaussian centered at the barrier maximum which would bias sampling towards the barrier region but not affect regions far away from the barrier. The optimal biasing function would be the negative of the biased free energy surface. The resulting effective energy function would then be completely flat and the simulated motions on that surface would be barrierless and entirely diffusive. This approach is taken in multi-canonical simulations (*78*). One problem with this methodology is that the entire conformational landscape is usually not known beforehand and some form of adaptive sampling is usually required to build up the biasing function. Another problem is that a completely flat energy surface is not necessarily desirable because random diffusion in high-dimensional space can be less efficient than sampling on a sloped surface with moderate kinetic barriers that guides sampling to a small subspace of the configurational space as for example in protein folding (*79*, *80*).

A variant of adaptive umbrella sampling is the so-called metadynamics (*81*) method which is closely related to the earlier idea of conformational flooding (*82*). In this method, small Gaussians are placed successively at locations previously visited in a simulation thereby effectively raising the energy of an extensively sampled minimum to the height of the surrounding barriers until barrier crossing becomes feasible. Then the next minimum is filled up with Gaussian and so on until all of the low-energy conformational states have been visited.

Closely related to the meta-dynamics approach is a method called accelerated MD (*83*), which raises the energy of low-energy states irrespective of whether those states have been visited previously or not instead of lowering the kinetic barrier. This is accomplished by modifying the potential $V(r)$ below a given threshold E as follows:

$$V_{effective}(\mathbf{r}) = \begin{cases} V(\mathbf{r}) & V(\mathbf{r}) \geq E \\ V(\mathbf{r}) + \dfrac{(E - V(\mathbf{r}))^2}{\alpha + (E - V(\mathbf{r}))} & V(\mathbf{r}) < E \end{cases} \tag{13}$$

where a is an adjustable parameter that determines the shape of the function at the minima.

Another related technique is steered MD (SMD) (*84*) where a force in a certain direction is applied to overcome kinetic barriers. Either a constant-force or a constant-velocity scheme is applied. SMD simulations are more physically intuitive and constant-velocity simulations correspond directly to atomic force microscopy experiments. However, in contrast to umbrella sampling simulations, SMD simulations are non-equilibrium simulations thereby complicating the extraction of thermodynamic quantities. It is possible, though, to calculate thermodynamic quantities and in particular PMFs from an ensemble of non-equilibrium trajectories (*85*) but convergence is often more challenging than with umbrella sampling techniques.

Replica Exchange Simulations

Replica exchange simulations improve upon the idea of umbrella sampling through the parallel coupling of multiple biased simulations (*86*). The coupling provides an opportunity to exchange conformations between different biasing conditions (or equivalently exchange conditions) at frequent intervals. A typical replica exchange simulation involves N separate simulations, called replicas, in each of which a different biasing potential is applied. For example, a harmonic potential according to Eq. 7 may be applied with different values of ξ_0. At fixed intervals, typically on the order of 1 ps, the energies of neighboring replicas i and j are compared and an exchange is accepted in an MC-like fashion according to the probability:

$$P = \min\left(1, e^{-\left(\Delta U_{ij} + \Delta U_{ji}\right)/k_B T}\right) \qquad (14)$$

where $\Delta U_{ij} = U_i(\xi_j) - U_i(\xi_i)$ and $U_i(\xi_j)$ is the biasing potential with the value of ξ_0 for the j-th replica evaluated for the value of ξ at the i-th replica. This criteria essentially tests to what extent the conformation in the i-th replica is representative of the distribution under the biasing function of the j-th replica and vice versa.

Given sufficient overlap between neighboring replicas for exchanges to be accepted frequently, multiple replicas then contribute to the sampling for a given biasing potential thereby greatly improving convergence. Unbiased probability distributions are obtained from replica exchange simulations in the same fashion as from multiple sequential umbrella sampling runs.

The replica exchange scheme described so far is called Hamiltonian replica exchange since it involves a modification of the Hamiltonian operator for each replica. There are many variants of this scheme. E.g. it is possible to scale part of the interaction potential in different replicas instead of adding a biasing potential (*87*, *88*). It is also possible to use temperature as a means for biasing simulations (*89*). In temperature replica exchange simulations, each replica runs at a different temperature with the lowest temperature often corresponding to the temperature of interest. In temperature replica exchange simulations the criteria for accepting exchanges becomes:

$$P = \min\left(1, e^{-\left(1/kT_j - 1/kT_i\right)\left(E(q_i) - E(q_j)\right)}\right) \qquad (15)$$

where T_i and T_{ij} are the temperatures and $E(q_i)$ and $E(q_j)$ are the potential energies of adjacent replicas i and j.

The higher temperature replicas enhance the crossing of barriers in a non-specific manner simply by providing extra kinetic energy. This scheme is very attractive for studying previously uncharacterized systems since it does not require the identification of a suitable reaction coordinate for a specific biasing potential. As a result, temperature replica exchange simulations are widely used and are in fact the first type of replica exchange simulations that were developed. However, replica exchange simulations also have drawbacks. The most serious limitation is that elevated temperatures do not only accelerate the crossing of kinetic barriers but modify the entire free energy surface to favor high entropy states. This is exemplified by replica exchange simulations of proteins that will denature at high temperatures. Because the resulting unfolded conformations are statistically irrelevant for native state dynamics, replicas above the folding temperature essentially do not contribute anymore to the sampling of low-temperature replicas after equilibrium has been reached thereby limiting the effectiveness of the replica exchange scheme (*90*). Another limitation is that fluctuations in the energy difference in Eq. 15 decrease with increasing system sizes. As a result, densely spaced replicas are required to achieve sufficient overlap in systems with many degrees of freedom such as simulations using explicit solvent.

Multi-Scale Methods

Multi-scale methods aim at adapting the model resolution to the minimum level of detail that is required to address a given question, thereby avoiding unnecessarily detailed and computationally expensive models. Multi-scale schemes may involve representations at different scales for different parts of a given system, or low- and high-resolution representations may be alternated dynamically (*73*, *91–94*).

An example of a comprehensive multi-scale modeling approach is depicted in Fig. 1. Here, a chemical reaction center is represented quantum mechanically, surrounding areas are modeled classically in atomic detail, far away parts of the molecular complex are modeled at a coarse-grained level, and the surrounding solvent is modeled implicitly as a dielectric continuum. Such a model would be well suited to study the reaction center in the context of the entire solvated molecule. In contrast, this model would be less appropriate to investigate, for example, conformational changes of the entire molecule in response to different solvent conditions because neither the solute as a whole nor the solvent are represented in sufficient detail.

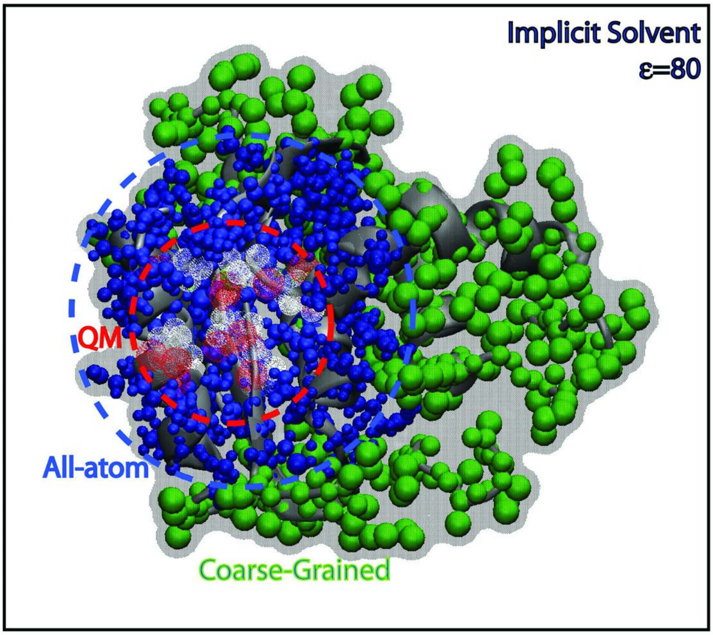


Figure 1. Multi-scale modeling scheme.

While a complete multi-scale model as shown in Fig. 1 poses technical challenges and requires a careful design to be effective, simplified schemes can be used more readily and are enjoying increasing popularity. One example is given by QM/MM simulations. Another popular combination is the use of an implicit representation of the environment in conjunction with all-atom models of a given solute. Finally, coarse-grained models of the solute either with explicit coarse-grained solvent or with implicit solvent are attractive for rapid sampling of very large complexes or of smaller systems over very long time scales. While such models often only provide qualitative insight, quantitative information can be obtained by reconstructing all-atom models from representative coarse-grained structures and re-evaluating those conformations with an all-atom energy function. Implicit solvent models and coarse-graining methods are described in more detail in the following.

Implicit Descriptions of Solvent Environments

Solvated molecular solutes often interact intimately with the surrounding solvent environment. The essential role of solvent in prescribing the structure and dynamics of biological macromolecules is well known (*80*, *95*), but solute-solvent interactions are equally important for many other types of solutes (*96*). It is therefore critical that solute-solvent interactions are represented accurately in molecular simulations (*97*). The canonical approach is to explicitly include solvent molecules in periodically replicating systems (*3*). In order to provide a bulk-like environment and avoid periodicity artifacts it is usually necessary to include at least three layers of solvent which translates into a significant number of solvent molecules. As a result, in most simulations with explicit solvent more computational time is spent on solvent-solvent interactions than on solute-solute and solute-solvent interactions. The cost for calculating solvent interactions becomes comparable to the cost for calculating solute interactions only for very large, globular solutes. Therefore, the computational cost of a molecular simulation can be reduced significantly with an implicit mean-field formalism that depends only on the conformational state and properties of a given solute and does not require any explicit solvent molecules. Of course, such formalism needs to preserve a sufficient level of realism and needs to be computationally efficient by itself since the advantage of an implicit solvent scheme would otherwise be lost.

The typical approach is to estimate the solvation free energy and add that term to the (vacuum) solute interaction potential to form an effective, implicit solvent interaction potential:

$$V_{effective}(\mathbf{r}) = V_{solute}(\mathbf{r}) + \Delta G_{solvation}(\mathbf{r}) \tag{16}$$

The solvation free energy arises due to electrostatic and non-polar interactions and often those contributions are estimated separately (*98*). Except for very hydrophobic environments, the electrostatic contribution is by far the largest contribution to the solvation free energy.

There are a number of different ways how the electrostatic solvation free energy may be estimated. They range from relative crude but inexpensive dielectric screening of electrostatic interactions (*99–101*) to empirical approaches based on the solvent-accessible surface areas of different residues and atom types (*102*) and physically more rigorous models that rely on a dielectric continuum (*103*, *104*) or fluctuating dipole approximation of the solvent environment (*105*). Implicit solvent models based on dielectric continuum theory are most widely used today and will be described in more detail in the following.

The basic idea of dielectric continuum models is to maintain a fully atomistic representation of the solute and model the environment as a dielectric medium with a given dielectric function ε to reflect the polarizability of the solvent. Because the dielectric response is nearly constant over the frequency range corresponding to molecular fluctuations, only the static, constant contribution to the dielectric function is considered. Such a model is rigorously described by the Poisson equation:

$$\nabla\left[\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r})\right] = -4\pi\rho(\mathbf{r}) \tag{17}$$

where the electrostatic potential $\phi(\mathbf{r})$ throughout space is related to an explicit charge distribution $\rho(\mathbf{r})$ and the distribution of the dielectric constant. For a solute in aqueous solvent, $\varepsilon(\mathbf{r})$ would typically have a value of 1 inside the solute cavity where the explicit charges are present and a value of 80 everywhere else. A higher interior dielectric constant is sometimes used when structures are not thermalized appropriately as in minimized structures or average structures from experiments (*106*).

Eq. 17 can be solved for the electrostatic potential using grid-based, iterative finite difference techniques (*107*, *108*). The electrostatic solvation free energy can then be readily calculated from the electrostatic potential. However, this route is computationally not very attractive because convergence of finite difference methods is poor and many iterations and fine grids are required to obtain accurate results (*109*, *110*). In practice, the much less costly Generalized Born (GB) approximation is therefore used instead (*111*, *112*):

$$\Delta G_{elec} = -\frac{1}{2}\left(1-\frac{1}{\varepsilon}\right)\sum_{i,j}\frac{q_i q_j}{\sqrt{r_{ij}^2+\alpha_i\alpha_j e^{-r_{ij}^2/F\alpha_i\alpha_j}}} \qquad (18)$$

where q_i are partial atomic charges of the solute from a given force field, r_{ij} are pair-wise atomic distances, F is an adjustable parameter, ε is the dielectric constant of the environment, and the generalized Born radii α_i are calculated essentially as a function of the solute density surrounding a given charge site i.

The non-polar contribution to the free energy of solvation consists of van der Waals solute-solvent interactions and the cost associated with cavity formation. The cost of cavity formation is approximately proportional to the solvent accessible surface area or the solvent excluded volume (*113–115*). Formalisms for implicit van der Waals interactions have also been proposed (*116*), but often the entire non-polar contribution to the solvation free energy is combined in a single term.

Implicit solvent models based on the GB formalism and a solvent-accessible surface area based non-polar term can be surprisingly effective. There are now numerous example where implicit solvent simulations produced essentially equivalent dynamics to fully explicit solvent simulations. However, it is clear that implicit solvent treatments are not appropriate for all types of applications, especially when very specific solvent interactions are at play. In these cases, fully explicit solvent or hybrid implicit/explicit solvent schemes cannot be avoided.

Implicit solvent models can be extended to low-dielectric environments (*117*) and heterogeneous environments, such as membranes. Membrane environments can be represented as layered implicit solvents with different dielectric constants in the membrane interior (*118*, *119*) or different empirical solvation contributions (*120*).

Coarse-Graining

In coarse-grained models, the solute itself is approximated by combining several atomic interactions sites into a coarse-grained particle. A new interaction potential is then designed at the coarse-grained level. Coarse-grained models can be explored with either MD or MC simulations. There are a great variety of such models ranging from very low-resolution models where a coarse-grained particle may correspond to an entire molecule or even multiple closely interacting molecules to models at near-atomic resolution where only some atoms (e.g. C-H groups) are combined while others remain in full atomic detail. The level of coarse-graining that is chosen in practice depends on the best comprise between computational speed and model accuracy for a given application.

Except for the highest-resolution models, coarse-grained interaction potentials consist primarily of empirical terms which limit transferability of such models. Coarse-grained interaction potentials may either apply generally to a certain type of molecules, thereby conferring some degree of transferability, or they may encode specific interactions of a given system. An example of a more general coarse-grained interaction potential is the well-studied H-P model of proteins which consists of a single particle per residue (*121*). The coarse-grained particles are either hydrophobic (H) or polar (P) based on the amino acid that they represent. A contact-based potential is then defined for H-H, H-P, and P-P interactions. Such models have been used extensively for fundamental studies of protein folding (*122*). More sophisticated variants have been used in the context of protein structure prediction (*123*, *124*). Recently, new models have been proposed that rely less on empirical terms and thereby improve transferability (*125–127*). This latest generation of coarse-grained models will likely be more widely applicable and lead to broader use of coarse-grained models.

System-specific coarse-grained models directly reflect knowledge about intra-molecular interactions from structural data, all-atom simulations, or other experiments. Such models generally lack transferability to other systems and they may have limited predictive abilities beyond what is already known about a given system. Nevertheless such models have become popular recently because they offer the most direct route to very long time-scale simulations of large systems that are otherwise intractable (*128*). A common approach for designing such models involves interaction sites that are connected with harmonic springs to reflect either bonded or non-bonded interactions. In the simplest form, the connectivity is determined from structural data and a common spring constant is used for all interactions (*129*). More sophisticated schemes may obtain pairwise or higher-order interaction potentials from molecular dynamics simulations through Boltzmann inversion (*128*) or force matching (*130*).

A hybrid between the first and second types is the so-called Go model which combines general interaction parameters with a term that strongly biases towards native-like intramolecular contact formation (*131*). Go-models are used widely in the study of protein folding since they practically guarantee that the native state is reached in folding simulations (*132*).

Summary

Molecular simulations have come a long way since the first simulations were published many decades ago. It is becoming possible to routinely reach microseconds even for relatively large systems and we will likely see the first millisecond simulations in the near future. Furthermore, the current generation of interaction potentials is much more realistic than earlier versions. Nevertheless, the application of simulations to real problems remains challenging. There is still room for improving the accuracy of fixed charge force fields and the routine inclusion of polarizability in simulations of non-trivial systems is only just becoming reality. Another issue is that despite ever increasing computer power simulations are still often too short to achieve full convergence. Much progress has been made in the development of enhanced sampling protocols to accelerate the crossing of barriers, but even those methods are often not sufficient to fully sample the relevant dynamics in a complex molecular system. A promising direction is the use of reduced representations either for the solvent or solute part of a given molecular system. Reduced representations are becoming standard tools for meeting the desire to simulate ever longer time scales and larger system sizes.

References

1. Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: San Diego, CA, 2002.
2. Leach, A. *Molecular Modelling: Principles and Applications*, 2nd ed.; Prentice Hall: 2001.
3. Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; 1st ed.; Oxford University Press: New York, 1987.
4. Mackerell, A. D. *J. Comput. Chem.* **2004**, *25*, 1584–604.
5. Lennard-Jones, J. E. *Proc. Phys. Soc.* **1931**, *43*, 461–82.
6. Rahman, A. *Phys. Rev. A* **1964**, *136*, 405–11.
7. McCammon, J. A.; et al. *Nature* **1977**, *267*, 585–90.
8. Verlet, L. *Phys. Rev.* **1967**, *159*, 98–103.
9. Yasri, A.; et al. *Protein Eng.* **1996**, *9*, 959–76.
10. Ryckaert, J. P.; et al. *J. Comput. Phys.* **1977**, *23*, 327–41.
11. Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24–34.
12. Zhou, R. H.; et al. *J. Chem. Phys.* **2001**, *115*, 2348–58.
13. Izaguirre, J. A.; et al. *J. Chem. Phys.* **1999**, *110*, 9853–64.
14. Izaguirre, J. A.; et al. *J. Chem. Phys.* **2001**, *114*, 2090–8.
15. Adelman, S. A.; Brooks, C. L. *J. Phys. Chem.* **1982**, *86*, 1511–24.
16. Hoover, W. G.; et al. *Physica D* **2004**, *187*, 253–67.
17. Nose, S. *Mol. Phys.* **1984**, *52*, 255–68.
18. Feig, M. *J. Chem. Theory Comput.* **2007**, *3*, 1734–48.
19. Zagrovic, B.; Pande, V. *J. Comput. Chem.* **2003**, *24*, 1432–6.
20. Levy, R. M.; et al. *Chem. Phys. Lett.* **1979**, *65*, 4–11.
21. Watanabe, H.; Kobayashi, H. *Phys. Rev. E* **2007**, *75*, 040102.
22. Nicholas, M.; et al. *J. Chem. Phys.* **1953**, *21*, 1087–92.
23. Bortz, A. B.; et al. *J. Comput. Phys.* **1975**, *17*, 10–8.
24. Yamashita, H.; et al. *Chem. Phys. Lett.* **2001**, *342*, 382–6.
25. MacKerell, A. D., Jr.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586–616.
26. Sato, F.; et al. *J. Phys. Chem. A* **2003**, *107*, 248–57.
27. Cornell, W. D.; et al. *J. Am. Chem. Soc.* **1995**, *117*, 5179–97.
28. Jorgensen, W. L.; et al. *J. Am. Chem. Soc.* **1996**, *118*, 11225–36.

29. Oostenbrink, C.; et al. *J. Comput. Chem.* **2004**, *25*, 1656–76.
30. Foloppe, N.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2000**, *21*, 86–104.
31. Feller, S. E.; MacKerell, A. D. *J. Phys. Chem. B* **2000**, *104*, 7510–5.
32. Taylor, J.; et al. *Biochim. Biophys. Acta, Biomembr.* **2009**, *1788*, 638–49.
33. Tessier, M. B.; et al. *Mol. Simul.* **2008**, *34*, 349–63.
34. Kamath, G.; et al. *J. Chem. Theory Comput.* **2008**, *4*, 765–78.
35. Kuttel, M.; et al. *J. Comput. Chem.* **2002**, *23*, 1236–43.
36. Kirschner, K. N.; et al. *J. Comput. Chem.* **2008**, *29*, 622–55.
37. Jorgensen, W. L.; et al. *J. Chem. Phys.* **1983**, *79*, 926–35.
38. Hermans, J.; et al. *Biopolymers* **1984**, *23*, 1513–8.
39. MacKerell, A. D., Jr.; et al. *J. Am. Chem. Soc.* **2004**, *126*, 698–9.
40. Feig, M.; Pettitt, B. M. *J. Phys. Chem. B* **1997**, *101*, 7361–3.
41. MacKerell, A. D., Jr.; et al. *J. Comput. Chem.* **2004**, *25*, 1400–15.
42. Feig, M.; et al. *J. Phys. Chem. B* **2003**, *107*, 231–6.
43. Best, R. B.; et al. *Biophys. J.* **2008**, *95*, L7–L9.
44. Graf, J.; et al. *J. Am. Chem. Soc.* **2007**, *129*, 1179–89.
45. Hornak, V.; et al. *Proteins* **2006**, *65*, 712–25.
46. Feig, M. *J. Chem. Theory Comput.* **2008**, *4*, 1555–64.
47. Price, D. J.; Brooks, C. L., III. *J. Comput. Chem.* **2002**, *23*, 1045–57.
48. Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236–42.
49. Rick, S. W.; et al. *J. Chem. Phys.* **1994**, *101*, 6141–56.
50. Kaminski, G. A.; et al. *J. Comput. Chem.* **2002**, *23*, 1515–31.
51. Patel, S.; et al. *J. Comput. Chem.* **2004**, *25*, 1504–14.
52. Yang, Z. Z.; Zhang, Q. *J. Comput. Chem.* **2006**, *27*, 1–10.
53. Xie, W. S.; et al. *J. Chem. Theory Comput.* **2007**, *3*, 1878–89.
54. Ren, P. Y.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–47.
55. Lamoureux, G.; et al. *J. Chem. Phys.* **2003**, *119*, 5185–97.
56. Kim, B. C.; et al. *J. Phys. Chem. B* **2005**, *109*, 16529–38.
57. Yang, Z. Z.; et al. *J. Theor. Comput. Chem.* **2008**, *7*, 697–705.
58. Patel, S.; Brooks, C. L. *Mol. Simul.* **2006**, *32*, 231–49.
59. Gao, J. L.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467–505.
60. Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–49.
61. Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–229.
62. Mura, C.; McCammon, J. A. *Nucleic Acids Res.* **2008**, *36*, 4941–55.
63. Monticelli, L.; et al. *J. Comput. Chem.* **2008**, *29*, 1740–52.
64. Freddolino, P. L.; et al. *Biophys. J.* **2008**, *94*, L75–L7.
65. Stone, J. E.; et al. *J. Comput. Chem.* **2007**, *28*, 2618–40.
66. Larson, R. H.; et al. *2008 IEEE 14th International Symposium on High Peformance Computer Architecture* **2008**, 303–14.
67. Ewald, P. P. *Ann. Phys.* **1921**, *64*, 253–87.
68. Darden, T. A.; et al. *J. Chem. Phys.* **1993**, *98*, 10089–92.
69. Kudin, K. N.; Scuseria, G. E. *Chem. Phys. Lett.* **1998**, *283*, 61–8.
70. Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–200.
71. Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469–73.
72. Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–50.
73. Feig, M.; et al. *J. Mol. Graphics Modell.* **2004**, *22*, 377–95.
74. Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578–81.

75. Beveridge, D. L.; Dicapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431–92.
76. Kumar, S.; et al. *J. Comput. Chem.* **1992**, *13*, 1011–21.
77. Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*.
78. Okamoto, Y. *J. Mol. Graphics Modell.* **2004**, *22*, 425–39.
79. Brooks, C. L.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11037–8.
80. Bryngelson, J. D.; et al. *Proteins* **1995**, *21*, 167–95.
81. Micheletti, C.; et al. *Phys. Rev. Lett.* **2004**, *92*.
82. Grubmuller, H. *Phys. Rev. E* **1995**, *52*, 2893–906.
83. Hamelberg, D.; et al. *J. Chem. Phys.* **2004**, *120*, 11919–29.
84. Balsera, M.; et al. *Biophys. J.* **1997**, *73*, 1281–7.
85. Park, S.; Schulten, K. *J. Chem. Phys.* **2004**, *120*, 5946–61.
86. Murata, K.; et al. *Chem. Phys. Lett.* **2004**, *385*, 1–7.
87. Liu, P.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13749–54.
88. Su, L.; Cukier, R. I. *J. Phys. Chem. B* **2007**, *111*, 12310–21.
89. Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–51.
90. Zheng, W. H.; et al. *J. Phys. Chem. B* **2008**, *112*, 6083–93.
91. Amato, F.; et al. *Biomed. Microdevices* **2006**, *8*, 291–8.
92. Heath, A. P.; et al. *Proteins* **2007**, *68*, 646–61.
93. Arkhipov, A.; et al. *Biophys. J.* **2008**, *95*, 2806–21.
94. Sherwood, P.; et al. *Curr. Opin. Struct. Biol.* **2008**, *18*, 630–40.
95. Helms, V. *ChemPhysChem* **2007**, *8*, 23–33.
96. Roccatano, D. *Curr. Protein Pept. Sci.* **2008**, *9*, 407–26.
97. Kollman, P. A. *Acc. Chem. Res.* **1996**, *29*, 461–9.
98. Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.
99. Krol, M. *J. Comput. Chem.* **2003**, *24*, 531–46.
100. Schaefer, M.; et al. *Theor. Chem. Acc.* **1999**, *101*, 194–204.
101. Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133–52.
102. Wesson, L.; Eisenberg, D. *Protein Sci.* **1992**, *1*, 227–35.
103. Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–43.
104. Feig, M.; Brooks, C. L., III. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–24.
105. Papazyan, A.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 11254–64.
106. Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301–32.
107. Gilson, M. K.; et al. *J. Comput. Chem.* **1987**, *9*, 327–35.
108. Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671–9.
109. Feig, M.; et al. *J. Comput. Chem.* **2004**, *25*, 265–84.
110. Zhou, Y. C.; et al. *J. Comput. Chem.* **2008**, *29*, 87–97.
111. Still, W. C.; et al. *J. Am. Chem. Soc.* **1990**, *112*, 6127–9.
112. Feig, M.; Brooks, C. L., III. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–24.
113. Sitkoff, D.; et al. *J. Phys. Chem.* **1994**, *98*, 1978–88.
114. Tan, C.; et al. *J. Phys. Chem. B* **2007**, *111*, 12263–74.
115. Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8331–6.
116. Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–99.
117. Feig, M.; et al. *J. Chem. Phys.* **2004**, *120*, 903–11.
118. Im, W.; et al. *Biophys. J.* **2003**, *85*, 2900–18.

119. Tanizaki, S.; Feig, M. *J. Chem. Phys.* **2005**, *122*, 124706.
120. Lazaridis, T. *Proteins* **2003**, *52*, 176–92.
121. Dill, K. A. *Biochemistry* **1985**, *24*, 1501–9.
122. Dill, K. A.; et al. *Protein Sci.* **1995**, *4*, 561–602.
123. Kolinski, A.; Skolnick, J. *Proteins* **1994**, *18*, 338–52.
124. Hinds, D. A.; Levitt, M. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 2536–40.
125. Marrink, S. J.; et al. *J. Phys. Chem. B* **2004**, *108*, 750–60.
126. Monticelli, L.; et al. *J. Chem. Theory Comput.* **2008**, *4*, 819–34.
127. Moritsugu, K.; Smith, J. C. *Biophys. J.* **2008**, *95*, 1639–48.
128. Arkhipov, A.; et al. *Structure* **2006**, *14*, 1767–77.
129. Jernigan, R. L.; Bahar, I. *Curr. Opin. Struct. Biol.* **1996**, *6*, 195–209.
130. Zhou, J.; et al. *Biophys. J.* **2007**, *92*, 4289–303.
131. Taketomi, H.; et al. *Int. J. Pept. Protein Res.* **1975**, *7*, 445–59.
132. Shea, J.-E.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 12512–7.

Chapter 9

Quantum Mechanical Modeling of Sugar Thermochemistry

Joshua Engelkemier[1] and Theresa L. Windus[1,2,*]

[1]Department of Chemistry, Iowa State University, Ames, IA 50011
[2]Ames Laboratory, Ames, IA 50011
[*]Theresa@fi.ameslab.gov

The recently developed homodesmotic hierarchy for hydrocarbons is extended to include oxygen so that accurate thermochemical quantities for sugars and sugar polymers may be computed with relatively small computational cost. In particular, the method will allow for the determination of heats of formation, which can be used to determine bond strengths important in the decomposition of sugars in, for example, the pyrolysis of biomass. This chapter includes a brief review of the current methodology for calculating thermodynamic properties using electronic structure methods and a description of the proposed extensions. Preliminary results using the lowest members of the hierarchy give a standard heat of formation value of β-D-glucopyranose-gg to be approximately 250 to 260 kcal/mol. These results are promising, and future work will include the calculation of highly accurate building blocks on which this method is based.

Many reports describe the multiple challenges associated with the composition of biomass to useful fuels (*1*). Of critical importance to the conversion is understanding the decomposition of lignocellulose whose main components – cellulose, hemicellulose and lignin – are difficult to break into constituent sugar components because of the polymeric nature of the material that acts to "harden" the material and prevent its decomposition. Even when the sugars are released from the biomass, there is still the challenge of converting the sugar to fuel (primarily ethanol and biodiesel) in an energetically and

In Computational Modeling in Lignocellulosic Biofuel Production; Nimlos, M., et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2010.

environmentally conservative manner (i.e., using the least amount of energy to accomplish the conversion in a way that will not produce more environmental issues). To understand the decomposition of the source materials requires a detailed understanding of the thermodynamics and kinetics of each of the building blocks (sugars and hydrocarbons) that the source material is composed of. Additionally, the energy involved in bond breaking during decomposition must be understood to predict product formation and to "disrupt" the current process in such a way as to produce more of the desired products. All of these issues point to the need for accurate thermochemistry for these quite large systems. Unfortunately, there is a distinct lack of such information for these systems from both experimental and computational sources. While some specific species have been examined (*2*), to the best of our knowledge, a systematic examination of the sugars and their decomposition pathways has not been undertaken. Because of the computationally intense nature of many of the current methods, this is not a trivial undertaking.

We propose an extension to the recent hydrocarbon homodesmotic hierarchy (*3*) to include oxygen. While our motivation is the examination of sugars and products of biomass conversion, the proposed extensions apply to many oxygen-containing species. The next section discusses the two main quantum mechanical methods, composite and balanced reactions, for obtaining accurate thermochemical quantities. It includes a detailed description of the homodesmotic hierarchy that extends to include oxygen important in sugars, which is discussed in the following section. The section after that reviews the application of the hierarchy to obtain the heat of formation of β-D-glucopyranose-gg. Finally, conclusions are given.

Quantum Mechanical Methodology

The chemical community has long been concerned with calculating accurate thermochemical quantities such as heats of formation, ionization energies, proton affinities, and dissociation energies because these are the basic building blocks for understanding the stability of reactants and products and their reactions. In addition, these quantities are of interest to experimentalists and can be directly compared to experimental values when they are available. Therefore, much effort has been put into evaluating these quantities to obtain "chemical accuracy." For example, heats of formation are one of the most fundamental quantities because many other properties (such as bond dissociation energies) can be derived from them. Chemical accuracy for this property usually refers to being within 1 kcal/mol of the actual experimental value. Unfortunately, very accurate experimental values are not always available, making benchmarking methods a challenge. However, using available experimental values allowsthe theoretical community to develop several different methodologies that can roughly be separated into two different categories: additive or composite methods and balanced equation methods. Both of these methods and the common theories in each category are described below.

Composite Methods

Calculating molecular energies is the first step in determining heats of formation to chemical accuracy and requires extremely accurate calculations, usually at great computational expense. These molecular energies can then be used with atomic information to determine the overall heats of formation. However, for anything other than very small systems, calculating the molecular energy accurately is challenging. To overcome this, multiple methods have been determined to either add together information of many lower-level computations or to extrapolate to a molecular energy that includes both complete basis sets and high levels of electron correlation. Because the idea is to approximate the results of a very high theory level with a lower level, the pieces in the composite method should be affordable on the current generation of computer architectures. However, computer architectures are continuing to advance, so we can either use a higher level of theory for a smaller molecule or apply the lower levels of theory with a larger molecule. In general, researchers prefer the latter approach because the larger molecular systems usually do not have very accurate experimental thermochemical data available. However, as will be described below, there is still much activity around using higher levels of theory to obtain even sub-chemical accuracy (~0.1 kcal/mol error from the correct value). Because several reviews are available (*4*), only a brief description of the most-used methods are described here.

Perhaps the most well-known composite method is the "Gaussian" models of Curtiss and coworkers (*5*). The fundamental versions of these are denoted as Gn, where n=1-4. Because several of the other composition methods use similar concepts, and because the method is often used, an example using G2 theory (*5c*) is described here. In G2, the first step is to compute a geometry at the MP2(FU)/6-31G(d) level (*6*, *7*), where FU means using all of the electrons in the system (i.e., not using a frozen core). This geometry is used for all of the subsequent steps in the calculation. The second step is to calculate an energy at the baseline, higher level calculation, which in this case is MP4/6-311G(d,p) – denoted E[MP4/6-311G(d,p)]. Next, a correction is made for diffuse functions, which are especially important for anions and molecules with extensive wavefunctions, ΔE(+), and is given by:

$$\Delta E(+) = E[MP4/6-311+G(d,p)] - E[MP4/6-311G(d,p)] \tag{1}$$

The fourth step is the addition of extra polarization functions on heavy atoms (all but hydrogen), Δ*E*(2*df*):

$$\Delta E(2df) = E[MP4/6-311G(2df,p)] - E[MP4/6-311G(d,p)] \tag{2}$$

The fifth step is a correction for an additional d function on heavy elements and a p function on hydrogen, Δ:

$$\begin{aligned}\Delta = E[MP2/6-311+G(3df,2p)] - E[MP2/6-311G(2df,p)] \\ - E[MP2/6-311+G(d,p)] + E[MP2/6-311G(d,p)]\end{aligned} \tag{3}$$

Up to this point, all of the corrections have improved the basis set limits. The next one, the sixth step, improves the electron correlation from MP4 to QCI (*8*), $\Delta E(QCI)$:

$$\Delta E(QCI) = E[QCI/6-311G(d,p)] - E[MP4/6-311G(d,p)] \qquad (4)$$

The seventh step is the addition of a higher level correction (HLC) that takes into account other basis set errors:

$$HLC = -0.00481 * n_{\beta} - 0.00019 * n_{\alpha} \qquad (5)$$

where n_β and n_α are the number of beta and alpha electrons in the valence on the molecule, respectively. The first coefficient (-0.00481) was optimized to give a zero mean deviation of the calculated atomization energies of 55 molecules from well-known experimental values. The second coefficient (-0.00019) is a correction in the energy associated with the hydrogen atom.

For the eighth step, the zero-point energy must be included to get the total energy, E_0. Because there is a specific scaling relationship with Hartree-Fock calculations (*9*), the frequencies from a HF/6-31G(d) calculation are scaled by 0.893, $E(ZPE)$.

Finally, the total energy is obtained by adding equations (1)−(5) and $E(ZPE)$ to the base MP4/6-311G(d,p) energy:

$$\begin{aligned} E_0 &= E[MP4/6-311G(d,p)] + \Delta E(+) + \Delta E(2df) + \Delta \\ &\quad + \Delta E(QCI) + HLC + E(ZPE) \end{aligned} \qquad (6)$$

Once the calculated molecular energy is available, it can be combined with additional information to calculate important thermodynamic information. Using a test set (denoted the G2 test set) of 125 molecular systems with accurate experimental data for dissociation energies, ionization energies, electron affinities, and proton affinities developed by the G2 authors, the G2 method was able to obtain a mean average deviation of 1.21 kcal/mol compared to experiment – very close to the goal of chemical accuracy. In fact, one of the important contributions from the Gn work is the development of a series of test sets where accurate experimental information is available and can be used for multiple thermodynamic calculational methodologies (*10*). Subsequent to G2, various improvements were made to the composite method. The most recent version, G4 (*5m*), uses the latest G3/05 test set (*11*) containing 454 experimental energies. It delivers an average absolute deviation of 0.83 kcal/mol.

In similar research by DeYonker, Cundari, Wilson and co-workers, the correlation-consistent composite approach (ccCA) (*12*) uses the G3B method (*13*) as a foundation for additional changes to improve the overall accuracy. The G3B method is similar to G2, except the B3LYP (*14*) density functional method is used to optimize geometries and determine the zero-point energies. Compared to the G3B method, one of the main differences is that the Pople-style basis sets, such as 6-311G(d,p), are replaced with the correlation-consistent basis set of Dunning and co-workers (*15*). Because these basis sets have many well-known extrapolations to the complete basis set limit (*16*), the authors used several of them to obtain accurate one-electron energies at the MP2 theory level. Based on their results, the extrapolations developed by Peterson and co-workers (*16b*) and by Wilson and Dunning (*16e*) gave the best results for the overall thermochemical data. With this extrapolation in place, and after research by others revealed that the triples excitations in the MP4 formalism can cause large electron correlation errors (*17*), the MP2 theory level using the extrapolated basis results was chosen as the starting single-point level instead of the MP4 in G3B. In addition, the ccCA includes a correction for core-valence correlation in the basis and does *not* include the HLC. As with the Gn methods, higher levels of electron correlation are included either through QCISD(T) or the coupled-cluster method with singles, doubles, and approximate triples, CCSD(T), with the latter generally being preferred. In later versions of the method, relativistic effects with the Douglas-Kroll-Hess Hamiltonian (*18*) and first-order spin-orbit coupling were added. For elements in the second row and below, a correction for tight core functions was also added. Using the G3/99 test set, the ccCA method obtains a 0.96 kcal/mol mean absolute deviation, essentially the accuracy of the G3X model, while avoiding the MP4 calculation as well as involving no empirical parameters.

While having the same goals of decreasing the overall errors of the complete energy, Petersson and co-workers take a different approach toward this extrapolation in their complete basis set CBS-4, CBS-Q, and CBS-QCI/APNO methods (*19*). These methods are similar to the G2 method in both approach and cost. The major difference is that the models use nonlinear pair natural orbital extrapolations to the complete basis set limit. However, nonlinear extrapolations will not produce calculations that are size consistent unless the corrections are applied to localized quantities. Size consistency means that one calculation including non-interacting molecules (i.e., molecules at infinity from one another) is the same as the sum of separate energy calculations on each molecule. This, of course, is an important property for thermochemical accuracy. In the CBS method, the Pipek and Mezey localization method is used to localize populations to correct for this issue (*20*). In addition, these methods also have an empirical correction that is specific for each model, but it is based on overlaps to obtain a size-consistent generalization of the correction used in Gn theories. The root mean square errors for the 125 chemical energy differences of the G2 test set are 2.5, 1.3, and 0.7 kcal/mol for CBS-4, CBS-Q, and CBS-QCI/APNO, respectively.

While the methods described above are readily applied to moderately sized molecular systems, some recent computationally intensive composite methods aim for less than 0.25 kcal/mol (or ~ 1 kJ/mol) accuracy and are currently only available for the smallest of molecules. These "calibration accuracy" methods generally fall into two different categories: 1) the Weizmann-n (Wn) (*21*) methods of Martin and co-workers, and 2) the "high accuracy extrapolated ab initio thermochemistry," HEAT (*22*), method developed by a multi-national effort. Both of these methods take advantage of the Active Thermochemical Tables (ATcT) (*23*), which is a novel source of highly accurate thermochemical data from the best experimental and theoretical data available. Additionally, it includes error bar analysis. In the ATcT approach, thermodynamic networks are used to solve for the thermodynamic quantities of interest (e.g., heats of formation). This approach also determines species that are not well known. In addition, the error analysis is propagated through the system with smaller weights in the network to less well characterized data (i.e., trying to minimize the error in each of the individual thermodynamic quantities). This provides a highly accurate database of information for which to compare theoretical results.

In the Wn methods, all elements of the Hamiltonian that can contribute at the kJ/mol level are included in the overall scheme, basis set convergence is determined at each level of theory used (i.e. SCF, CCSD(T)), the smallest basis set possible (still quite large) is used to obtain the target accuracy, and no parameters from fits to experimental data are used. Concerning the last point, several of the Wn methods use a parameter to improve the basis set extrapolations. In addition to the extensive basis set extrapolation and extrapolations for higher-order operators in the coupled cluster series, each level in the Wn hierarchy includes scalar relativistic effects through one-electron Darwin and mass-velocity terms. The W4 version includes an estimation of the $\hat{T}^4$ and $\hat{T}^5$ terms in the coupled cluster operator, inner core correlation, and atomic spin-orbit coupling, as well as the diagonal Born-Oppenheimer correction. For a series of 30 molecules, an average accuracy of 0.1 kcal/mol in atomization energies are obtained, compared to those of the ATcT.

The HEAT method has many similarities to the Wn method (some developed before the Wn methods and some developed after). The first significant difference is that the geometries and harmonic zero-point energies in the HEAT method are determined at a very high CCSD(T)/cc-pVQZ level of theory correlating all electrons in the molecule (including the core). The second major difference is that the corrections between the triple and quadruple contributions for the coupled cluster expansion are calculated exactly (i.e., not by approximation) using the double zeta correlation consistent basis set extrapolations. Other terms are similar to those in the Wn methods. The focus of this work is to obtain highly accurate total energies and not atomization energies. While including this data, the research group specifically chose to examine reactions with respect to the elements in their standard states. However, some modifications to the scheme are required because elements such as graphite are prohibitive at this computational level. In this case, the authors chose to use CO_2 as a "substitute" and used appropriate reactions. For the 31 sample molecules chosen from the ATcT, all of the reported heats of formation, with the exception of one, fall within 0.5 kJ/mol of the value given by ATcT. The exception, H_2O_2, fell within 1.0 kJ/mol of the expected value.

In other research, the "focal point" approach (*24*) of Wheeler and co-workers provides a general strategy for obtaining very accurate thermochemical information using correlation-consistent basis sets to obtain systematic dual one- and n-particle expansions and includes electron correlation through second-order perturbation and coupled-cluster methods. Combining this method with additional corrections for the anharmonic zero-point energy, the diagonal Born-Oppenheimer contribution, and scalar relativistic effects, these authors have obtained accurate thermochemical information for key small intermediates in soot formation.

In all of the methods described above, there is an assumption that a single reference is sufficient for describing the molecular system. In many molecular systems, this assumption is appropriate. However, it is clear that there are systems where multiple references are required (*25*), such as those that include electronic excited states or complex bond formation or dissociation. Sølling and co-workers have developed a multi-reference equivalent to the G2 and G3 methods to deal with these types of systems (*26*). In addition, very accurate potentials and *vibrational levels* (a true challenge!) have been obtained for several diatomic and triatomic systems by Bytautus and Ruedenberg through a novel configuration interaction extrapolation method (*27*).

Balanced Equations/Homodesmotic Reactions

As described above, the composite methods have the goal of obtaining very accurate overall energies that can then be used to compute thermodynamic quantities. The next type of methods rely heavily on cancellation of error in reaction calculations using several different types of reaction definitions. The advantage of these methods is that relatively low levels of theory can be used to obtain accurate thermochemical data, even on large chemical systems. Again, researchers have introduced several different methodologies. This chapter describes only the commonly used methods.

Pople et al. developed the first electronic-structure-based method, and it defines an isogyric reaction as one that leaves the number of unpaired electron spins unchanged (*28*). Dissociation energies for simple hydrides were developed using a hydrogen molecule (e.g., $BH_2 \rightarrow BH + H$ or $BH_3 + H \rightarrow BH_2 + H_2$). Combining these results with the exact dissociation energy of H_2 gives the desired result. Researchers used these types of reactions to determine singlet-triplet separations, ionization energies, and enthalpies of formation for a series of small molecules where experimental data was available. Using MP4 methods with small basis sets to calculate the total energies, the authors obtained ionization energies within 0.1 eV ≈ 2.3 kcal/mol for most cases.

Since those initial isogyric schemes, there have been many additional definitions of balanced equations, including homodesmotic (equal bonds) (*29*), isodesmic (*28a*), and hyperhomodesmotic (*30*), to name only a few. All of them depend on balancing different parts of the chemical equation – bond types, hybridization of the atoms, etc. All of these methods rely on having very accurate heats of formation available for a set of relatively small molecules or "standard" molecules for which the reactions can be built. Therefore, the composition methods are of critical importance in this effort as well. While several of these balanced equation methods are fairly well defined, xtensive confusion in the literature exists between the different methods.

Wheeler and colleagues clearly define a standard homodesmotic hierarchy for unstrained hydrocarbons (*3*) to clarify the balanced reaction types. In this work, the authors also describe the relationship of their work to other methods in the literature. Their scheme contains five progressive classes of homodesmotic reactions, in which the higher orders are systematically more rigorously balanced, yet also more expensive, than the classes below them. Indeed, what was a product in the previous class, becomes a reactant in the next highest one. (The reactions are formed in such a manner that the species of interest, usually larger than any of the standard species, is combined with small reactants and broken down into substituents that are larger than the reactants, but smaller than the original molecule.) The following list paraphrases their definitions:

RC1 (Isogyric): the number of total electron pairs is balanced.
RC2 (Isodesmic): (a) the number of total electron pairs and (b) the number of carbon-carbon single, double, and triple bonds are balanced.
RC3 (Hypohomodesmotic): (a) the number of carbon atoms in sp^3, sp^2, and sp hybridization and (b) the number of carbon atoms with zero, one, two, or three attached hydrogens are balanced.
RC4 (Homodesmotic): (a) the number of each combination of two, bonded, separately hybridization-specific carbon atoms and (b) the number of sp^3, sp^2, and sp hybridized carbon atoms with zero, one, two, or three attached hydrogens are balanced.
RC5 (Hyperhomodesmotic): (a) the number of each combination of two, bonded carbons each with zero, one, two, or three attached hydrogens and connected by a single, double, or triple bond and (b) the number of sp^3, sp^2, and sp carbon atoms with zero, one, two, or three attached hydrogens are balanced.

Using these definitions, they were also able to define all possible elemental reactants and products necessary to satisfy the requirements of each reaction class for any hydrocarbon. The visual interpretation of this system is especially helpful in understanding the inherent logic. The hypohomodesmotic definition can be thought of as breaking the molecule into smaller products based on each non-terminal carbon center. For example, a carbon attached to three carbons by two single bonds and a double bond is represented on the product side by 2-methylpropene. To counter-balance the extra terminal sp^3 and sp^2 carbons ($-CH_3$, $=CH_2$) now in the products, ethane and ethylene must be added to the reactants. The homodesmotic definition is identical except in cases when the study molecule has a conjugated pi system. RC4 preserves conjugating pairs of pi bonds in the products, which Wheeler and colleagues found to produce significant improvement over RC3 calculations. The hyperhomodesmotic definition is analogous to the hypohomodesmotic one, but it involves *two* non-terminal carbon atoms, the bond between them, and then the remaining, truncated bonds of each to any other carbons as in RC3. Thus, RC4 behaves like RC5, but only in situations with conjugation. As Wheeler and colleagues mentioned in their publication, the hierarchy could theoretically be expanded even further so that each product molecule contained the bonds of each three atoms and so forth, until the reactants and products were exactly identical. However, in this extreme case, the technique ceases to be useful.

Of particular interest in this work are the results of calculations of 38 hydrocarbons containing five or six carbon atoms. The RC4 and RC5 reactions give bond separation enthalpies with errors consistently less than 0.4 kcal/mol using several levels of theory including HF, DFT, MP2, and CCSD(T). Even RC2 and RC3 results are consistently below 8 kcal/mol and 2 kcal/mol, respectively. These results are promising in that highly expensive calculations can be avoided except in computing the constituent standard molecules, which are small in comparison to the systems of interest. Enthalpies of formation for large polyynes were also computed ($C_{10}H_2$-$C_{26}H_2$) using DFT, showing the use of these methods to obtain accurate values for large molecular systems.

This chapter adapts these standards to include oxygen-based functional groups, namely alcohols, ethers, ketones, and the combinations of them. Because oxygen is the most common heteroatom in organic molecules and sugars, the usefulness of homodesmotic reactions is greatly enhanced by the capability to include oxygen. The current research was tested with β-D-glucopyranose-gg (Figure 1).

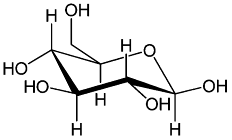


Figure 1. β-D-glucopyranose-gg

Extension for Sugars

Just as the hydrocarbon homodesmotic hierarchy is self-consistent in that the higher classes' parameters are expansions, not revisions, of the lower ones, so too must be a hierarchy that includes oxygen. Because carbon and oxygen are large atoms compared to hydrogen, it is reasonable to give oxygen equal priority to carbon, not merely as a substituent of it. We show that any sp^3 or sp^2 carbon atom from Wheeler and coworker's elemental molecules may be substituted by a sp^3 or sp^2 oxygen atom, respectively.

Oxygen-oxygen bonding was not considered at this time because oxygen significantly differs from carbon in its ability to form long chains of bonds. However, oxygen can still be part of a primarily carbon chain in the form of ethers. The lack of side branching in ethers structurally differentiates them from interstitial carbons. Likewise, an alcohol, (primary, secondary, or tertiary) is analogous to a terminal carbon, whether it is bonded to a primary, secondary, or tertiary carbon. They are both heavy atoms connected to the appropriate number of hydrogens and just one carbon.

Concerning the sp^2 hybridized functionalities, a ketone is structurally like an alkene except with an oxygen as one of the pi-bonding pairs, rather than a carbon. A notable difference in this situation is that the ketone is necessarily terminal, whereas the alkene may be in the middle of a carbon chain. The ketone may, however, still be involved in conjugation. Oxygens are rarely, if ever, species? hybridized at equilibrium conditions, so there is no suitable comparison with alkynes.

Substitution of a carbon atom with a similarly hybridized oxygen atom does not significantly change the molecule's basic form, except when it would cause fragmentation as noted above. Therefore, the patterns that Wheeler and colleagues have already established can be modified by this principle to include oxygen. The adapted definitions are:

RC1: the number of total electron pairs is balanced.

RC2: (a) the number of total electron pairs and (b_1) the number of carbon-carbon single, double, and triple bonds and (b_2) the number of oxygen-carbon single and double bonds are balanced.

RC3: (a_1) the number of carbon atoms in sp^3, sp^2, and sp hybridization and (a_2) the number of oxygen atoms in sp^3 and sp^2 hybridization and (b_1) the number of carbon atoms with zero, one, two, or three attached hydrogens and (b_2) the number of oxygen atoms with zero or one attached hydrogen are balanced.

RC4: (a) the number of each combination of two, bonded, separately hybridization-specific carbon atom and carbon-or-oxygen atom pairs and (b_1) the number of sp^3, sp^2, and sp hybridized carbon atoms with zero, one, two, or three attached hydrogens and (b_2) the number of sp^3 and sp^2 hybridized oxygen atoms with zero or one attached hydrogen are balanced.

RC5: (a) the number of each combination of two, bonded carbon atom and carbon-or-oxygen atom pair each with zero, one, two, or three attached hydrogens and connected by a single, double, or triple bond and (b_1) the number of sp^3, sp^2, and sp hybridized carbon atoms with zero, one, two, or three attached hydrogens and (b_2) the number of sp^3 and sp^2 hybridized oxygen atoms with zero or one attached hydrogen are balanced.

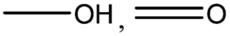
Figure 2 includes a chart of all possible elemental reactants and products of this redefined homodesmotic hierarchy that are *in addition* to the purely hydrocarbon sets proposed by Wheeler and colleagues.

Obviously, there are many more elemental pieces in the oxygen-inclusive homodesmotic hierarchy. There are about 150 in all, more than 100 of which are RC5-only. Carboxylic acids first appear as RC3 products, but esters only manifest in a few RC5 products. Many of these theoretical combinations seem highly unstable, which helps explain why experimental data is often scarce. Some molecules that look absurd, however, may seem more reasonable when reassembled into the study molecule. Gem polyols, for example, could represent one or more ether linkages after being broken down according to the requirements of the reaction class. Even so, the sheer number of possibilities, already ignoring stereochemistry, makes the task of finding suitably accurate enthalpies of formation for all the elements quite daunting. It's useful to note here that this scheme does not include radical species that are certainly of interest in the decomposition of sugars (especially in pyrolysis). This deficiency will be addressed in future work.

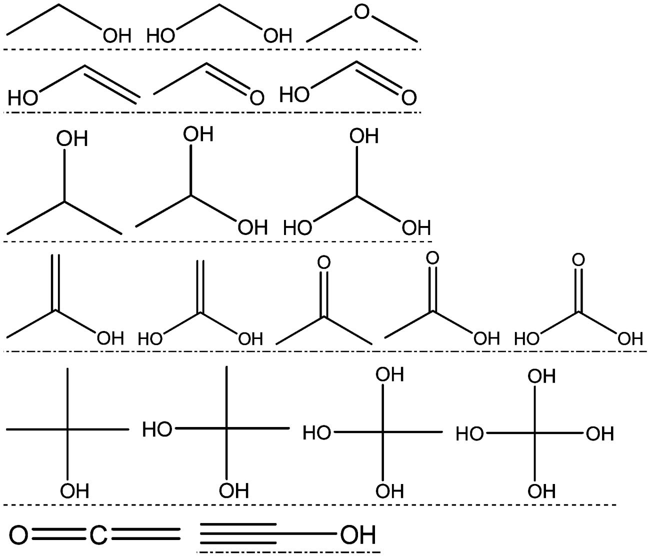
RC1 Products/RC2 Reactants

H_2O

RC2 Products/RC3 & RC4 Reactants

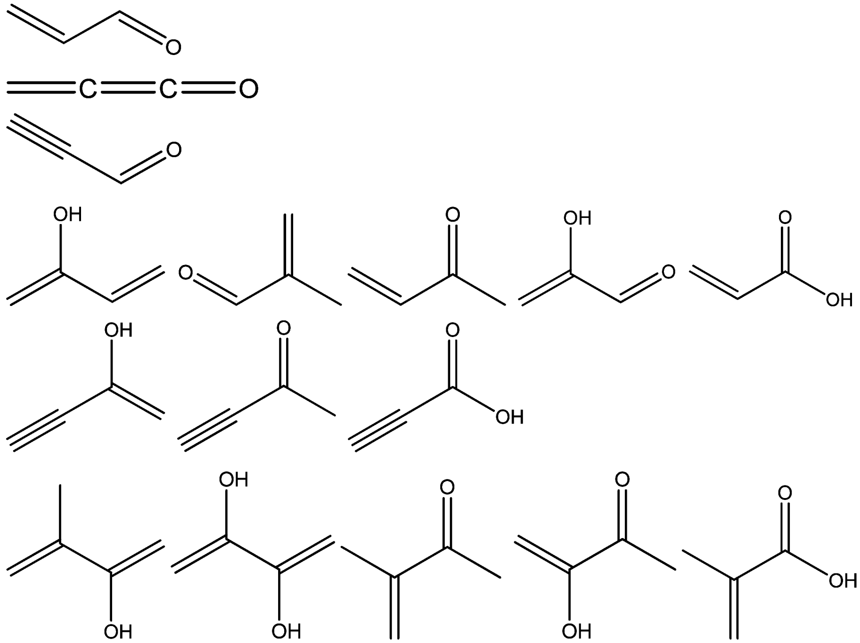


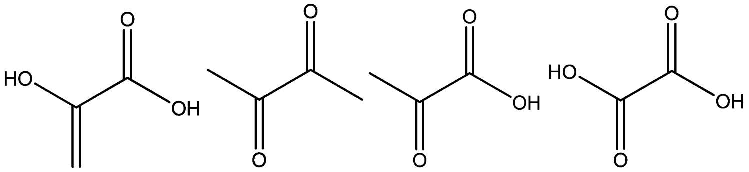
RC3 & RC4 Products/RC4 & RC5 Reactants*



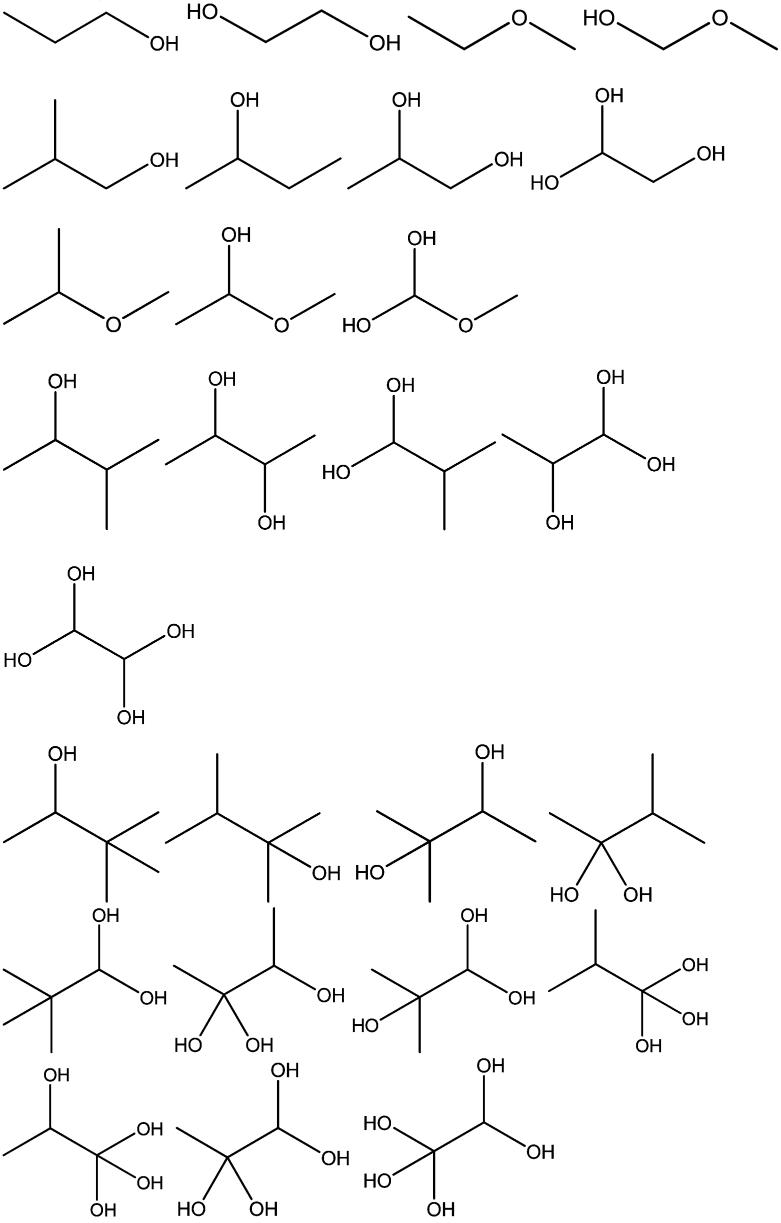
*dotted underline indicates that this molecule is a product in RC4; mixed underline indicates that this molecule is both a product and reactant in RC4

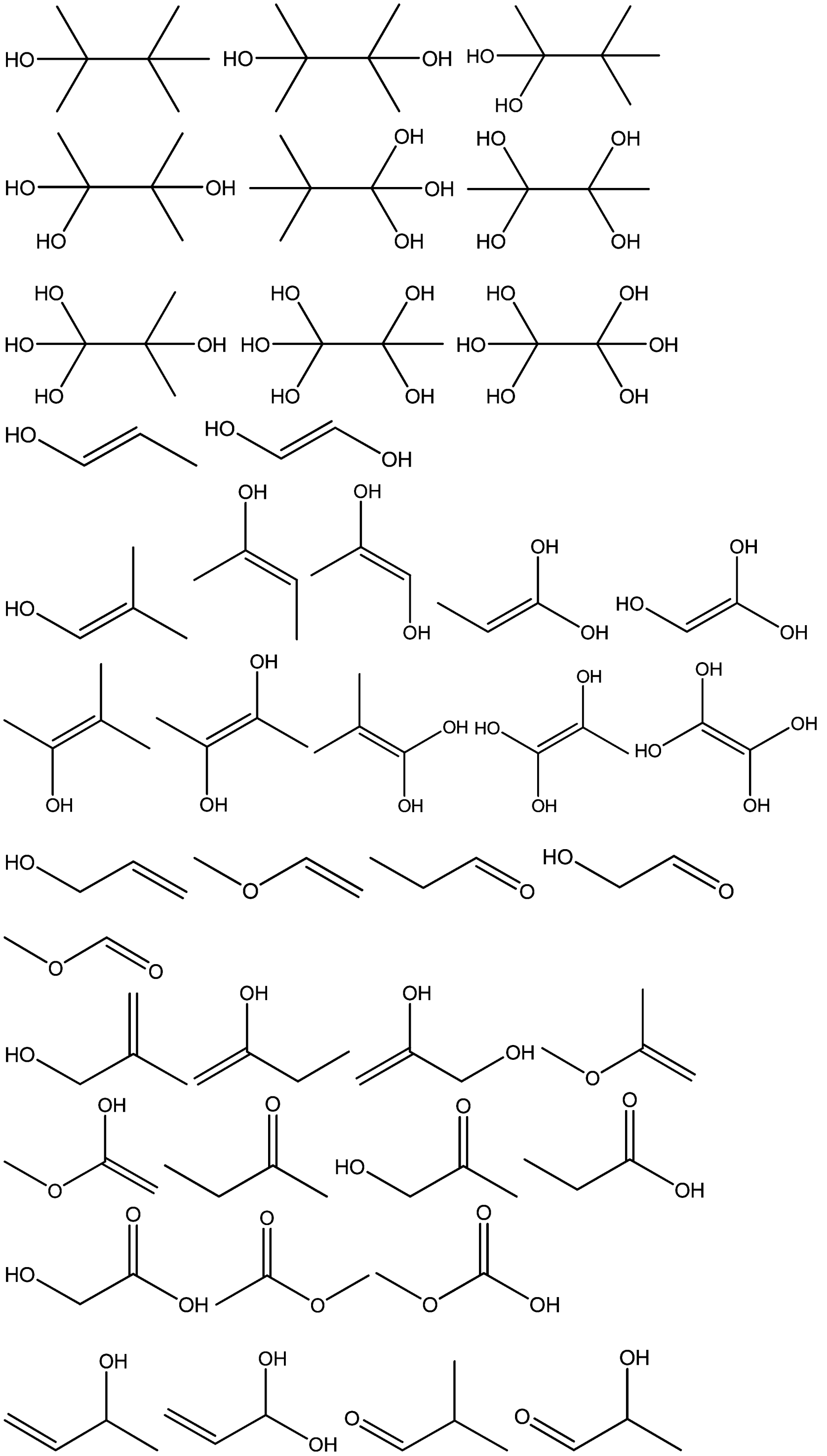
RC4 & RC5 Products

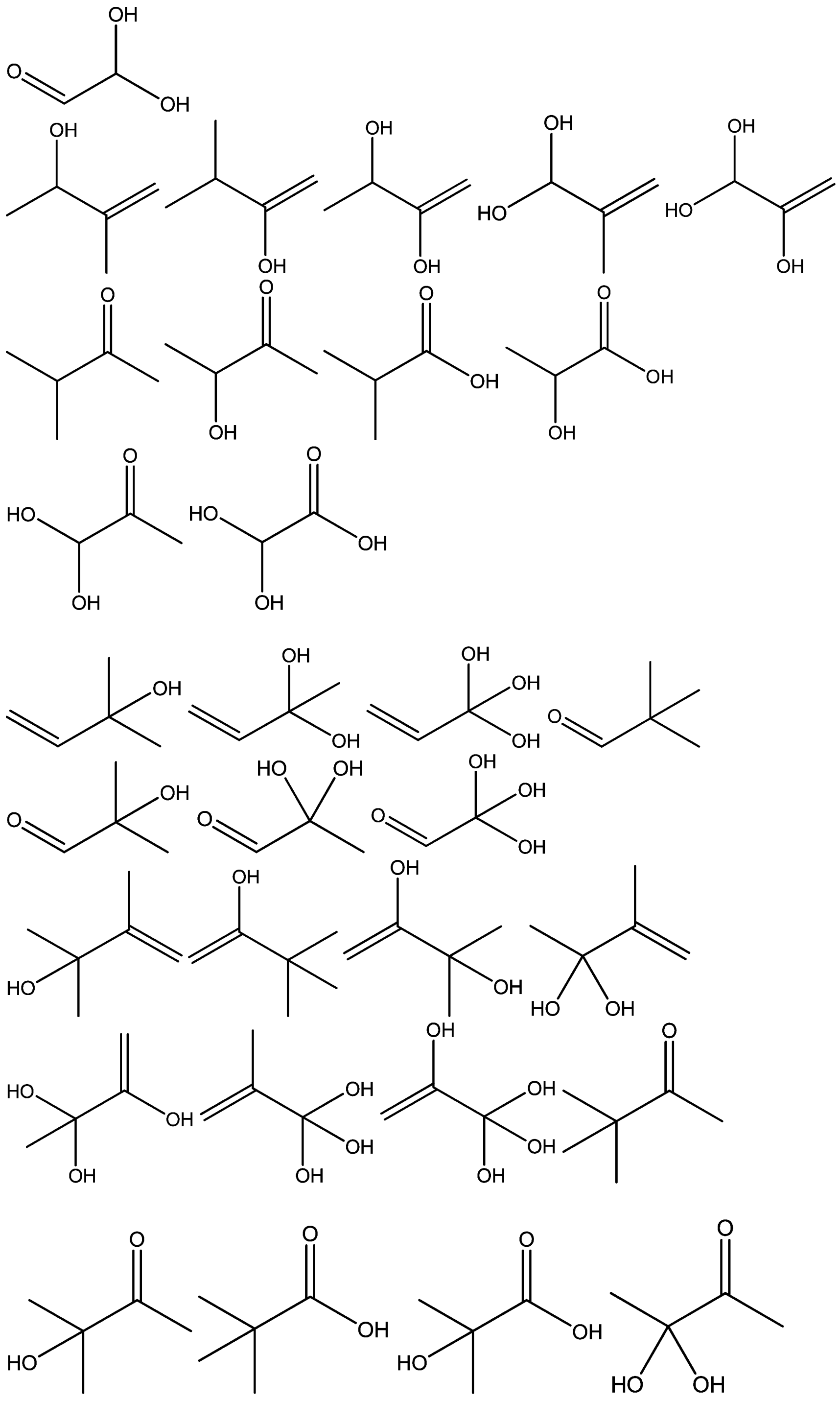




RC5-only Products







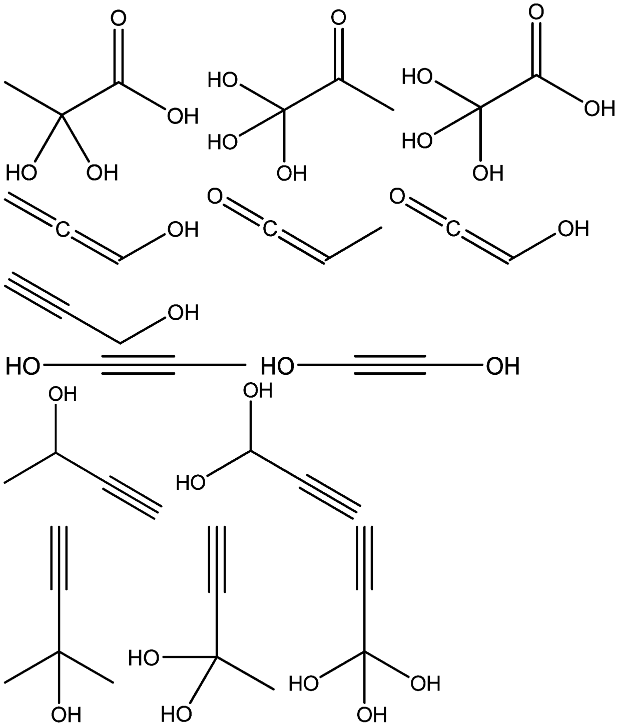
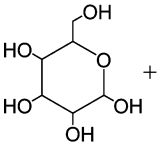


Figure 2. Schematics of the additional reactants and products needed at each level of the hierarchy for oxygen containing species.

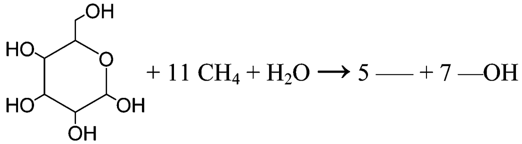
Initial Results

Wheeler and co-workers showed that even density functional methods with a modest basis set level can be used to give excellent results, especially for levels RC3 and above. β-D-glucopyranose-gg, and all of the elemental reactants and products needed to satisfy the five classes of the hierarchy, were prepared using the graphical user interface Ecce (*31*) and energetically optimized by NWChem (*32*). We used restricted density functional theory (B3LYP) and TZVP DFT orbitals without coulomb or exchange fitting. The resulting balanced reactions, (7) through (10), are given below:

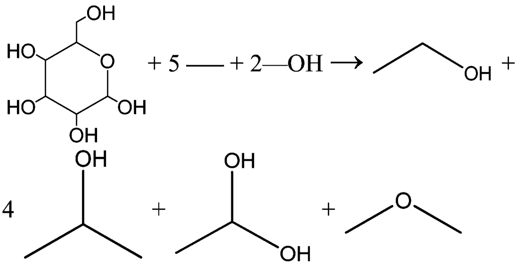


$$+ 12\ H_2 \longrightarrow 6\ CH_4 + 6\ H_2O$$

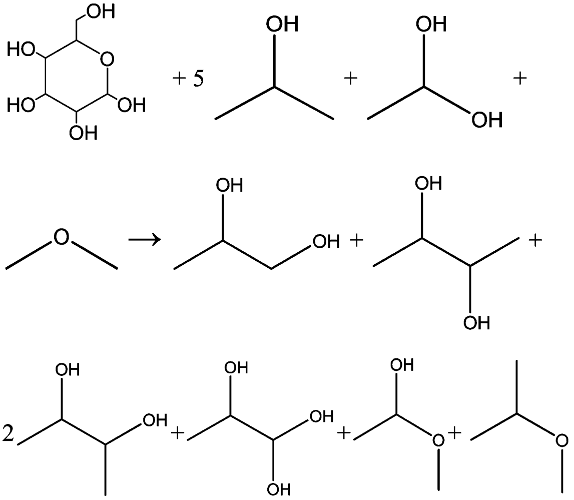
RC1 (7)



RC2 (8)



RC3 & RC4 (9)



RC5 (10)

The zero-point corrected total energy values are used to calculate the enthalpy of each reaction according to the general equation:

$$\Sigma(\text{Energy}_{\text{products}}) - \Sigma(\text{Energy}_{\text{reactants}}) = \Delta_f H^\circ_{\text{rxn}} \qquad (11)$$

Standard enthalpy of formation values (*33*) were then used to calculate the enthalpy of formation of the study molecule, β-glucopyranose-gg, for reactions (7) and (8) according to the following equation:

$$\Delta_f H^\circ_{\text{study}} = \Sigma(\Delta_f H^\circ_{\text{prod}}) - \Sigma(\Delta_f H^\circ_{\text{other react}}) - \Delta_f H^\circ_{\text{rxn}} \qquad (12)$$

At this time, accurate values for the standard enthalpies of formation of all the elemental molecules for reactions (9) and (10) are unavailable. We are in the process of determining accurate enthalpies of formation for these species using the composite methods described earlier.

Table 1. Thermodynamic data from computation and literature

	Total Energy + ZPVE (E_h)	*$\Delta_f H°_{(g)}$ (kJ/mol) (33)*	*$\Delta_f H°_{(g)}$ (E_h)*
β-D-glucopyranose gauche gauche	-687.227422055		
Hydrogen	-1.1691918625	0	0
Methane	-40.4913470200	-74.6	-0.02841
Water	-76.4379808374	-241.8	-0.092097
Ethane	-79.7856244897	-84.0	-0.03199
Methanol	-115.717518491	-201.0	-0.076557

Table 2. Derived thermodynamic data from calculation

	ΔE_{rxn} (E_h)	*$\Delta_f H°_{(g)}$ (E_h)*	*$\Delta_f H°_{(g)}$ (kcal/mol)*
RC1	-0.318242739	-0.4048	-254.0
RC2	-0.119468227	-0.4107	-257.7

The necessary information for reactions (7) and (8) are given in Table 1. Using this data and equations (11) and (12), the reaction energy and the heats of formation can be calculated as in Table 2.

It is also worth noting that the RC1 and RC2 values from Table 2 for the standard enthalpy of formation of β-D-glucopyranose-gg are reasonably in agreement with each other. The reaction energy approaches zero going up the hierarchy, which makes sense because the difference between reactants and products is also decreasing.

Conclusions

In this work, we presented an extension to the hydrocarbon hierarchy of homodesmotic reactions developed by Wheeler and co-workers to include oxygen. More than 150 new elements need to have accurate heats of formation available before each hierarchical level can be fully exploited. Obviously, much work remains to be done. Neither our research nor Wheelers took stereochemical effects, such as rotamers of butane-2,3-dione (one of the products of the RC5 reaction for β-D-glucopyranose-gg), into consideration. This is problematic because preliminary DFT calculations indicate the difference can be quite significant – in excess of 1 kcal/mol. To verify the soundness of the adapted homodesmotic scheme presented here, more calculations at higher levels of theory need to be performed, as well as comparisons with experimental data. In addition, incorporating this data into an ATcT scheme would improve the overall accuracy of the developed information. If the current scheme is successful, room exists for the hierarchy to expand, to include radicals and more heteroatoms, such as nitrogen.

It is too early to determine if the theories presented here are successful. As noted above, there are many barriers to implementing this technique in a practical manner; but the possibilities are great, and the early results are promising. Even with the crude RC1 and RC2 methods, general agreement between the two gave the standard heat of formation value of β-D-glucopyranose-gg to be approximately 250 to 260 kcal/mol. This is not precise enough to use in experiment, but it is hopeful that RC3, RC4, and RC5 will provide even better results. It will be especially interesting to see if RC5 is dramatically better than RC4, because it has the capability to preserve more complex oxygen functional groups, such as esters.

In particular, however, a relatively low level of theory is required to examine large molecular systems such as the sugars. Using the obtained heats of formation aquired through new extensions to the homodesmotic hierarchy, accurate bond dissociation energies and other critical thermochemical data may help us understand the processes involved in the decomposition of sugars and biomass.

Acknowledgments

Work at the Ames Laboratory was supported by the Department of Energy Office of Biological and Environmental Research (BER) and Office of Advanced Scientific Computing Research (ASCR) divisions as part of the Scientific Discovery through Advanced Computing (SciDAC) under Contract No. DE-AC02-07CH11358. In addition, the authors thank Ames Laboratory and Iowa State University for the computing resources used in this work.

References

1. For example (a) National Biofuels Action Plan. Available at http://www1.eere.energy.gov/biomass/pdfs/nbap.pdf. (b) U.S. Economic Impact of Advanced Biofuels Production: Perspectives to 2030. Available at http://bio.org/ind/EconomicImpactAdvancedBiofuels.pdf
2. (a) Huynh, L. K.; Violi, A. *J. Org. Chem.* **2008**, *73*, 94−101. (b) El-Nahas, A. M.; Navaroo, M. V.; Simmie, J. M.; Bozzelli, J. W.; Curran, H. J.; Dooley, S.; Metcalfe, W. *J. Phys. Chem. A* **2007**, *111*, 3727−3739. (c) Tewari, Y. B.; Lang, B. E.; Decker, S. R.; Goldberg, R. N. *J. Chem. Thermodyn.*, **2008**, *40*, 1517−1526.
3. Wheeler, S. E.; Houk, K. N.; Schleyer, P. v. R.; Allen, W. D. *J. Am. Chem. Soc.* **2009**, *131*, 2547–2560.
4. (a) Martin, J. M. L. *Annual Reports in Computational Chemistry*; Elsevier: New York, NY, 2005; Vol. 1, pp 31–43. (b) *Quantum Mechanical Prediction of Thermochemical Data*; Cioslowski, J.; Kluwer: Dordrecht, 2001.
5. See for instance (a) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. *J. Chem. Phys.* **1989**, *90*, 5622−5629. (b) Curtiss, L. A.; Jones, J.; Trucks, G. W.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1990**, *93*, 2537−2545. (c) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople,

J. A. *J. Chem. Phys.* **1991**, *94*, 7221−7230. (d) Curtiss, L. A.; Carpenter, J. E.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1992**, *96*, 9030−9034. (e) Curtiss, L. A.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1993**, *98*, 1293−1298. (f) Curtiss, L. A.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1995**, *103*, 4192−4200. (g) Curtiss, L. A.; McGrath, M. P.; Blaudeau, J.-P.; Davis, N. E.; Binning, R. *J. Chem. Phys.* **1995**, *103*, 6104−6113. (h) Curtiss, L. A.; Redfern, P. C.; Smith, B. J.; Radom, L. *J. Chem. Phys.* **1996**, *104*, 5148−5152. (i) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063−1079. (j) Blaudeau, J.-P.; McGrath, M. P.; Curtiss, L. A.; Radom, L. *J. Chem. Phys.* **1997**, *107*, 5016−5021. (k) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764−7776. (l) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **2001**, *114*, 108−117. (m) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *126*, 084108.

6. Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.
7. (a) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650−654. (b) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639−5648. (c) Blaudeau, J.-P.; McGrath, M. P.; Curtiss, L. A.; Radom, L. *J. Chem. Phys.* **1997**, *107*, 5016−5021. (d) Curtiss, L. A.; McGrath, M. P.; Blandeau, J.-P.; Davis, N. E.; Binning, Jr. R. C. ; Radom, L. *J. Chem. Phys.* **1995**, *103*, 6104−6113. (e) Glukhovstev, M. N.; Pross, A.; McGrath, M. P.; Radom, L. *J. Chem. Phys.* **1995**, *103*, 1878−1885.
8. Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968–5978.
9. Pople, J. A.; Schlegel, H. B.; Krishnan, R.; Defrees, D. J.; Binkley, J. S.; Frisch, M. J.; Whiteside, R. A.; Hour, R. F.; Hehre, W. J. *Int. J. Quantum Chem. Symp.* **1981**, *15*, 269–274.
10. Available on-line at http://www.cse.anl.gov/Catalysis_and_Energy_Conversion/Computational_Thermochemistry.shtml.
11. Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2005**, *123*, 124107.
12. (a) DeYonker, N. J.; Grimes, T. V.; Yockel, S. M.; Dinescu, A.; Mintz, B.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2006**, *125*, 104111/1. (b) DeYonker, N. J.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2006**, *124*, 114104/1. (c) Ho, D. S.; DeYonker, N. J.; Wilson, A. K.; Cundari, T. R. *J. Phys. Chem. A* **2006**, *110*, 9767−9770. (d) Grimes, T. V.; Wilson, A. K.; DeYonker, N. J.; Cundari, T. R. *J. Chem. Phys.* **2007**, *127*, 154117/1. (e) DeYonker, N. J.; Ho, D. S.; Wilson, A. K.; Cundari, T. R. *J. Phys. Chem. A* **2007**, *111*, 10776−10780. (f) DeYonker, N. J.; Mintz, B.; Cundari, T. R.; Wilson, A. K. *J. Chem. Theor. Comput.* **2008**, *4*, 328−334.
13. Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **1999**, *110*, 7650–7657.
14. Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
15. (a) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007−1023. (b) Dunning, T. H., Jr.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244−9253. (c) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; Dunning, T. H., Jr. *J.*

Chem. Phys. **1999**, *110*, 7667−7676. (d) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796−6806. (e) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 2975−2988. (f) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1995**, *103*, 4572−4585. (g) Peterson, K. A.; Dunning, T. H., Jr. *J. Chem. Phys.* **2002**, *117*, 10548−10560.

16. (a) Xantheas, S. S.; Dunning, T. H., Jr. *J. Phys. Chem.* **1993**, *97*, 18−19. (b) Peterson, K. A.; Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 7410−7415. (c) Martin, J. M. L. *Chem. Phys. Lett.* **1996**, *259*, 669−674. (d) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243−252. (e) Wilson A. K.; Dunning, T. H., Jr. *J. Chem. Phys.* **1997**, *106*, 8718−8726.
17. (a) Dunning, T. H., Jr.; Peterson, K. A. *J. Chem. Phys.* **2000**, *113*, 7799−7808. (b) Leininger, M. L.; Allen, W. D.; Schaefer, H. F.; Sherrill, C. D. *J. Chem. Phys.* **2000**, *112*, 9213−9222. (c) Handy, N. C.; Knowles, P. J.; Somasundram, K. *Theor. Chim. Acta* **1985**, *68*, 87−100.
18. (a) Douglas, M.; Kroll, N. M. *Ann. Phys.* **1974**, *82*, 89−155. (b) Hess, B. A. *Phys. Rev.* **1986**, *A33*, 3742−3748.
19. (a) Montgomery, J. A., Jr.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1994**, *101*, 5900−5909. (b) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A., Jr. *J. Chem. Phys.* **1996**, *104*, 2598−2619. (c) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822−2827. (d) Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. W. *J. Chem. Phys.* **2000**, *112*, 6532−6542.
20. Pipek, J.; Mezey, P. G. *J. Chem. Phys.* **1989**, *90*, 4916–4926.
21. (a) Martin, J. M. L.; de Oliveira, G. *J. Chem. Phys.* **1999**, *111*, 1843−1856. (b) Parthiban, S.; Martin, J. M. L. *J. Chem. Phys.* **2001**, *114*, 6014−6029. (c) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kállay M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 4129−4141. (d) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. *J. Chem. Phys.* **2006**, 125, 144108.
22. Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. *J. Chem. Phys.* **2004**, *121*, 11599.
23. (a) Ruscic, B.; Pinzon, R. E.; Morton, M. L.; von Laszewski, G.; Bittner, S.; Nijsure, S. G.; Amin, K. A.; Minkoff, M.; Wagner, A. F. *J. Phys. Chem. A* **2004**, *108*, 9979−9997. (b) Ruscic, B.; Pinzon, R. E.; von Laszewski, G.; Kodeboyina, D.; Burcat, A.; Leahy, D.; Montoya, D.; Wagner, A. F. *J. Phys.: Conf. Ser.* **2005**, *16*, 561−570. (c) Ruscic, B.; Pinzon, R. E.; Morton, M. L.; Srinivasan, N. K.; Su, M.-C.; Sutherland, J. W.; Michael, J. V. *J. Phys. Chem. A* **2006**, *110*, 6592−6601.
24. (a) Wheeler, S. E.; Robertson, K. A.; Allen, W. D.; Schaefer, H. F.; Bomble, Y. J.; Stanton, J. F. *J. Phys. Chem. A* **2007**, *111*, 3819−3830. (b) Wheeler, S. E.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2004**, *121*, 8800−8813. (c) Császár, A. G.; Allen, W. D.; Schaefer, H. F., III. *J. Chem. Phys.* **1998**, *108*, 9751−9764.
25. Schmidt, M. W.; Gordon, M. S. *Annu. Rev. Phys. Chem.* **1998**, *49*, 233–266.
26. Solling, T. I.; Smith, D. M.; Radom, L.; Freitag, M. A.; Gordon, M. S. *J. Chem. Phys.* **2001**, *115*, 8758–8772.

27. (a) Bytautas, L.; Ruedenberg, K. *J. Chem. Phys.* **2004**, *121*, 10905. (b) Bytautas, L.; Ruedenberg, K. *J. Chem. Phys.* **2004**, *121*, 10919.

28. (a) Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. *J. Am. Chem. Soc.* **1970**, *92*, 4796−4801. (b) Radom, L.; Hehre, W. J.; Pople, J. *J. Am. Chem. Soc.* **1971**, *93*, 289−300. (c) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley-Interscience: New York, NY, 1986.

29. (a) George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M. *Theor. Chem. Acc.* **1975**, *38*, 121−129. (b) George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M. *Tetrahedron* **1976**, *32*, 317−323. (c) George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M. *J. Chem. Soc., Perkin Trans.* **1976**, *2*, 1222−1227.

30. Hess, B. A., Jr.; Schaad, L. J. *J. Am. Chem. Soc.* **1983**, *105*, 7500–7505.

31. Black, G.; Daily, J.; Elsethagen, T.; Feller, D.; Gracio, D.; Jones, D.; Keller, T.; Matsumoto, S.; Palmer, B.; Peterson, M.; Schuchardt, K.; Stephan, E.; Sun, L.; Swanson, K.; Taylor, H.; Vorpagel, E.; Windus, T.; Winters, C. *ECCE, A Problem Solving Environment for Computational Chemistry*, Software Version 5.1;Pacific Northwest National Laboratory: Richland, WA, U.S.A., 2009.

32. (a) Bylaska, E. J.; de Jong, W. A.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Apra, E.; Windus, T. L.; Hammond, J.; Nichols, P.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Wu, Q.; Van Voorhis, T.; Auer, A. A.; Nooijen, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Fruchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers*, Version 5.1; Pacific Northwest National Laboratory: Richland, WA, U.S.A., 2007. (b) Kendall, R. A.; Apra, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. *Computer Phys. Commun.* **2000**, *128*, 260−283.

33. *CRC Handbook of Chemistry and Physics*; Lide, D. R., Ed.; 89th ed. (Internet Version 2009); CRC Press/Taylor and Francis: Boca Raton, FL, 2009.

Chapter 10

Development of Detailed Kinetic Models for the Thermal Conversion of Biomass via First Principle Methods and Rate Estimation Rules

Hans-Heinrich Carstensen* **and Anthony M. Dean**

Chemical Engineering Department, Colorado School of Mines, Golden, Colorado, USA
***hcarsten@mines.edu**

Electronic structure methods have matured to a point that they can be routinely used to calculate rate expressions for kinetic mechanisms. However, given the size of modern reaction sets, it is not feasible to perform high-level calculations for every reaction found in a kinetic model. Furthermore, chemically accurate calculations can only be done for moderately small species. Therefore we propose to derive kinetic expressions from first principle calculations on a series of small reactants for a given reaction class and use the data to create rate estimation rules. Those are then used for all members of the reaction class. In this contribution we discuss four selected example systems related to the thermal conversion of biomass to illustrate this approach or to show its limitations. The reaction classes include H abstraction and water elimination reactions from alcohols, retro-Diels-Alder reactions and the initial unimolecular decomposition step in phenyl ethers.

Introduction

Detailed chemical kinetic modeling has proven to be a valuable tool to optimize the operation conditions of complex chemical systems. Examples include the successful application of atmospheric and tropospheric chemistry models to understand the ozone formation and destruction cycles (*1*, *2*); ignition and combustion models to predict and adjust heat release profiles in internal combustion engines such as the homogeneous charge compression ignition engine

In Computational Modeling in Lignocellulosic Biofuel Production; Nimlos, M., et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2010.

(*3*) and to prevent soot formation (*4*); microkinetic NO_x models to analyze and improve three-way catalysts (*5*); and many others. Given these successes, it is clear that the availability of a reaction model, composed of elementary reaction steps, for the biomass gasification process would be of tremendous value to engineers who try to make this technique economically feasible. However, despite many decades of research in biomass pyrolysis, such a model has yet to emerge. Part of the reason is the complexity and varieties of components found in biomass material (*6*). Even if only the three major components cellulose, hemicellulose and lignin are considered, biomass is still not well defined because the relative compositions of these three different components vary substantially in different biomass sources. Furthermore the components themselves are not well-defined molecules but macromolecules with varying degrees of (co-)polymerization and crystallization. A second major obstacle in developing a detailed chemical model for the thermochemical conversion of biomass is that reactions proceed simultaneously in the condensed phase (solid, melted or solution) and in the gas phase, and the degree to which reactions in each phase contribute depends on difficult to control parameters such as mineral content, acidity, prior treatment of the biomass, heat transfer within chunks of biomass and so on. Given these difficulties it seems unlikely that a comprehensive kinetic model completely based on elementary reactions can be developed in the foreseeable future.

One solution to this dilemma is to focus on gas phase reactions and to substitute the major biomass components with a set of model compounds. Generation of an elementary reaction mechanism for such a system is still a daunting task since it will likely contain hundreds of species and thousands of reactions. However, there are good reasons to believe that such a mechanism can be developed in the near future. First, even though oxygenated species formed via biomass volatilization react certainly differently than hydrocarbon species, we expect that the same reaction types will be important. This means we can learn from the decades long experience of kineticists in creating hydrocarbon pyrolysis and oxidation (combustion) mechanisms. Related to this, there is significant evidence that suggests that the biomass pyrolysis leads to secondary and tertiary products that are less oxygenated (*6–12*) and hence pure hydrocarbon chemistry will form an important subset of any biomass model. Second, rate constants for many reaction classes have been shown to be very consistent with respect to a homologous series of reactants. This allows one to develop rate estimation rules for such reaction classes. These rate rules simplify the assignment of rate constants and lead to internal consistency of the rate expressions in a mechanism. Third, the increase of CPU power, data storage capacity and the improvement of algorithms have made theoretical calculations of thermodynamic data and rate expressions even for moderately large molecules feasible. This not only enables one to calculate rate constants for a desired reaction but also to provide kinetic data sets to develop the aforementioned rate estimation rules. Finally, sub-mechanisms for a number of model compounds are already known in the literature. For example, Sendt et al. (*13*) developed a furan pyrolysis mechanism, Horn et al. provide a reaction scheme for phenol (*11*), Pecullan et al. (*14*) describe a kinetic set for anisole pyrolysis, Britt and coworkers (*15*) studied the reaction pathways of phenethyl phenyl ether, and so on. Such sub-mechanisms can be incorporated either directly or with moderate adaptations into a biomass mechanism and thus significantly reduce the mechanism development time.

In this chapter we will discuss our approach to develop a biomass gas phase mechanism. A significant emphasis will be on the development and use of rate estimation rules. The simple rules are expressed in modified Arrhenius form

$$k(T) = n_{sites} \cdot A \cdot T^{n} \cdot \exp(-E/RT)$$

For each reaction class the rate rule defines a common pre-exponential factor ($A \bullet T^{n}$), normalized on a per site basis. The E value, to which we will refer in the following as the activation energy or barrier even though it strictly speaking differs from the activation energy by nRT, is either a constant value, or it is obtained from the heat of reaction, $\Delta_R H$, via an Evans-Polanyi relationship,

$$E = E^{ref} + m \cdot (\Delta_R H - \Delta_R H^{ref}).$$

Application of rate rules in the mechanism development process requires knowledge of their transferability. In other words, one needs to know how closely related a reaction has to be to the test set (reaction class) to be sure that the rule can still be applied with acceptable accuracy. For example: Can a rate rule that is valid for simple alcohols also be used for polyols or alcohols with other functional groups? This question will be addressed in two examples: (1) H abstraction from oxygenates by H and CH_3 radicals and (2) the elimination of molecular water from alcohols. Both examples provide insight in how far the influence of a functional group extends to neighboring molecular sites. With respect to the abstraction reactions, we will investigate if and how the reactivity of C-H bonds in the α–, and β–positions are altered due to the OH group. The impact of neighboring functional groups on the reactivity will also be addressed in the second reaction class (H_2O elimination from alcohols). In addition it will serve as a basis to discuss the reliability of different calculation methods to yield accurate kinetic parameters. Since biomass material is mainly made up of large polymers, it is desirable to study model compounds as large as feasible to explore non-next-neighbor effects. However, large molecules (e.g., cellubiose and beyond) cannot be treated by high-level electronic structure methods, and the question arises as to whether lower-level methods provide acceptable results. A third aspect addressed using the water elimination reaction as an example deals with the role of water molecules in the gas phase. The gas phase in a thermal biomass gasifier contains large amounts of water. Therefore there might be a possibility that water can act as a catalyst in the elimination process.

Not all reactions in a given mechanism can be described by rate estimation methods. In order to provide examples for reaction systems of expected importance in biomass gas phase chemistry that do not seem to be following such rules we will discuss some results for retro-Diels-Alder reactions involving anhydro-glucose and levoglucosenone. Finally, we present calculations related to the initial decomposition reactions of phenyl ethers with the focus on the possibility to simplify the calculations by either replacing the large phenyl group by a smaller kinetically similar vinyl group or by performing the calculations with a faster electronic structure method.

The remainder of this chapter is organized as follows. First we will outline the calculation methods used in this work with an emphasis of special problems associated with H-bonds found in many oxygenates (biomass model compounds). Next, we discuss results for the four reaction systems mentioned previously, followed by a discussion that will attempt to provide a more comprehensive picture of the role and limitations of rate rules. In conclusion, we summarize the major results of the calculations (a) with respect to generalized rate expressions and (b) in terms of the importance for biomass gas-phase chemistry.

Calculation Methodology

The methodology used in this work consists of three main steps. First, electronic structure calculations are performed to determine an optimized geometry and the lowest electronic energy. Molecular parameters obtained in the first step are then used in a statistical thermodynamics analysis that provides thermodynamic properties. Finally, transition state calculations using these thermodynamic data as input yield the rate expressions. In the following we provide more details of this approach.

Electronic Structure Calculations

The choice of the calculation method is generally determined by two opposing factors: the desire for highly accurate results and the need to obtain these in a reasonable time frame on the available computing platform. Among the many methods that emerged in the past two decades to fulfill both conditions the CBS-QB3 method by Peterson and coworkers (*16*, *17*) has gained particular popularity. It provides accurate energetics (for most species of the G2/97 test set the error is within about 1 kcal/mol (*17*), but this test set contains mainly small molecules and a larger average error is expected for species of increased size (*18*)) and can be applied to moderately large molecules. More specifically, CBS-QB3 calculations for molecules with up to 10-12 non-hydrogen atoms are feasible on supercomputers with sufficient scratch space and memory. Furthermore, recent studies by several groups have shown that the CBS-QB3 method also provides accurate results for transition states (*19*–*21*); hence, this method is an obvious choice to investigate reactions of biomass model compounds. On the other hand, since the disk storage, memory and CPU time requirements increase exponentially with the number of (heavy) atoms, its usefulness will possibly even in the future be limited to molecules that can only partly been regarded as model compounds for biomass.

Most of the calculations described in this chapter have been performed at the CBS-QB3 level of theory as implemented in the Gaussian 03 suite of programs (*22*). It is a well-defined multi-step calculation procedure that determines geometries and frequencies at the B3LYP/CBSB7 level of theory (*23*–*25*) and combines several higher level energy calculations and extrapolations to obtain a “complete basis set” approximation of the electronic energy that accounts for a large portion of electron correlation. To explore the possibility of using lower level calculations applicable to larger molecules, we compare CBS-QB3 results in the study of the H_2O elimination from alcohols with data obtained with the less expensive CBS-4M (*17*, *26*) and the widely used B3LYP/6-31G(d) method. The CBS-4M method is also a composite method that involves several calculation steps at different levels of theory while, in contrast, B3LYP/6-31G(d) energies are obtained from single self-consistent field calculations. The CBS-4M method was also used in the phenyl ether study.

Thermodynamic Properties

The primary results from electronic structure calculations are the electronic energies, geometries and frequencies of a molecule. We use the atomization energy method (*27*) to convert electronic energies into heats of formation. An analysis by Petersson et al. has shown that atomization energies at the CBS-QB3 level contain systematic bond errors that can be corrected to improve the heats of formation (*18*). In this work we are mainly interested in relative energies and therefore do not apply such corrections. However, such a correction would be necessary if the calculation results were to be used to build a thermodynamic database. Statistical mechanics methods are used to calculate the entropy at 298 K and heat capacities as a function of temperature from the geometry (rotational constants) and frequency data. All harmonic frequencies are scaled by a factor that depends on the method used for the frequency calculation (*28*) (e.g., a factor of 0.99 is used to scale frequencies obtained with the CBS-QB3 method). Except for those vibrations that resemble a rotation around a single bond (hindered rotation), the analysis relies on the harmonic oscillator rigid rotor assumption.

Internal rotations are treated separately for several reasons: (1) Those modes are often associated with low-frequency vibrations. Small errors in these difficult to calculate frequency values have a profound impact on the entropy and heat capacity results, and replacing those with the hindered rotor treatment generally improves the accuracy of the data. (2) Each frequency contributes at high temperatures an amount of R to the heat capacity (Cp) while internal rotors contribute only R/2 to Cp. Hence, a vibration resembling a rotation should be treated as such to make sure that the correct high temperature Cp value is reached. (3) Analyzing the internal rotations in a molecule by calculating the corresponding hindrance potentials provides a convenient way to detect lower energy conformers – provided the hindrance potential calculation is done at a sufficiently high level of theory.

The first step of the hindered rotor treatment is the aforementioned calculation of the hindrance potential. This is generally done via a relaxed potential energy surface scan (PES scan) in which the dihedral angle corresponding to the selected rotation is varied in steps of 10 degrees until a full rotation (360 degrees) is achieved. At each value of the dihedral angle all other degrees of freedom are allowed to change until the energy is minimized. In some cases, however, we fix additional coordinates to ensure that the final energy and geometry after 360-degree rotation are the same as those of the starting point. The requirement to return to the same geometry and energy after a full rotation is normally easily achieved for pure hydrocarbons, but this is not necessarily the case for oxygenates such as sugars. The reason is that oxygen containing functional groups can act as a donor or acceptor of hydrogen bonds. Since these relatively strong bonds may be broken during the rotation (e.g., of a hydroxyl group), the geometry might change during this rotation to an extent that another local minimum position on the PES is reached. In Figure 1 we present hindrance potential results for a rotating OH group in ethylene glycol, obtained at the B3LYP/6-31G(d) level of theory to illustrate this problem. As can easily be seen, the 360 degrees rotation around the C-OH bond of the left hydroxyl group leads to a different geometry and energy (the hydrogen bond is lost). This new geometry belongs to a local minimum but not to the original global minimum. Since the geometry is not restored, this is not a pure rotation. In addition, clockwise and counter clockwise rotations lead to different hindrance potentials. Figure 1 also shows that a small part of both (cw and ccw) hindrance potentials overlap. We can use this overlap region to construct an approximate potential by combining parts of both scans (the full line in Figure 1). This leads to more plausible hindrance potentials than using either of the original potentials, even though a sometimes arbitrary decision has to be made in terms of how to combine the two scans. The combined potential is then fitted to a Fourier series prior to its use in the hindered-rotor calculation.

Besides the hindrance potential the hindered rotor evaluation requires a value for the effective moment of inertia. For symmetric rotors this value is easily calculated since the rotating axis coincides with the actual bond (*29*), but for asymmetric rotors the calculation is more demanding. Kilpartick and Pitzer (*30*) solved this problem for the general case and East and Radom (*31*) provided practical approximate methods to calculate effective moment of inertias for internal rotations. We use their $I^{(2,3)}$ method in this work.

With the potential and effective moment of inertia in hand we can solve the Schrödinger equation for an one-dimensional rotor and use the energy eigenvalues to calculate the partition function and contributions of this mode to the thermal energy, entropy and heat capacity.

The outlined method is used to calculate thermodynamic properties for reactants, products as well as transition state structures. In the latter case the imaginary frequency is ignored. Thermodynamic data are calculated for a temperature range of 300K to 2500K, fitted to NASA polynomials, and stored in a database file.

Rate Expressions

The thermodynamic properties calculated as outlined in the previous paragraphs serve as input data needed to calculate the rate constant via transition state theory (TST) (*32*, *33*):

$$k(T) = \kappa(T) \cdot k_B \cdot T / h \cdot V_m^{(n-1)} \cdot \exp(-\Delta G^{\#}/RT)$$

Here, $\kappa(T)$ is the tunneling correction factor, V_m the molar volume at standard pressure ($V_m = R \bullet T/p$, with p = 1atm), n is the molarity of the reaction (e.g., n=1 for unimolecular, n=2 for bimolecular), and $\Delta G^{\#}$ is the difference in Gibbs Free Energy between the transition state geometry (ΔG^{TS}) and the reactant(s) (ΔG^{Reac}),

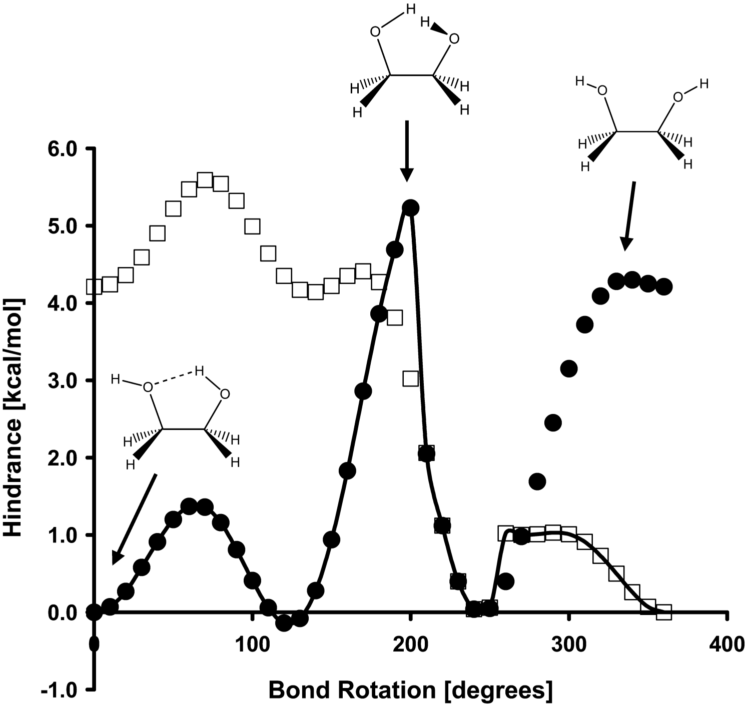


Figure 1. Hindrance potential for the internal rotation of the left OH group in ethylene glycol. Open squares: clockwise rotation, filled circles: counterclockwise rotation, solid line: "combined" hindrance potential. The drawn structures are schematic representations of some geometries.

$$\Delta G^{\#} = \Delta G^{TS} - \Delta G^{Reac} = \Delta H^{TS} - T\cdot\Delta S^{TS} - \Delta H^{Reac} + T\cdot\Delta S^{Reac}$$

As usual, the ΔG^{TS} term does not include contributions from the reaction path mode (imaginary frequency). It is calculated from enthalpy (ΔH) and entropy (ΔS) contributions stored in the thermodynamic database. The remaining symbols in the TST rate expression represent common physical constants or variables. One should note that this formulation of the transition state theory is equivalent to the original and more commonly known formulation in terms of partition functions,

$$k(T) = \kappa(T) \cdot k_B\cdot T / h \cdot Q^{TS}/Q^{Reac} \cdot \exp(-E/RT).$$

In this equation, Q^{TS} refers to the total partition function for the transition state (with contributions from translation, rotation, vibration, internal rotation and so on, but again with omission of the reaction coordinate) and Q^{Reac} is the total partition function for the reactant(s). Note that the $\Delta G^{\#}$ term in the exponential part is replaced by the barrier height E, since entropic contributions are now accounted for in the partition functions. We prefer the "ΔG version" of the transition state theory because it allows us to directly use the thermodynamic database to retrieve the input data and therefore ensures thermodynamic consistency in all steps of the mechanism development process.

The temperature dependent transmission factor κ(*T*), which accounts for contributions from quantum mechanical tunnelling, is obtained from asymmetric Eckart potentials (*34*). The correction factors obtained in this way differ in most cases only marginally from previous calculations in which we used the simpler correction formula by Wigner (*35*). However differences become more severe for reactions with small barriers and in those cases Eckart tunnelling corrections are more reliable. The choice of the tunnelling method is mainly based on the fact that this method can be applied automatically without additional calculations. Other more sophisticated treatments such as the small curvature method (*36*) require substantial additional efforts while probably changing the results by less than a factor of two under biomass pyrolysis conditions. Given that the focus is on the development of large biomass mechanisms, such an additional effort seems not warranted at this time.

Results

Rate Rules for H Abstraction Reactions from Alcohols by H Atoms and CH_3 Radicals

During the biomass gasification process volatile primary pyrolysis products are for a short time (about one to ten seconds) exposed to temperatures on the order of 800°C. Under such reaction conditions gas phase radical reactions are expected to play an important role in the cracking process. Consequently this reaction type has to be incorporated into any comprehensive biomass gasification mechanism. Two of the most dominant reactive species are H atoms and methyl radicals, which will participate in H atoms abstraction reactions to produce molecular hydrogen and methane. Since most of the C-H bonds are weaker than the bonds in H_2 and CH_4 these reactions are slightly exothermic. On the other hand, O-H bond strengths generally exceed those in H_2 and methane and those abstraction reactions are endothermic. Theoretical studies of H abstraction reactions by H and CH_3 from pure hydrocarbons have shown that individual rate constants can - with high accuracy - be replaced by generic rate expressions (rate rules) (*20*, *37–39*). We expect that similar rate rules can also be formulated for abstraction reactions involving oxygenated species. Since the hydroxyl moiety is the most dominant functional group in carbohydrates we focus on abstraction reactions related to this group. First, we discuss the applicability of rate rules for H abstraction reactions by H atoms from a set of simple alcohols. Then we briefly present analogous results for methyl radicals as the abstracting reactant. This will demonstrate that methyl radicals behave similarly to H atoms even though the kinetic parameters are different. We will also provide evidence that the rate expressions for abstractions from C-H bonds that are two or more C atoms away from the OH moiety converge to those for simple alkanes. All calculations presented here are performed at the CBS-QB3 level of theory for the most stable conformer at the DFT level. Although we will only selectively provide detailed comparisons with experimental data, we note here that the calculated abstraction rate constant for methanol plus hydrogen atom forming molecular hydrogen and the hydroxyl methyl radical agrees to better than a factor of two with rate expressions found on the NIST chemical kinetics website (*40*). A similar comparison for the second channel leading to methoxy and hydrogen is not meaningful because no reliable experimental data are available. Nevertheless we expect, also based on previous work (*20*), that our calculations are equally accurate for other reaction pathways and larger reactants.

H-Abstraction by H Atoms from the Hydroxyl Group in Alcohols

We calculated rate constants for the abstraction of the hydroxyl hydrogen by H atoms,

$$\mathrm{H} + \mathrm{ROH} \rightarrow \mathrm{H_2} + \mathrm{RO}\bullet,$$

for the following alcohols (see Figure 2): methanol (CH3OH), ethanol (C2H5OH), n-propanol (CCCOH), i-propanol (C2COH), n-butanol (CCCCOH), i-butanol (C2CCOH), s-butanol (CCC(C)OH), and t-butanol (C3COH). The names in parenthesis refer to our naming nomenclature in the figures, which removes some hydrogens and avoids subscripts to make the legends more readable (the product names in the plots should be self-explanatory). This set contains primary, secondary and tertiary alcohols and provides information on the impact of the R group on the abstraction rate constant. The calculated rate constants are plotted in Figure 3 as a function of temperature. One can see that the rate constants are so similar that individual rate constants cannot be recognized, except at low temperature conditions, at which, however, all rate constants are sufficiently small to make the reaction in practice unimportant. Since the R group has seemingly little impact on the reactivity of the hydroxyl hydrogen it is possible to describe the entire set of individual rate constants with a single generic rate expression without introducing a substantial error:

$$k(T) = n_H \cdot 3.6 \cdot 10^6\,\mathrm{cm^3 mol^{-1} s^{-1}} \cdot T^{2.11} \cdot \exp(-9.83\ \mathrm{kcal/mol}/(R\cdot T))$$

Here, n_H is the number of hydroxyl groups found in a molecule and for this particular test set its value is always one. This rate expression ("rule") was obtained in two steps: First we averaged the temperature exponents n of all individual rate constants, and refitted the calculated rate constants k(T) to a new modified Arrhenius equation by using the averaged n-value as a constant parameter. Then we averaged the A-factors and the activation energies to obtain the final values of the rate rule. This two-step procedure takes into account that the three parameters in modified Arrhenius expressions are strongly correlated. It yields rate rules with lower errors than one would get if the three parameters were independently averaged. Of course, a more rigorous procedure would have been to subject all rate constants to a least-square optimization, but we don't expect significant improvements in the results. Also, while the data for the A, n, and E parameters are given to 2 or 3 digits, this should not imply that these fitting values are known with such a high accuracy. They obviously depend on the number of test reactions, the accuracy of each individual rate constant as well on the validity of the assumption that the reaction rate constants of a given reaction class can be generalized. A better measure of the accuracy of the rate rule expression is to compare the generic rate constant to the individual TST constants at a typical temperature of interest. For H abstraction from the hydroxyl group we found that at 1000 K for all reactions of the test set the agreement with the rate rule value is better than 20%. We also calculated the rate constant for the reaction of H atoms with ethylene glycol,

$$H + HOCH_2CH_2OH \rightarrow H_2 + HOCH_2CH_2O\bullet.$$

As mentioned in "Calculation Methods", intramolecular hydrogen bonding makes the rate constant calculation for this reaction more difficult because the hindered rotor treatment breaks the hydrogen bond. As a result one would anticipate a larger uncertainty in particular in the pre-exponential factor. Thus it is remarkable that the generic rate constant and the TST calculation yield the same value of k = 5.5E10 $cm^3mol^{-1}s^{-1}$ at 1000 K.

Before continuing with other abstraction reactions, it is worthwhile to point out mechanistic implications of this reaction class. Alkoxy radicals, which are formed as products, can easily undergo β-scission reactions (*41*, *42*) to either produce aldehydes and H atoms,

$$RCH_2O\bullet \rightarrow RCH{=}O + H,$$

or formaldehyde and an alkyl radical,

$$RCH_2O\bullet \rightarrow R\bullet + CH_2{=}O.$$

Taking the ethoxy radical as an example, both channels are only moderately endothermic:

$$CH_3CH_2O\bullet \rightarrow CH_3CH{=}O + H \qquad \Delta_RH^{298} = 15.3 \text{ kcal/mol}$$

$$CH_3CH_2O\bullet \rightarrow CH_2{=}O + CH_3 \qquad \Delta_RH^{298} = 11.5 \text{ kcal/mol.}$$

The activation energy of these β-scission reactions is about 6.4 kcal/mol above the endothermicity (*41*). Although these data derived from CBS-QB3 energies differ somewhat from MP2 and QCISD calculation results by Caralp et al. (*41*, *42*) who report heats of reaction of 13.2 kcal/mol and 9.5 kcal/mol for above channels, respectively, both studies agree that they differ by about 4 kcal/mol. Using literature data for the heats of formation of formaldehyde (-27.7 kcal/mol (*43*)), acetaldehyde (-39.2 kcal/mol (*44*)), methyl (34.8 kcal/mol (*43*)), and hydrogen (52.1 kcal/mol (*43*)) we obtain with our heats of reaction results for ethoxy radical heat of formation values of -4.0 and -4.3 kcal/mol, respectively. These results are in excellent agreement with the literature value of -3.7+/-0.8 kcal/mol (*45*). In contrast, the value reported by Caralp et al. is 0+/-1 kcal/mol. This implies that the CBS-QB3 based heats of reaction are likely more accurate. In general, we expect uncertainties in calculated heats of reaction to be around 1 kcal/mol since some calculation errors will cancel out, but derived heats of formation can be larger.

Returning to the decomposition of ethoxy: while the C-C bond scission is dominant at low temperatures (< 500 K) (*41*), the only slightly more endothermic

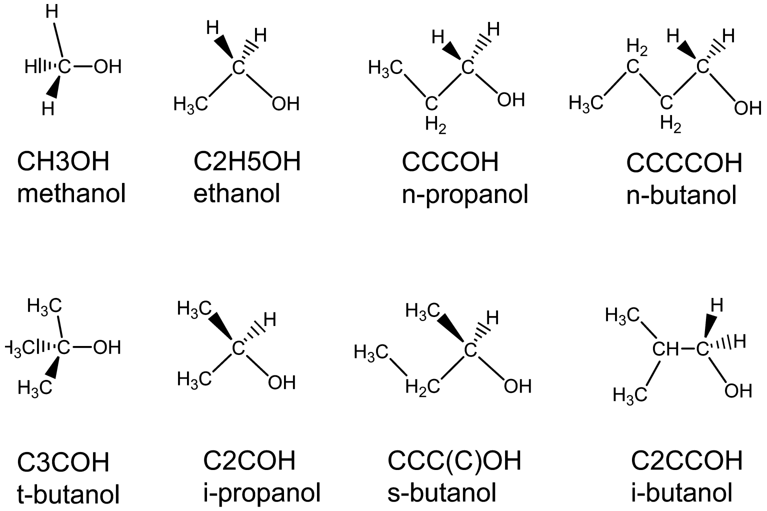


Figure 2. Structures and nomenclature for alcohols used in this study

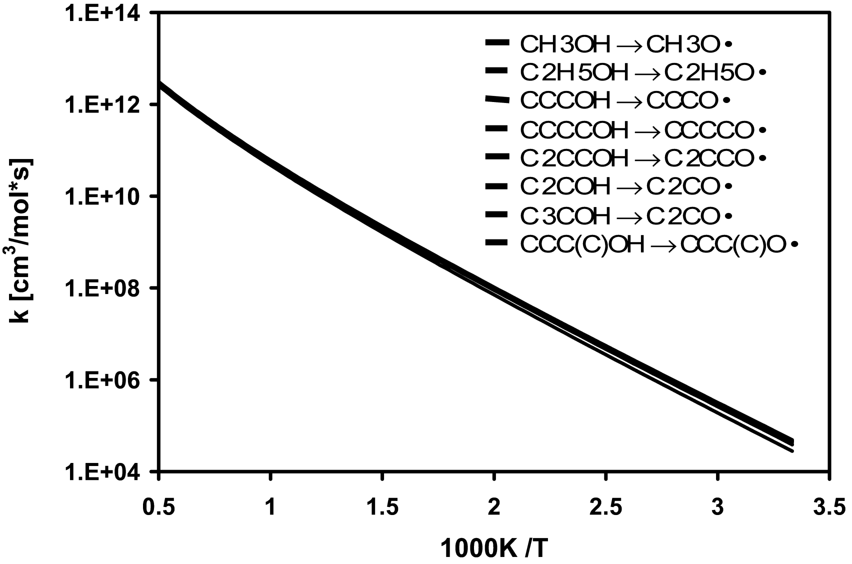


Figure 3. Calculated rate constants (per H site) for the H abstraction from simple alcohols ROH by H atoms

C-H bond cleavage will become increasingly competitive as the temperature rises. Because aldehydes are more reactive than alcohols and the formed radical sustains the abstraction process, we expect that this reaction sequence will increase the reactivity of the mixture. In addition, the most likely fate of formaldehyde is its oxidation to CO, one of the desired products of the gasification process.

H-Abstraction by H Atoms from the C-H Bond in α-Position to the Hydroxyl Group in Alcohols

We have calculated rate constants for the H abstraction by H atoms from C-H bonds in the α-position to the hydroxyl group,

$$H + R_xH_{2-x}CHOH \rightarrow H_2 + R_xH_{2-x}C\bullet OH\text{, with } x = 0,1\text{, or } 2,$$

for the same set of alcohols described in the previous section, except obviously for t-butanol. The results are presented in Figure 4. As expected, we find that the rate constants depend on the nature of the C-H bond as is evident from the fact that the entire set of rate constants group into three sub-categories. Methanol is the only molecule with a primary α-C-H group and forms its own sub-set in this context. The second sub-set consists of the reactions by ethanol, propanol, n-butanol, and i-butanol. Finally the alcohols i-propanol and s-butanol containing tertiary α-C-H bonds create the third group. The rate constants within the sub-sets vary only slightly, so that it is reasonable to define two representative generalized rate constants for secondary and tertiary α-C-H abstraction reactions by H atoms from alcohols:

$$RCH_2OH\text{:}\quad k(T) = n_H \cdot 2.7 \cdot 10^6\,cm^3mol^{-1}s^{-1} \cdot T^{2.16} \cdot \exp(-4.14\text{ kcal/mol}/(R\cdot T))$$

$$RR'CHOH\text{: } k(T) = n_H \cdot 2.4 \cdot 10^7\,cm^3mol^{-1}s^{-1} \cdot T^{1.88} \cdot \exp(-2.46\text{ kcal/mol}/(R\cdot T))$$

n_H represents the number of α-C-H bonds in the reacting alcohol. Note again that the values for the activation energy are fitting results and the number of digits does not reflect the accuracy of these values. The quality of these generic rate constants is good and deviations for a typical reaction temperature of 1000 K are again within 20% or better. We also tested the applicability of the rate rule for the reaction of ethylene glycol and found an agreement to within 35% at 1000 K. Such a good agreement was not expected because one would think that the strength of the internal H bond changes during the reaction. This should have an impact on the rate constant and therefore distinguish the hydrogen abstraction in ethylene glycol from simple alcohols. Obviously this effect appears to be small.

Similar to the H abstraction reactions from hydroxyl groups, the products formed via abstraction of α-C-H's of alcohols are very reactive and can easily undergo β-scission reactions. One channel leads to H atoms and an aldehyde:

$$RCH\bullet OH \rightarrow RCH{=}O + H.$$

Alternative pathways produce unsaturated alcohols and either hydrogen or an alkyl radical:

$$R'CH_2CH\bullet OH \rightarrow H + R'CH{=}CHOH$$

$$R'CH_2CH\bullet OH \rightarrow R'\bullet + CH_2{=}CHOH$$

Because α-hydroxyalkyl radicals are about 10 kcal/mol more stable than alkoxy radicals (e. g., the CBS-QB3 atomization energy $\Delta_{atom}H^{298}(CH_3CH_2CH\bullet OH)$ = -17.5 kcal/mol mol is 9.7 kcal/mol lower than $\Delta_{atom}H^{298}(CH_3CH_2CH_2O\bullet)$ =

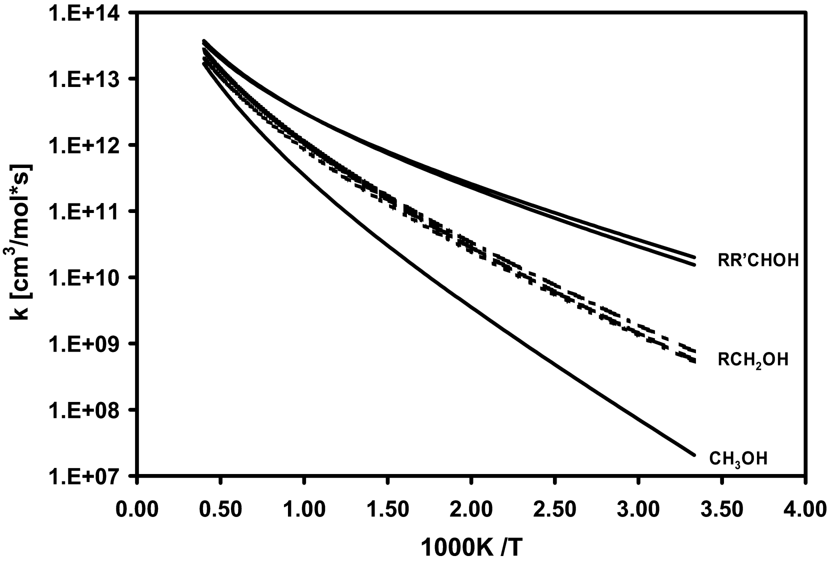


Figure 4. Rate constants for the H abstraction from C-H groups in α-position to the hydroxyl moiety of simple alcohols.

-7.8 kcal/mol) these subsequent β-scission reactions are more endothermic than those discussed earlier. For example, the calculated heats of reaction for the $CH_3CH_2CH{\bullet}OH$ radical are

$$CH_3CH_2CH{\bullet}OH \rightarrow CH_3CH_2CH{=}O + H \qquad \Delta_RH^{298} = 25.3\ \text{kcal/mol}$$

$$CH_3CH_2CH{\bullet}OH \rightarrow CH_3CH{=}CHOH + H \qquad \Delta_RH^{298} = 34.0\ \text{kcal/mol}$$

$$CH_3CH_2CH{\bullet}OH \rightarrow CH_2{=}CHOH + CH_3 \qquad \Delta_RH^{298} = 23.4\ \text{kcal/mol.}$$

Even with a small additional barrier of a few kcal/mol, the available thermal energy under typical gasification conditions should be sufficient to allow fast β-scission reactions of RCH•OH radicals to occur. We expect that these subsequent reactions accelerate the overall conversion process because the products are more reactive and the radicals necessary for the H abstraction reaction are regenerated. Therefore this reaction class and the chemistry of the products need to be incorporated in biomass mechanisms.

The higher stability of RCH•OH radicals compared to $RCH_2O{\bullet}$ radicals also explains why the H abstraction reactions from α-CH bonds are faster than those from the alcoholic hydroxyl group: a higher stability corresponds to a lower bond strength and therefore a lower barrier. The fitted barriers E (in kcal/mol) for α-C-H abstraction are with 2.5, 4.1 and 5.9 (for methanol) clearly smaller than the value of 9.8 that was found for abstraction from RO-H. The same conclusion holds for Arrhenius activation energies Ea (Ea = E + nRT) since all n values are similar. The calculations also yield higher A-factors for the abstractions from α-C-H bonds compared to those for the RO-H bond abstraction, but the reason for this is less obvious.

H-Abstraction by H Atoms from the C-H Bond in β-Position to the Hydroxyl Group in Alcohols

Finally we present in Figure 5 results for H abstraction reactions from the β-carbon position in alcohols:

$$H + R'_yH_{2-y}CHCH_xR_{2-x}OH \rightarrow H_2 + R'_yH_{2-y}C\bullet CH_xR_{2-x}OH, \text{ with } x,y = 0,1, \text{ or } 2.$$

As before, the rate constant is mainly determined by the nature of the C-H bond as the three distinct groups of rate constants show, and the reactivity order follows the general expectation: tertiary C-H bonds are more reactive than secondary C-H, which in turn react faster than primary C-H bonds. Using the two-step procedure described earlier to derive generic rate rules we obtain the following three rules from fitting to the TST rate constants:

$$\text{prim. } \beta\text{-C-H:} \quad k(T) = n_H \cdot 4.4 \cdot 10^6\ cm^3mol^{-1}s^{-1} \cdot T^{2.12} \cdot \exp(-8.98\ kcal/mol/(R{\cdot}T))$$

$$\text{sec. } \beta\text{-C-H:} \quad k(T) = n_H \cdot 7.7 \cdot 10^6\ cm^3mol^{-1}s^{-1} \cdot T^{2.05} \cdot \exp(-6.15\ kcal/mol/(R{\cdot}T))$$

$$\text{tert. } \beta\text{-C-H:} \quad k(T) = n_H \cdot 6.6 \cdot 10^6\ cm^3mol^{-1}s^{-1} \cdot T^{2.08} \cdot \exp(-4.18\ kcal/mol/(R{\cdot}T))$$

The rate constants are smaller than those obtained for α-C-H bonds, which is mainly due to higher barriers. This can be rationalized by the fact that β-C-H groups do not experience the stabilizing mesomeric effect of the OH group (RC•OH Û RC=O•H) found for α-C-H sites. The barriers of the three rate rule expressions for H abstraction from β-C-H bonds follow a linear Gibbs Free Energy relationship and the slope of 0.93 is essentially identical to the Evans-Polanyi slope found for alkanes. This is an indication that the influence of the functional OH group ceases in the β-position and the rate rules for oxygenates converge to those for hydrocarbons. We will come back to this point in the next section where we discuss H abstraction rate rules for methyl radicals.

H Abstraction Reactions from Alcohols by Methyl Radicals

Hydrogen abstraction by methyl radicals leads to a series of analogous reaction classes. In general we expect similar results compared to abstraction by H atoms because of the nearly identical bond strengths in methane and the hydrogen molecule. However, there are also at least three clear differences between these reactants: (1) Hydrogen atoms are smaller and lighter than methyl radicals and should therefore exhibit less steric interactions. (2) H atoms use their spherical 1s orbital in abstraction reactions while the reacting orbital of methyl radicals is of p-type. Since p-orbitals are oriented in space, this should add to the expected increased steric sensitivity of abstraction reactions by methyl radicals. (3) The hydrogen atom loses only translational degrees of freedom upon entering the reaction while the methyl radical loses translational as well as rotational degrees of freedom. These differences should manifest themselves in a larger spread of the calculated rate constants. While this argument suggests that the abstraction rate constants for methyl are less uniform by nature, the calculation results will also be less accurate, because the heavier methyl radical produces to two low-frequency bending modes in the transition state. Frequencies depend inversely on the square root of the reduced mass; hence heavier abstracting reactants generate bending modes of lower frequency (if all other parameters are kept the same). Unavoidable small errors in low-frequency vibrational modes lead to large errors in the pre-exponential factor and might create fluctuations in the rate constants of the test set. This makes it difficult to assign deviations of individual rate constants to real reactivity differences.

In Figure 6 and Figure 7 we present the results for H abstraction from the hydroxyl, and α-CH and β-CH groups by methyl radicals. The rate constants, especially for the reactions of the OH and β-C-H sites show the expected impact from steric interactions. For example, the rate constants for iso- and tertiary butanol in Figure 6a are clearly smaller than those for the remaining alcohols. Deviations are seen over the entire temperature range but they are largest at the lower temperature end. This indicates that both the pre-exponential factors and the activation energies are influenced by steric interactions. Despite these deviations, the corresponding rate rules at 1000 K are still within a factor of two of the TST results as Figure 6b reveals for abstraction reactions of the OH group. The following rate rule expressions for CH_3 as abstracting radical have been obtained:

ROH: $k(T) = n_H \cdot 2.5 \cdot 10^2\,cm^3mol^{-1}s^{-1} \cdot T^{2.93} \cdot \exp(-7.50\ kcal/mol/(R{\cdot}T))$

prim. α-CH: $k(T) = n_H \cdot 5.7 \cdot 10^1\,cm^3mol^{-1}s^{-1} \cdot T^{3.25} \cdot \exp(-9.26\ kcal/mol/(R{\cdot}T))$

sec. α-CH: $k(T) = n_H \cdot 1.1 \cdot 10^2\,cm^3mol^{-1}s^{-1} \cdot T^{3.12} \cdot \exp(-7.60\ kcal/mol/(R{\cdot}T))$

tert. α-CH: $k(T) = n_H \cdot 6.6 \cdot 10^2\,cm^3mol^{-1}s^{-1} \cdot T^{2.90} \cdot \exp(-5.73\ kcal/mol/(R{\cdot}T))$

prim. β-CH: $k(T) = n_H \cdot 1.8 \cdot 10^2\,cm^3mol^{-1}s^{-1} \cdot T^{3.00} \cdot \exp(-10.9\ kcal/mol/(R{\cdot}T))$

sec. β-CH: $k(T) = n_H \cdot 4.6 \cdot 10^2\,cm^3mol^{-1}s^{-1} \cdot T^{2.91} \cdot \exp(-8.79\ kcal/mol/(R{\cdot}T))$

tert. β-CH: $k(T) = n_H \cdot 1.2 \cdot 10^3\,cm^3mol^{-1}s^{-1} \cdot T^{2.87} \cdot \exp(-7.11\ kcal/mol/(R{\cdot}T))$

Since the OH moiety has its biggest impact on the reactivity of the α-CH bond, we selected this reaction class to compare the Evans-Polanyi relationship for CH_3 radicals with that obtained for H atoms (see Figure 8). Both slopes are identical within the error margins and larger than unity. In general one would expect a value between 0 and 1 for the slope, but larger values have been reported for other reactions (*46*). The large value might be understood in terms of an additional stabilization of the transition states (relative to the products) by hyperconjugation with the alkyl groups. This explanation is equally valid for both, H atoms and CH_3 radicals as abstracting species. Further we notice a shift of the straight lines by 4 kcal/mol. This reflects that, similar to H abstraction reactions involving pure hydrocarbons, the barriers for abstractions by methyl groups are higher than those

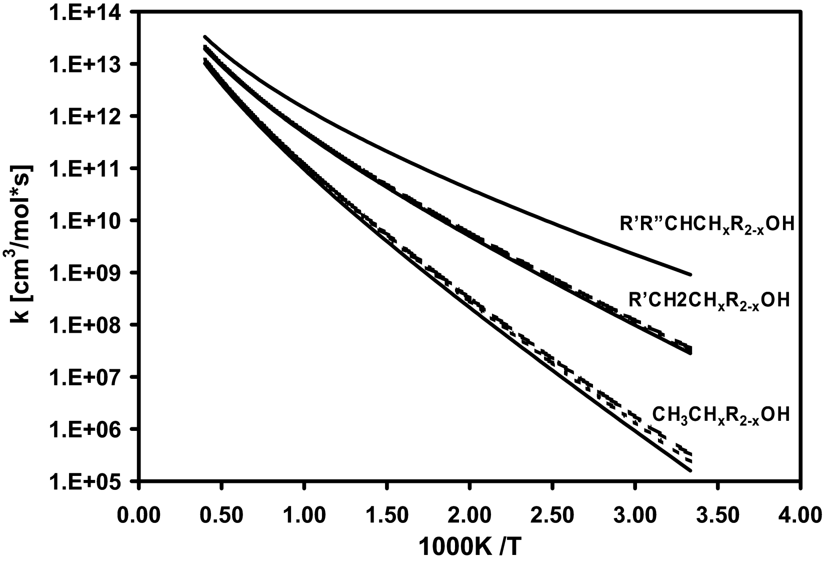


Figure 5. Rate constants for the H abstraction (per H) from C-H groups in β-position to the hydroxyl moietyfor simple alcohols.

for H atoms. Since different orbitals are engaged in the reactions of H atoms and CH_3 radicals, finding different barrier heights for these reaction classes is plausible.

Finally we compare in Figure 9 the generic rate constants for H abstraction by methyl radicals from the β-C-H groups in alcohols with those in pure hydrocarbons. Obviously there is very little difference among the corresponding rate expressions. This shows that the impact of the OH group on the reactivity of neighboring C-H bonds vanishes at the β-position. In other words, except for the immediate vicinity of the OH group we can apply generic rate constants for pure hydrocarbons to estimate the reactivity of a C-H group. Unpublished results for other oxygenates support this conclusion.

In summary, rate rules can be used confidently for H abstraction reactions from different sites in alcohols. The OH group per se has only a notable reactivity altering effect on C-H bonds in the α-position. It enhances the abstraction rate constants of these C-H groups significantly, which makes alcohols more reactive than alkanes. Steric effects are observed, especially if CH_3 is the abstracting species, but these effects are small and can to a first approximation be ignored. While the rate estimation rules work well for the test set investigated, future work needs to focus on the transferability of these rules to multi-functional molecules with OH groups. Initial calculations with ethylene glycol as reactant indicate that the rules might also be applicable if more than one functional group is present.

Rate Rules for the Gas-Phase Elimination of Water from Alcohols

An important reaction in biomass pyrolysis is dehydration. Although probably most of the chemically bound water is released in the condensed phase during the vaporization and depolymerization process, the question arises as to

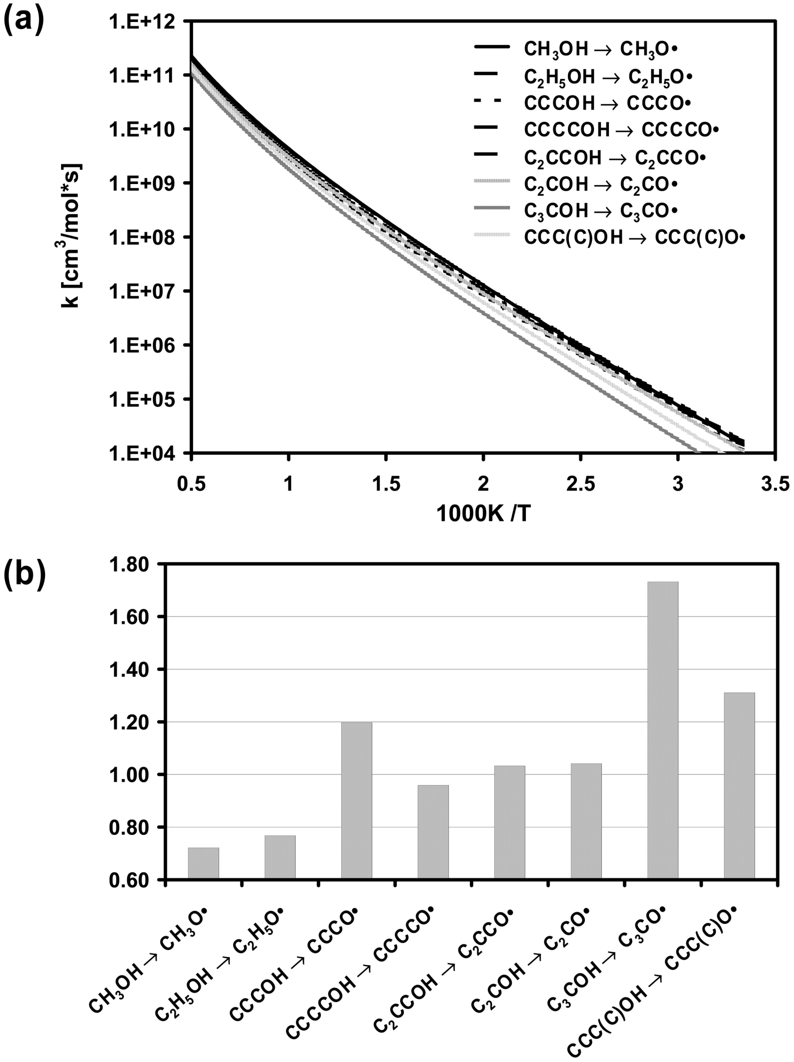


Figure 6. H abstraction rate constants per H from the OH moiety in alcohols by CH_3 radicals. (a) TST calculated rate constants (b) ratios between estimated and calculated rate constants at 1000 K

whether gas phase reactions contribute to the dehydration of the primary pyrolysis products of biomass. If so, which reactions are responsible? In the absence of oxygen, the most likely mechanism for water formation is its elimination from hydroxyl groups containing molecules such as sugars, (poly) alcohols and other multifunctional primary pyrolysis products. In order to investigate the possible role of this reaction type we performed a systematic study of the rate constants - similar to the H abstraction reactions discussed in the previous sections. The objectives of this part are, however, broader: (1) We are interested to see if rate estimation rules also work for unimolecular elimination reactions. (2) If

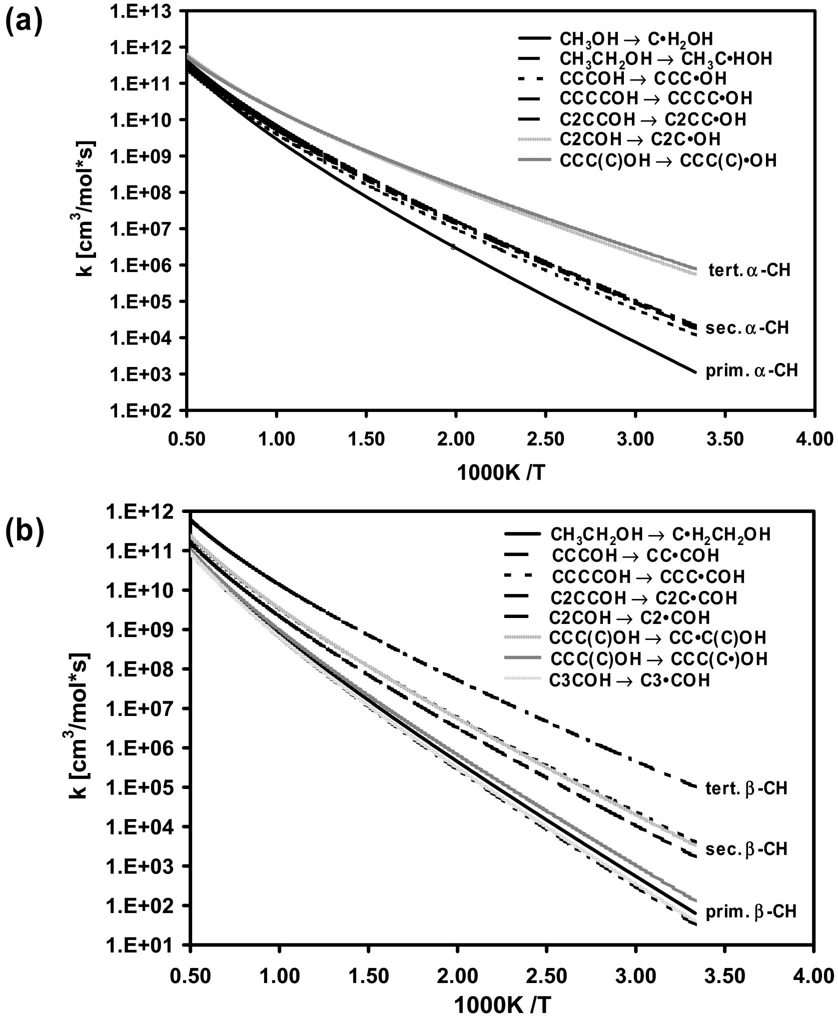


Figure 7. H abstraction rate constants on a per H atom basis from α- and β-C-H sites in alcohols by CH_3 radicals.

so, then the question of transferability of the rate rule(s) to related reaction classes becomes important. (3) We will investigate the sensitivity of our rate constant results to the calculation method by comparing the performance of two lower-level methods with CBS-QB3. Should a lower level method be able to produce acceptable results, one could use it with confidence to investigate the reactivity of larger biomass model compounds. (4) Related to the molecular water elimination reaction, we will also briefly discuss a 'water assisted' bimolecular elimination pathway in the gas phase.

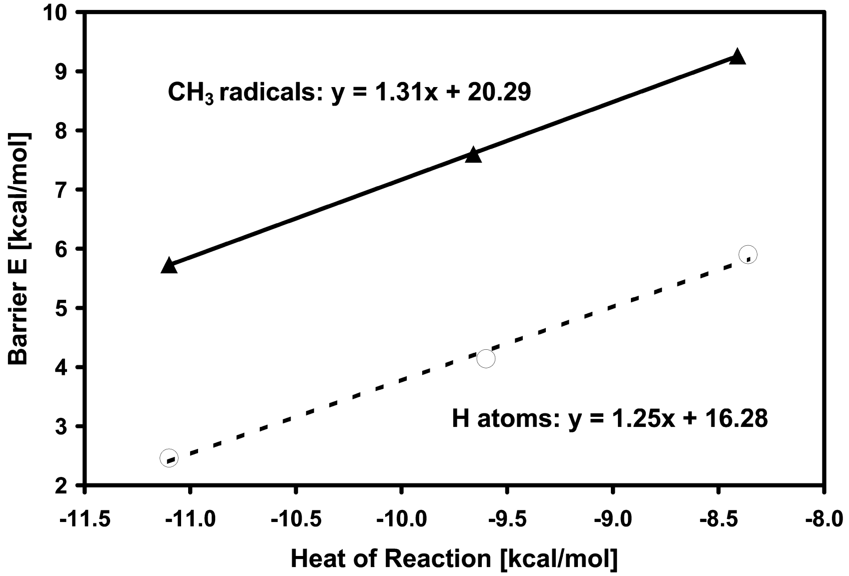


Figure 8. Evans-Polanyi relationships for H abstraction by H and CH_3, respectively, from α-C-H in alcohols. The heats of reaction values are obtained from CBS-QB3 calculations.

1. Rate Rule Development

To study the elimination of water from alcohols the following test set has been chosen: ethanol, n-propanol, i-propanol, n-butanol, i-butanol, s-butanol, and t-butanol (see Figure 2). The rate constants calculated using the same procedure described previously are presented in Figure 10. Within the temperature range of 300 K – 2000 K the rate constants increase by more than 36 orders of magnitude and reach values on the order of 1 x 10^6 s^{-1} at the highest temperature. This strong temperature dependence is caused by a high activation energy (about 65 kcal/mol). All individual rate constants appear to group well together but the large range of rate constants covered in Figure 10 makes an assessment of how well the rate constants agree difficult. Hence we plot in Figure 11 the energies E from restricted modified Arrhenius fits (as described for H abstraction reactions) against the calculated heats of reaction.

All reactions are endothermic (see Table 1) but the degree of endothermicity varies between 6.6 and 14.2 kcal/mol. It correlates reasonably well with the nature of the C-H bond of the leaving hydrogen. The nature of the C–OH group on the other hand seems to have little impact on the thermochemistry. For example, water elimination from C_2CCOH and from C_3COH yield the same olefin i-butene ($C_2C=C$). Since the products are the same, the heats of reaction are determined by the reactants. The reaction of C_2CCOH is the least endothermic reaction (6.6 kcal/mol) even though a primary C-OH bond is broken. In contrast, the endothermicity for C_3COH with a tertiary OH bond is the highest of this set (14.2 kcal/mol). The next highest endothermic reactions involve CCC(C)OH, in which the OH is

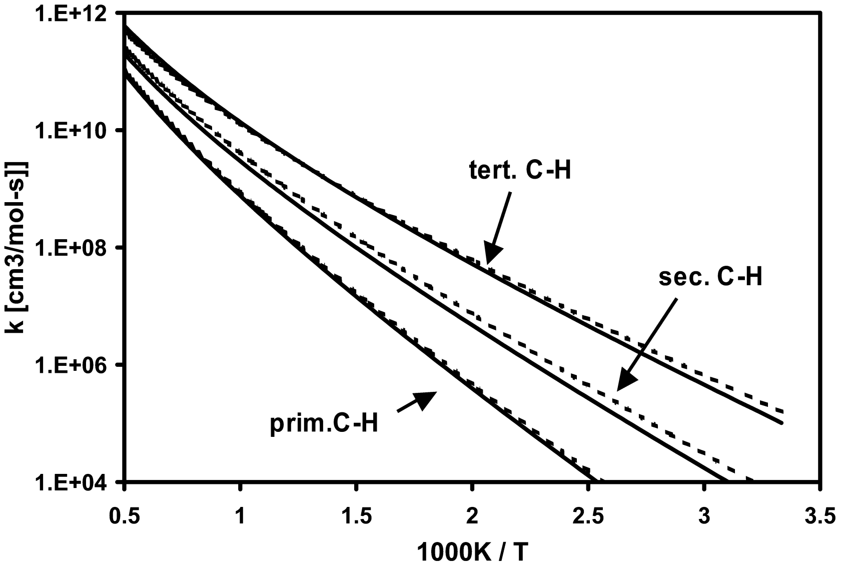


Figure 9. Comparison of the rate constants from rate rules for H abstraction by CH_3 radicals from C-H bonds in β-position of alcohols (solid lines) with those of pure alkanes (broken lines)

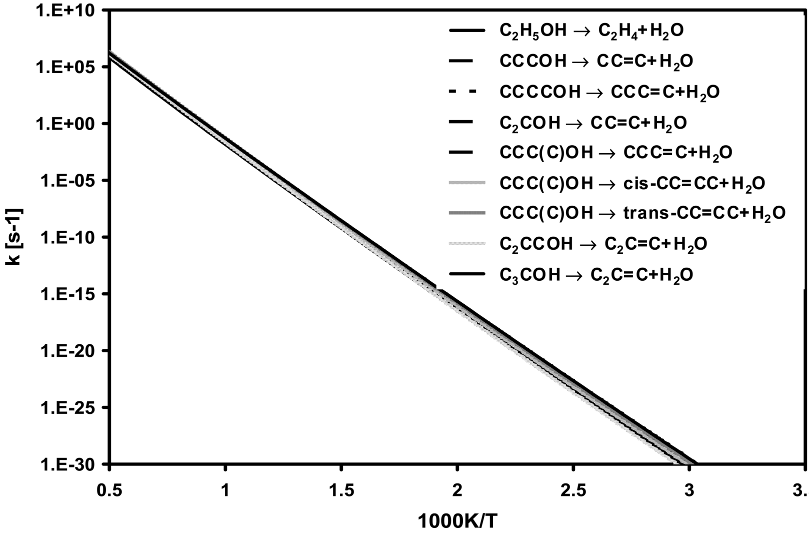


Figure 10. Calculated unimolecular rate constants for the elimination of water from simple alcohols.

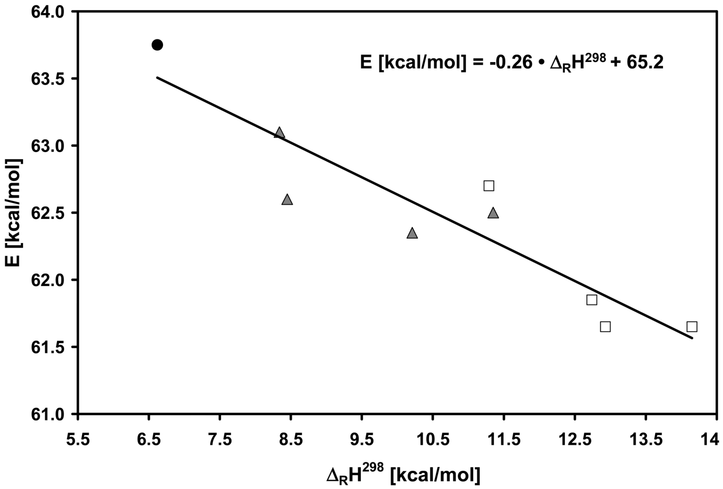


Figure 11. Evans-Polanyi plot for the elimination of molecular water from alcohols. The different symbols distinguish the nature of the reacting C-H bond: open squares: H from CH_3 groups; triangles: H from CH_2 groups; filled circle: H from a CH group.

bound to a secondary carbon but a primary C-H bond is broken. Hence the type of the C-H bond dominates the endothermicity. The most surprising aspect of the Evans-Polanyi plot in Figure 11, however, is its slope. Typically, the barrier E decreases with decreasing endothermicity but in this case the energetically more favorable reactions have the highest barriers. The three lowest barriers belong to alcohols in which the OH group is bound to either a tertiary or a secondary carbon and the leaving H atom is connected to a CH_3 group. This indicates that the type of OH site and possible steric effects are stabilizing factors for the transition state. Hence this unusual plot can be explained by different stabilization mechanisms for the reactants and/or products and the transition states. The argumentation made for E will also hold for the activation energy Ea (=E+nRT), since the E values were obtained for a constant value of n.

Since the Evans-Polanyi plot reveals a small dependence of the barrier on the heat of reaction, we use this information to generate for this reaction class a rate rule with a variable barrier height. The recommended rate constant is:

$$k(T) = n_H \cdot 1.6 \cdot 10^4\ s^{-1} \cdot T^{2.64} \cdot \exp(-E/(R \cdot T)); \qquad E = 65.2\ \text{kcal/mol} - 0.26 \cdot \Delta_R H^{298}$$

Here n_H is the number of equivalent C-H bonds that are available for the water elimination. At T = 1000 K the use of this generic rate rule reproduces the original rate constants within a factor of two except for ethanol, for which the deviations are somewhat larger (factor of 2.4). Given the very strong temperature dependence of this reaction class this agreement is sufficient for mechanism development purposes.

Using an average E value of 62.4 kcal/mol, the rate rule evaluates at 1000 K to $3.1 \cdot 10^{-2}\ s^{-1}$. A look at Figure 10 shows that the rate constants for simple alcohols

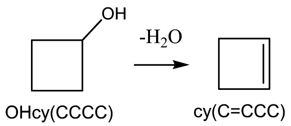
Table 1. Bond dissociation energies (BDE), heats of reaction ($\Delta_R H^{298}$) and barriers (E) for the water elimination from alcohols. All values are derived from CBS-QB3 calculations. The estimated uncertainty is 1kcal/mol

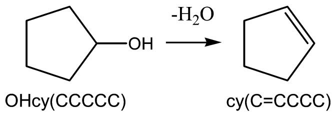
Reaction	*BDE C-OH*	*BDE C-H*	$\Delta_R H^{298}$	*E*
	kcal/mol	*kcal/mol*	*kcal/mol*	*kcal/mol*
CCOH→C=C+H_2O	95.2	102.8	11.6	62.7
CCCOH→CC=C+H_2O	95.5	99.9	8.3	63.1
CCCCOH→CCC=C+H_2O	95.3	100.1	8.5	62.6
C2COH→CC=C+H_2O	96.6	103.0	12.2	61.9
C2CCOH→C2C=C+H_2O	96.0	98.0	6.6	63.8
C3COH→C2C=C+H_2O	97.8	103.1	14.2	61.7
CCC(C)OH→CCC=C+H_2O	96.9	103.2	12.9	61.7
CCC(C)OH→c-CC=CC+H_2O	96.9	100.1	11.4	62.5
CCC(C)OH→t-CC=CC+H_2O	96.9	100.1	10.2	62.4

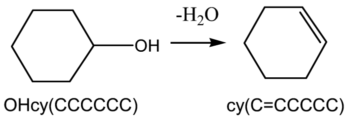
exceed the value of 1 s^{-1} between 1100 K and 1150 K. Considering a residence time of several seconds in the thermal cracker, this reaction type could play a role at the highest gasification temperatures.

2. *Transferability of Rate Rules*

Since the rate rule for water elimination is based on a test set containing only simple alcohols, the question arises as to whether this rule is also valid for reactants containing additional functional groups. We addressed this question by calculating TST rate constants for OH group containing molecules that are structurally related to carbohydrates, such as diols, hydroxyaldehydes, cyclic alcohols and hemiacetals. We also investigated water elimination from the C_2 position in levoglucosan. Some selected results are presented in Figure 12. Figure 12a compares predicted and calculated rate constants for the following three cyclic alcohols:

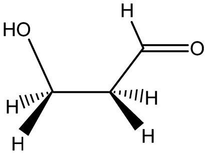




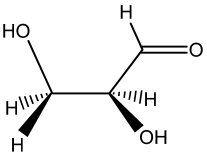


The solid lines represent the TST results and the dotted lines are rate rule estimates from above using the three heats of reaction. The three estimates are close together because the calculated heats of reaction for the three reactions are similar and therefore the estimated E values range only between 61-63 kcal/mol. On the other hand, the individually calculated constants vary clearly more than the estimates, especially at low temperatures. Cylcopentanol reacts faster that cyclohexanol and cyclobutanol, and the barriers of the modified Arrhenius fits show substantially more variation than the predicted ones (between 60 and 67 kcal/mol). This suggests that the Evans-Polanyi relationship for linear alcohols does not capture the differences in cyclic alcohols well. At higher temperatures the differences between the estimated and the directly calculated rate constants are relatively small and the rate rule yields reasonably close predictions.

Significantly larger deviations are observed for substituted aldehydes when water is eliminated from the α,β-position. This is demonstrated in Figure 12b for the reactants 3-hydroxy propanal and glyceraldehyde.



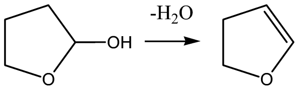
3-hydroxy propanal
HOCCCH=O



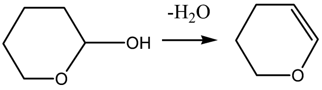
Glyceraldehyde
HOCC(OH)CH=O

The calculated rate constants for both reactants are clearly larger than predicted by the rate rule. This is not surprising, because the rate rule was developed for saturated alcohols yielding olefins with an isolated double bond. In contrast, these two reactions yield products with conjugated double bonds. In other words, the aldehyde moiety influences the reactivity by stabilizing the transition state and the products via a stabilization mechanism (the mesomeric effect) that is different from simple aliphatic alcohols. Furthermore, the inverse Evans-Polanyi relationship found for simple alcohols leads to a severe overestimation of the barrier. This explains the discrepancy between the estimated and calculated rate constants. In Figure 12b we also include the rate constant for water elimination from ethylene glycol (HOCCOH, see Figure 1 for the structure). This rate constant is very close to the estimates. Similar to the observation for H abstraction reactions we can conclude that the rate rule for water elimination works satisfactorily for ethylene glycol and probably other polyols as long as the hindered rotor analysis treats intra-molecular hydrogen bonds adequately.

The third part of Figure 12 shows the results of applying the rate rule to two cyclic hemiacetals



HOcy(COCCC) cy(OC=CCC)



HOcy(COCCCC) cy(OC=CCCC)

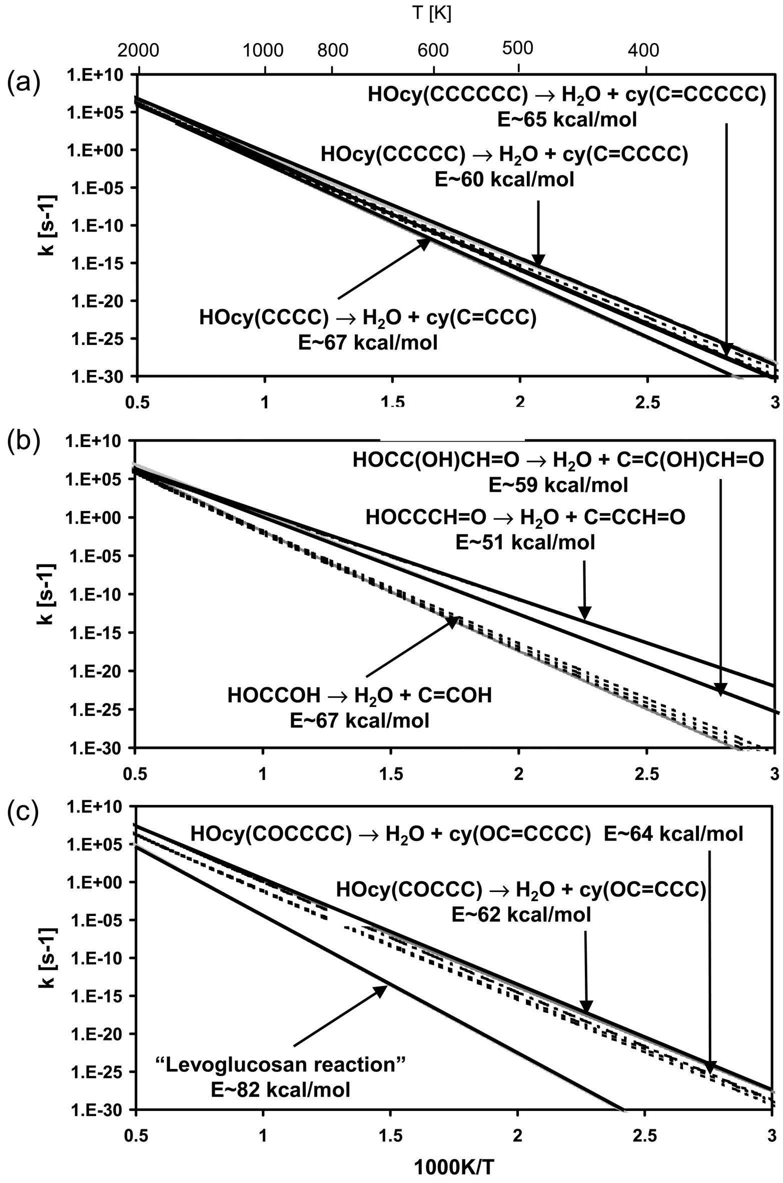
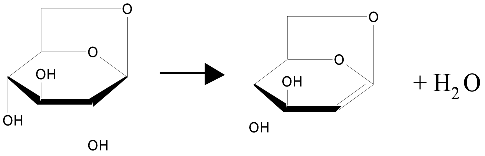


Figure 12. Comparison of calculated rate constants for the water elimination from multifunctional alcohols to estimated rate constants using the rate rule. The barriers E are approximate values. See text for details.

and to levoglucosan. The following water elimination reaction of levoglucosan has been considered:



The calculations predict the rate constants for the two hemiacetals to be very close at high temperature (they converge to a constant pre-exponential factor), but the 5-member ring reacts at low temperatures clearly faster than the 2-hydroxy-tetrahydropyran (6-member ring structure). The magnitude of this difference is not captured by the rate rule (see the two dotted lines in Figure 12), although it predicts a slightly higher reactivity for 2-hydroxy- tetrahydrofuran, HOcy(COCCC), than for 2-hydroxy-tetrahydropyran, HOcy(COCCCC). Since the Evans-Polanyi relationship was developed for linear alcohols, it does not capture ring-strain effects. The rate rule agrees qualitatively with the convergence of both hemiacetal rate constants at higher temperatures, but the predicted value is too small.

The rate constant for the elimination reaction of levoglucosan is significantly lower than rate constants for the hemiacetals. This is also obviously related to the ring strain, which increases significantly in the product. The large endothermicity of this reaction of more than 44 kcal/mol confirms this point. The inverse Evans-Polanyi relationship (Figure 11) predicts that the barrier should be significantly lower compared to those for simple alcohols and hemiacetals. More specifically, it predicts a barrier of 53.6 kcal/mol for the levoglucosan reaction, while the predicted barriers for the hemiacetals are 60.6 kcal/mol and 61.4 kcal/mol. This predicted low barrier for the levoglucosan reaction is obviously in contradiction with the results shown in Figure 12 and also with the common wisdom that reactions leading to sterically constrained products should have higher barriers than similar reactions that do not lead to strained products. Therefore, the rate rule cannot be applied to this reaction and we did not even include the estimated rate constant in Figure 12c.

The example of water elimination from levoglucosan together with the reactions leading to conjugated double bonds show that chemical intuition can often *a priori* determine when a rate rule is likely to fail. Those identified reactions would require an individual kinetic analysis. In all other cases, for which no obvious reasons for a dramatic failure of the rate rule exist, the estimates are reasonably close to the 'correct' values. Since rate and sensitivity analysis tools in modeling software are able to identify crucial reactions in a kinetic model, one can use rate rules to quickly create a reasonably complete reaction set without worrying too much about high accuracy. At a later stage those reactions that are shown to be important under certain given conditions can be reinvestigated in more detail.

3. Performance of Lower Levels of Theory

We repeated the rate constants calculations for water elimination from simple alcohols of the test set with two lower level methods: the CBS-4M composite model and the B3LYP/6-31G(d) DFT method. The main motivation to use these lower level theories is to investigate their reliabilities. The hope is to identify one of these methods to be capable to produce kinetic data with acceptable accuracy and reliability. This would allow one to extend calculations to larger and more biomass-like model compounds. On the other hand, finding out that neither method is capable of producing rate constants with an acceptable quality is also a valuable result because this would support our strategy to focus on small to medium sized representative molecules and to use these results to develop generalized rate expressions.

Our reasons for selecting the two above mentioned methods for this part of the study are outlined in the following. The hybrid DFT method B3LYP is widely used and known to produce in the majority of cases accurate geometries and, after application of a scaling factor, reasonably accurate frequencies. B3LYP geometries do not depend strongly on the size of the basis set, so even the 6-31G(d) set works well. Hence the entropy and heat capacities calculated from B3LYP results are generally quite accurate. Note, however, that some exceptions to this general assessment exist. The main issue with the B3LYP calculation is the reliability of predicted energies (*47*). In contrast, the CBS-4M method is designed to provide rather accurate electronic energies at a low cost of CPU time and moderate demands for memory and disk storage space. To do so it relies among other things on less reliable geometry and frequency calculations at the Hartree Fock and MP2 level of theory with small basis sets. If the optimized geometries deviate significantly from the correct structure, the will likely result in inaccurate entropies. Thus the choice of these two calculation methods includes one method that has its strength in the geometry optimization aspect and a second one with advantages in the energy calculation. Both methods are fast enough that they can be applied to molecules that are significantly larger than glucose.

The rate constants obtained with both calculation methods are shown in the top panel of Figure 13. Each calculation set shows some spread in the rate constants at low temperatures but they seem to converge to a common pre-exponential factor at high temperatures. As indicated in the figure, the set of rate constants from CBS-4M data is systematically below that obtained with the B3LYP method. This becomes clearer from an examination of the resulting rate rules shown in the bottom part of Figure 13. The rate rules based on individual rate constant calculations with these two models were derived in the same manner as described earlier for the CBS-QB3 calculations. The B3LYP rate rule agrees very well with the CBS-QB3 result, which is taken as reference, while the supposedly more accurate CBS-4M data deviate clearly from the solid reference line. To further investigate this matter, we compare in Figure 14 the heats of formation for the reactants and transition states. These results are presented relative to the CBS-QB3 results, which again serve as benchmark. One can see that the B3LYP/6-31G(d) energies for both the stable species and the transition states are consistently between 7-9 kcal/mol above the CBS-QB3 values. (The

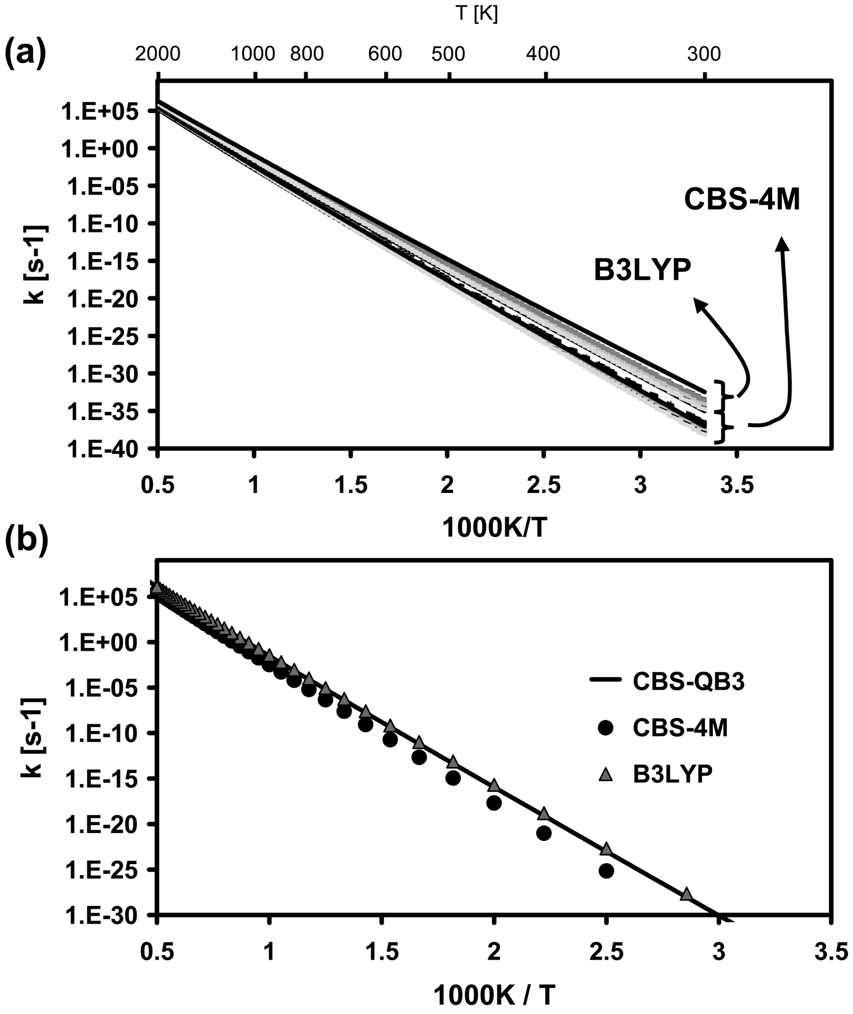


Figure 13. Rate constants for the elimination of water from simple alcohols. (a) individual TST rate constants calculated at the B3LYP/6-31G(d) and CBS-4M levels of theory, (b) comparison of the rate rules generated from the data in (a).

heat of formation for C_3COH is about 11 kcal/mol higher, which causes the rate constant for this reaction to deviate somewhat from the rest). Since the barrier height depends on the difference between transition state and reactant energies, the rather large individual errors cancel out. This explains the unexpected good rate constant predictions by the B3LYP method. On the other hand, all individual CBS-4M results are much closer to the CBS-QB3 data compared to the B3LYP method. However, the deviations are not constant but instead vary between –1 and –4 kcal/mol for the set of reactants and between +2 and –1 kcal/mol for the transition states. Consequently there is no fortunate error cancellation as in the B3LYP case. The barriers are predicted to be on average about 3 kcal/mol

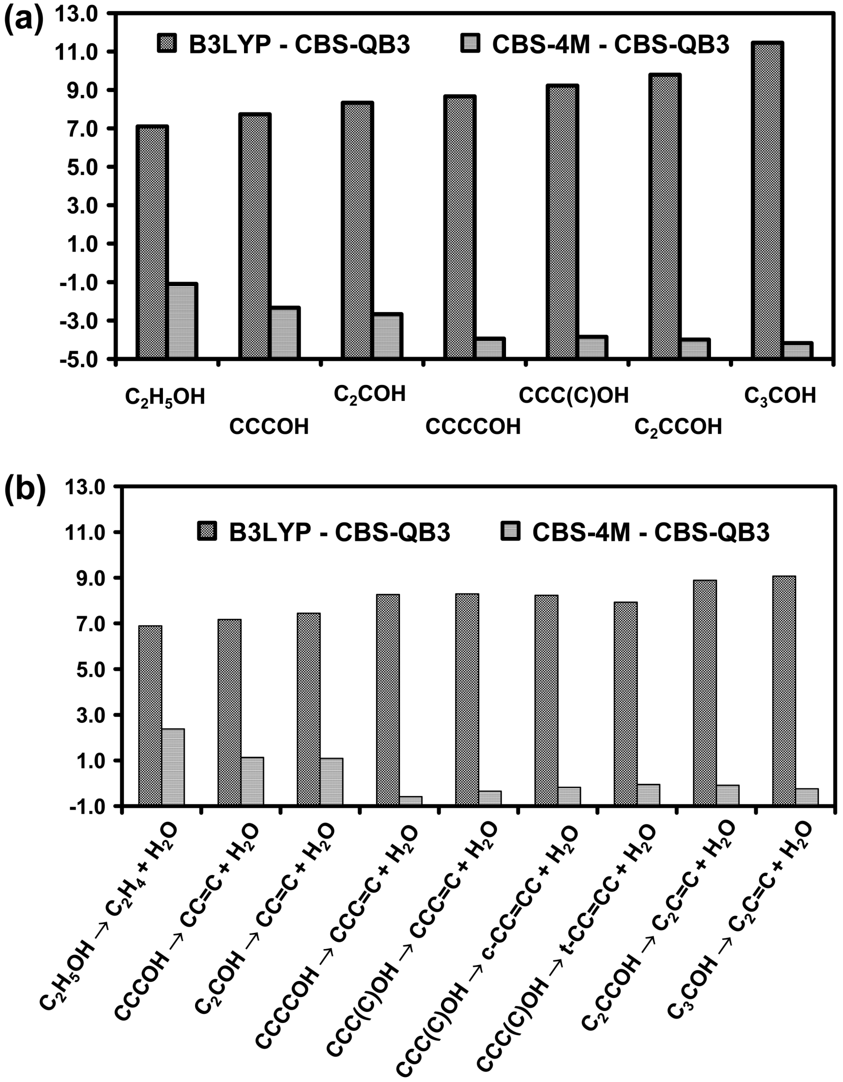


Figure 14. Comparison of the heats of formation for (a) reactants and (b) transition states obtained at the B3LYP/6-31G(d) and CBS-4M levels of theory. The results are presented relative to the CBS-QB3 data, which serve as benchmarks.

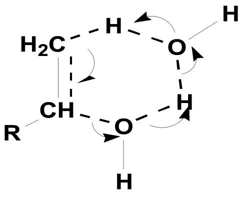
higher than the CBS-QB3 barriers and this leads to the predictions of smaller rate constants at lower temperatures.

At higher temperatures, meaning under conditions where the rate constant is dominated by the pre-exponential factor, all three methods yield quite similar results. This suggests that the geometry and frequency information obtained with the CBS-4M method for this reaction class is sufficiently accurate (not notably worse that the DFT based data). A possible explanation for this finding is that the transition states are tight and the reactive centers are therefore well defined on the PES. This leads to large frequencies for the vibrational modes of the reactive groups, which even if they contain some uncertainties, do not introduce notable errors in the partition function part (or equivalently in the entropy contribution). Errors in the pre-exponential factor are mainly caused by low-frequency modes and internal rotations, if these modes differ in the transition state and the reactant(s).

The conclusion that the apparent accuracy of the B3LYP calculations is due to a fortunate cancellation of errors leads to the question as to whether this is also the case for the reverse reaction. A requirement for yielding accurate rate constants for the reverse process is that the heats of reaction calculated at the B3LYP level are in agreement with those obtained with CBS-QB3 theory. This is not the case as can readily be seen from Figure 15. The B3LYP/6-31G(d) results are consistently higher (5-9 kcal/mol) than the CBS-QB3 values. On the other hand, the CBS-4M values are remarkably close to the CBS-QB3 benchmark values. Therefore, among the two methods tested the CBS-4M method is probably the better choice for systems too large for CBS-QB3, with the recognition that calculated rate constants will be more uncertain.

4. Is the 'Water Assisted' Elimination of Water from Alcohols Competitive?

The rate constants discussed in the previous sections are very small except at temperatures well above 1000 K. From experimental studies we know that water is formed at much lower temperatures (*48–50*). This suggests that the molecular elimination from hydroxyl groups carrying molecules is likely unimportant. However, it could be possible that steam in the gas phase is able to catalyze the elimination process and hence accelerate it. The following schematic representation of a transition state structure shows how a H_2O molecule can act as a catalyst in the elimination process:



The water molecule in the upper right part of the scheme abstracts the leaving H atom from the alcohol while simultaneously donating one of his H atoms to the leaving OH moiety. The breaking C-OH and C-H bonds of the alcohol form the new double bond of the alkene product molecule.

We investigated the potential of this reaction by calculating rate constants for a few alcohols. The results are presented in Figure 16 (top). All rate constants group nicely together indicating that a rate rule will work well for this reaction class. However, the magnitude of the rate constants for this bimolecular reaction is very small. To demonstrate this we converted the second order rate constant to a first

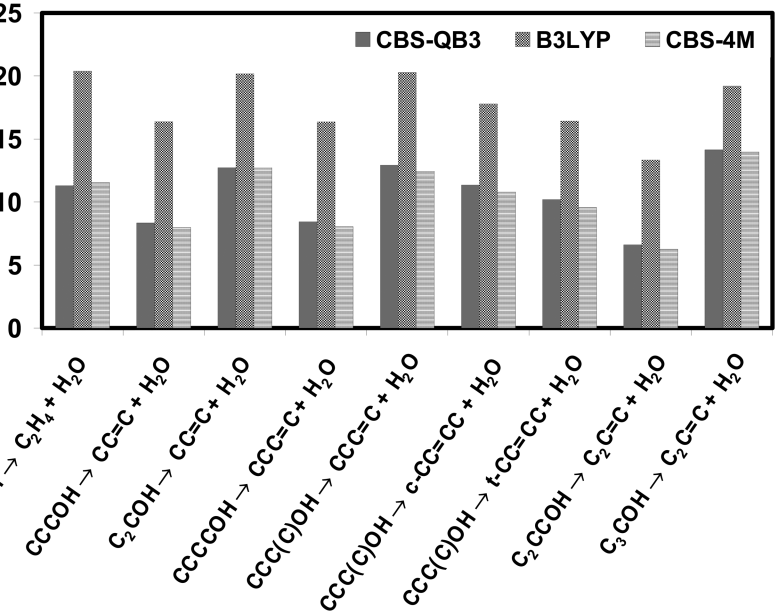


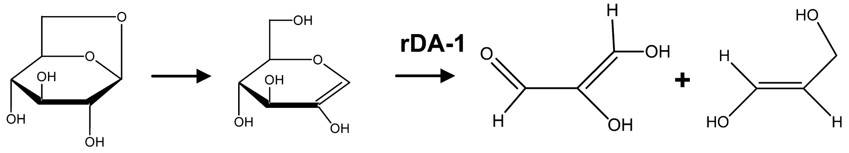
Figure 15. Comparison of calculated heats of reaction for the elimination of water from alcohols at the B3LYP/6-31G(d), CBS-4M and CBS-QB3 level.

order rate expression assuming a water content of 40% and 1 atm pressure. The obtained pseudo-first order rate constants are shown in the bottom panel of Figure 16. It becomes immediately clear that the water-catalyzed reaction is several orders of magnitude slower than the direct unimolecular channels (see Figure 10). Given the big difference, this means that increasing the pressure by an order of magnitude will still not make the bimolecular reaction competitive towards the unimolecular elimination pathway. This leads to the conclusion that the water assisted elimination of water from alcohols in the gas phase is unlikely to make a significant contribution to the biomass pyrolysis chemistry.

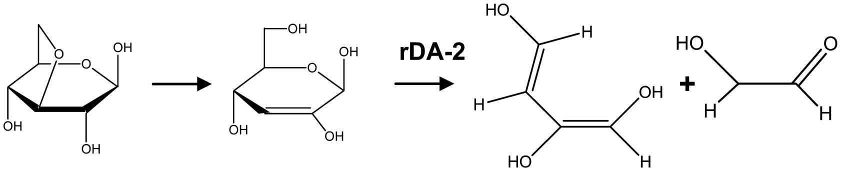
Retro-Diels-Alder Reactions: A Rapid Reaction Path to Small Pyrolysis Products?

A large fraction of biomass material contains moieties with 6-member ring structures, such as the pyranosic rings in cellulose and hemicellulose. If the first decomposition step involves the loss of water or an alcohol, the possibility of forming an unsaturated 6-member ring exists. Those species should be able to undergo a retro-Diels-Alder (rDA) reaction. To test whether this reaction class is fast enough to be important in the rapid gasification process, we studied two such reactions involving derivatives of levoglucosan. The following reaction sequences were considered:

Sequence 1:



Sequence 2:



The initial reaction step in both reaction schemes involves the formation of a double bond and a new hydroxyl group. TST calculations yielded rate constants on the order of 10 s^{-1} at 1073 K and confirmed that this reaction is fast enough at high biomass conversion temperatures to proceed. (During a typical residence time of several seconds in the thermal cracker all of the levoglucosan would undertake the olefin formation step if no other reaction pathways were competitive). Once the unsaturated products are formed they can undergo the retro-Diels-Alder reaction. We are interested in the rate constants for these reactions or generally this reaction class.

Electronic structure calculations at the CBS-QB3 method followed by the TST calculations yielded the following rate constants and heats of reaction for the retro-Diels-Alder reaction (given in simple Arrhenius format):

$$\text{rDA-1:}\quad k(T) = 2.3\cdot10^{13}\,s^{-1}\cdot\exp(-44.6\ \text{kcal/mol}\,/(R\cdot T))\ ;\ \Delta_R H^{298} = 16.3\ \text{kcal/mol}$$

$$\text{rDA-2:}\quad k(T) = 2.1\cdot10^{15}\,s^{-1}\cdot\exp(-55.3\ \text{kcal/mol}\,/(R\cdot T))\ ;\ \Delta_R H^{298} = 31.8\ \text{kcal/mol}$$

Even though both reactions belong to the same reaction class (it is the reverse reaction of a Diels-Alder reaction, a 4π+2π cycloaddition reaction), the obtained kinetic parameters differ significantly. This is not only true for the barriers but also for the pre-exponential factors, which differ by two orders of magnitude. In this concrete example, the reasons are at least partially clear: the reactant of rDA-2 is approximately 7.5 kcal/mol more stable than the anhydro-glucose species of rDA-1, while the products of the first rDA reaction are about 8 kcal/mol more stable than those of r-DA2. This makes rDA-2 more than 15 kcal/mol more endothermic and explains the significantly higher barrier. The clearly different A-factors can be explained in a similar way: Even though the reactants of rDA-1 and rDA-2 look quite similar, the entropy of the reactant in rDA-2 is found to be 5.7 units lower at 298K than the entropy of the reactant in rDA-1. With respect to transition states, the entropy for rDA-2 is 2.7 units higher than that for rDA-1. Thus the large difference in the A-factor, which depends on the entropy differences in transition states and reactants, becomes understandable.

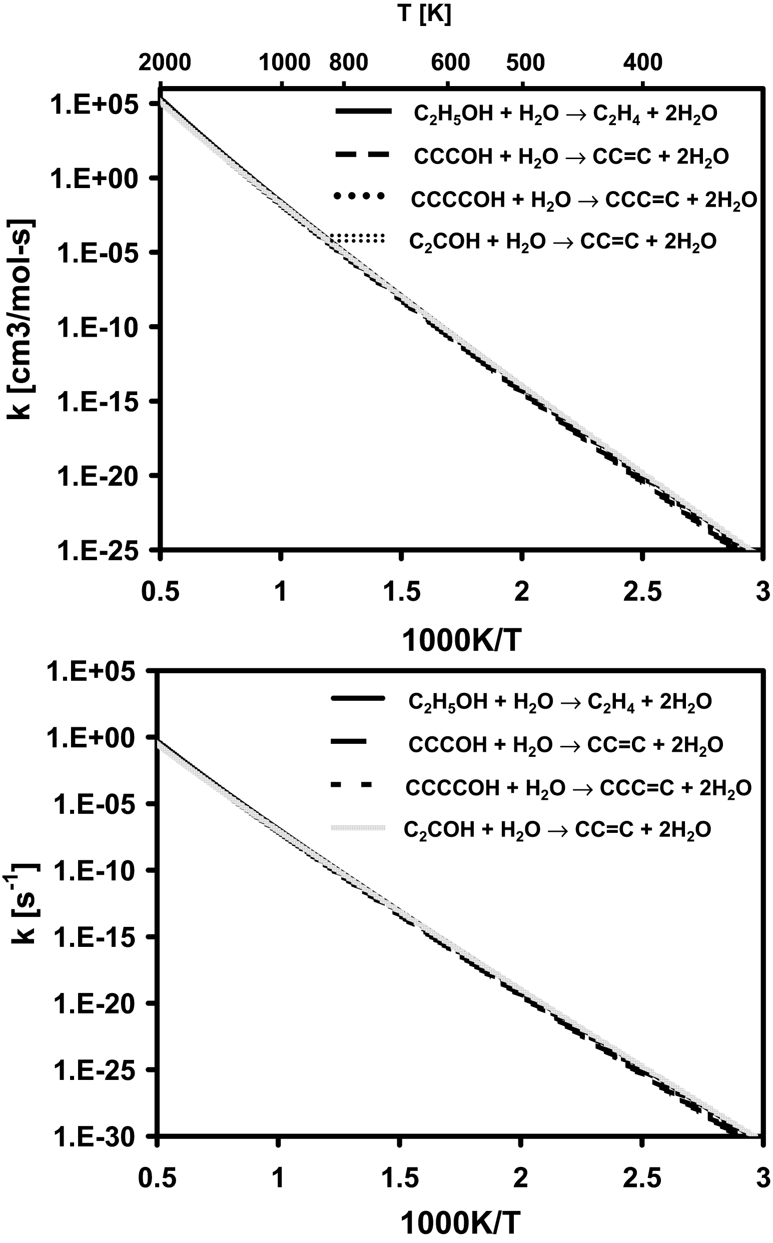
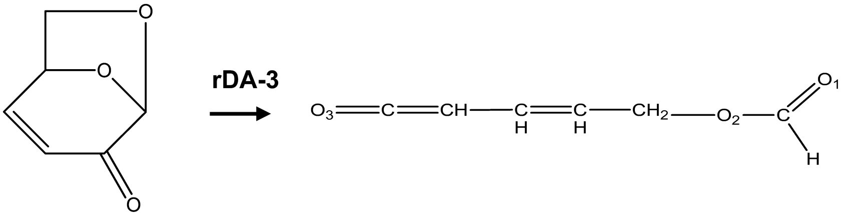


Figure 16. Calculated rate constants for the water assisted elimination of water from alcohols. (top: second order rate constants; bottom: converted to first order assuming 40% steam and P = 1 atm)

In a more general sense cycloaddition reactions, and therefore also the reverse reactions, are known to be very sensitive to electronic effects of substituents on the participating dienes and dienophiles (substituted olefin) (*51*). The rDA-1 reaction produces an unsaturated aldehyde as diene and the dienophile is an enol. The opposite product distribution is found in rDA-2: the diene is an enol and the dienophile product is an aldehyde. Such different product combinations either stabilize or destabilize the transition state of Diels-Alder reactions and reverse Diels-Alder reactions. More details can be found in textbooks on mechanisms in organic chemistry (*51*). In this context, the important point to recognize is that the reactive center is not a narrowly defined moiety but expands over a number of atoms and is profoundly influenced by neighboring electron-donating or electron-abstracting groups. Therefore retro-Diels-Alder reactions are an example for a reaction class that cannot easily be characterized by simple rate rules. Interestingly, although the kinetic parameters of the two reactions differ substantially, both rate constants are similar at 1073 K (800 °C): 16000 vs. 9800 s^{-1}. This shows that retro-Diels-Alder reactions in the gas phase are fast and they should be considered in the biomass mechanism development process.

To confirm our conclusion that retro-Diels-Alder reactions proceed rapidly in the gas phase under biomass gasification conditions, we extended the investigation to levoglucosenone, which is an important product in the cellulose pyrolysis and might serve as starting material for the synthesis of valuable products (*52*). The following retro-Diels-Alder reaction is possible:



Due to the bicyclic structure of levoglucosenone only one product is formed and because of the steric constrains one would expect that the rate constant for this specific retro-Diels-Alder reaction is smaller than those for the two previously discussed reactions. However, the rate constant

$$\text{rDA-3: } k(T) = 1.9 \cdot 10^{14}\, s^{-1} \cdot \exp(-53.0 \text{ kcal/mol} / (R \cdot T)) \text{ ; } \Delta_R H^{298} = 4.4 \text{ kcal/mol}$$

is predicted to be 3000 s^{-1} at 1073 K, which is comparable to the rate constants for the previous two reactions at this temperature. This demonstrates that the retro-Diels-Alder reaction of levoglucosenone is also fast despite the steric interactions. The large E value of the rate expression and small heat of reaction shows that the barrier height in retro-Diels-Alder reactions does not seem to correlate with the heat of reaction.

In Figure 17 we compare the three discussed retro-Diels-Alder rate constants with results for additional (simpler) reactants (cyclohexene, 3,4-dihydro-2H-pyran, and 3,6-dihydro-2H-pyran). These plots confirm that retro-Diels-Alder rate constants do not group together and thus cannot be represented in terms of a simple generic rate estimation rule.

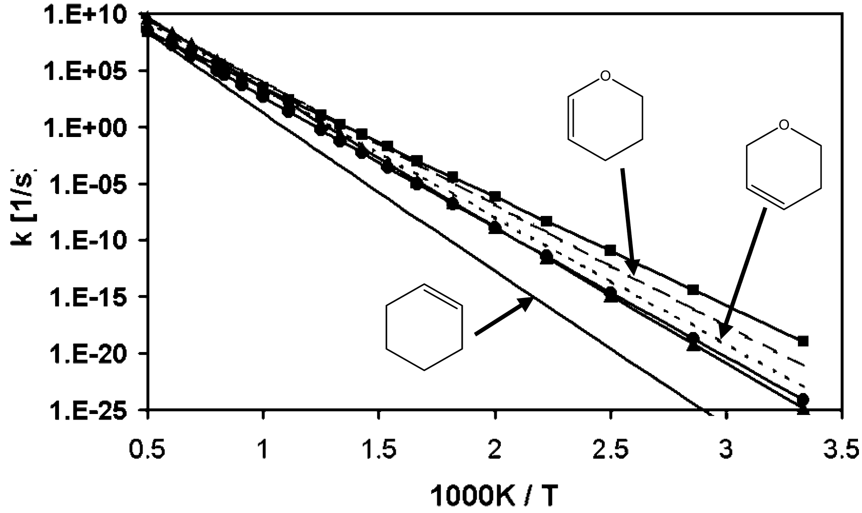


Figure 17. Comparison of retro-Diels-Alder rate constants (rDA) for cyclohexene and related reactants. Filled squares: rDA-1, filled triangles: rDA-2, filled circles: rDA-3, straight line: rDA of cyclohexene, long dashed line: rDA of 3,4-dihydro-2H-pyran, short dashed line: rDA of 3,6-dihydro-2H-pyran.

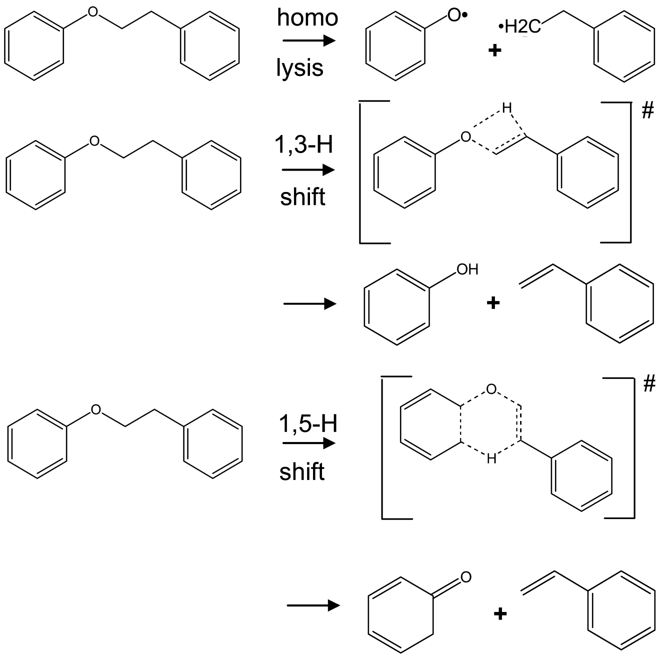


Figure 18. Reaction pathways for the initial decomposition of phenethyl phenyl ether.

In summary, retro-Diels-Alder reactions represent an interesting reaction class that might be important in the gasification process. The rate constants appear large enough to compete with radical chemistry. The three presented examples show comparable reactivities at 1073 K but the kinetic expressions differ drastically. This demonstrates that rate rules of the type used throughout this study do not work for this reaction class. Instead each reaction needs to be investigated individually or more complex rate rules are necessary.

What Is the Initial Decomposition Step of Lignin Model Compounds?

The three reaction classes discussed above are related to the chemistry of cellulose and hemicellulose model compounds. Therefore, we have chosen as the final example a reaction system related to lignin chemistry. Klein et al. (*53*) recommended phenethyl phenyl ether (PPE) as a model compound for the study of lignin decomposition and Britt et al. have extensively studied this molecule and its derivatives (*15*, *54*, *55*). Britt suggests that the decomposition of PPE proceeds mainly via a radical mechanism. The initial radical producing step is believed to involve breakage of the -O-CH2- bond to form the phenoxy and the 2-phenyl ethyl radicals. However, this homolysis reaction is not the only possible initiation reaction. Other reaction pathways that warrant consideration are a 1,3-H shift reaction that leads to phenol and styrene and a 1,5-H shift (retroene reactions) which yields 2,4-cyclohexadienone and styrene. The three reactions are illustrated in Figure 18. A theoretical investigation of this reaction at the CBS-QB3 level of theory is not possible, because the reactant and transition states species are too large (15 heavy atoms exceeds the CPU time and storage capacities available to us). This leaves two options: either to use a lower level method to study this reaction or to retain the well-tested CBS-QB3 method and replace PPE with a smaller molecule of assumed comparable reactivity. Possible choices for smaller model compounds are vinyl vinylethyl ether (VVE), ethoxybenzene, phenyl vinylethyl ether (PVE) and others. In VVE the two phenyls are replaced by vinyl moieties. This is motivated by the anticipation that the reactivity of a vinyl group resembles that of a phenyl group. In the same spirit, the substitution of only one phenyl group (the not actively reacting one) with vinyl leads to PVE. Finally, if the non-reacting phenyl group is replaced with a hydrogen atom we obtain the possible model compound ethoxybenzene. We have tested both options and used the CBS-4M level of theory on all four compounds (including PPE) and compare the potential energy surfaces (PES) of the three smaller reactants with those obtained at the CBS-QB3 level. Qualitatively all PES are similar and thus we show in Figure 19 only those for VVE and PVE. The PESs reveal that the 1,5-H shift reaction is the pathway with the lowest barrier. In the case of VVE, the assumed barrierless homolysis (i.e., no activation energy other than the energy needed to overcome the reaction endothermicity) is energetically slightly lower than the barrier for the 1,3-H shift while the opposite is true for PVE. This difference makes sense, because in VVE the 1,3-H shift reaction produces vinylalcohol, which is the high-energy tautomer of acetaldehyde. The corresponding reaction for PVE yields the most stable tautomer phenol. Hence one would expect that the transition states leading to phenol experiences more stabilization than the corresponding transition state

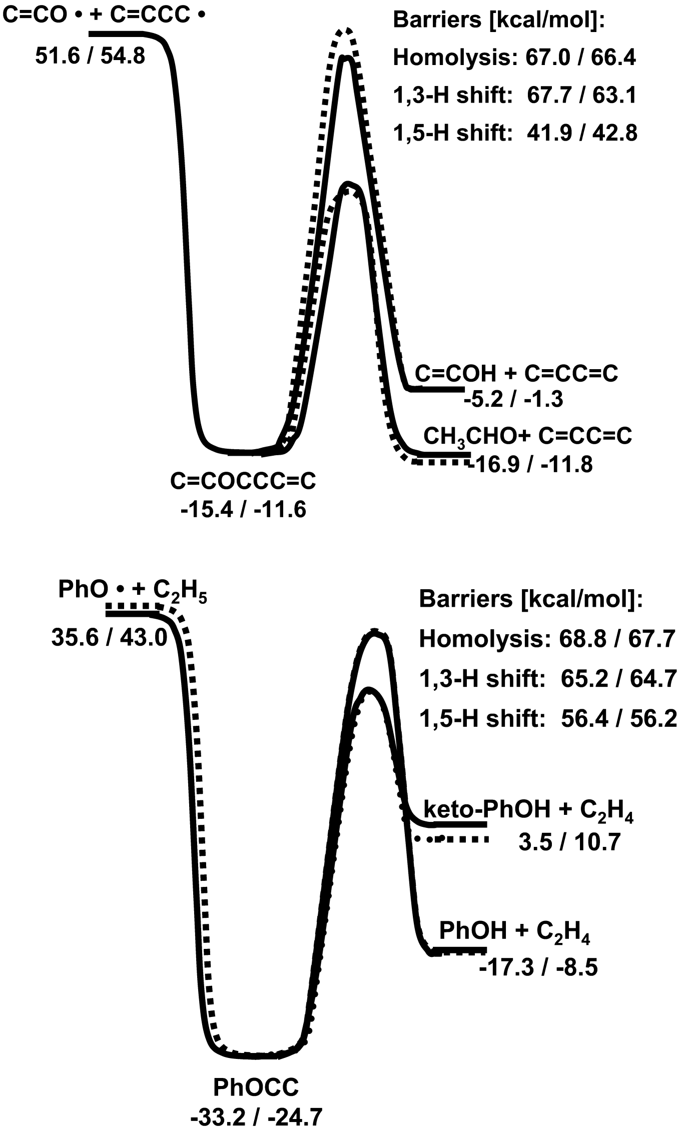


Figure 19. The PES for vinyl vinylethyl ether (top) and phenyl vinylethyl ether (bottom): comparison of CBS-4M (dotted line) and CBS-QB3 (solid line) results. The transition states and product channels are shown relative to the reactant energies. Absolute uncorrected heats of formation or reaction and barrier heights are provided in kcal/mol units (CBS-4M / CBS-QB3) .

for VVE. Figure 19 also reveals that while the relative energies (barriers at 298K and heat of reactions) found with the CBS-M and CBS-QB3 are very close, this is not the case for absolute heat of formation values. The deviations are particularly large in the PVE system, which indicates that the error increases with the size of the molecule. This also serves as a reminder that the often quoted expected accuracy of a model chemistry, which is derived from test sets that include mainly small molecules, is not necessarily realistic for large species without further corrections.

A plot of the barriers for all four reactions is provided in Figure 20. Obviously the barriers show clear structure dependence, and the assumption that the phenethyl phenyl ether can be investigated using smaller model compounds appears not to be valid. On the other hand, except for one barrier we see good agreement between the barriers obtained with CBS-4M and CBS-QB3. This suggests that for these reaction types the lower-level CBS-4M method may be used with some confidence.

The barrier heights alone are not sufficient to decide which reaction channel might be the most important one. Homolysis reactions have generally large A-factors, while both hydrogen shift reactions proceed via tight transition states, which lead to probably several orders of magnitude smaller pre-exponential factors. In Table 2 we present calculated rate constants for the three model compounds vinyl vinylethyl ether, ethoxybenzene, phenyl vinylethyl ether. Since the 1,3-H shift reactions tie up 2 rotors less than the 1,5-H shift reaction, we consistently see a more than one order of magnitude higher pre-exponential factors for the 1,3-H migration reactions even though its 4-member cyclic transition state structure should contain additional strain.

Even though the A-factor for the 1,5-H migration reaction is the smallest of the three pathways, this channel will probably dominate at low temperatures due to the large differences in the barriers. This does not mean that this reaction is dominant in bulk systems, since once radicals are formed their reactions are likely faster than the unimolecular reactions discussed here.

Finally, one might be tempted to conclude from the rate expressions in Table 2 that the choice of calculation method has a large impact on A-factors. Solely based on these parameters, agreement between both calculation methods is poor. However, one needs to take into account that these Arrhenius parameters are derived from fits to the simple Arrhenius equation and part of the reason for the high variability of the A-factors is that the rate constants do not follow Arrhenius behavior. Therefore, at this stage, one should only look at the rate parameters in the context described above, that is to see them as a reminder of the correlation between the number of lost internal rotors and the magnitude of the A-factor.

Discussion

The starting point of this chapter was the realization that the development of detailed kinetic biomass models presents a challenging task, even if only gas phase reactions are considered, but that there is also reason to believe that this challenge can be met in the near future. We presented four examples to illustrate that in particular the enormous advances in computational chemistry and the possibility to generate generalized rate constants of reaction classes provide powerful tools to meet this challenge. The first example, H abstraction reactions from alcohols not only demonstrated how well rate estimation rules work, but it also showed how seamlessly these rate rules for oxygenates converge to those of pure hydrocarbons. More specifically, the influence of hydroxyl groups on

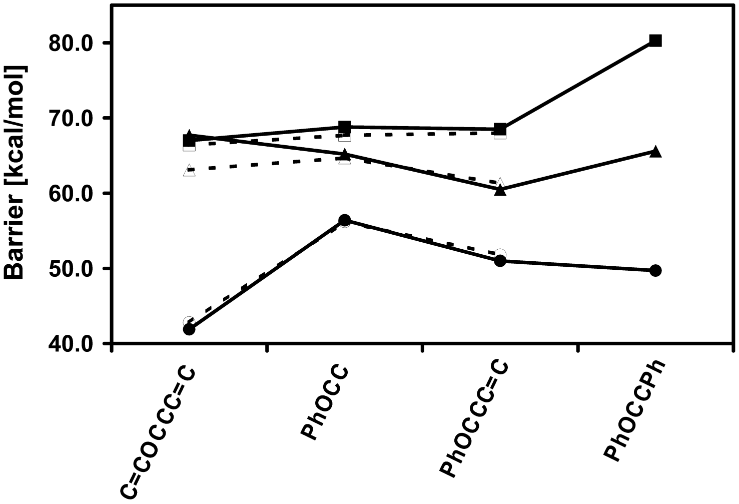


Figure 20. Variations of the barriers for the homolysis, 1,3-H shift and 1,5-H shift reactions with model compound. Solid lines with filled symbols: CBS-4M data; dotted lines with open symbols: CBS-QB3; squares: homolysis, triangles: 1,3-H shift, circles: 1,5-H shift reaction

the reactivity of neighboring moieties vanishes at the β-position. H-abstraction reactions are especially suitable for rate rules since the reactive centre is generally well localized and steric effects are often minor. Therefore it is important to verify that the rate rule concept is also applicable to other reaction types. We selected the unimolecular elimination of water from alcohols to address this issue. The result: As long as the rate rule was applied to the type of reactants for which it was developed, it performed reasonably well. The accuracy is certainly sufficient for its use in mechanism development since a rate or sensitivity analysis can later identify particularly important reactions, which then can be studied in greater detail using more sophisticated and time-consuming methods. However, when the water elimination rate rule derived for simple alcohols was applied to reactants with more than one functional group or cyclic species, some of the observed deviations are large. But none of these “failures” were surprises. In fact, the cases for which the rate rules do not work well could easily have been predicted by chemical intuition. For example, water elimination from the α,β-position of an aldehyde group leads to a product with conjugated double bonds. This resonance (mesomeric) effect is not considered in rate rules for saturated reactants. A new rate rule capturing those effects is required and the failure of the original rule is not necessarily an indication that rate rules are not reliable. In the same spirit, it is well known that steric constrains alter the reactivity. Hence it is not surprising to find that constrained moledules do not follow the estimates made with rate rules developed for unstrained reactants. Again, since these “failures” are predictable or even expected, the rate rule concept remains viable.

Table 2. Comparison of Arrhenius rate expressions for the three mentioned PPE model compounds. Normal font: CBS-4M, bold font: CBS-QB3. All rate constants are in s^{-1} units. Rate constants for the homolysis reaction cannot be calculated with the methods used in this study

	C=COCCC=C		*PhOCC*		*PhOCCC=C*	
Reaction Type	*A*	*E*	*A*	*E*	*A*	*E*
Homolysis	N/A		N/A		N/A	
1,3-H shift	1.2E+14	67.1	3.6E+13	64.1	9.1E+12	59.9
	1.2E+13	**63.4**	**2.5E+14**	**64.4**	**1.1E+14**	**61.6**
1,5-H shift	2.3E+11	41.7	4.0E+12	56.6	8.7E+11	51.2
	6.7E+11	**43.6**	**9.2E+12**	56.9	**4.1E+12**	52.9

While the establishment of rate estimation rule is an important tool that allows for a quick assignment of thermodynamically consistent rate constants, it cannot be used for all reaction systems. Retro-Diels-Alder reactions represent a reaction class for which simple rate rules fail, because these reactions are very sensitive to the electronic states (HOMO, LUMO) involved in this process. Substituents bound to the reactive center can drastically change the electronic properties and change the reactivity. Interestingly, the rate constants of the three reactions studies are still remarkably similar at 1073 K (within a factor of 5), but this is probably a coincidence.

The basic idea behind the use of rate rules is that we can use highly accurate methods to study a given reaction class for small reactants and then transfer this knowledge to larger systems. The approach of the first two examples reflected this: we first generalized rate rules and then tested those on more complicated reactants. The final reaction system presented in this work takes a slightly different approach. Instead of starting small and going up, we begin with the molecule of interest (phenethyl phenyl ether, PPE) that is too large to be treated at the preferred level of theory. The question then arises of how can one simplify the problem without sacrificing accuracy. We considered two non-exclusive options: (a) to replace phenyl groups with vinyl groups (or even H atoms) and assume that the reactivities of the smaller reactants are similar and transferable, and (b) to use a lower level theory, validate its performance compared to a benchmark method to ensure its reliability for model species of interests and then apply it to the target molecule phenethyl phenyl ether, hoping that the quality of the calculation does not change. The presented PES calculations for four phenyl and vinyl ether showed that a phenyl group cannot simply be replaced by a vinyl group without changing the reactivity. A part of the results can be rationalized by chemical intuition but, for example, it was a surprise that the replacement of the phenyl in the phenethyl moiety by vinyl led to a noticeable change in some barrier heights when compared to PPE. The main conclusion therefore is that rate expressions obtained for ‘simplified’ model compounds should only be transferred to larger system with caution and if possible double-checked by lower-level theories.

Rate estimation rules have been used for a long time to develop reaction mechanisms and recent developments have been reviewed by Sumathy et al. (*56*). However, the motivation and way rate rules are used has changed significantly in the last one to two decades. In the past such rules were developed based on sparse sets of experimental data, which were often only available for a narrow range of conditions. This made it in many cases difficult to formulate very accurate rules. The rate rules were applied out of necessity because no other resource of kinetic data was available for many reaction classes. This has changed when highly accurate electronic structure calculations became feasible. We now have the luxury to base rate estimation rules of large sets of rate constants, obtained from systematic studies over an extensive temperature range. The highest levels of calculations yield rate constants that agree to within a factor of two or better with the best experimental data. One of these well performing (*19*, *21*) model chemistries is the CBS-QB3 method, which is used in this work. Therefore we are now in a position to generate reliable rate rules for clearly defined reaction classes (*20*, *37–39*). Further, these rate rules are not necessarily used because it is the only source of data (it is just a question of time to run a calculation at a suitable level of theory), but it is now a resource of choice, because their use has several advantages over other resources. (1) Rate constants from rate rules are easy to calculate and amendable to automated mechanism generating codes. (2) Mechanisms based on these rules are easy to maintain and thermodynamically consistent. (3) Rate rules can serve as a quick test to ensure that rate constants from other sources are reasonable. (4) Rate rules also help to identify anomalies and therefore improve our understanding of the factors that influence reactivity.

The increasing role of electronic structure calculations in the development of reaction mechanisms demands prudence and constant reevaluation of the quality of the predictions. We presented two example systems to demonstrate this point. The popular B3LYP method seems to be very suitable in predicting rate constants for the elimination of water from alcohols. However, theoretical methods are always subject to the danger of obtaining ”the right results for the wrong reasons”. We showed that this is the case for the B3LYP results as well. Fortunate error cancellations lead to rate constants predictions that are very close to the CBS-QB3 benchmark data, but the rate constants for the reverse reaction will be orders of magnitude off, because the heats of formation contain large errors. On the other hand, the phenyl ether example provided evidence that less expensive CBS-4M calculations yield PES data comparable to the CBS-QB3 method and that the use of this level of theory is more reliable than substituting the PPE molecule with a smaller model compound. The general conclusion is that electronic structure methods need to be chosen and critically evaluated for each system to which they are applied.

Summary

We have presented selected results for four reaction systems to discuss the advantages and limitations of rate estimation rules. The results show that such rules can play an important role in the development process of reaction mechanisms as long as they are applied carefully to reactions for which the rules were developed. The use of rate rules makes the mechanism development process faster (e. g., via automatic mechanism generation approaches), it ensures consistency and frees up time for the kineticist to study those special reactions that cannot be described by general rules. Some of the presented results also point out that care is needed when choosing a calculation method. The B3LYP method was shown to be unreliable while the CBS-4M method seems to work well for the PES calculations of phenyl ether compounds.

The long-term motivation of this work is to better understand thermochemical biomass conversion. We demonstrated that one promising approach is to generate appropriate rate rules. But the results also led to other observations, such as that H abstraction reactions from alcohols by radicals are fast and they yield products that will increase the gas phase reactivity. We also found that retro-Diels-Alder reactions are fast enough to be potentially able to contribute to the chemistry in the gasifier. In contrast, the molecular elimination of water from hydroxyl group containing species is likely too slow to play a significant role even at high conversion temperatures. This conclusion is also true for some multifunctional species even though a neighboring aldehyde group would increase the rate constant by several orders of magnitude if it survived the volatilization step.

While most of the results are relevant for cellulose and hemicellulose chemistry, the final example considers proposed lignin model compounds. The two main results of this part are: (1) It appears to be more advantageous to rely on a lower level of theory (CBS-4M) than to replace a phenyl groups by a smaller substituent. (2) The barrier for the 1,5-H shift reaction is significantly lower that the homolysis and the 1,3-H shift reaction. This indicates that the molecular channel might play a role in the initial decomposition process at lower temperatures even though radical chemistry will dominate once a radical pool is formed.

Acknowledgments

The authers thank Prof. G. Barney Ellison from the University of Colorado for helpful and stimulating discussions. This work was supported by NREL via subcontract ZFT-8-88565-01 (program manager Dr. Mark R. Nimlos). We further acknowledge access to the NREL high performance computer LESTER.

References

1. Ariya, P. A.; Sander, R.; Crutzen, P. J. *J. Geophys. Res.* **2000**, *105*, 17721.
2. Poisson, N.; Kanakidou, M.; Crutzen, P. J. *J. Atmos. Chem.* **2000**, *36*, 157.
3. Westbrook, C. K.; Pitz, W. J.; Herbinet, O.; Curran, H. J.; Silke, E. J. *Combust. Flame* **2009**, *156*, 181–199.
4. Blanquart, G.; Pepiot-Desjardins, P.; Pitsch, H. *Combust. Flame* **2009**, *156*, 588–607.
5. Chatterjee, D.; Deutschmann, O.; Warnatz, J. *Faraday Discuss.* **2001**, *119*, 371.

6. Evans, R. J.; Milne, T. A. *Energy Fuels* **1987**, *1*, 123.
7. Morf, P.; Hasler, P.; Nussbaumer, T. *Fuel* **2002**, *81*, 843.
8. Jablonski, W.; Gaston, K. R.; Nimlos, M. R.; Carpenter, D. L.; Feik, C. J.; Phillips, S. D. *Ind. Eng. Chem. Res.* **2009**, ASAP.
9. Vasiliou, A.; Nimlos, M. R.; Daily, J. W.; Ellison, G. B. *J. Phys. Chem. A* **2009**, *113*, 8540–8547.
10. Organ, P. P.; Mackie, J. C. *J. Chem. Soc., Faraday Trans.* **1991**, *87*, 815.
11. Horn, C.; Frank, P. *High Temperature Pyrolysis of Phenol*; Vol. 2000.
12. Arends, I. W. C. E.; Louw, R.; Mulder, P. *J. Phys. Chem.* **1993**, *97*, 7914.
13. Sendt, K.; Bacskay, G. B.; Mackie, J. C. *J. Phys. Chem. A* **2000**, *104*, 1861.
14. Pecullan, M.; Brezinsky, K.; Glassman, I. *J. Phys. Chem. A* **1997**, *101*, 3305.
15. Beste, A.; Buchanan, A. C.; Britt, P. F.; Hathorn, B. C.; Harrison, R. J. *J. Phys. Chem. A* **2007**, *111*, 12118.
16. Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822.
17. Montgomery, J. A., Jr.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **2000**, *112*, 6532.
18. Petersson, G. A.; Malick, D. K.; Wilson, W. G.; Ochterski, J. W.; Montgomery, J. A., Jr.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 10570.
19. Vandeputte, A. G.; Sabbe, M. K.; Reyniers, M.-F.; Van Speybroeck, V.; Waroquier, M.; Marin, G. B. *J. Phys. Chem. A* **2007**, *111*, 11771.
20. Carstensen, H.-H.; Dean, A. M. *J. Phys. Chem. A* **2009**, *113*, 367–380.
21. Carstensen, H.-H.; Naik, C. V.; Dean, A. M. *J. Phys. Chem. A* **2005**, *109*, 2264.
22. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision E.1; Gaussian, Inc: Pittsburgh, PA, 2003.
23. Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
24. Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
25. Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
26. Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A., Jr. *J. Chem. Phys.* **1996**, *104*, 2598.
27. Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063.

28. Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502.
29. Pitzer, K. S.; Gwinn, W. D. *J. Chem. Phys.* **1942**, *10*, 428.
30. Kilpatrick, J. E.; Pitzer, K. S. *J. Chem. Phys.* **1949**, *17*, 1064.
31. East, A. L. L.; Radom, L. *J. Chem. Phys.* **1997**, *106*, 6655.
32. Evans, M. G.; Polanyi, M. *Trans. Faraday Soc.* **1935**, *31*, 875.
33. Eyring, H. *J. Chem. Phys.* **1935**, *3*, 107.
34. Eckart, C. *Phys. Rev.* **1930**, *35*, 1303.
35. Wigner, E. *Z. Phys. Chem.* **1932**, *19B*, 203.
36. Skodje, R. T.; Truhlar, D. G.; Garrett, B. C. *J. Phys. Chem.* **1981**, *85*, 3019.
37. Zhang, S.; Truong, T. N. *J. Phys. Chem. A* **2003**, *107*, 1138.
38. Sumathi, R.; Carstensen, H.-H.; Green, W. H., Jr. *J. Phys. Chem. A* **2001**, *105*, 6910.
39. Sumathi, R.; Carstensen, H.-H.; Green, W. H., Jr. *J. Phys. Chem. A* **2001**, *105*, 8969.
40. Manion, J. A.; Huie, R. E.; Levin, R. D.; Burgess, D. R., Jr.; Orkin, V. L.; Tsang, W.; McGivern, W. S.; Hudgens, J. W.; Knyazev, V. D.; Atkinson, D. B.; Chai, E.; Tereza, A. M.; Lin, C.-Y.; Allison, T. C.; Mallard, W. G.; Westley, F.; Herron, J. T.; Hampson, R. F.; Frizzell, D. H. NIST Chemical Kinetics Database, NIST Standard Reference Database 17, Version 7.0 (Web Version); National Institute of Standards and Technology: Gaithersburg, Maryland, 2009; Release 1.4.3, Data version 2008.12.
41. Curran, H. J. *Int. J. Chem. Kinet.* **2006**, *38*, 250.
42. Caralp, F.; Devolder, P.; Fittschen, C.; Gomez, N.; Hippler, H.; Mereau, R.; Rayez, M. T.; Striebel, F.; Viskolcz, B. *Phys. Chem. Chem. Phys.* **1999**, *1*, 2935.
43. Chase, M. W., Jr. *J. Phys. Chem. Ref. Data* **1998**, *9*, 1.
44. Wiberg, K. B.; Crocker, L. S.; Morgan, K. M. *J. Am. Chem. Soc.* **1991**, *113*, 3447.
45. Berkowitz, J.; Ellison, G. B.; Gutman, D. *J. Phys. Chem.* **1994**, *98*, 2744.
46. Carstensen, H.-H.; Dean, A. M.; Deutschmann, O. *Proc. Combust. Inst.* **2007**, *31*, 149.
47. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A* **2007**, *111*, 10439.
48. Pouwels, A. D.; Eukel, G. B.; Boon, J. J. *J. Anal. Appl. Pyrolysis* **1989**, *14*, 237.
49. Piskorz, J.; Radlein, D.; Scott, D. S. *J. Anal. Appl. Pyrolysis* **1986**, *9*, 121.
50. Antal, M. J., Jr.; Varhegyi, G. *Ind. Eng. Chem. Res.* **1995**, *34*, 703.
51. Sykes, P. *A Guidebook to Mechanism in Organic Chemistry*; Longman Singapore Publishers (Pte) Ltd., 1985.
52. Shafizadeh, F.; Furneaux, R. H.; Stevenson, T. T. *Carbohydr. Res.* **1979**, *71*, 169.
53. Klein, M. T.; Virk, P. S. *Ind. Eng. Chem. Fundam.* **1983**, *22*, 35.
54. Britt, P. F.; Kidder, M. K.; Buchanan, A. C., III. *Energy Fuels* **2007**, *21*, 3102.
55. Beste, A.; Buchanan, A. C., III.; Harrison, R. J. *J. Phys. Chem. A* **2008**, *112*, 4982.
56. Sumathi, R.; Green, W. H., Jr. *Theor. Chem. Acc.* **2002**, *108*, 187.

Chapter 11

Multiscale/Multiphysics Modeling of Biomass Thermochemical Processes

Sreekanth Pannala,[1,*] Srdjan Simunovic,[2] and George Frantziskonis[2]

[1]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831
[2]Civil Engineering and Engineering Mechanics, University of Arizona, Tucson, AZ 85721-0072
[*]pannalas@ornl.gov

Computational problems in simulating biomass thermochemical processes involve coupled processes that span several orders of magnitude in space and time. Computational difficulties arise from a multitude of governing equations, each typically applicable over a narrow range of spatiotemporal scales, thus making it necessary to represent the processes as the result of the interaction of multiple physics modules, termed here as multiscale/ multiphysics (MSMP) coupling. Predictive simulations for such processes require algorithms that efficiently integrate the underlying MSMP methods across scales to achieve prescribed accuracy and control computational cost. In addition, MSMP algorithms must scale to one hundred thousand processors or more to effectively harness new computational resources and accelerate scientific advances. In this chapter, we discuss state-of-the-art modeling of macro-scale phenomena in a biomass pyrolysis reactor along with details of shortcomings and prospects in improving predictability. We also introduce the various multiphysics modules needed to model thermochemical conversion at lower spatiotemporal scales. Furthermore, we illustrate the need for MSMP coupling for thermochemical processes in biomass and provide an overview of the wavelet-based coupling techniques we have developed recently. In particular, we provide details about the compound wavelet matrix (CWM) and the dynamic CWM (dCWM) methods and show that they are highly efficient in transferring information among multiphysics models across multiple temporal and spatial scales. The algorithmic gain is in addition to the parallel spatial scalability from traditional domain decomposition methods. The CWM algorithms are serial in time and limited by the smallest-system time-scales. To relax this algorithmic constraint, we have recently coupled time parallel (TP) algorithms to CWM, thus yielding a novel approach termed tpCWM. We present preliminary results from the tpCWM technique, indicating that we can accelerate time-to-solution by two to three orders of magnitude even on 20-processors. These improvements can potentially constitute a new paradigm for MSMP simulations. If such improvements in simulation capability can be generalized, the tpCWM approach can lead the way to predictive simulations of biomass thermochemical processes.

Motivation and the Scales Encountered in Biomass Thermochemical Conversion

One of the most pressing problems facing the world today is finding sustainable, cost-effective, and ecologically friendly energy sources combined with efficient utilization of energy. In this chapter, we describe some mathematical and simulation tools that can be used to accelerate energy solutions not only to keep up with the current demand, but also to meet the increasing demand, which is expected to double by 2050 and triple by the end of the century. Rigorous multiscale coupling tools can significantly improve the predictability of computational simulations of the new biomass-based energy technologies. The formalism proposed in this chapter can provide a framework for integrating laboratory-scale experiments, various science and engineering disciplines, and industrial-scale solutions.

Modeling the thermochemical processes in biomass conversion typically employs a variety of approximate representations – physics-based as well as mathematics-based – spanning from atomistic to mean-field (quantum-level, classical atomistic, statistical mechanics, continuum, reduced-order, and mean-field models, etc.). Figure 1 shows various representative levels of the system for heterogeneous chemical reactors for biomass. In this figure, only representations from kinetic Monte Carlo (KMC) and beyond are shown with the typical applicability range of time and length scales. At the atomistic level, KMC (*1*) resolves the chemical reactions, surface diffusion, adsorption, and desorption with rate parameters determined by experiments or atomistic-level simulations. The next higher scale, commonly referred to as the intermediate, mesoscopic scale of approximation, involves modeling the fluid flow around the particles by the Boltzmann Equations (*2–8*), while considering the mean-field equations for reactions on the surface. One possible option for solving the fluid flow is to use the Lattice Boltzmann Method (LBM - e.g. see the review by Chen et al. (*9*)). In LBM, the Boltzmann equations are solved on lattices using pre-defined velocity

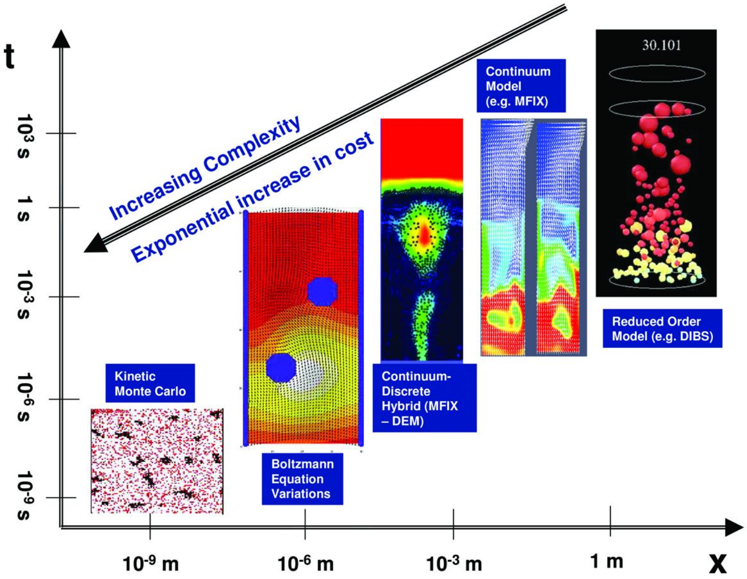

t
10^3 s
1 s
10^{-3} s
10^{-6} s
10^{-9} s
Increasing Complexity
Exponential increase in cost
Kinetic Monte Carlo
Boltzmann Equation Variations
Continuum-Discrete Hybrid (MFIX – DEM)
Continuum Model (e.g. MFIX)
30.101
Reduced Order Model (e.g. DIBS)
10^{-9} m
10^{-6} m
10^{-3} m
1 m
X

Figure 1. Schematic of the various representation levels for a heterogeneous chemical reactor.

directions; and in the incompressible limit, the solutions correspond to that of Navier-Stokes equations. The next level of approximation involves modeling the main phase using continuum Navier-Stokes equations where particles are modeled in a discrete fashion (*10–14*) through various closures for heat and mass transfer to the particles as well as chemical reactions. Subsequent levels of approximation introduce the assumption of interpenetrating continuum for different phases (*15*). At the overall systems level, we can construct reduced-order models of the underlying dynamics based on a physical interpretation of the system (*16–18*) or by using techniques like proper orthogonal decomposition (*19*).

Multiphysics Components for Biomass Thermal Conversion

In this section, we describe the multiphysics models for the biomass thermochemical processes from the microscopic to the continuum scales. We start with a multiphase computational fluid dynamics (CFD) model for the continuum representation and provide a computational example to demonstrate the complex interactions between the hydrodynamics, heat and mass transfer, and chemistry. Then we explore multiphysics methods, namely the Lattice Boltzmann Method, and Kinetic Monte Carlo that can improve the physical fidelity of the various subcomponents used in the CFD models. After that, we describe a wavelet-based, multiscale coupling method that can be applied to various multiphysics components to develop fully integrated and predictable models for biomass thermochemical processes. We conclude with a summary of the features and offer some remarks for future work.

Continuum Level Computational Fluid Dynamics for Biomass Thermochemical Processes

Multiphase reacting flows are ubiquitous in the thermochemical conversion of biomass and pervade many other applications across the entire energy cycle including:

- Fuel production and processing: catalytic crackers, H2 production, S removal, coal gasification, clean-up (SOx, NOx, Hg, CO2), biomass (cellulosic) pyrolysis and gasification, nuclear fuel production
- Energy production: fuel cells, coal and biomass combustion, nuclear reactors and separation, silicon production and coating for photovoltaic applications, novel combustion technologies like oxycombustion and chemical looping combustion
- Energy Utilization and Efficiency: polymerization reactors, catalytic reactors, multiphase flow reactors used in most energy intensive industrial processes.

To build predictive simulation capabilities for these flows, it is important to accurately model the nonlinear, tightly-coupled interactions between the various phases.

Despite the wide use of multiphase flow reactors, most of their development and design have been primarily based on experiments because of the complex, multiscale nature of the processes that control heat, mass, and momentum transport, and the flow interactions with chemical reactions. Our current understanding of these multiscale/multiphysics processes is limited, and direct measurements are difficult because of the dense and erosive flow environment. Even when diagnostic tools are available, they are often intrusive, so that they alter the natural dynamics of the devices. Predictive computational tools can fill the gap between the available experiments and the actual dynamics of the multiphase flow reactors. Two primary approaches exist for highly spatially resolved macroscale time-dependent simulations of multiphase problems (*13*):

a) Discrete Element Method (DEM) (also referred to as Eulerian-Lagrangian treatment)
b) Two-fluid or Multi-fluid Model (also referred to as Eulerian-Eulerian Treatment, Continuum model, etc.).

In the DEM, the gas-phase is modeled using single-phase equations (Navier-Stokes equations for momentum, energy, and species conservation equations), and the discrete phase is modeled as a collection of particles (*12*, *20*, *21*). These particles represent either an individual particle or a parcel of particles. The particle trajectories are updated using Newton's equations of motion accounting for in-particle heat/mass transfer to the surrounding fluid and particles along with explicit accounting of momentum gain and loss through collisions. Typical reactors of interest can have billions of particles, making this method computationally prohibitive.

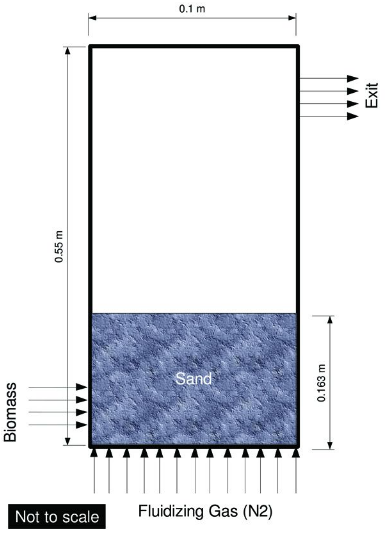


Figure 2. Schematic of fluidized bed reactor for biomass pyrolysis.

The multi-fluid approach is based on the assumption that various phases can be represented as an inter-penetrating continuum and properties corresponding to any particular phase are accounted through an averaged (spatial, temporal, or ensemble) fraction occupied by that phase. Under this assumption, we can use extensions of the single-phase formulation to solve for the multiphase system (*22–28*). The biggest challenge in modeling these systems is constructing closure relations for the interfacial processes such as drag, heat and mass transfer, the granular stresses, and the surface and bulk reaction rates. In this section, we use standard correlations available from experiments to model a biomass pyrolysis process. In later sections, we outline a method for obtaining the closure relationships through more detailed simulations and rigorous upscaling. The results reported here are obtained using the multiphase reacting flow software MFIX (Multiphase Flow with Interphase eXchanges, https://mfix.netl.doe.gov/) developed primarily at the National Energy Technology Laboratory (NETL) with partners at Oak Ridge National Laboratory (ORNL) and other institutions. For brevity, we do not elaborate on the formulation as it is reasonably standard and the interested reader is referred elsewhere (*29*, *30*) for details. In summary, the continuity equation and momentum/ energy/species conservation equations are solved for all phases using appropriate closures for granular stresses, interfacial drag, and heat/mass transfer coupled with the chemical reactions. The reactor simulations provide transient spatially varying (1D, 2D, or 3D) field data of pressure, phase volume-fractions, velocity, temperature, granular energy/temperature, reaction rates, species mass fractions, and any other derived quantities.

Demonstration Problem

We simulate a biomass pyrolysis reactor similar to that of Lathouwers and Bellan (*31*) to demonstrate the state-of-the-art in macroscale continuum simulations of biomass thermochemical processes and identify the areas that can improve the model predictability. The appeal of this problem is that extensive experiments have been done with the biomass used for this pyrolysis reactor, and all the chemical kinetics and thermodynamic properties needed for simulating this system are available. The geometry (2D Cartesian) and flow conditions are similar as in the cited paper. Figure 2 shows the schematic for this type of reactor configuration. The original problem is simplified to have one single inlet for the biomass instead of two, and the inlet/outlets dimensions are matched to the grid resolution used in this simulation. In this setup, the high temperature fluidizing gas enters the sand bed and pyrolyzes the biomass feedstock introduced through the side inlet. Inert fluidization gas and pyrolysis products exit the domain at the outlet. The configuration and flow parameters are given in Table 1.

The chemical reaction mechanism, reaction rates, and transport/physical properties used in these simulations are the same as those of Lathouwers and Bellan (*31*). The biomass in the reactor is decomposed into cellulose (*c*), hemicellulose (*h*), and lignin (*l*). The chemical reactions for this system are:

$$virgin\ (s) \xrightarrow{K_1} active\ (s)$$
$$active\ (s) \xrightarrow{K_2} tar\ (g) \xrightarrow{K_4} gas\ (g)$$
$$active\ (s) \xrightarrow{K_3} X\ char\ (s) + (1 - X)\ gas\ (g)$$

The rate constants and activation energy for this biomass pyrolysis kinetics scheme is provided in Table 2.

Spatial and Temporal Distributions of the Various Field Data

The macroscale simulations provide detailed field data in space and time. In Figure 3, the gas volume fraction is plotted along with gas velocity vectors at two different instants (~1s and ~3s) in a 5-second simulation. The qualitative features include:

- Fluidization gas levitates particles and biomass.
- The reactor operates in the bubbling bed regime.
- The gas undergoes local acceleration and deceleration depending on the flow of the solids.
- The flow of solids and gas is transient and highly dynamic.
- The reactor geometry causes recirculation near the left-hand top corner of the reactor (see the velocity vectors).
- The fluidizing and product gas leave the domain through the exit.

Table 1. Important configuration and flow parameters. Geometric details are given in Figure 1

Model Parameter	Value	Units
Sand particle diameter	500	μm
Sand particle density	2200	kg m^{-3}
Biomass particle diameter	500	μm
Biomass particle density	500	kg m^{-3}
Coefficient of restitution	0.8	
Angle of internal friction	30	
Resolution – dx	0.005	m
Resolution – dy	0.01	m
Fluidizing Gas – N2 @ 700K and 0.5 m/s		
Biomass (Cellulose – 0.36, Hemicellulose – 0.47 and Lignin – 0.17) is injected @ 5 Kg/s, 400K		
For all other properties see Lathouwers and Bellan, 2001		

Table 2. Reaction rates for the various biomass pyrolysis chemical reactions

Reaction	*A (1/s)*	*E (J/kmol)*
K_1^c	2.8 x 10^{19}	242.4 x 10^6
K_2^c	3.28 x 10^{14}	196.5 x 10^6
K_3^c	1.3 x 10^{10}	150.5 x 10^6
K_1^h	2.1 x 10^{16}	186.7 x 10^6
K_2^h	8.75 x 10^{15}	202.4 x 10^6
K_3^h	2.6 x 10^{11}	145.7 x 10^6
K_1^l	9.6 x 10^8	107.6 x 10^6
K_2^l	1.5 x 10^9	143.8 x 10^6
K_3^l	7.7 x 10^6	111.4 x 10^6
K_4	4.28 x 10^6	108 x 10^6

The char formation ratios for reaction K_3 are: $X^c = 0.35$, $X^h = 0.6$, and $X^l = 0.75$. For more information about the setup, please refer to Lathouwers and Bellan (*31*).

The gas flow is closely coupled to the spatiotemporal distribution of the biomass, chemical reactions, and heat and mass transfer. In the following figures, we illustrate the nature of this coupling. Figure 4 shows the biomass mass distribution along with the gas velocity vectors at two different instants corresponding to those in Figure 3. Here are some of the qualitative features of this distribution:

- The gas flow is perturbed slightly to accommodate the injection of the biomass.
- Fluidization gas and the bed particles exert stream-wise force on the incoming biomass into the reactor.

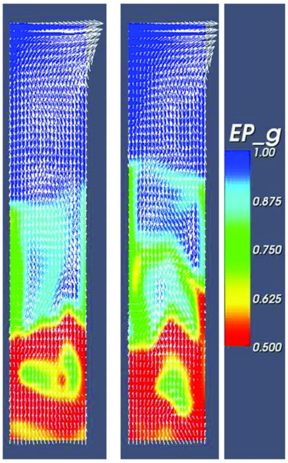


Figure 3. Void fraction with gas velocity vectors at two different instants (at ~1s and ~3s).

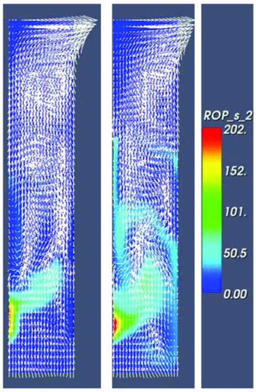


Figure 4. Biomass mass distribution along with gas velocity vectors at the same instants as those in Figure 3.

- The biomass accumulates close to the inlet but quickly disperses and also undergoes pyrolysis in contact with the high-temperature gas and solids.

The overall biomass distribution in the reactor is also closely coupled to the biomass species. Figure 5 shows the mass fraction of the hemicellulose and cellulose in the biomass at the two instants corresponding to Figure 4. The fast-reacting hemicellulose is primarily found near the inlet, while the slower reacting cellulose makes a significant portion of the biomass all throughout the reactor. The distribution of the various species is related to the physical/thermodynamic/transport properties along with the chemical reactivity as functions of local temperature. This in turn influences the variation of the product gas in both space and time. Figures 6a and 6b show the spatial variation of the tar gas and the product gas mass fraction, respectively. The figures clearly

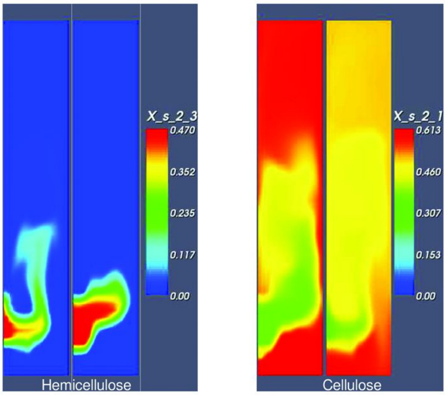


Figure 5. Biomass Composition (Note: Mass fraction of the biomass is shown here, and that can be misleading in regions where the total mass of the biomass is low; however, it instantly shows the relative composition of the biomass components.)

show that at the later times, fluidizing gas bypasses through the bed to the outlet and product species have higher residence time in the reactor. Figure 6c shows the temporal variation of the product and the tar gases at the outlet. The product and tar gases are still in the transient regime after 5 seconds of biomass injection; this is a function of reaction kinetics, temperature distribution, and solids and gas interaction.

In summary, the current state-of-the-art macroscale continuum simulations can provide spatiotemporal variations of solids, biomass, gas, solid species, and reaction rates. Such simulations have been recently employed to study detailed 3D simulations of coal gasifiers by using hundreds or even thousands of supercomputer processors (*32*).

However, reaction kinetics and heat/mass transfer formulations in continuum multiphase flow models need to be determined experimentally or by multiscale upscaling of lower length- and time-scale models. These models can include the DEM for accurately resolving particle-particle collisions, LBM for obtaining the critical parameters related to the drag, heat and mass transfer, KMC for obtaining the chemical kinetics (see Figure 1). In the following sub-sections, we will describe how we can use different methods for deriving these closure relations. We use the LBM for resolving the flow over the solid particles and obtaining the heat and mass transfer closures, and we use the KMC for the chemical reactions occurring on the surfaces and inside the pores of a biomass particle for closures of the local chemical reactions. In the next section, we describe how these quantities can be upscaled and how macroscale simulations can be linked to lower length- and time-scale models where scale separation is not possible.

Lattice Boltzmann Method (LBM)

The LBM solves the Boltzmann equations on discrete lattices using pre-defined velocity directions and magnitudes for the fluid particles. In the incompressible limit, the LBM solutions correspond to the solutions of the Navier-Stokes equations. Because of its discrete representation, the LBM is applicable in cases where the Navier-Stokes differential equations for fluid flow fail under conditions of large Knudsen numbers (ratio of mean free molecule path and the characteristic length scale) (*33*). Such conditions are generally quite common in energy conversion devices and in systems where the mean free molecular path is similar to the geometric constraints. The classical LBM is most applicable to large Knudsen number flows with low Mach numbers (ratio of the particle speed to the speed of sound in the medium). Extension to large Mach numbers is possible through LBM modifications or by using dissipative pseudo-particle dynamics methods.

A brief introduction to the LBM follows. More details are available in the review by Chen et al (*9*). The LBM solves the lattice Boltzmann equations given in the following format (*34*): $\frac{\partial f_i}{\partial t} + \bar{c}_i \frac{\partial f_i}{\partial x} = -\frac{1}{\tau}(f_i - f_i^{eq})$, where $f_i(\vec{x},t) \equiv f_i(\vec{x},\vec{v} = \vec{c}_i,t)$, *i=1*, and *b* is the probability distribution of finding a fluid particle at lattice site $\vec{x}$ at time *t*, moving along the lattice direction given by the discrete speed $\vec{c}_i$. Here, *b* is the number of discrete populations associated with each node P of the computational grid. The left side of the above equation corresponds to the molecular streaming, while the right side represents molecular collisions through a single-time relaxation towards local equilibrium f_i^e on a typical time scale *τ*. This local equilibrium can be represented in terms of lattice sound speed, and the fluid density ($\rho = \sum_i f_i$) and fluid velocity ($\vec{u}_i = \sum_i \vec{c}_i f_i / \rho$) are calculated. To recover fluid dynamic properties, t mass, momentum, and energy must be conserved. LBE discretized with a particular choice of time-difference scheme will yield a finite-volume formulation, which can be solved for each node at every time step. Wall boundary and inlet-outlet conditions are applied to simulate problems of interest. Details of discrete speeds, discretization and boundary conditions are available in the cited references and are not discussed here.

Chemical species can be modeled by scalar equations and convection, diffusion, and reaction terms. A common approach is to track population densities corresponding to each species. This ensures seamless integration of the LBM fluid-flow simulations with those of the species. However, this also burdens the computational effort tremendously as the number of chemical species increases. The recently developed Lax–Wendroff scheme can be used to model multicomponent fluid transport (and reaction) within an LBM simulation framework (*5–7*). Extending the LBM to thermal flows is not straightforward as the number of kinetic moments to be included to accurately model heat fluxes is very high (*35*). Alternative approaches (*36*) have been proposed to eliminate the need for carrying an additional distribution function for temperature, thereby avoiding the computation of higher-order moments.

The LBM is inherently computationally expensive both in terms of floating point operations and storage. The method requires a small time increment for maintaining computational stability and High Performance Computing (HPC), which is usually a prerequisite for realistic LBM simulations. The LBM consists of two main steps: (i) propagation (left side of LBM equations), where

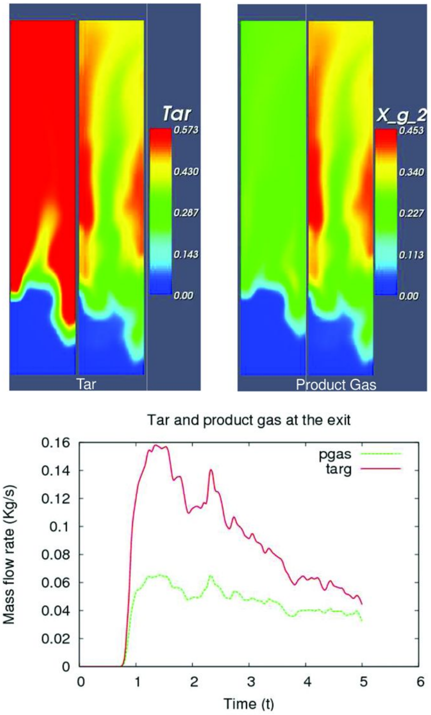


Figure 6. a) Tar gas mass fraction, b) Product gas mass fraction and c) temporal variation of the product and the tar gases at the outlet.

fluid-particle distribution moves along the lattice bonds to the neighboring lattice nodes; and (ii) collision and forcing terms (right side), where fluid particles on the same node collide and adjust velocities to conserve mass and momentum. The interaction between the nodes in the lattice is required only in the propagation step. In HPC implementations, when different parts of the computational domain are assigned to separate computer processors, the propagation step can move fluid packets between the processors and, therefore, require message passing or update of the shared memory location. Previous HPC research on implementing the LBM has indicated that minimizing the computational load imbalance was more important than minimizing the communication imbalance. The parallel implementations to date take advantage of the special locality for updating. They also employ standard static domain decomposition strategies, such as, slice, shaft, box, and Orthogonal Recursive Bisection (ORB), with addition of ghost layers at the sub-domain boundaries. For many practical applications, such an approach is not sufficient, as the number of impermeable nodes, which do not contribute to the computational load, can be significant. Sparse representation of the conductive phase of the lattice (*37*) and graph partitioning approaches are then necessary to reduce the overall memory requirements and distribute the computational load evenly among processors.

The LBM method described here, coupled with heat and mass transfer, can be used to generate better closures for the heat and mass transfer just as

LBM has been extensively used for constructing drag correlations (*38*, *39*). Full Navier-Stokes of granular assemblies based on Immersed Boundary Methods (*40*) also offer alternate means to obtain such detailed data to construct more accurate correlations for drag and heat and mass transfer.

Kinetic Monte Carlo

Monte Carlo (MC) (Stochastic) methods have been used in a wide variety of scientific and non-scientific disciplines, such as materials science, nuclear physics, chemical reactions, sintering, financial markets, and traffic flow, to name a few. MC methods approximate solutions to mathematical problems by statistically sampling computational experiments. The MC method has been widely used in materials science to determine equilibrium structures or thermodynamic properties. However, MC methods have also found application in non-equilibrium and kinetic phenomena. For any kinetic phenomenon, the phase space must be explored along a Markov chain such that each state is accessible from the preceding state along the chain to preserve balance. Thus, the two main steps in a KMC algorithm are: (i) identification of all the possible events that can occur and (ii) determination of the probabilities at which these events can occur.

A short description of the KMC method for a simple 1st order reversible reaction system ($A \leftrightarrow B$) is shown below for illustration (additional details can be found elsewhere (*41*, *42*). In this case, two possible reaction events $A \rightarrow B$ and $B \rightarrow A$ (R_1 and R_2) can happen with reactions rates k_1 and k_2, respectively. In the case we have n events, we have $R_i, i = 1,2,.....,n$. The system contains N_A molecules of A and N_B molecules of B, with $N = N_A + N_B$. We define r_i as the rate at which an event of type R_i takes place. In this case, $r_1 = k_1 N_A$ and $r_2 = k_2 N_B$. Let τ be the time such that no reaction occurs between time $t=0$ and $t=\tau$. One can now construct a stochastic differential equation whose solution is

$$\tau = \frac{1}{\sum_{i=1}^{n} r_i} \ln\left(\frac{1}{\alpha_1}\right)$$

where α_1 is a uniformly distributed random number between zero and unity. Once a reaction event occurs, the probability that it is of type R_i is

$$P(R_i) = \frac{r_i}{\sum_{i=1}^{n} r_i}$$

The reaction selection algorithm can be easily implemented by introducing another uniform random number (α_2) in the interval (0, 1) and dividing it into i segments. An event R_i will take place when,

$$P(R_{i-1}) \leq \alpha_2 < P(R_i), \text{ with } P(R_0) = 0$$

.

In summary, at any given state all the reaction rates must be calculated to yield τ based on the first uniform random number. An event i occurs based on the probability of each event and the second random number and yields the concentrations of the new state. This procedure is repeated for a desired time. The reaction rates (k_i) and reaction steps can be provided by more detailed atomistic simulations (*43–49*).

Multiscale/Multiphysics Coupling Methods with Examples

We now describe a general formulation for interfacing two different length scale and physics methods for simulating reactive flows. The first, the microscale KMC reaction model, deals with microscopic-length scales and is restricted to operating on the boundary of the macroscopic, LBM flow model. The microscopic KMC domain has one less dimensionality than the macroscopic LBM space. Both the KMC and LBM methods are based on probability density functions and the local rules and thereby have common structure over which the coupling can be implemented in a consistent and formal way. The fine-scale information captured in the KMC simulations is transferred to the thermohydrodynamics of the LBM through a modified non-reflecting Neumann boundary condition (BC), similar to the overlapped Schwarz alternating method. The coarse-scale LBM can act as a Dirichlet BC constraint for the micro-scale KMC simulation. For isothermal LBM, the method is simplified, as the Dirichlet BCs can be used on both sides. The approach can be extended to coupling LBM with molecular dynamics (MD) and requires additional processing of the MD statistics to match the probability density functions in both schemes. The aforementioned multiphysics/multiscale method falls into the general class of heterogeneous multiscale methods (HMM) (*50–53*).

The CWM for Coupling Multiscale/Multiphysics Components

The CWM method can couple different physics models operating at different spatiotemporal scales (*54–57*). The formulation is applicable to various methods in many scientific and engineering areas, and we describe it here within the context of coupling from small spatiotemporal scales (KMC) to larger ones (LBM). Within a multiscale framework, we are typically interested in macroscopic or mesoscopic processes described through state variables such as concentrations of reactants, products, intermediaries, denoted separately or collectively as U. A global macroscopic grid-based scheme is used for calculating U. Macroscopic data can be estimated from a microscopic model, where the state variables corresponding to U are denoted as u (*50–53*).

In this case, the CWM method can fulfill two purposes: (i) to project u to U as needed by the LBM method and determine how the union of the spatiotemporal scales modeled by the KMC and LBM affects the entire process; and (ii) to make predictions on and recommendations for improving the efficiency of the studied process. Purpose (ii) is addressed in a subsequent subsection.

The CWM method is based on wavelet transform mathematical theory. In one dimension (extendable to higher dimensions), the wavelet *ψ(x)* transforms a function *f(x)* according to

$$W_f(a,b) = \int_{-\infty}^{\infty} f(x)\psi_{a,b}(x)dx$$

The two-parameter family of functions, $\psi_{a,b}(x) = (1/\sqrt{a})\psi(\frac{xb}{a})$ is obtained from a single one, *ψ*, through dilatations by the factor *a* and translations by the factor *b*. Given the wavelet coefficients *W*$_f$*(a,b)* associated with *f*, it is possible to construct the representation of *f* at a range of scales between s_1 and s_2 ($s_1 \leq s_2$) through the inversion formula

$$f_{s_1,s_2}(x) = \frac{1}{c_\psi}\int_{s_1}^{s_2}\int_{-\infty}^{\infty} W_f(a,b)\psi_{a,b}(x)db\frac{da}{a^2}$$

(c_ψ being a constant). By setting $s_1 \rightarrow 0$, $s_2 \rightarrow \infty$, *f* is reconstructed. It is this representation between s_1 and s_2 that allows projections from *u* to *U* and from *U* to *u* (both needed for purpose (ii)).

Figure 7 shows the schematic of the CWM construction process. The wavelet transform of a state variable *u*/*U* in two dimensions, i.e., *x*, *t*, for KMC, and *X*, *T* for LBM, includes a transform in the *x*/*X* direction, a transform in the *t*/*T* direction, and one in the *x*-*t*/*X*-*T* direction (*58*). Because the scales of *U*, *X*, and *T* in LBM are larger than the scales of *u*, *x*, and *t* in the KMC, the wavelet transform of *U* fills in different sub-matrices of the CWM as compared to the wavelet transform of KMC. Before describing the process of constructing the CWM in further detail, we emphasize that the CWM method is general enough so that it does not require an independent interaction along scales in a HMM sense. In fact, the CWM could be used effectively even if the LBM and KMC were used independently of each other. The scale interaction provided by the CWM is analogous to global error propagation and control in Schwarz methods using the combination of fine and coarse solvers.

CWM for Projection Operations in Time and Space

As illustrated schematically in Figure 7, state variables *U* from LBM are transformed using wavelets (red-dotted line) and the resulting wavelet transform coefficients fill in parts of the CWM operator, i.e., those parts that correspond to the scales used in LBM scales or the fine scales of the CWM (lower left corner), as pointed out by the black arrows. Similarly, corresponding state variables *u* from KMC are transformed using wavelets (blue-dotted line); and the resulting wavelet coefficients fill in the rest of the CWM corresponding to the fine scales of the system (all except the lower left corner), as indicated by the black arrows. More importantly, for specific times (snapshots) the CWM allows for spatial projections along scales (*59*), or, when time is a variable, for spatiotemporal projections along scales (*54*). In the former case, for studying a boundary as in

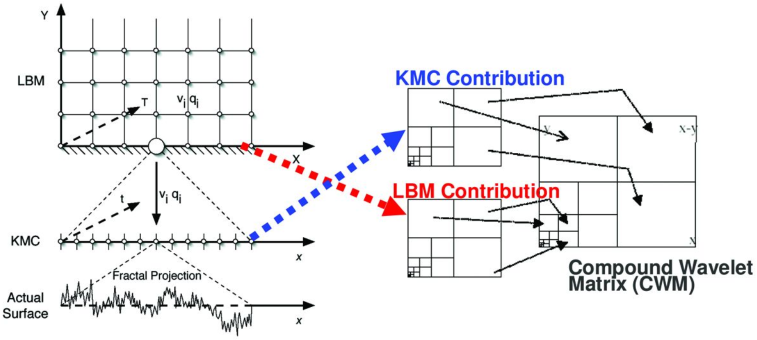


Figure 7. Schematic detailing the coupling of KMC with LBM.

Figure 7, the CWM is the result of wavelet transforms in 1-D (*x*/*X*, *u*/*U*, scale), while in the latter case it is the result of wavelet transform in 2-D (*x*/*X*, *t*/*T*, *u*/*U*, scale). If the boundary is two-dimensional, i.e. *x*, *y* in KMC and *X*, *Y* in LBM, the dimensionality of the CWM increases accordingly. In all cases, the CWM is indeed a wavelet transform, yet not constructed from a single data base, but compounded from two data bases, one resulting from LBM and one from KMC. The construction process is statistical (transfer of statistics) and the outputs from LBM and KMC in the form of probability densities fit the CWM construction process ideally. Inverting the CWM is possible under certain conditions of stationarity or quasi-stationarity (*56*, *57*).

The KMC and LBM processes are related to each other, thus an upscaling projection operator denoted by Q is such that $Qu = U$. Similarly, for downscaling, a projection operator denoted by R is such that $RU = u$. If the complete CWM is given, Q and R can be easily obtained by straightforward application of the wavelet transforms, given above. For implementing the HMM process, a projection operation $\bar{Q}u = U$ can be used, where $\bar{Q}$ denotes an approximation of Q. Various alternatives for obtaining $\bar{Q}$ can be employed: (a) using the complete CWM from previous time steps; (b) using an approximation of CWM constructed from current time-step data of KMC and previous time-step data of LBM; or (c) using locally large enough KMC simulations to allow representation at the spatial scales of the LBM. For efficiency of the projections and error control, several spatiotemporal scales between those of the KMC and LBM should overlap (*54*, *59*).

Using CWM for Making Predictions and Recommendations

The CWM allows us to either study *U* at macro spatiotemporal scales or even zoom in and observe “flashes” of relevant microscopic processes. The predictive capability remains at the final product, i.e., the macro scales, unless a mathematical method is devised for the concurrent description of state variables at all scales involved; and the CWM is efficient in that. Predictions have been reported for different physical problems (*54*, *59*).

The CWM method concurrently represents the state variables of interest at the spatial and temporal scales that are the union of those handled by the LBM and the KMC. Its predictive capabilities are inherently multiscale, because it contains information about the studied process at all available scales. This allows us to study interactions along scales, both spatially and temporally, and thus predict how to improve the efficiency of the process by altering parts of the micro or meso parts as needed. For example, in reactive flow problems, it is not clear how the microscale chemical reaction rates affect the transport of reactant species at the meso/macro scales. The CWM (with time scaling included) can provide the mathematical details of such interactions, e.g., by studying the energy of the wavelet transform, available through the CWM, or other information measures on the concentration of reactants. If such measures peak at small spatial and temporal scales, it indicates that the process is inefficient – reactants do not “diffuse” to large scales effectively. If the energy of the wavelet transform peaks at large scales, the process is again inefficient because reaction rates at microscales are not used to their full capability in terms of “diffusing” reactants to the large scales. A more or less constant distribution of the wavelet transform energies (or any other relevant information measure) would imply process efficiency. Moreover, the CWM method would indicate how this can be achieved, e.g., by formulating better catalysts to change the chemical activity or by altering the macroscopic flow conditions. This technique can potentially provide a model-based design tool for catalysts rather than the present day combinatorial techniques used by catalyst manufacturers.

Time Integration of KMC-LBM Coupling

The LBM and the KMC operate on widely disparate time scales. Thus, the entries in the CWM need to be evaluated at sub-intervals of pre-determined duration and location so that the right-hand side of the LBM governing equations can be updated as few times as possible. The implicit assumption is that the chemical reactions do not drastically affect the evolution of the macroscopic solution. Separation of processes into fast and slow allow one to work within the framework of Fast-Slow Splitting variation of the Multiple Time Stepping (MTS) and Spectral Deferred Correction (SDC) methods (*60*). The separation of the time scales (processes) for a large class of problems can be formally written as a summation of forcing terms on the RHS in the LBM corresponding to slow and fast dynamics, i.e.,

$$\dot{y} = f^{[SLOW]}(y) + f^{[FAST]}(y)$$

The fast dynamics are evaluated by the KMC method, whereas the slow dynamics correspond to the classical LBM terms. In the MTS framework, the time integration can then be formally written as a composition:

$$\left(\Phi^{[slow]}_{h/2}\right)^{*} \circ \left(\Phi^{[fast]}_{h/N}\right)^{N} \circ \Phi^{[slow]}_{h/2}$$

where $\Phi^{[slow]}_{h/2}$ and $\Phi^{[fast]}_{h/N}$ denote numerical integrators of slow and fast processes, respectively. This approach is known as the Impulse Method because the slow modes are evaluated in an impulse fashion only at the end of integration intervals, whereas the fast modes are evaluated in many sub-intervals in between. The MTS methods are quite elegant when the Hamiltonian of the system can be explicitly defined and the solution method can exploit its symplecticity. In cases when the combined methods include automata, a universal description (i.e., Hamiltonian) is not compatible with the character of the underlying methods. The LBM is not derived from a discretization of the Navier-Stokes differential equations. Instead, it is a space-, momentum- and time-discretized automata version of the Boltzmann transport equations with an objective of incorporating the statistical physics nature of fluids into the hydrodynamics solution. The collision rules do not perform Newton dynamics simulations, and they are only constrained by the local conservation and by the requirement of rotational symmetry (isotropy).

In the KMC/LBM coupling, the natural partition into fast and slow processes already exists along the line of methods separation (i.e., LBM and KMC), and the CWM mapping can be explored for the definition of further refinement of the process separation within KMC. The separation depends on the physical processes involved. It has been shown that increasing the sampling frequency N leads to "artificial resonances," so increasing the N does not necessarily lead to a higher accuracy. The optimal choice of the integration increments (N sampling) depends on the problem characteristics and the splitting strategy. Necessary conditions for stability and symmetry need to be maintained.

The SDC methods improve the accuracy of the numerical solutions by reducing the errors from the operator splitting (i.e. decoupling) and integration by iterative approximations of the solutions and applying error corrections during the integration process. In the case of two explicit time integration methods, the corrections involve restarting each method with different constraints and integration of source terms. The evolution of solution space associated with dominant CWM factors can be used to restrict the extent of computational effort, which is particularly useful for problems with multiple species and reactions. The CWM matrix from the previous time step can be used to interpolate provisional solutions and error correction. Numerical experiments are necessary to determine the optimum integration strategies.

Fractal Projection for Catalytic Surface

Reactive processes on catalytic surfaces are influenced by many local surface topological and chemical morphology-related factors such as limitations on mobility of adsorbed molecules, localization of reactions, adsorbate-induced surface restructuring, surface transport, to name a few (*61–63*). It has been shown that fractal surfaces can affect the dynamics of reactions and the steady state (*64*) of the system. Processes that occur on smaller length scales can have larger reaction probability. Residence and reaction times also depend on the distribution of active sites and their conformity to the surface morphology. Fractals on the surface are bounded below by interatomic distances, and the upper cutoff, ξ, is determined by surface processing. This allows us to consider the fractal surface as consisting of repeating units of size proportional to the upper self-similarity cutoff. This, in turn, replaces self-similarity by the translational invariance of order ξ.

Time Acceleration with Time-Parallel Compound Wavelet Matrix Method

Even though spatial domain decomposition, parallel computing, and CWM can provide significant computational gains, the sequential nature of time still mandates that the time integration evolve in consecutive time increments. This constraint is compounded by high spatial resolution of lower scales and a commensurate reduction of the stable time increments through Courant criteria. Time parallel (TP) methods have been introduced to alleviate this problem. The key idea of the TP algorithms is to use different time propagators distributed across phase space and iterate on their evolution until convergence. In a case of two propagators, termed here "coarse" and "fine," in each global iteration, one obtains a temporally coarse solution of the problem; and then at several temporal "nodes" along the coarse solution independently instantiates fine-grained temporal simulations. The fine simulations correct the coarse counterpart, and the process is repeated, in a predictor-corrector sense, until convergence is achieved.

The method is very efficiently parallelizable and conceptually simple, clear advantages for utilizing supercomputing resources. Figure 8 shows a schematic of the TP solution process; the fine solutions are iterated in coordination with the coarse one, in a predictor corrector sense, until satisfactory convergence occurs. The TP algorithms are described in detail (*65–67*) and references cited therein.

One obvious drawback of this method is that the coarse and fine simulations interact only at the temporal "nodes" where the fine simulations instantiate, in effect localizing the errors. In addition, the fine simulation has to run for the entire duration of the assigned time interval. These may impede convergence and miss features of the response that depend strongly on the initial and boundary conditions. A newly proposed framework, the Time Parallel Compound Wavelet Matrix (tpCWM) method (*68*) combines the TP and the CWM methods to make multiscale/multiphysics simulations computationally scalable in time and space realms. To illustrate the main tpCWM idea, we consider the CWM method as a strictly time-scaling method within one time interval of the TP method. Figure 9 shows the CWM operating on temporal scales within the context of "coarse" and "fine" methods as in the TP method. By wavelet compounding of the small-scale information from the fine solution and the large-scale information from the coarse method, an improved temporal response is obtained that, in addition to "correcting" the coarse trajectory, also incorporates small-scale information from the fine method.

In tpCWM, a coarse scale solution is iteratively corrected by the CWM, which compounds the fine and coarse solutions for the considered time interval. The computational savings, over the full fine solution (in terms of the real time required to perform the simulations) can reach several orders of magnitude, depending on the number of parallel processors available, the number of iterations required for convergence, and the efficacy of the CWM process. The computational savings compared to the conventional TP method without a CWM upscaling can also reach several orders of magnitude, depending on the number of iterations required by

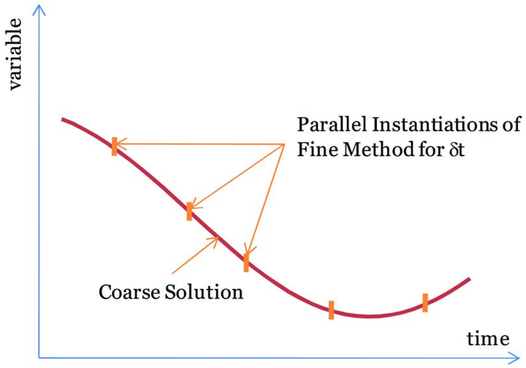


Figure 8. Schematic of the TP method. The fine method instantiates at several temporal "nodes" typically for a period δt that covers time until the next node.

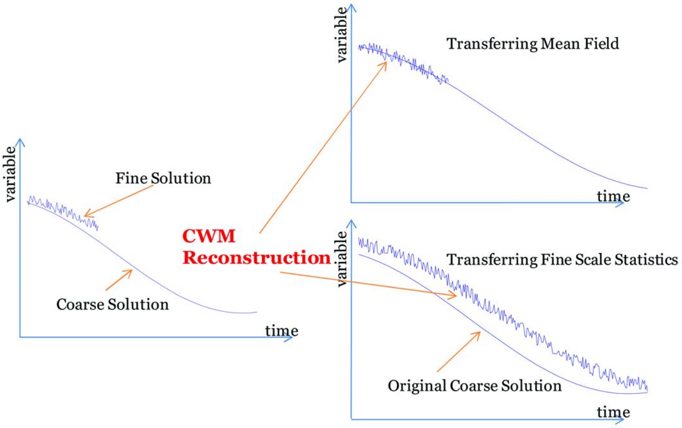


Figure 9. Schematic of temporal CWM. The fine method is employed for a fraction of the coarse method, and the CWM reconstruction updates the temporal statistics as well as the mean field.

the TP method. Given the current computer processor design path, it is likely to have millions of processor cores on exascale parallel computers, which makes the tpCWM algorithm even more appealing for accelerating multiscale simulations.

The tpCWM Method for a Simple Chemical Process

We now apply the tpCWM method to a simple chemical process with oscillatory trajectory and analyze the method's convergence and scalability in time. Applying the tpCWM method to a specific biomass thermomechanical process requires proper simulation of the underlying processes, but the overall tpCWM approach remains conceptually the same as for this simple case. In TP as well as in CWM, the coarse method can be a deterministic solution at coarse discretization of the relevant rate equation; and the fine simulations can be a KMC method. The benchmark solution, for testing the method, can then be the solution of the KMC that was run over the entire interval of interest.

Let a and b denote two time-dependent concentrations of the two intermediates of interest for a system. At steady-state, the respective concentrations are a_0,b_0, respectively, and deviations from steady state are denoted as $A = a - a_0$, $B = b - b_0$, respectively. Processes governed by (1) are examined, where coefficients κ_{ij} are rate constants. Processes (1) have applications in many fields, including chemical, biological, biochemical systems, heat flow, and membrane vibrations. [Noyes and Field (*69*) and references therein for example].

$$\begin{aligned}\frac{dA}{dt} &= \kappa_{11}A + \kappa_{12}B \\ \frac{dB}{dt} &= \kappa_{21}A + \kappa_{22}B\end{aligned} \tag{1}$$

Two cases of (1) have been examined. In the first case, $-\kappa_{11} = \kappa_{12} = -\kappa_{21} = \kappa_{22} = \kappa = const.$, and (1) yields an exponential decay/increase solution for A and B (e.g., modeling a unimolecular reversible chemical reaction). The second case, $\kappa_{11} = \kappa_{22} = 0$, $-\kappa_{21} = \kappa_{12} = \kappa = const$ yields oscillatory solutions for A, B. In the following, the tpCWM is presented with respect to the oscillatory case. The formulation and results for the exponential case are qualitatively similar, yet the tpCWM converges to the correct solution even faster than for the oscillatory case.

The fine model uses the KMC algorithm for solving the kinetic evolution (1) for the deviations from the steady state, i.e., A, B. The rate constant κ (inverse time (t) units) is taken to be equal to 0.001 (sec^{-1}), and the times required for one unit change in the value of A, B for the oscillatory case are expressed as:

$$\begin{aligned}t_1 &= -\frac{1}{\kappa|A|}\ln(1-R_1) \\ t_2 &= -\frac{1}{\kappa|B|}\ln(1-R_2)\end{aligned} \tag{2}$$

where R_1 and R_2 are independent, uniformly distributed random numbers between zero and unity.

At any time in the simulation, an event that requires the least time is the one that will occur. Thus, at every KMC iteration step, two random numbers are generated, i.e., R_1,R_2, and t_1,t_2 are evaluated based on (2). The minimum of t_1,t_2 is the time increment associated with the selected unit change event. For initial value for A equal to zero and for B equal to 10,000, Figure 10 shows the time evolution of species A obtained by KMC for the given time interval and will serve here as a benchmark solution to the reaction problem. The closed-form solution to (1) for A and the values given above is $A(t) = 10000sin(\kappa t)$. This solution is the same as the ensemble average of the benchmark, but lacks any short-scale oscillations.

The coarse model uses a deterministic algorithm for solving the kinetic evolution for A, B. The oscillatory case for finite difference discretization using the first order Eulerian scheme yields, with Δ denoting finite difference,

$$\Delta A = \kappa B \Delta t, \qquad \Delta B = -\kappa A \Delta t \tag{3}$$

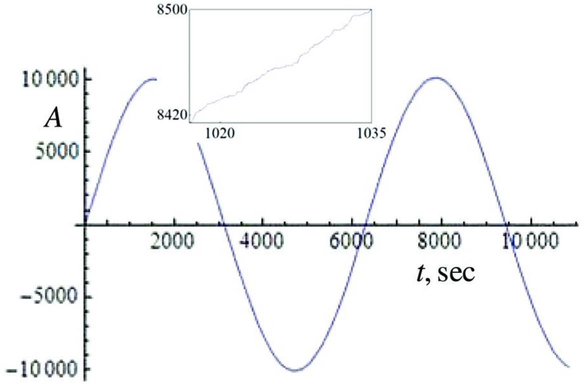


Figure 10. Benchmark solution of the time evolution of species A obtained by KMC.

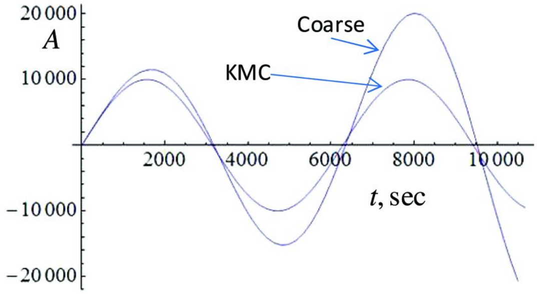


Figure 11. Time evolution of species A obtained from the coarse solution and comparison to the benchmark solution.

We intentionally selected the first-order Eulerian scheme because its finite difference error increases as time *t* increases and, eventually, diverges. Yet, despite the divergence, it will be shown that the tpCWM will be able to converge to the correct solution quickly. The difference equations are solved iteratively, and *A* and *B* are updated for each time step. Of course, the purpose here is to examine the stability of the tpCWM; thus, large time increments are used to obtain a relevant solution from the coarse model. In an attempt to keep the number of time steps in the coarse method small, a time increment of 175s is used. The corresponding plot is shown in Figure 11; a total of 60 time steps are used in the coarse method. In other words, 60 TP processes (or parallel processors) are used in each time-integration case.

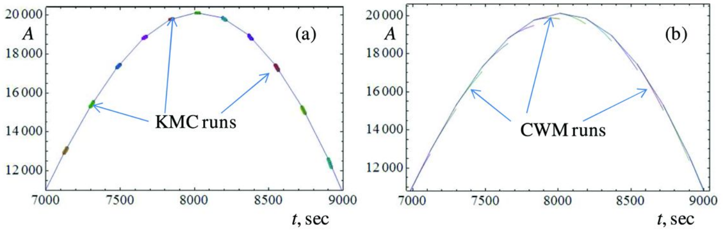


Figure 12. (a) KMC, (b) CWM runs, n_p=60, at each node for the first iteration. Part of the time domain is shown for clarity.

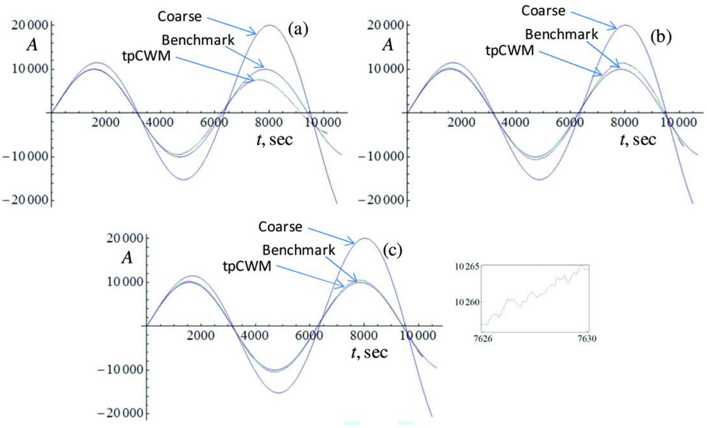


Figure 13. tpCWM solution, n_p=60, at iterations 1 (a), 2 (b), 3 (c). The CWM, for a particular time interval is shown in (c) (inset) depicting the relevant fluctuations. Also, the benchmark solution is shown, towards which the tpCWM iterations converge.

tpCWM Solution

A tpCWM algorithm to this reaction problem calls for instantiation of CWM solutions at the beginning of each of the time steps from the coarse method, called nodes. For the present case, the number of nodes is considered to be equal to the number of TP processes (number of processors) denoted as n_p. Each KMC simulation runs, for each iteration, for a portion of (tpCWM) or the entire (TP) time period between nodes, equaling the time increment of the coarse method. The details of the TP iteration algorithm can be found elsewhere (*65*).

The most important difference between the TP and the tpCWM is that the CWM compounding of the coarse and fine solutions within time intervals allows for the fine solution to only run for a fraction of the time until the next node. This fraction, denoted as *f* was herein chosen to be 1/16. Figure 12a shows the KMC results at each node for the first iteration. The same value for *f*, 1/16, is used at

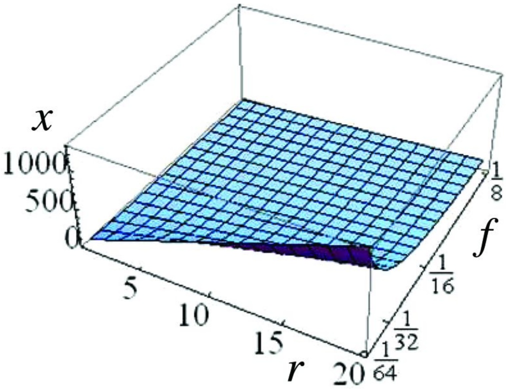


Figure 14. Factor of computational savings, X, as a function of the ratio r and the fraction f.

the each iteration, and the KMC results are used in forming the CWM for the each node and iteration. Results for the first iteration are shown in Figure 12b.

Figure 13 shows results from the tpCWM process at three iteration steps, where, similarly to the TP, it can be seen that it only takes a few iterations for this problem to converge.

Computational Efficiency of the tpCWM Method

The tpCWM method offers significant computational savings compared to the classical TP. In the TP, the KMC would run for the entire time interval for every TP iteration, and the computational savings over the complete KMC solution would come solely from the use of parallel processors. In the tpCWM, however, the KMC runs for only a fraction of time interval. This saving in computational time occurs at every TP iteration, so that, for example, for f=1/16 and seven iterations, the computational savings of the tpCWM over TP is approximately 7*16~112 times. The CWM calculations in the tpCWM reduce this factor but not considerably. In the examples presented previously, the actual saving factor was about 95 instead of 112.

Let n_i denote the number of iterations required for convergence. The ratio r of TP processes over n_i, denoted as

$$r = \frac{n_p}{n_i}$$

is approximately 60/3, for n_p=60, of the tpCWM example presented above.

Figure 14 shows the factor of computational savings X, defined as the ratio of computational time required for the benchmark method (pure KMC) over time required for the tpCWM, as a function of r (number of processors/number of iterations) and f (fraction of KMC time used in each assigned time interval). Three orders of magnitude in X can be achieved by r in the range of 20 and f in the order of 1/64.

It is important to determine how the ratio r changes as the number of TP processes n_p increases. If r remains the same, i.e., is independent of n_p, then increasing n_p merely implies that efficiency is increasing proportionally to r. However, if r increases with n_p, this will imply that the efficiency of tpCWM increases further with an increasing number of processors, a clear trend in parallel processing machines. Based on results from using 30, 60, and 90 TP processes n_p, the required number of iterations was tracked, and they were 6, 4, and 3, showing that the number of iterations remains low and even becomes lower as n_p increases. Clearly, the tendency for r to increase with increasing n_p, or even stay constant and equal to a small number furnishes an advantage of the tpCWM.

The tpCWM can be extended to other multiphysics/multiscale problems, especially to the biomass thermochemical conversion processes. The tpCWM can incorporate different implementations of fine and coarse methods described in this chapter and efficiently couple surface reactions in the porous biomass, the heat and mass transfer from the heterogeneous biomass material, with the overall macroscale transport.

Conclusions and Future Work

Biomass thermochemical processes are inherently multiscale phenomena, both spatially and temporally, yet the lack of adequate multiscale methods led to an extensive use of loosely coupled, macroscopic continuous methods that do not capture the details necessary to model the process in a predictive manner. Incorporating fine-scale simulations in the macroscopic models is a challenging goal, and many hurdles need to be overcome. The CWM method can help overcome some of the difficulties.However, it may not be enough to bridge the gap between microscopic and macroscopic methods that operate at industrial timescales. The combined TP and CWM method offers a viable alternative to incorporating fine-scale information into the coarse spatiotemporal scales useful to the biomass conversion industries. Simulation at scales of the industrial interest calls for (a) detailed simulation of thermochemical processes at fine scales that is consistent with macroscopic models to some acceptable degree; (b) identification of process intervals, spatial and temporal, over which the tpCWM method can be applied; and (c) effective computational acceleration with suitable load-balancing strategies using contemporary massively parallel computers. Even though the above goals may be achievable in the future, we note that coupling of scales unexplored previously may reveal new issues not captured by models used for lower scales. This would require renewed modeling at either larger scales or smaller, or both, in a coupled fashion. Such a predictive multiscale tool can dramatically alter how future biomass thermochemical devices are designed as it would allow us to optimize the processes at the microscale while accounting for all the macroscopic effects of the device.

References

1. Hu, R.; Huang, S. P.; Liu, Z. P.; Wang, W. C. *Appl. Surf. Sci.* **2005**, *242* (3−4), 353–361.
2. Fox, R. O. *J. Comput. Phys.* **2008**, *227* (12), 6313–6350.
3. Fox, R. O.; Laurent, F.; Massot, M. *J. Comput. Phys.* **2008**, *227* (6), 3058–3088.
4. Succi, S.; Filippova, O.; Smith, G.; Kaxiras, E. *Comput. Sci. Eng.* **2001**, *3* (6), 26–37.
5. Succi, S.; Gabrielli, A.; Smith, G.; Kaxiras, E. *Eur. Phys. J.: Appl. Phys.* **2001**, *16* (1), 71–84.
6. Gabrielli, A.; Succi, S.; Kaxiras, E. *Comput. Phys. Commun.* **2002**, *147* (1-2), 516–521.
7. Succi, S.; Smith, G.; Kaxiras, E. *J. Stat. Phys.* **2002**, *107* (1-2), 343–366.
8. Pannala, S.; Simunovic, S.; Daw, C. S.; Nukala, P.; Frantziskonis, G.; Mishra, S. K.; Muralidharan, K.; Deymier, P.; Fox, R. O.; Gao, Z. *Micro-mesoscopic modeling of heterogeneous chemically reacting flows (MMM-HCRF) over catalytic/solid surfaces: 2007 annual progress report*; Technical report; Oak Ridge National Laboratory, 2007.
9. Chen, S.; Doolen, G. D. *Annu. Rev. Fluid Mech.* **1998**, *30*, 329–364.
10. Boyalakuntla, D. S.; Pannala, S.; Daw, S. C.; Benyahia, S.; O'Brien, T.; Syamlal, M. Unpublished, Cincinnati, OH, United States, 2005.
11. Tanaka, T.; Kawaguchi, T.; Tsuji, Y. *Int. J. Mod. Phys. B* **1993**, *7* (9−10), 1889–1898.
12. Tsuji, Y.; Kawaguchi, T.; Tanaka, T. *Powder Technol.* **1993**, *77* (1), 79–87.
13. Tsuji, Y.; Tanaka, T.; Yonemura, S. *Powder Technol.* **1998**, *95* (3), 254–264.
14. Tsuji, Y. *Powder Technol.* **2000**, *113* (3), 278–286.
15. Pannala, S.; Daw, C. S.; Finney, C. E. A.; Boyalakuntla, D.; Syamlal, M.; O'Brien, T. J. *Chem. Vap. Deposition* **2007**, *13* (9), 481–490.
16. Pannala, S.; Daw, C. S.; Halow, J. *Int. J. Chem. React. Eng.* **2003**, *1* (A20).
17. Pannala, S.; Daw, C. S.; Halow, J. S. *Chaos* **2004**, *14* (2), 487–498.
18. Blomgren, P.; Palacios, A.; Bing, Z.; Daw, S.; Finney, C.; Halow, J.; Pannala, S. *Chaos* **2007**, *17* (1), 13120–13121.
19. Cizmas, P. G.; Palacios, A.; O'Brien, T.; Syamlal, M. *Chem. Eng. Sci.* **2003**, *58* (19), 4417–4427.
20. Campbell, C. S. *Annu. Rev. Fluid Mech.* **1990**, *22*, 57–92.
21. Bertrand, F.; Leclaire, L. A.; Levecque, G. *Chem. Eng. Sci.* **2005**, *60* (8−9), 2517–2531.
22. Prosperetti, A. In *Computational Methods for Multiphase Flow*; Prosperetti, A., Tryggvason, G., Eds.; Cambridge University Press: Cambridge, 2007.
23. Zhang, D. Z.; Prosperetti, A. *Phys. Fluids* **1994**, *6* (9), 2956–2970.
24. Zhang, D. Z.; Prosperetti, A. *J. Fluid Mech.* **1994**, *267*, 185–219.
25. Anderson, T. B.; Jackson, R. *Ind. Eng. Chem. Fundam.* **1967**, *6* (4), 527–&.
26. Jackson, R. *Chem. Eng. Sci.* **1997**, *52* (15), 2457–2469.
27. Edward, J. T. *J. Chem. Educ.* **1970**, *47* (4), 261–&.
28. Gidaspow, D. *Multiphase Flow and Fluidization: Continuum and Kinetic Theory Descriptions*; Academic Press: Boston, 1994.

29. Syamlal, M.; O'Brien, T. J. *AIChE J.* **2003**, *49* (11), 2793–2801.

30. Syamlal, M.; Rogers, W.; O'Brien, T. J. Report No. DOE/METC-94/1004 (DE94000087), 1993.

31. Lathouwers, D.; Bellan, J. *Int. J. Multiphase Flow* **2001**, *27* (12), 2155–2187.

32. Pannala, S.; Guenther, C.; Galvin, J.; SyamlalM.; Gel, A. In *Computational Fluid Dynamics in Chemical Reaction Engineering V*; Whistler, BC, Canada, 2008.

33. Nieminen, R. M. *J. Phys.: Condens. Matter* **2002**, *14* (11), 2859–2876.

34. Ubertini, S.; Bella, G.; Succi, S. *Phys. Rev. E* **2003**, *68* (1), 016701-1–016701-10.

35. Succi, S. *The Lattice Boltzmann Equation*; Oxford University Press: Oxford, 2001.

36. D'Orazio, A.; Succi, S. *Future Generation Computer Systems* **2004**, *20* (6), 935–944.

37. Pan, C. X.; Prins, J. F.; Miller, C. T. *Comput. Phys. Commun.* **2004**, *158* (2), 89–105.

38. van der Hoef, M. A.; Annaland, M. V.; Deen, N. G.; Kuipers, J. A. M. *Annu. Rev. Fluid Mech.* **2008**, *40*, 47–70.

39. Van der Hoef, M. A.; Beetstra, R.; Kuipers, J. A. M. *J. Fluid Mech.* **2005**, *528*, 233–254.

40. Wachmann, B.; Schwarzer, S.; Hofler, K. *Int. J. Mod. Phys. C* **1998**, *9* (8), 1361–1371.

41. Binder, K.; Heermann, D. W. *Monte Carlo Simulation in Statistical Physics*, 4th ed.; Springer: Berlin, 2002.

42. Battaile, C. C.; Srolovitz, D. J. *Annu. Rev. Mater. Res.* **2002**, *32*, 297–319.

43. Neurock, M. *J. Catal.* **2003**, *216* (1−2), 73–88.

44. Mei, D. H.; Hansen, E. W.; Neurock, M. *J. Phys. Chem. B* **2003**, *107* (3), 798–810.

45. Hansen, E.; Neurock, M. *J. Phys. Chem. B* **2001**, *105* (38), 9218–9229.

46. Hansen, E. W.; Neurock, M. *J. Catal.* **2000**, *196* (2), 241–252.

47. Hansen, E. W.; Neurock, M. *Surf. Sci.* **2000**, *464* (2−3), 91–107.

48. Hansen, E.; Neurock, M. *Surf. Sci.* **1999**, *441* (2−3), 410–424.

49. Hansen, E. W.; Neurock, M. *Chem. Eng. Sci.* **1999**, *54* (15−16), 3411–3421.

50. Engquist, W. E. B.; Li, X. T.; Ren, W. Q.; Vanden-Eijnden, E. *Communications in Computational Physics* **2007**, *2* (3), 367–450.

51. Sun, Y.; Engquist, B. *Multiscale Model. Simul.* **2006**, *5* (2), 532–563.

52. Engquist, B.; Tsai, Y. H. *Math. Comput.* **2005**, *74* (252), 1707–1742.

53. Weinan, E.; Engquist, B.; Huang, Z. Y. *Phys. Rev. B* **2003**, *67* (9), 092101-1–092101-4.

54. Frantziskonis, G.; Deymier, P. *Phys. Rev. B* **2003**, *68* (2), 024105-1–024105-8.

55. Frantziskonis, G.; Mishra, S. K.; Pannala, S.; Simunovic, S.; Daw, C. S.; Nukala, P.; Fox, R. O.; Deymier, P. A. *Int. J. Multiscale Comput. Eng.* **2006**, *4* (5−6), 755–770.

56. Mishra, S. K.; Muralidharan, K.; Pannala, S.; Simunovic, S.; Daw, C. S.; Nukala, P.; Fox, R.; Deymier, P. A.; Frantziskonis, G. N. *Int. J. Chem. React. Eng.* **2008**, *6*.

57. Muralidharan, K.; Mishra, S. K.; Frantziskonis, G.; Deymier, P. A.; Nukala, P.; Simunovic, S.; Pannala, S. *Phys. Rev. E* **2008**, *77* (2), 026714-1–026714-14.

58. For example, the wavelet transform of a 1024x1024 matrix consists of three 512x512 matrices, three 256x256 matrices, and so on; each decomposition level is at half the resolution of the previous one.

59. Frantziskonis, G.; Deymier, P. A. *Modell. Simul. Mater. Sci. Eng.* **2000**, *8* (5), 649–664.

60. Hairer, E.; Lubich, C.; Wanner,G.; *Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations*; Springer-Verlag, 2002.

61. Benavraham, D.; Considine, D.; Meakin, P.; Redner, S.; Takayasu, H. *J. Phys. A: Math. Gen.* **1990**, *23* (19), 4297–4312.

62. Provata, A.; Turner, J. W.; Nicolis, G. *J. Stat. Phys.* **1993**, *70* (5−6), 1195–1213.

63. Oshanin, G.; Blumen, A. *J. Chem. Phys.* **1998**, *108* (3), 1140–1147.

64. Radoev, B. P.; Tenchov, B. G. *J. Phys. A: Math. Gen.* **1987**, *20* (3), L159–L162.

65. Baffico, L.; Bernard, S.; Maday, Y.; Turinici, G.; Zerah, G. *Phys. Rev. E* **2002**, *66* (5), 057701-1–057701-4.

66. Garrido, I.; Lee, B.; Fladmark, G. E.; Espedal, M. S. *Math. Comput.* **2006**, *75* (255), 1403–1428.

67. Gander, M. J.; Vandewalle, S. *Siam Journal on Scientific Computing* **2007**, *29* (2), 556, 578.

68. Frantziskonis, G.; Muralidharan, K.; Deymier, P.; Simunovic S.; Pannala, S. In arXiv:0804.0017v1 (also submitted to Physical Review E. Rapid Communications, 2008).

69. Noyes, R. M.; Field, R. J. *Annu. Rev. Phys. Chem.* **1974**, *25*, 95–119.

Chapter 12

Computational Fluid Dynamics Modeling of Biomass Gasification and Pyrolysis

P. Pepiot,* C. J. Dibble, and T. D. Foust

National Renewable Energy Laboratory, Golden CO, 80401
***perrine.pepiot@nrel.gov**

Biomass thermochemical conversion holds great promise for producing biofuels and will play a determining role in displacing petroleum-based fuel consumption toward renewable sources. Empirical approaches have shown severe limitations in their capability to understand and control the conversion processes. However, without the ability to accurately predict and optimize thermochemical conversion performance, large-scale commercialization of these systems is severely compromised. In this context, Computational Fluid Dynamics (CFD) appears as an essential tool to better comprehend the complex physical and chemical processes involved, paving the way toward efficient control and design strategies. After a brief description of the numerical models needed to simulate biomass gasification and pyrolysis, the contributions of CFD to process design and optimization are detailed. Finally, the state of the art in terms of numerical models for the dense, reactive particulate flows typically found in conversion processes are reviewed. Shortcomings of existing CFD simulations, especially in terms of validation and predictability, are examined; and directions for future research based on the progress of CFD in other fields are suggested.

Introduction

Currently, crude oil is almost exclusively used for producing transportation fuels worldwide. In fact, more than 97% of transportation fuel needs are met with crude oil (*1*). To reduce our sole dependence on crude oil to meet transportation needs, and to limit the environmental impact of crude oil usage such as greenhouse gas emissions, many countries and regions are rapidly developing and deploying biofuels and have set some very aggressive goals for near-term deployment. For example, the EU has mandated that biofuels account for 10% of transportation fuel use by 2020 (*2*). Furthermore, the United States has set both a near-term goal of a 20% reduction in 2007 gasoline usage by 2017, to be met predominantly with increased biofuels production (*3*), as well as a long term "30x30" goal to displace 30% of the 2004 gasoline demand with biofuels by 2030 (*4*).

Although many countries are rapidly deploying biofuels, this first wave of development focuses almost exclusively on first-generation biofuels technologies that utilize food- or feed-based feedstocks. Brazil and the United States are rapidly moving forward with developing and deploying ethanol technology, with Brazil using sugarcane as the feedstock and the United States using corn. Brazilian sugar cane ethanol is generally regarded as having little to no impact on primary food supplies and prices, because Brazil has increased its sugar cane production to more than offset the amount of sugar diverted to ethanol production. However, food supply and price concerns have been raised about corn ethanol production in the United States (*5*), because corn grain is an important food and animal feed commodity. The EU, the largest biodiesel producer, uses rapeseed oil as its main feedstock and again concerns about fats and oils supplies and prices have been raised over the diversion of rapeseed oil to biodiesel production.

Because of these concerns and the overall limitations of first-generation biofuels technology primarily due to feedstock restrictions, advanced or second-generation biofuels technologies, based on sustainable, non-food sources of feedstocks, will be required to meet aggressive volume goals for biofuels deployment (*6*). Several different technologies exist (*6–8*) for converting cellulosic biomass to biofuels. The predominant differentiation between the conversion options is the primary catalysis system (*9*). Biochemical conversion routes rely on biocatalysts, such as enzymes and microbial cells, in addition to heat and chemicals, to convert biomass first to an intermediate mixed sugar stream and then to ethanol or other fermentation-produced biofuel. Conversely, thermochemical conversion technologies rely on heat and/or physical catalysts to convert biomass to an intermediate gas or liquid, followed by an additional conversion step to transform that intermediate to a biofuel. Thermochemical conversion processes will play a determining role in the development and deployment of second generation biofuels, because they offer significant advantages, such as the ability to robustly handle a wide range of feedstocks, and they are capable of producing various types of transportation fuels. Biofuels production via themochemical approaches shows great promise for being economically competitive with conventional petroleum derived gasoline and diesel (*4*) in both the near and long term. Additionally, the economics of biofuels production via thermochemical approaches compare favorably with the economics of other biomass-to-biofuels conversion routes, such as biochemical approaches (*10*).

Thermochemical conversion technologies for producing transportation fuels can be categorized as either gasification or pyrolysis (*11*). Gasification is a complete depolymerization of biomass with limited oxygen at high temperatures, typically > 850°C, to a gaseous intermediate synthesis gas (syngas) consisting of H_2 and CO. A review of the existing types of gasifiers and their relative advantages and disadvantages for transportation fuel production is provided by Spath and Dayton (*12*). Pyrolysis, on the other hand, is the milder depolymerization of biomass producing a liquid intermediate (pyrolysis oil or "bio-oil") in the absence of added oxygen at lower temperatures, typically in the range of 400°C to 650°C. Detailed reviews of pyrolysis techniques and their current technical status are provided by Bridgwater and Peacocke (*12*) and Czernik and Bridgwater (*13*). Although there are a number of gasification and pyrolysis processes under development (*14*), fluidized bed processes are attractive for converting biomass (*15*) because they are easily scalable, very robust, and do not require significant size reduction of the feedstock, which can be problematic for biomass.

Current State of Development and Research Needs

Driven by a worldwide desire to develop second-generation biofuels and the high potential offered by thermochemical conversion technologies, considerable progress has been achieved for both gasification and pyrolysis routes for biofuels production (*16*). Despite these advances, some technical challenges still need to be solved to enable large-scale industrialization of these processes (*17*).

One of the big challenges associated with either fluidized-bed gasification or pyrolysis is the high variability in reactor performances, noticeably increasing the risks associated with the development of industrial-scale facilities. For example, Figure 1 shows the hydrogen-to-carbon monoxide ratio as a function of the operating temperature reported in the literature for several existing biomass gasification fluidized-bed reactors. Hydrogen-to-carbon monoxide ratio is a critical output parameter for liquid fuel synthesis, and still, variations of nearly an order of magnitude are observed, which cannot be explained by the current state of understanding of these systems.

Another challenge is the issue of undesirable tar production. In this context, the term "tar" refers to the complex mixture of organic compounds that are produced by either biomass gasification or pyrolysis. In fuel synthesis operations, tars are especially problematic, because if they are not fully reacted to product gases or removed, they can condense and rapidly foul downstream equipment, significantly decreasing the overall efficiency of the conversion process. The ultimate nature of the tar produced from gasification or pyrolysis is a combination of primary tar formation and secondary tar reactions that alter both the total amount and composition of tars produced (*17*). Although a considerable amount of work has focused on addressing the tar production problem, the overwhelming majority of this previous work tends to fall in two distinct categories: the use of catalysts to prevent or reduce tar formation in the fluidized bed (*24*, *25*), or the use of downstream tar-reforming catalysts to reform the tars into additional syngas (*26*, *27*). However, a fundamental understanding of the chemical and physical processes responsible for tar formation is essential to develop efficient and viable control strategies.

Computational fluid dynamics (CFD) has been recognized as a powerful design and development tool in many industrial areas. For example, spectacular

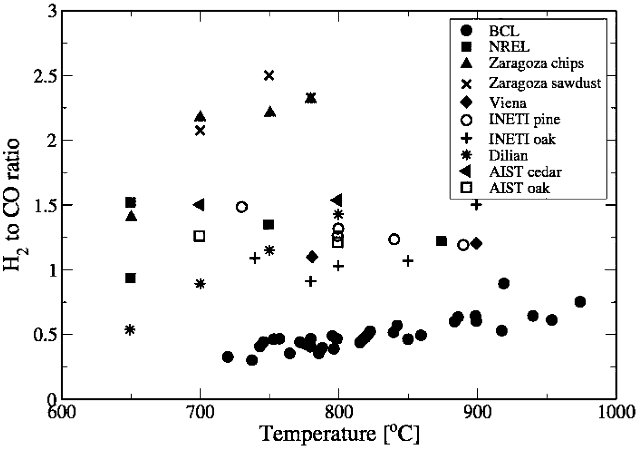


Figure 1. Reported hydrogen-to-carbon monoxide ratios from existing fluidized-bed gasifiers. (sources: (18–23))

progress has been made in aircraft design that enabled a deep understanding of crucial processes such as laminar to turbulent transitions, and fluid/structure interactions for external flows, or turbulence/chemistry interactions, flame stability and pollutant formation for engines, leading to the computational exploration of novel, more efficient designs (*28–30*). These advances in numerical techniques, applied to biomass thermochemical conversion systems, provide a unique opportunity to improve CFD predictive capabilities and move beyond the strictly empirical strategies that have shown severe limitations in terms of cost, flexibility, reactor scale-up, and optimization of reactor design and operating conditions.

This chapter will focus on the modeling of the fluidized-bed reactor itself, disregarding the subsequent tar-reforming and fuel-synthesis processes. Performance during this first stage is crucial to the overall process, as it determines the extent and cost of the treatments required to clean and condition the synthesis gas prior to power generation or fuel synthesis. The fluidized-bed reactor combines most of the possible phenomena involved in thermal conversion, including hydrodynamics, chemical processes and heat release, in a single unit operation. Therefore, progress in this area will help develop comprehensive models that encompass the additional units necessary for the gasification or pyrolysis of biomass. For instance, in the case of indirect gasification, modeling heat production through char combustion in an annex reactor can directly take advantage of the multiphase reactive models developed for the gasifier. Accurate, validated modeling of biomass behavior in fluidized-bed reactors is the crucial next step for advancing computational modeling in biomass gasification and pyrolysis.

The sections in this review include a description of biomass gasification processes and existing numerical approaches that may be used to model them, the role of CFD in reactor and process design and optimization, and the steps necessary to translate descriptive CFD into predictive models.

Biomass Thermochemical Conversion Processes and Modeling

Physical and Chemical Characteristics of Biomass Gasification

The first step of biomass gasification occurs when pelletized biomass enters a fluidized-bed reactor (Figure 2). A fluidized-bed reactor consists of a bed of inert material such as olivine sand, which is fluidized by injecting a gaseous medium from the bottom. The bubbling or turbulent motion of the bed ensures good mixing properties and uniform heat distribution. Superheated steam is usually used to create fluidization when gasification is desired, as it actively contributes to conversion to syngas. Biomass is injected into the reactor, either at the bottom or top of the bed, and is quickly heated by the hot flowing gas and through collisions with the hot sand.

The first stages of gasification are similar to those of pyrolysis. First, water contained in the biomass evaporates. Then, volatiles are released, producing permanent gases and primary tar, leaving behind fixed carbon in the form of char. Primary products undergo further decomposition both inside the biomass particle and in the gas phase. The composition of the gas released from the biomass as well as the amount of carbon left in the char is highly dependent on both the intra-particle thermal and chemical processes and the coupling with the external flow. Gas flow in the reactor and particle collisions provide heat and carry away devolatilization products, driving the biomass conversion to syngas. To emphasize the coupling between all these processes, Figure 3 shows the structural changes occurring in the plant cell walls when subjected to intense external heat as encountered in fluidized beds. While the heat-exposed side (left) suffers extensive deformation, the interior walls (right) remain nearly intact. Clearly, heat and mass transfer inside the particle plays a crucial role in biomass conversion, and need to be taken carefully into account in CFD to correctly describe gasification and pyrolysis.

Some phenomena, similar to those briefly outlined above, can be found in other engineering applications in which CFD might be at a more advanced stage and, therefore, have been already studied extensively. Others are specific to biomass conversion, and the available work mostly concentrates on local, detailed modeling that has been rarely ported into CFD codes. Therefore, as illustrated in Figure 4, capturing the complex dynamics of the multiphase reacting flows encountered in a fluidized-bed gasifier requires combining a significant number of models from various backgrounds and levels of accuracy and identifying and handling potential interactions that may exist between these models. Next, we give a short description of each of the processes and existing models that can be applied in the context of fluidized beds.

Hydrodynamics of Particle-Laden Flows

Solid Phase Description

Gas-solid fluidized beds, characterized by their excellent mixing and heat transfer properties, are used in a wide variety of chemical and engineering industrial processes (*31*). The overall dynamic of these systems is dominated by

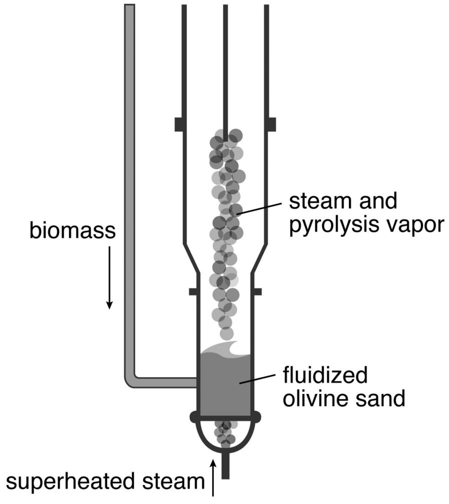


Figure 2. Schematic of a fluidized bed reactor

large-scale structures such as bubbles and recirculation regions along the walls, which, in turn, are controlled by interactions at the particle scale. Although robust and accurate numerical methods have been developed over the years for single-phase flows involving a gas or a liquid, such methods are still in the development stage for multiphase flows or flows involving a particulate phase. In a recent review of existing numerical models for gas-solid fluidized beds, Van der Hoef et al. (*32*) identified five major classes of methods based on the strategy, either Eulerian or Lagrangian, used to describe the gas phase and the solid phase, respectively. Among them, the two-fluid model (TFM) and the unresolved discrete particle model (DPM) have been the topics of a very large number of studies. Both approaches usually make the assumption of constant size, spherical particles.

TFM (*33*) follows a Euler-Euler approach that considers both phases as continuous interpenetrating media described using Navier-Stockes-type conservation equations. However, although easily implemented in pre-existing CFD codes and relatively computationally inexpensive, this method requires all processes at the particle scale, such as drag, collision and friction forces, and heterogeneous chemistry to be included as phase-interaction terms into the governing equations. Closures are obtained using the kinetic theory of granular flows (*34*) or experimental correlation, for example for the fluid-solid drag force (*33*, *35*).

Two-fluid methods can be extended for polydisperse, particle-laden flows involving particle size distributions and evolving particle sizes by using the more general and rigorous framework of the methods of moments. Equations are derived for a particle number density and its moments, conditioned on size, density or velocity (*36*). Although mostly applied to dilute solid-phase systems such as sprays (*37*), recent developments focused on the dense granular flows found in fluidized beds (*38*).

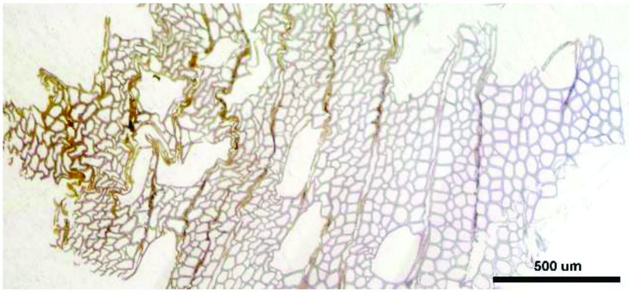


Figure 3. Structural changes in a heated biomass particle. (courtesy of Drs. T. Haas and B. Donohoe, National Renewable Energy Laboratory, Golden, CO)

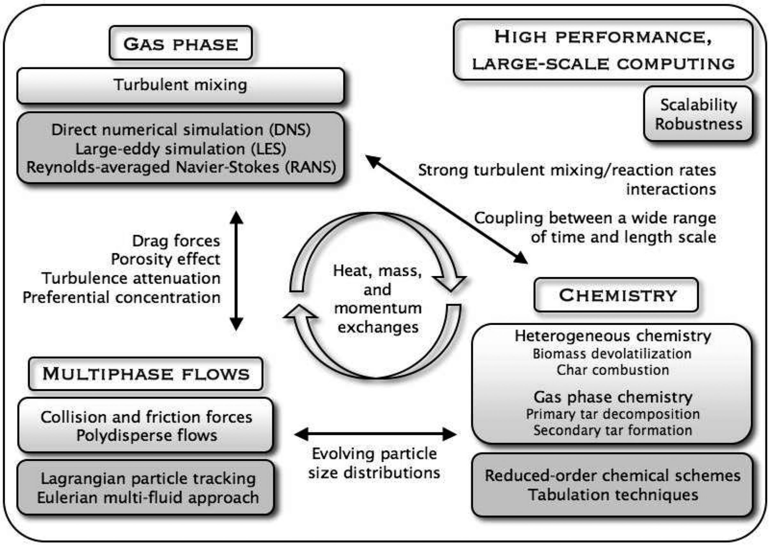

GAS PHASE
Turbulent mixing
Direct numerical simulation (DNS)
Large-eddy simulation (LES)
Reynolds-averaged Navier-Stokes (RANS)
HIGH PERFORMANCE, LARGE-SCALE COMPUTING
Scalability
Robustness
Strong turbulent mixing/reaction rates interactions
Coupling between a wide range of time and length scale
Drag forces
Porosity effect
Turbulence attenuation
Preferential concentration
Heat, mass, and momentum exchanges
CHEMISTRY
Heterogeneous chemistry
Biomass devolatilization
Char combustion
Gas phase chemistry
Primary tar decomposition
Secondary tar formation
MULTIPHASE FLOWS
Collision and friction forces
Polydisperse flows
Lagrangian particle tracking
Eulerian multi-fluid approach
Evolving particle size distributions
Reduced-order chemical schemes
Tabulation techniques

Figure 4. Modeling of physical and chemical processes interactions in biomass thermochemical conversion.

DPM methods, on the other hand, use Lagrangian particle tracking to describe the solid phase, in which each particle is transported following Newton's laws of motion (Figure 5). Particles are typically very small compared to the gas flow computational cells. Therefore, details of the flow around the particle are not resolved, and drag forces need to be modeled in a similar way as in TFM. However, collisions between particles are considered explicitly using,for example, hard-sphere (*39*) or soft-sphere (*40*, *41*) collision models. Porosity is taken into account by introducing the void fraction in the gas-phase governing equations. Due to the very large number of particles involved in fluidized beds, DPM approaches may become prohibitively expensive and are applied most often in two-dimensional configurations.

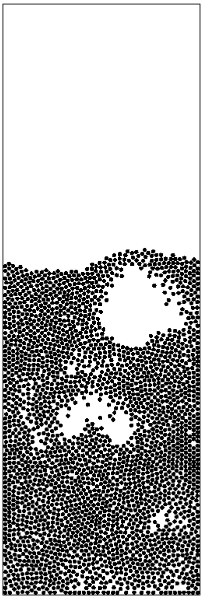


Figure 5. DPM simulation of bubble formation in a fluidized bed.

Gas Phase Coupling

In most practical applications, the gas flow is fully turbulent and must be treated adequately to obtain meaningful results. Mature techniques developed for aeronautics or internal engine flows may be extended to chemical reactors. Depending on the required resolution of the solution and the computational resources, Reynolds average equations (*42*) (RANS) or Large Eddy Simulation (*43*) (LES) may be used. The former solves for the mean quantities of the flow while modeling the fluctuations; whereas the latter can be seen as a filter operation, resolving the large, energy-containing scales and modeling the smallest scales. RANS methods are computationally affordable, as they usually require only a few additional evolution equations, compared to laminar configurations, for the kinetic energy and eddy dissipation rate. LES simulations are more expensive because they demand a higher grid resolution, but provide a much more accurate description of the unsteady, large-scale processes.

The interactions between gas and particles can have a significant impact on the turbulence intensity of the flow through, and increase the apparent gas viscosity (*44*). Precisely modeling the gas-particle couplings is important to accurately capture the mixing properties of the bed, which are expected to have a predominant impact on the overall reactor performances. A detailed review of the numerical aspects associated with gas-solid coupling in particle-laden flows is given in Curtis *et al.* (*45*).

Reactive Particles

To simulate biomass conversion, a description of the complex chemical processes needs to be combined with the above models for the hydrodynamics of fluidized-bed reactors. The chemistry of biomass conversion has been the topic of many experimental and modeling studies reviewed by Di Blasi (*46*) among others. Most studies focus on the heterogeneous reactions associated with biomass devolatilization. They often rely on kinetically controlled experiments providing weight mass loss as a function of time or temperature (*47–50*). The resulting data are used to derive kinetic rates for global schemes involving only a few representative compounds and reactive steps. Kinetic control, however, is not representative of the conditions found in realistic reactors and might even be difficult to ensure in the experiments (*46*). Instead, heat transfer to and inside the reactive particle by conduction, convection, and radiation is often the limiting process. However, the details of heat transfer inside the reactor are much more difficult to quantify experimentally. To understand heat transfer better, kinetic data can be coupled to one-dimensional partial differential equations describing the temperature evolution inside the particle. Most of these simulations are done on one single particle in an infinite domain, and the results largely depend on the kind of assumptions introduced (*51–54*). When more realistic configurations are considered, heat transfer is often modeled using correlations based on a few characteristic dimensionless numbers such as the Nusselt (ratio of conductive to convective heat transfer) and Sherwood (ratio of convective to diffusive mass transfer) numbers (*55*, *56*). Although very convenient to use, these models have been shown to be inadequate, especially for dense granular flows (*57*).

Comparatively few systematic studies into the nature or kinetics of gas-phase tar formation from biomass devolatilization products have been conducted. Taralas and Kontiominas (*58*) looked at secondary pyrolysis of vaporized unsaturated hydrocarbons in the presence of water vapor and oxygen using toluene and benzene as model compounds. Morf *et al.* (*59*) performed a mechanistic study of primary and secondary tar reactions and concluded that secondary tar reactions become important at temperatures higher than 650°C. These studies have just begun to address the important issue of tar formation, and more detailed work needs to be performed in this important area.

CFD Contribution to Process Design and Optimization

CFD appears as a cost-effective option to explore various configurations and operating conditions directly at the industrial scale to find the optimal configuration depending on the project specifications. However, CFD results are only valid if appropriate chemical and physical models are developed and validated using lab-scale experiments and robust numerical methods are used. In the case of indirect gasification, the optimal configuration corresponds to minimum tar content in the syngas; whereas in the case of pyrolysis, the tar output needs to be maximized. This translates into different operating temperatures and gas residence time inside the reactor. However, it must be emphasized that biomass pyrolysis and gasification in a fluidized-bed involve the same small-scale phenomena. Therefore, both types of conversion can be studied simultaneously and effectively using CFD, providing a versatility that cannot be found in experimental approaches. The advantages of using CFD in conjunction with well-designed validation experiments are given below, namely a global understanding of the large-scale impact of local phenomena, the identification of the most sensitive parameters, and guidelines for the industrial scale-up of new reactors.

Numerical Experiments

As mentioned earlier, a current limitation of biomass gasification in a fluidized-bed reactor is the high tar level in the product gas, causing fouling and hazardous waste. Costly cleaning treatments to control the tar content of the exiting gas are usually applied downstream. An appealing approach would be to directly optimize the gasifier design and operating conditions to the biomass properties for minimal tar production. Tar formation involves a lot of different steps occurring at different time and length scales. Primary tars are formed through biomass devolatilization inside the bed, which is highly sensitive to the local heat transfer between the sand and biomass particles and gas flow around the particles. Primary tar molecules are then transported across the reactor, where their decomposition and further reactions to form secondary tars depend on the temperature profile, residence time in the gas phase, and turbulence impact on reaction rates. Developing efficient control strategies requires understanding in detail each of these steps and how they are coupled with the larger scale flow phenomena happening in the reactor. However, the types of measurement that can be done on a full-scale reactor are considerably restricted, because visual access inside a three-dimensional fluidized-bed reactor is very limited, and the dense gas-solid mixture is difficult to sample without disruption. In most cases, only global time-averaged data are available, such as exit gas composition, flow rate, and temperature. To access internal details, some small-scale processes have been extensively studied in well-controlled environments. They include biomass devolatilization, non-reactive particle collisions, and gas-phase chemical reactions. CFD offers a unique opportunity to incorporate these fundamental results into a larger scale framework, in which process interactions can be studied.

Parameterization of the simulated configurations enables sensitivity analysis studies, normally an expensive and highly time-consuming process when performed experimentally. Sensitivity analysis should be conducted with two objectives in mind: identifying the limiting steps, and providing guidelines for further detailed modeling, as significant progress will be achieved if the most important and sensitive processes are correctly and accurately modeled first. Identifying the limiting processes and focusing on developing accurate models for them are crucial steps toward optimizing biomass gasification, as they will determine the exit gas composition and the amount of char left behind.

The CFD capabilities outlined above require that every component of the whole CFD model be extensively validated, which will be detailed in a later section.

Reactor Scale-Up

CFD models based on first principles should be able to predict accurately the gas-solid hydrodynamics at all scales, provided that a thorough validation of the code has been performed. A fundamental challenge faced when trying to develop new industrial-scale reactors based on laboratory results is the scale-up problem. Scaling laws based on conserving non-dimensionless numbers from one scale to another are often too simplistic or overly constraining to be reliable. The limited success of the scaling approach comes from the fact that the hydrodynamic of fluidized-bed reactors is not scale-similar because they involve scale-related phenomena, e.g., bubble dynamics and wall effects, which promote better mixing in small reactors (*60*). Instead, scale-up is usually done using one-dimensional models that incorporate simple descriptions of the mixing processes and heat transfer along the reactor. However, a pilot scale is often required between the lab and full-scale reactors, as these simplified models have little predictability (*60*). Minimizing or skipping pilot-plant validation of scale up through strategic use of CFD models would represent both capital cost and time-to-market savings. Also, the risk associated with developing a new industrial facility would be significantly reduced, making it more attractive to investors.

Several initial attempts have been made to use CFD to help with scale up. Lathouwers and Bellan (*61*) numerically studied the impact of reactor scale-up for biomass pyrolysis in a fluidized bed. Although no experimental validation was presented, they showed that increasing the bed size negatively impacted tar production and that shallow fluidized beds with high fluidization velocity had better scalability characteristics. Following a different approach, Van Ommen *et al.* (*62*) investigated the performances of different sets of scaling rules based on dimensionless numbers using CFD. They reported large variations in the void fraction and pressure data and noted that none of the scaling laws tested led to complete similarity between the two reactor sizes considered. No experimental validation was provided to corroborate the results. Even without experimental validation, these models suggest an interesting direction for pilot-scale work. In their review about fluidized-bed scale-up, Knowlton *et al.* (*60*) recognized that CFD is a promising approach, but given the state-of-the-art in numerical methods, experimental work is still required for successful scale up. With the right combination of computational advances and experimental validation, this may not always be the case. Limitations to that vision will be highlighted in the next section.

From Descriptive to Predictive CFD

The potential of CFD techniques is now widely acknowledged. However, the models are not advanced enough yet to be considered as a reliable and useful engineering tool for biomass gasification or pyrolysis. Two reasons for this are the lack of predictability and limited computational resources. In the following section, we give a brief overview of the different attempts to develop comprehensive CFD gasification and pyrolysis models, highlighting the strong points and shortcomings of each method. Then, the obstacles that need to be overcome to get truly predictive results will be discussed.

Existing Models

Comprehensive CFD simulations of biomass gasification or pyrolysis are scarce. Although the challenges faced in coal conversion are slightly different from those encountered in biomass conversion, the numerical frameworks remain close, and tools developed for one type of fuel are expected to be applicable for the other with only minor changes of the numerical methods required. Therefore, advances in both domains will be reported here.

Simplified Models

The pressing need for predictive tools that are more sophisticated than engineering correlations to optimize or scale up existing technologies has led to the development of numerous simplified models that incorporate most physical and chemical phenomena into a streamlined system of one-dimensional differential equations. These equations represent the evolution of the biomass particles as they move along the reactor. For example, Gobel *et al.* (*63*) developed a mathematical model for a fixed-bed coal gasifier that included conservation of mass and energy and reaction kinetics in the gas phase and char based on chemical equilibrium and Langmuir–Hinshelwood correlations. Individual components of the model were either based on first principles or determined from thermo-gravimetric experiments. The approach was validated and implemented in a real plant. A slightly more detailed approach was followed by Radmanesh *et al.* (*64*) to model a bubbling fluidized-bed reactor under isothermal conditions. Three stages were simulated successively. The various product yields of biomass pyrolysis, including a tar pseudo-component, served as initial conditions in the one-dimensional simulation of the bed using a countercurrent back-mixing model describing the evolution of the bubble (gas) and emulsion (mixture of gas and solids) phases. This was further combined with a gas-phase chemistry model describing the tar conversion to permanent gases and refractory tar. Maistrenko et al. (*65*) developed a one-dimensional unsteady model for polydisperse combustion of coal in a fluidized bed with heterogeneous chemistry involving only a few permanent gas species.

The conservation equations along the reactor axis may be coupled to a one-dimensional description of the evolution of the reactive particle, usually assumed to be spherically symmetric. An example of such pseudo two-dimensional differential systems is found in Luo et al. (*66*), who studied wood-flash pyrolysis in a fluidized-bed reactor. All variables were assumed to be homogeneous in the bed radial and azimutal directions. Very recently, Pierucci *et al.* (*54*) combined the semi-detailed chemistry model for biomass pyrolysis and gas-phase reactions of Ranzi et al. (*67*) with a one-dimensional model for moving bed gasifiers using a multi-zone description of the biomass particles. Validation was performed using gas composition at the exit of a lab-scale reactor, and results showed acceptable agreement between simulated and experimental data.

Although some of the models above have been used successfully in the context of real industrial-scale processes, clearly the assumptions and simplifications made in their development prevent them from being extended to different conditions and configurations, and they provide only limited insight on the details of the small-scale physical phenomena responsible for some major behavior such as the amount of tar in the exit gas. Therefore, a more detailed consideration of those phenomena is essential.

CFD in Realistic Configurations

The next step toward predictive CFD tools in various configurations is to rely on first principles and directly solve the coupled conservation equations for all variables. The increased resolution comes with an increase in computational cost, and most studies can afford only two-dimensional domains, even for cold-flow configurations in which no energy or species conservation equation is solved. Numerical simulations of cold fluidized beds are numerous. A two-fluid approach is employed in the recent papers on segregation in poly-disperse fluidized beds by Gera *et al.* (*68*), Huilin *et al.* (*69*), and Fan and Fox (*38*), or in the work of van *Wachem et al.* (*70*), Patil *et al.* (*71*, *72*), or Papadikis et al. (*73*) on closure formulation and validation. Others are based on a discrete particle approach (e.g., see Xu and Yu (*74*), Patankar *et al.* (*44*), Goldschmidt *et al.* (*75*) for the soft-sphere model; von Wachem *et al.* (*76*) for the hard-sphere model; Snider (*77*) for a volumetric approach to particle collisions; or the review paper of Deen *et al.* (*78*)). However, very few studies have tackled the coupled gas-solids hydrodynamic/reactive particle problem characteristic of biomass or coal thermochemical conversion systems. The following description of some of the most significant advances in this field gives a sense of the challenges still faced before CFD can become a predictable and reliable tool.

Fletcher et al. (*55*, *79*) performed a three-dimensional simulation of coal combustion in an entrained flow biomass gasifier using the CFX package (*80*). Although the entrained flow reactor only involves a dilute particle-laden flow where collisions have been neglected, the simulation includes most of the important processes found in biomass gasification in fluidized beds and is worth mentioning here. Biomass particles are tracked down the reactor using a Lagrangian approach. A particle-size distribution is prescribed, and the corresponding initial particle diameters are assumed to remain constant during the simulation. Chemistry is included using a global kinetic model involving only a few species: CH_4, H_2, CO, CO_2, H_2O, N_2, and O_2, but considers devolatilization, heterogeneous char conversion, and gas-phase chemistry. Gas-phase turbulence is modeled using a RANS approach, and the effect of turbulence on both the particles and the chemical reactions is taken into account. Very limited validation was carried out, with only the qualitative exit gas composition compared with experimental data.

Lathouwers and Bellan (*61*, *81*) proposed a comprehensive mathematical model to describe the dynamics of dense, reactive gas-solid mixtures and applied it to the simulation of biomass pyrolysis in a fluidized bed. Ensemble average equations are derived for each of the gas and various solids phases and appropriate closure models are formulated. Because the equations are derived using general moment methods, they do not require certain properties that are assumed in other two-fluid models such as equi-partition of granular energy among the particle classes. Chemistry is included in the form of the kinetic model for biomass pyrolysis developed by Miller and Bellan (*51*). The biomass particles are assumed to have a constant diameter throughout the simulation, while the porosity of the solids phase increases. No comparison with experimental data is presented; however, the qualitative response of the reactor in terms of tar yield as a function of parameters such as temperature is recovered. Parametric simulations are performed and demonstrate the ability of CFD models to identify optimal reactor operating conditions and assist in the scale-up process.

Zhou et al. (*56*, *82*) coupled a soft-sphere discrete particle method with LES to describe coal combustion in a bubbling fluidized-bed reactor. The effect of particles on sub-grid scale gas flow and the turbulent gas-particle interaction force were taken into account. The kinetic model for char combustion is based on a distributed activation energy approach function of the heating rate. NO_x formation being a major concern for combustion systems, nitrogen species are considered along with CO, CO_2, H_2O, and O_2. The burning char evolves following a shrinking-core approach, in which particle density remains constant while their diameter decreases as combustion proceeds. Particles are assumed to be isothermal. Parametric simulations were conducted to study the heating and subsequent combustion of the initially cold coal particles introduced in the hot sand bed. Excess temperatures of the burning particles were found to be in agreement with values reported in the literature. Also, the simulations showed that the presence of large reactive particles significantly affected the particle flow structure and that momentum and energy were mainly exchanged through collisions between particles rather than through the gas phase.

More recently, Yu et al. (*83*) developed a two-fluid model based on the kinetic theory of granular flow and coupled it with a multi-step chemistry model to study coal gasification in a two-dimensional bubbling fluidized-bed gasifier. Turbulence is captured using a k-ε RANS model, and the competition between kinetics and turbulent mixing is considered. The coal particles are assumed to be monodisperse spheres of constant density. Evolution of the major permanent gas species are included, as well as C_2H_6, C_6H_6, and H_2S. Calculated mole fractions of the major species in the exit gas were shown to be in good agreement with experimental data for different running conditions. A comparative summary of these studies is given in Table 1.

Table 1. Comparison of the models used in comprehensive CFD studies of biomass or coal conversion devices

Authors	*Lathouwers et al.*	*Zhou et al.*	*Yu et al.*	*Fletcher et al.*
Reference	*(61, 81)*	*(56, 82)*	*(83)*	*(55, 79)*
Fuel	Biomass	Coal	Coal	Biomass
Application	Pyrolysis in fluidized-bed	Combustion in fluidized-bed	Gasification in fluidized-bed	Gasification in downdraft gasifier
Dimensions	2	2	2	3
Multiphase	Eulerian	Lagrangian	Eulerian	Lagrangian
Turbulence	RANS	LES	RANS	RANS
Chemistry	Global, 3-component wood, tar, char, gas	Multi-step, NOx species	Multi-step, CO, CO_2, O_2, H_2O, H_2, CH_4, char	Multi-step, CO, CO_2, O_2, H_2O, H_2, CH_4, char
Validation with experiments	None	None	Major species composition of exit gas	Very limited (exit gas composition)

An attempt to integrate a more comprehensive description of the complex chemistry taking place in the gas phase into a CFD code is given in Gerun *et al.* (*84*). They studied the impact of tar formation on the temperature and velocity patterns in the oxidation zone of a two-stage downdraft gasifier using a semi-detailed mechanism for a tar model compound, phenol, in a two-dimensional axisymmetric domain. No solids were considered, and turbulence was modeled using a RANS approach coupled with an eddy dissipation model. A very partial comparison with experiments was done for the average gas temperature profile inside the reactor and tar concentration after the oxidation zone.

Model Verification and Validation

One of the major obstacles preventing a wider use of CFD tools at both the research and industrial level is the lack of thorough verification and validation of the existing models. Verification involves confirming the correct implementation of the model from a numerical point of view, while validation aims to assess the ability of the model to represent the actual physical process considered (*85*). Grace and Taghipour (*86*) provide a critical analysis of the current standards for fluidized-bed CFD model validation, highlighting the fact that virtually none of the existing models, although claimed to be validated with experimental data, have enough credibility to be applied beyond the model development stage. The validation process usually follows a hierarchical approach, with the building blocks of the CFD model being first tested in simple configurations involving a limited range of physical or chemical phenomena, or for which theory can provide analytical results. Examples of these simpler test cases include transient interface levels for batch sedimentation of particles (*87*), Reynolds number at minimum fluidization conditions to validate the drag force model (*88*), or hopper discharge rate to evaluate the particle friction model (*89*).

On a more global scale, most experimental data on the hydrodynamics of fluidized beds are obtained in pseudo-two-dimensional gas-fluidized beds (*41*, *75*, *76*, *90*, *91*). It must be noted that because of the chaotic nature of fluidized-bed reactors, only flow statistics are meaningful enough to compare between simulations and experiments. Images from these experiments are either used for visual qualitative comparison or post-treated to extract quantitative data such as average void fraction or porosity across the bed, bed height, and bubble average diameter and velocity. Additional measurements can be done for bed expansion (*92*, *93*) or pressure fluctuations (*70*). Fluidization and segregation has been studied extensively for binary mixtures (*94*) and continuous particle size distributions (*95*). Measures have been developed to quantify the extent of segregation in a fluidized bed (*93*). To measure the gas turbulence in the freeboard of a fluidized bed being induced by bubble bursting at the surface of the bed, Solimene *et al.* (*96*) recently developed laser diagnostics to study the evolution of vortices generated by a single bubble bursting at the surface of the bed. These experiments provide quantitative measurements of vortex displacement and concentration of a tracer species. As illustrated above, the hydrodynamics of a fluidized bed are experimentally characterized only on a global scale, often for simplified, two-dimensional systems. Gas and particle velocity fluctuations throughout the bed or three-dimensional measurements of the bubble dynamics are not available, nor are systematic studies over ranges of fluidization conditions, for non-spherical particles or evolving particle size and density distributions, both characteristic of biomass systems.

The situation is even more critical for reactive systems, for which very little detailed experimental data is available. Average exit gas flow rate and composition can be easily measured and compared to CFD results (*83*). Radmanesh *et al.* (*64*) measured the evolution of the gas composition along a fluidized-bed reactor during biomass gasification for various conditions in terms of bed temperature and equivalence ratio. Only total dry gas yield and major permanent gas species were provided. Van Paasen *et al.* (*97*) provide comprehensive tar measurements from biomass gasification in fixed and fluidized beds. However, gas samples were not taken inside the reactor, but after a cyclone, which might not be directly comparable with CFD simulations. Numerous studies of wood devolatilization have been conducted to measure devolatilization time and char yield for various feedstock, particle sizes and shapes in fluidized beds, by Sreekanth *et al.* (*98*), Di Blasi and Branca (*99*), Wang *et al.* (*100*) and de Diego *et al.* (*101*) for large wood cylinders or cuboids. Using one single model for devolatilization time, Sreekanth *et al.* (*98*) showed that those measurements were consistent with each other. Jand *et al.* (*102*) performed devolatilization of a finite number of smaller wood particles in a fluidized bed. Conditions were designed so that the devolatilization chemistry occurred inside the particle, therefore limiting the extra-particle and gas-phase conversion of volatiles to permanent gases and secondary tar. Char yield measurements, permanent gas compositions, and amount of tar in the exit gas were reported.

However, even if models describing multiple particles should remain an extension over single-particle models, the one-at-a-time particle feeding approach used in most of these experiments causes significant problems for CFD, preventing the corresponding experimental data from being used for rigorous model validation. On one hand, Eulerian techniques are not designed to handle a finite number of particles, as the average solid fraction on which the conservation equations are based becomes ill defined in the limit of very few particles. On the other hand, most Lagrangian methods assume a small particle diameter compared to the gas-phase computational grid. This assumption breaks down when large particles are gasified in moderate size fluidized beds. In that case, two-way coupling between gas and particles needs to be revised (*103*). Moreover, single particles injected on top of a bed can remain on top, sink to the bottom of the bed or be caught in a recirculation region. These different trajectories involve different types of physical processes and different heat transfer modes. Therefore, it is very unlikely that Euler-Lagrange CFD simulation of a single-particle gasification matches what happened in the experiments. A statistical treatment needs to be adopted, such as Monte-Carlo simulations, where the same process is simulated a large number of time from slightly different initial conditions (*104*). Also, very little is known about the evolution of biomass particle size, density, porosity and composition during gasification, even if those variables are expected to greatly impact both the hydrodynamics of the bed and products release as a function of time.

Another important aspect of CFD validation using reactive bed experimental data, such as exit gas composition, is the fact that these data are the results of a large number of different processes that cannot be distinguished from one another. Therefore, a good agreement for the output gas composition between a CFD simulation and some set of experimental data may be the result of error compensation. Again, a reasonable way to validate a CFD model is to proceed hierarchically from much simpler configurations involving only a few processes and considering a wide variety of experimental results. However, even this approach might prove difficult. For instance, many kinetic data on biomass devolatilization may not have been obtained under rigorous kinetic control, casting doubt on their validity (*46*).

Computational Resources and Modeling Challenges

Simulating industrial-size fluidized-bed reactors while considering all physical and chemical processes involved is a fantastic task, even supposing that all appropriate models have been developed. The multi-scale nature of the flow requires either a very fine mesh resolution or adequate sub-grid scale models. Cold Lagrangian simulations, originally developed for two-dimensional beds (*40*), have very rarely been applied to three-dimensional beds and usually involve only a few thousands particles, very far from the actual number found in realistic systems and not sufficient to derive good statistical data. To circumvent this computational limitation and move toward more realistic simulations, particles with similar characteristics may be grouped into parcels (*44*). Still, the problem is exacerbated by the wide particle-size distribution functions usually found in biomass samples.

To deal with particle-size distributions, Lagrangian methods require even more notional particles to correctly represent the distribution function, while Euler methods often consider separate conservation equations for each particle class. This approach quickly becomes complicated, as closure terms and interactions must be defined for each particle class and assumptions have to be made on particle velocity distributions (*105*). Simple binary mixtures were investigated and compared to experimental data using this approach (*106*). A very promising method called quadrature method of moments, or QMOM, represents the particle size distribution as a collection of weighted delta functions for which separate evolution equations can be solved (*107*). Particle mixing and segregation in a fluidized bed with a continuous particle size distribution were investigated and the results compared favorably with detailed DPM simulations (*38*).

Incorporating detailed chemistry into a CFD code is also very computationally expensive, as it introduces many additional conservation equations for each of the species considered, which in turn requires lengthy evaluation of the chemical source terms and may lead to space and time scales much smaller than those of the hydrodynamic processes. Once reactive time-scales dominate, the time steps and grid spacing are dramatically restricted. Virtually all CFD studies of reactive fluidized beds have considered only global conversion steps or multi-step kinetic schemes involving only a few major species. Very recently, Ranzi et al. (*67*) proposed the first semi-detailed kinetic model for biomass conversion containing several hundreds species and reactions. This model employed a lumped approach and a limited number of model compounds to describe biomass devolatilization where detailed mechanistic relationships are developed to describe the decomposition of the model compounds. However, the size of such mechanisms prevents them from being used directly in CFD. Several approaches to handle complex chemistry in CFD simulations have been developed in the context of hydrocarbon combustion systems that could potentially be transferred to gasification processes. The first one is to reduce *a priori* the detailed mechanism to a smaller size in terms of species and reactions based on homogeneous simulation results, so that only the chemical pathways relevant to the current study are retained (*108*). Then, computing the chemical source term can be optimized and accelerated using an on-the-fly storage and retrieval technique called in-situ adaptive tabulation, or ISAT (*109*). In each computational cell, the chemical source term and corresponding chemical state are stored as the simulation proceeds and are re-used instead of re-computed if a similar chemical configuration is found later in the simulation. The applicability of the method for reactive fluidized beds was demonstrated by Xie et al. (*110*) for silane pyrolysis.

Another tabulation technique consists of solving the evolution equations for a few variables only, chosen so that these variables map the entire chemical composition space with good accuracy. Detailed chemistry is tabulated *a priori* using these variables, which drastically reduces computational time. Tabulation methods have been developed and extensively validated for gas phase combustion systems (*111*), in which a mixture fraction linked to the local ratio of fuel and oxidizer, or a progress variable describing the local extent of the global combustion reaction are easily defined. Such techniques cannot be directly applied to gasification processes, but must be adapted to the specific chemistry occurring in the fluidized bed.

One of the major obstacles to predictive CFD capabilities of biomass gasification or pyrolysis is the intrinsic complexity of the biomass itself. Biomass composition varies widely depending on the feedstock, age, geographic location, and even time of year. It was also shown repeatedly that biomass is not just the sum of its major components; lignin, cellulose, and hemicellulose (*46*). Biomass particles are highly anisotropic (*112*), contain trace components such as metal that can act as chemical catalysts (*113*) and moisture that may delay gasification (*102*). Biomass gasification is, therefore, a perfect example of a multi-scale problem, in which phenomena localized at the smallest scales are responsible for large-scale behavior. Reactor optimization requires these processes to be fully understood and characterized. Some detailed modeling studies at the scale of the biomass or coal particle take into account their complex structures (*114*, *115*). Such investigations, coupled with experimental observations, are essential to develop larger scale statistical models suitable for use in CFD, because it is not currently conceivable to include that amount of detail in large-scale simulations.

Conclusions

Computational fluid dynamics methods, as an essential complement to experimental investigations, has a considerable potential to help meet our present and future needs for efficient energy production and conversion systems. However, we have shown that a lot of challenges still lay ahead in the near future for biomass thermochemical conversion processes to obtain reliable, predictive simulations that can be used as stand-alone tools for reactor design and optimization. Many areas of biomass conversion modeling can benefit from advances in other fields, such as combustion systems, in which CFD is a very active research topic. However, the modeling of biomass conversion systems combines two of perhaps the most challenging aspects of CFD, namely dense particulate flows and a complex and very specific chemistry. Until now, biomass chemical models have been based on global, over-simplified mechanisms obtained through experimental correlations. Because detailed and accurate chemistry is a key element in understanding and controling output efficiency and tar formation, only significant progress in this area, coupled with major advances in numerical methods for multiphase flows and viable verification and validation strategies, will enable us to seize the opportunities provided by the ever-growing computational resources.

References

1. Sandalow, D. *Freedom from oil: How the next president can end the United States' oil addiction*; McGraw Hill: 2008.

2. Trostle, R. *Global Agricultural Supply and Demand: Factors Contributing to the Recent Increase in Food Commodity Prices*; WRS-0801; USDA – A Report from the Economic Research Service, May 2008.
3. *US DOE Biomass Multi-Year Program Plan*; Office of the Biomass Program, Energy Efficiency and Renewable Energy, U.S. DOE; March 2008.
4. Foust, T. D.; Wallace, R.; Wooley, R.; Sheehan, J.; Ibsen, K.; Dayton, D.; Himmel, M.; Ashworth, J.; McCormick, R.; Hess, J. R.; Wright, C.; Radtke, C.; Perlack, R.; Mielenz, J.; Wang, M.; Synder, S.; Werpy, T. *A National Laboratory Market and Technology Assessment of the 30 X 30 Scenario*; Technical Report, NREL/TP-510-40942; March 2007.
5. Mitchell, D. *A Note on Rising Food Prices*; Policy Research Working Paper 4682; The World Bank Development Prospects Group, July 2008.
6. Farrell, A. E.; Plevin, R. J.; Turner, B. T.; Jones, A. D.; O'Hare, M.; Kanman, D. M. Ethanol can contribute to energy and environmental goals. *Science* **2006**, *311*, 506–509.
7. Rammamorth, R.; Kastury, S.; Smith, W. H. *Bioenergy: Vision for the new millennium*. Science Publishers: Enfield, NH, 2000.
8. Huber, G. W.; Iborra, S.; Corma, A. Synthesis of transportation fuels from biomass: Chemistry, catalysts, and engineering. *Chem. Rev.* **2006**, *106*, 4044–4098.
9. Foust, T. D.; Ibsen, K. I.; Dayton, D. C.; Hess, J. R.; Kenney, K. E. Chapter 2: The Biorefinery. In *Biomass recalcitrance, deconstructing the plant cell wall for bioenergy*; Himmel, M. E., Ed.; Blackwell Publishing: 2008.
10. Wright, M. M.; Brown, R. C. Comparative economics of biorefineries based on the biochemical and thermochemical platforms. *Biofuels, Bioprod. Biorefin.* **2007**, *1*, 49–56.
11. McKendry, P. Energy production from biomass (part 2): conversion technologies. *Bioresour. Technol.* **2002**, *83* (1), 47–54.
12. Spath, P. L.; Dayton, D. C. *Preliminary Screening – Technical and economic assessment of synthesis gas to fuels and chemicals with emphasis on the potential for biomass-derived syngas*; NREL/TP-510-34929; National Renewable Energy Laboratory: Golden, CO, 2003.
13. Czernik, S.; Bridgwater, A. V. Overview of applications of biomass fast pyrolysis oil. *Energy Fuels* **2004**, *18* (2), 590–598.
14. Ciferno, J. P.; Marano, J. J. *Benchmarking biomass gasification technologies for fuels, chemicals, and hydrogen production*; U.S. Department of Energy/ National Energy Technology Laboratory (NETL) Report; June 2002.
15. Higman, C.; van der Burgt, M. Chapter 5: Gasification Processes. *Gasification*; Elsevier: 2003; pp 85−170.
16. Demirbas, A. Progress and recent trends in biofuels. *Prog. Energy Combust. Sci.* **2007**, *33* (1), 1–18.
17. Hamelinck, C. N.; Faaij, A. P. C. Outlook for advanced biofuels. *Energy Policy* **2006**, *34* (17), 3268–3283.
18. Feldmann, H. F.; Paisley, M. A.; Appelbaum, H. R.; Taylor, D. R. *Conversion of Forest Residues to a Methane-Rich Gas in a High-Throughput Gasifier*; PNL--6570; Pacific Northwest National Laboratory (formerly Pacific Northwest Laboratory), Department of Energy, 1988.

19. Herguido, J.; Corella, J.; Gonzalezsaiz, J. Steam gasification of lignocellulosic residues in a fluidized-bed at a small pilot scale - Effect of the type of feedstock. *Ind. Eng. Chem. Res.* **1992**, *31* (5), 1274–1282.
20. Hofbauer, H.; Rauch, R. In *Stoichiometric Water Consumption of Steam Gasification by the FICFB-Gasification Process*; Biomass Conversion, Proceedings of the International Conference, Innsbruck, Austria, 2000; Bridgwater, A. V., Ed.; Blackwell Science, Ltd.: Innsbruck, Austria, 2000; pp 199−208.
21. Franco, C.; Pinto, F.; Gulyurtlu, I.; Cabrita, I. The study of reactions influencing the biomass steam gasification process. *Fuel* **2003**, *82* (7), 835–842.
22. Wei, L. G.; Xu, S. P.; Liu, J. G.; Lu, C. L.; Liu, S. Q.; Liu, C. H. A novel process of biomass gasification for hydrogen-rich gas with solid heat carrier: Preliminary experimental results. *Energy Fuels* **2006**, *20* (5), 2266–2273.
23. Matsuoka, K.; Kuramoto, K.; Murakami, T.; Suzuki, Y. Steam gasification of woody biomass in a circulating dual bubbling fluidized bed system. *Energy Fuels* **2008**, *22* (3), 1980–1985.
24. Pecho, J.; Schildhauer, T. J.; Sturzenegger, A.; Biollaz, S.; Wokaun, A. Reactive bed materials for improved biomass gasification in a circulating fluidized bed reactor. *Chem. Eng. Sci.* **2008**, *63* (9), 2465–2476.
25. Tasaka, K.; Furusawa, T.; Tsutsumi, A. Biomass gasification in fluidized bed reactor with Co catalyst. *Chem. Eng. Sci.* **2007**, *62* (18−20), 5558–5563.
26. Pfiefer, C.; Hofbauer, H. Development of catalytic tar decomposition downstream from a dual fluidized bed biomass steam gasifier. *Powder Technol.* **2008**, *180* (1−2), 9–16.
27. Wang, T. J.; Chang, J.; Wu, C. Z.; Fu, Y.; Chen, Y. The steam reforming of naphthalene over a nickel-dolomite cracking catalyst. *Biomass Bioenergy* **2005**, *28* (5), 508–514.
28. Johnson, F. T.; Tinoco, E. N.; Yu, N. J. Thirty years of development and application of CFD at Boeing Commercial Airplanes, Seattle. *Comput. Fluids* **2005**, *34* (10), 1115–1151.
29. Kwak, D.; Kiris, C. CFD for incompressible flows at NASA Ames. *Comput. Fluids* **2009**, *38* (3), 504–510.
30. Pitsch, H.; Desjardins, O.; Balarac, G.; Ihme, M. Large-eddy simulation of turbulent reacting flows. *Progress in Aerospace Sciences* **2008**, *44* (6), 466–478.
31. Kunii, D.; Levenspiel, O. *Fluidization Engineering*; Butterworth-Heinemann: 1991.
32. van der Hoef, M. A.; Annaland, M. V.; Deen, N. G.; Kuipers, J. A. M. Numerical simulation of dense gas-solid fluidized beds: A multiscale modeling strategy. *Annu. Rev. Fluid Mech.* **2008**, *40*, 47–70.
33. Syamlal, M.; Rogers, W.; O'Brien, T. J. *MFIX Documentation: Theory Guide*; U.S. Department of Energy: Morgantown, WV, 1993.
34. Ding, J.; Gidaspow, D. A bubbling fluidization model using kinetic-theory of granular flow. *AIChE J.* **1990**, *36* (4), 523–538.
35. Wen, C. Y.; Yu, Y. H. Mechanics of fluidization. *Chem. Eng. Prog. Symp. Ser.* **1966**, *62*, 100–111.

36. Diemer, R. B.; Olson, J. H. A moment methodology for coagulation and breakage problems: Part 2 - Moment models and distribution reconstruction. *Chem. Eng. Sci.* **2002**, *57* (12), 2211–2228.
37. Desjardins, O.; Fox, R. O.; Villedieu, P. A quadrature-based moment method for dilute fluid-particle flows. *J. Comput. Phys.* **2008**, *227* (4), 2514–2539.
38. Fan, R.; Fox, R. O. Segregation in polydisperse fluidized beds: Validation of a multi-fluid model. *Chem. Eng. Sci.* **2008**, *63* (1), 272–285.
39. Hoomans, B. P. B.; Kuipers, J. A. M.; Briels, W. J.; vanSwaaij, W. P. M. Discrete particle simulation of bubble and slug formation in a two-dimensional gas-fluidised bed: A hard-sphere approach. *Chem. Eng. Sci.* **1996**, *51* (1), 99–118.
40. Cundall, P. A.; Strack, O. D. L. Discrete numerical model for granular assemblies. *Geotechnique* **1979**, *29* (1), 47–65.
41. Tsuji, Y.; Kawaguchi, T.; Tanaka, T. Discrete particle simulation of 2-dimensional fluidized-bed. *Powder Technol.* **1993**, *77* (1), 79–87.
42. Pope, S. B. *Turbulent Flows*; Cambridge University Press: 2000.
43. Ghosal, S.; Moin, P. The basic equations for the large-eddy simulation of turbulent flows in complex geometries. *J. Comput. Phys.* **1995**, *118* (1), 24–37.
44. Patankar, N. A.; Joseph, D. D. Modeling and numerical simulation of particulate flows by the Eulerian-Lagrangian approach. *Int. J. Multiphase Flow* **2001**, *27* (10), 1659–1684.
45. Curtis, J. S.; van Wachem, B. Modeling particle-laden flows: A research outlook. *AIChE J.* **2004**, *50* (11), 2638–2645.
46. Di Blasi, C. Modeling chemical and physical processes of wood and biomass pyrolysis. *Prog. Energy Combust. Sci.* **2008**, *34* (1), 47–90.
47. Yang, H. P.; Yan, R.; Chen, H. P.; Lee, D. H.; Zheng, C. G. Characteristics of hemicellulose, cellulose and lignin pyrolysis. *Fuel* **2007**, *86* (12−13), 1781–1788.
48. Antal, M. J.; Varhegyi, G.; Jakab, E. Cellulose pyrolysis kinetics: Revisited. *Ind. Eng. Chem. Res.* **1998**, *37* (4), 1267–1275.
49. Branca, C.; Di Blasi, C. Kinetics of the isothermal degradation of wood in the temperature range 528-708 K. *J. Anal. Appl. Pyrolysis* **2003**, *67* (2), 207–219.
50. Radmanesh, R.; Courbariaux, Y.; Chaouki, J.; Guy, C. A unified lumped approach in kinetic modeling of biomass pyrolysis. *Fuel* **2006**, *85* (9), 1211–1220.
51. Miller, R. S.; Bellan, J. A generalized biomass pyrolysis model based on superimposed cellulose, hemicellulose and lignin kinetics. *Combust. Sci. Technol.* **1997**, *126* (1−6), 97–137.
52. Galgano, A.; Di Blasi, C. Modeling wood degradation by the unreacted-core-shrinking approximation. *Ind. Eng. Chem. Res.* **2003**, *42* (10), 2101–2111.
53. Saastamoinen, J. J. Simplified model for calculation of devolatilization in fluidized beds. *Fuel* **2006**, *85* (17−18), 2388–2395.
54. Pierucci, S.; Ranzi, E. In *A general mathematical model for a moving bed gasifier*; 18th European Symposium on Computer Aided Process Engineering - ESCAPE 18, 2008; Elsevier B. V./Ltd.: 2008.

55. Fletcher, D. F.; Haynes, B. S.; Christo, F. C.; Joseph, S. D. A CFD based combustion model of an entrained flow biomass gasifier. *Applied Mathematical Modelling* **2000**, *24* (3), 165–182.
56. Zhou, H. S.; Flamant, G.; Gauthier, D. DEM-LES simulation of coal combustion in a bubbling fluidized bed Part II: coal combustion at the particle level. *Chem. Eng. Sci.* **2004**, *59* (20), 4205–4215.
57. Garg, R.; Tenneti, S.; Pai, M.; Subramanian, S., Heat transfer in ordered and random arrays of spheres at low Reynolds number. In *61st Annual Meeting of the APS Division of Fluid Dynamics*, San Antonio, TX, 2008.
58. Taralas, G.; Kontominas, M. G. Numerical modeling of tar species/VOC dissociation for clean and intelligent energy production. *Energy Fuels* **2005**, *19* (1), 87–93.
59. Morf, P.; Hasler, P.; Nussbaumer, T. Mechanisms and kinetics of homogeneous secondary reactions of tar from continuous pyrolysis of wood chips. *Fuel* **2002**, *81*, 843–853.
60. Knowlton, T. M.; Karri, S. B. R.; Issangya, A. Scale-up of fluidized-bed hydrodynamics. *Powder Technol.* **2005**, *150* (2), 72–77.
61. Lathouwers, D.; Bellan, J. Modeling of dense gas-solid reactive mixtures applied to biomass pyrolysis in a fluidized bed. *Int. J. Multiphase Flow* **2001**, *27* (12), 2155–2187.
62. van Ommen, J. R.; Teuling, M.; Nijenhuis, J.; van Wachem, B. G. M. Computational validation of the scaling rules for fluidized beds. *Powder Technol.* **2006**, *163* (1-2), 32–40.
63. Gobel, B.; Henriksen, U.; Jensen, T. K.; Qvale, B.; Houbak, N. The development of a computer model for a fixed bed gasifier and its use for optimization and control. *Bioresour. Technol.* **2007**, *98* (10), 2043–2052.
64. Radmanesh, R.; Chaouki, J.; Guy, C. Biomass gasification in a bubbling fluidized bed reactor: Experiments and modeling. *AIChE J.* **2006**, *52* (12), 4258–4272.
65. Maistrenko, A. Y.; Patskkov, V. P.; Topal, A. I.; Patskova, T. V. Numerical analysis of the process of combustion and gasification of the polydisperse coke residue of high-ash coal under pressure in a fluidized bed. *Journal of Engineering Physics and Thermophysics* **2007**, *80* (5).
66. Luo, Z. Y.; Wang, S. R.; Cen, K. F. A model of wood flash pyrolysis in fluidized bed reactor. *Renewable Energy* **2005**, *30* (3), 377–392.
67. Ranzi, E.; Cuoci, A.; Faravelli, T.; Frassoldati, A.; Migliavacca, G.; Pierucci, S.; Sommariva, S. Chemical Kinetics of Biomass Pyrolysis. *Energy and Fuels* **2008**in press.
68. Gera, D.; Syamlal, M.; O'Brien, T. J. Hydrodynamics of particle segregation in fluidized beds. *Int. J. Multiphase Flow* **2004**, *30* (4), 419–428.
69. Lu, H. L.; Zhao, Y. H.; Ding, J. M.; Gidaspow, D.; Wei, L. Investigation of mixing/segregation of mixture particles in gas-solid fluidized beds. *Chem. Eng. Sci.* **2007**, *62*, 301–317.
70. van Wachem, B. G. M.; Schouten, J. C.; van den Bleek, C. M.; Krishna, R.; Sinclair, J. L. Comparative analysis of CFD models of dense gas-solid systems. *AIChE J.* **2001**, *47* (5), 1035–1051.

71. Patil, D. J.; Annaland, M. V.; Kuipers, J. A. M. Critical comparison of hydrodynamic models for gas-solid fluidized beds - Part I: bubbling gas-solid fluidized beds operated with a jet. *Chem. Eng. Sci.* **2005**, *60* (1), 57–72.
72. Patil, D. J.; Annaland, A. V.; Kuipers, J. A. M. Critical comparison of hydrodynamic models for gas-solid fluidized beds - Part II: freely bubbling gas-solid fluidized beds. *Chem. Eng. Sci.* **2005**, *60* (1), 73–84.
73. Papadikis, K.; Bridgwater, A. V.; Gu, S. CFD modelling of the fast pyrolysis of biomass in fluidised bed reactors, Part A: Eulerian computation of momentum transport in bubbling fluidised beds. *Chem. Eng. Sci.* **2008**, *63* (16), 4218–4227.
74. Xu, B. H.; Yu, A. B. Numerical simulation of the gas-solid flow in a fluidized bed by combining discrete particle method with computational fluid dynamics. *Chem. Eng. Sci.* **1997**, *52* (16), 2785–2809.
75. Goldschmidt, M. J. V.; Beetstra, R.; Kuipers, J. A. M. Hydrodynamic modelling of dense gas-fluidised beds: comparison and validation of 3D discrete particle and continuum models. *Powder Technol.* **2004**, *142* (1), 23–47.
76. van Wachem, B. G. M.; van der Schaaf, J.; Schouten, J. C.; Krishna, R.; van den Bleek, C. M. Experimental validation of Lagrangian-Eulerian simulations of fluidized beds. *Powder Technol.* **2001**, *116* (2-3), 155–165.
77. Snider, D. M. An Incompressible Three-Dimensional Multiphase Particle-in-Cell Model for Dense Particle Flows. *Journal of Computational Physics* **2001**, *170*, 523–549.
78. Deen, N. G.; Annaland, M. V.; Van der Hoef, M. A.; Kuipers, J. A. M. Review of discrete particle modeling of fluidized beds. *Chem. Eng. Sci.* **2007**, *62* (1-2), 28–44.
79. Fletcher, D. F.; Haynes, B. S.; Chen, J.; Joseph, S. D. Computational fluid dynamics modelling of an entrained flow biomass gasifier. *Applied Mathematical Modelling* **1998**, *22* (10), 747–757.
80. CFX. *Flow solver user guide*; CFX International, AEA Technology, Harwell Laboratory: Didcot, Oxfordshire, U.K., 1996.
81. Lathouwers, D.; Bellan, J. Yield optimization and scaling of fluidized beds for tar production from biomass. *Energy Fuels* **2001**, *15* (5), 1247–1262.
82. Zhou, H. S.; Flamant, G.; Gauthier, D. DEM-LES of coal combustion in a bubbling fluidized bed. Part I: gas-particle turbulent flow structure. *Chem. Eng. Sci.* **2004**, *59* (20), 4193–4203.
83. Yu, L.; Lu, J.; Zhang, X. P.; Zhang, S. J. Numerical simulation of the bubbling fluidized bed coal gasification by the kinetic theory of granular flow (KTGF). *Fuel* **2007**, *86* (5−6), 722–734.
84. Gerun, L.; Paraschiv, M.; Vijeu, R.; Bellettre, J.; Tazerout, M.; Gobel, B.; Henriksen, U. Numerical investigation of the partial oxidation in a two-stage downdraft gasifier. *Fuel* **2008**, *87* (7), 1383–1393.
85. AIAA. *Guide for the verification and validation of computational fluid dynamic simulations*; American Institute of Aeronautics and Astronautics: 1998.

86. Grace, J. R.; Taghipour, F. Verification and validation of CFD models and dynamic similarity for fluidized beds. *Powder Technol.* **2004**, *139* (2), 99–110.

87. Davis, R. H.; Herbolzhiemer, E.; Acrivos, A. The sedimentation of polydisperse suspensions in vessels having inclined walls. *Int. J. Multiphase Flow* **1982**, *8* (6), 571–585.

88. Syamlal, M.; O'Brien, T. J. *The Derivation of a Drag Coefficient Formula from Velocity-Voidage Correlations*; 1987.

89. Benyahia, S. Validation Study of Two Continuum Granular Frictional Flow Theories. *Ind. Eng. Chem. Res.* **2008**, *47* (22), 8926–8932.

90. Kuipers, J. A. M. A two-fluid micro balance model of fluidized beds. University of Twente: Twente, 1990.

91. Boemer, A.; Qi, H.; Renz, U. Verification of Eulerian simulation of spontaneous bubble formation in a fluidized bed. *Chem. Eng. Sci.* **1998**, *53* (10), 1835+.

92. Hilligardt, K.; Werther, J. Local bubble-gas hold-up and expansion behavior of gas-solid fluidized beds. *Chem. Ing. Tech.* **1985**, *57* (7), 622–623.

93. Goldschmidt, M. J. V.; Link, J. M.; Mellema, S.; Kuipers, J. A. M. Digital image analysis measurements of bed expansion and segregation dynamics in dense gas-fluidised beds. *Powder Technol.* **2003**, *138* (2−3), 135–159.

94. Gilbertson, M. A.; Eames, I. Segregation patterns in gas-fluidized systems. *J. Fluid Mech.* **2001**, *433*, 347–356.

95. Grace, J. R.; Sun, G. Influence of particle-size distribution on the performance of fluidized-bed reactors. *Can. J. Chem. Eng.* **1991**, *69* (5), 1126–1134.

96. Solimene, R.; Marzocchella, A.; Ragucci, R.; Salatino, P. Laser diagnostics of hydrodynamics and gas-mixing induced by bubble bursting at the surface of gas-fluidized beds. *Chem. Eng. Sci.* **2007**, *62* (1−2), 94–108.

97. van Paasen, S. V. B.; Kiel, J. H. A. *Tar formation in a fluidised-bed gasifier - Impact of fuel properties and operating conditions*; 2004.

98. Sreekanth, M.; Sudhakar, D. R.; Prasad, B.; Kolar, A. K.; Leckner, B. Modelling and experimental investigation of devolatilizing wood in a fluidized bed combustor. *Fuel* **2008**, *87* (12), 2698–2712.

99. Di Blasi, C.; Branca, C. Temperatures of wood particles in a hot sand bed fluidized by nitrogen. *Energy Fuels* **2003**, *17* (1), 247–254.

100. Wang, X. Q.; Kersten, S. R. A.; Prins, W.; van Swaaij, W. P. M. Biomass pyrolysis in a fluidized bed reactor. Part 2: Experimental validation of model results. *Ind. Eng. Chem. Res.* **2005**, *44* (23), 8786–8795.

101. de Diego, L. F.; Garcia-Labiano, F.; Abad, A.; Gayan, P.; Adanez, J. Modeling of the devolatilization of nonspherical wet pine wood particles in fluidized beds. *Ind. Eng. Chem. Res.* **2002**, *41* (15), 3642–3650.

102. Jand, N.; Foscolo, P. U. Decomposition of wood particles in fluidized beds. *Ind. Eng. Chem. Res.* **2005**, *44* (14), 5079–5089.

103. Link, J. M.; Cuypers, L. A.; Deen, N. G.; Kuipers, J. A. M. Flow regimes in a spout-fluid bed: A combined experimental and simulation study. *Chem. Eng. Sci.* **2005**, *60* (13), 3425–3442.

104. Landau, D. P.; Binder, K. *A guide to Monte Carlo simulations in statistical physics*; Cambridge University Press: 2005.

105. Mathiesen, V.; Solberg, T.; Hjertager, B. H. Predictions of gas/particle flow with an Eulerian model including a realistic particle size distribution. *Powder Technol.* **2000**, *112* (1−2), 34–45.

106. Lu, H. L.; He, Y. R.; Dimitri, G.; Yang, L. D.; Qin, Y. K. Size segregation of binary mixture of solids in bubbling fluidized beds. *Powder Technol.* **2003**, *134* (1-2), 86–97.

107. Fan, R.; Marchisio, D. L.; Fox, R. O. Application of the direct quadrature method of moments to polydisperse gas-solid fluidized beds. *Powder Technol.* **2004**, *139* (1), 7–20.

108. Pepiot-Desjardins, P.; Pitsch, H. An efficient error-propagation-based reduction method for large chemical kinetic mechanisms. *Combust. Flame* **2008**, *154* (1-2), 67–81.

109. Pope, S. B. Computationally efficient implementation of combustion chemistry using in situ adaptive tabulation. *Combust. Theory Modell.* **1997**, *1* (1), 41–63.

110. Xie, N.; Battaglia, F.; Fox, R. O. Simulations of multiphase reactive flows in fluidized beds using in situ adaptive tabulation. *Combust. Theory Modell.* **2004**, *8* (2), 195–209.

111. Pierce, C. D.; Moin, P. Progress-variable approach for large-eddy simulation of non-premixed turbulent combustion. *J. Fluid Mech.* **2004**, *504*, 73–97.

112. Chan, W. C. R.; Kelbon, M.; Krieger, B. B. Modeling and experimental verification of physical and chemical processes during pyrolysis of a large biomass particle. *Fuel* **1985**, *64* (11), 1505–1513.

113. NikAzar, M.; Hajaligol, M. R.; Sohrabi, M.; Dabir, B. Mineral matter effects in rapid pyrolysis of beech wood. *Fuel Process. Technol.* **1997**, *51* (1−2), 7–17.

114. Mermoud, F.; Golfier, F.; Salvador, S.; Van de Steene, L.; Dirion, J. L. Experimental and numerical study of steam gasification of a single charcoal particle. *Combust. Flame* **2006**, *145* (1−2), 59–79.

115. Yamashita, T.; Fujii, Y.; Morozumi, Y.; Aoki, H.; Miura, T. Modeling of gasification and fragmentation behavior of char particles having complicated structures. *Combust. Flame* **2006**, *146* (1−2), 85–94.

Chapter 13

New Methods To Find Accurate Reaction Coordinates by Path Sampling

Gregg T. Beckham[1,*] and Baron Peters[2,3,*]

[1]National Bioenergy Center, National Renewable Energy Laboratory, Golden, CO 80401
[2]Department of Chemical Engineering, University of California, Santa Barbara, CA 93106
[3]Department of Chemistry and Biochemistry, University of California, Santa Barbara, CA 93106
[*]gregg.beckham@nrel.gov; baronp@engineering.ucsb.edu

Complex, high-dimensional systems are often characterized by dynamical bottlenecks, or rare events, that determine the rate of evolution of a given system. As the transition states through the dynamical bottlenecks are often difficult to capture experimentally, theory and computation are useful tools to elucidate transition states. This review describes a set of computational methods that enable the rigorous determination of mechanisms, free energy barriers, and rate constants for activated processes in complex, high-dimensional systems. The transition path sampling method for sampling reactive pathways and a subsequent methodological development, aimless shooting, are reviewed. Likelihood maximization, which is a method to extract the reaction coordinate of an activated process from path sampling data, is discussed in detail. In addition, the equilibrium path sampling approach and the earlier BOLAS approach for determining free energy barriers are examined. These techniques offer a means to access kinetically meaningful results from molecular simulation of activated processes in complex systems.

1. Introduction

For activated processes, or those that must overcome a free energy barrier to occur, systems of interest typically fluctuate for long times in metastable basins before passing through dynamical bottlenecks, or transition states. Transition states are often fleeting and therefore difficult to capture experimentally. However, details of these transition states often yield clues for rational design of effective catalysts (agents that reduce free energy barriers and thereby increase the rate of interest) or inhibitors (agents that increase free energy barriers and thereby reduce the rate of interest). Theory and computation allow researchers to test molecular-level hypotheses by characterizing dynamical bottlenecks in activated processes, thus representing essential tools in understanding and modifying activated processes.

Many computational approaches have been developed to understand the mechanisms of activated processes. For systems with few degrees of freedom, such as chemical reactions of small molecules in the gas phase, saddle-point methods have proven especially powerful as saddle points on the smooth potential energy surface typically reveal the mechanism (*1–11*). Conversely, for activated processes that involve many degrees of freedom, such as conformational changes in proteins and first-order phase transitions in molecular systems, the energy landscapes are “rough”, making saddle point methods unsuitable.

To address the issue of finding mechanisms on rough energy landscapes, Chandler and others pioneered the method of transition path sampling (TPS). TPS is designed to efficiently harvest pathways of activated processes without the need to assume a mechanism *a priori* (*12–18*). Since many research questions in both the construction and conversion processes of the plant cell wall, the subject of this book, are concerned with solvated biological systems (*e.g.*, hydrolysis of crystalline cellulose by enzymes) as well as solvated chemical systems (*e.g.*, chemical pretreatment of lignocellulosic biomass), both of which involve rough energy landscapes, the objective of this review is to discuss path sampling approaches to extract mechanisms and calculate free energy barriers in complex, molecular systems. The tools outlined here, primarily transition path sampling, aimless shooting, likelihood maximization, and equilibrium path sampling provide a rigorous methodology for extracting accurate kinetics from molecular simulation.

Free Energy and the Reaction Coordinate

Free energy diagrams for activated processes, such as that shown in Figure 1, are typically plotted as free energy (*F*) as a function of the reaction coordinate. The reaction coordinate is the 1-D variable that describes the progress along a reaction pathway from a reactant basin, denoted here as “basin A”, to a product basin, denoted as “basin B”. While many order parameters (OPs) tend to change with a reaction, we reserve the reaction coordinate designation for particular OPs that accurately project the 3N-dimensional dynamics onto 1-D.

The free energy barrier, which determines the rate of an activated process via transition state theory, is the difference in free energy from the transition state to the free energy of the reactant basin, or ($F‡ - F_A$). Accurate determination of the free energy barrier from molecular simulation is essential to make rate predictions from computational results. However, *to determine the free energy barrier in a kinetically meaningful way, it is essential to know the reaction coordinate of the process of interest.*

Figure 2 highlights the problems associated with assuming reaction coordinates. For the free energy surface in Figure 2, the pathway in blue denotes a typical route to go from basin A to basin B. However, one might assume that since q_1 changes between the reactant and the product, calculating the free energy with q_1 as the reaction coordinate will yield the correct free energy barrier. However, if other another collective variable, such as q_2 shown in Figure 2, is a significant component of the reaction coordinate, free energy sampling in both directions will yield severe hysteresis depending on the sampling direction 9 as shown in red.

Chandler and coworkers illustrated the problem with assuming reaction coordinates for several systems in which the choice of the reaction coordinate seemed obvious. For instance, in the dissociation reaction of a Na^+ and a Cl^- ion in aqueous solution, the first reaction coordinate of choice would be the distance between the ions. However, Geissler *et al.* showed that the distance between the ions for this reaction was a very inaccurate reaction coordinate (*19*). Instead they suggested solvent degrees of freedom as likely components of the correct reaction coordinate. Efforts in follow-up work to our knowledge have yet to definitively identify the appropriate reaction coordinate for this process (*20*).

Another example from the Chandler group in which the reaction coordinate turned out to be surprisingly complicated is the isomerization reaction of the alanine dipeptide, a frequent model system for validating methods in statistical mechanics. Decades of previous studies (*21–28*) assumed that the Ramachandran angles were accurate reaction coordinates, but Bolhuis and coworkers showed that an accurate reaction coordinate could not be constructed from Ramachandran angles alone (*29*). Later, Ma and Dinner showed that solvent degrees of freedom were important components of the reaction coordinate for the alanine dipeptide isomerization.

These two examples illustrate the danger in assuming the reaction coordinate even for seemingly simple systems. They also illustrate the danger of using coarse-grained models to understand the kinetics of biological reactions that occur in solution. The dynamically correct reaction coordinate in both of these systems involves complex solvent dynamics that would be lost in coarse-grained models with implicit solvent.

Clearly, the kinetics of complex reactions in solution are important in many aspects of biomass conversion. To obtain reliable insights into the mechanisms and rates of these processes from atomistic simulations, efficient algorithms are needed to identify accurate reaction coordinates. This review describes the development of path sampling methods that fulfill that need.

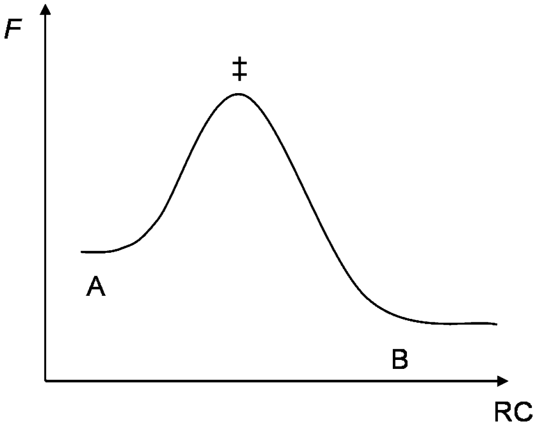


Figure 1. Free energy (F) as a function of the reaction coordinate. The reactant basin is denoted as A, the product as B, and transition state is denoted by ‡.

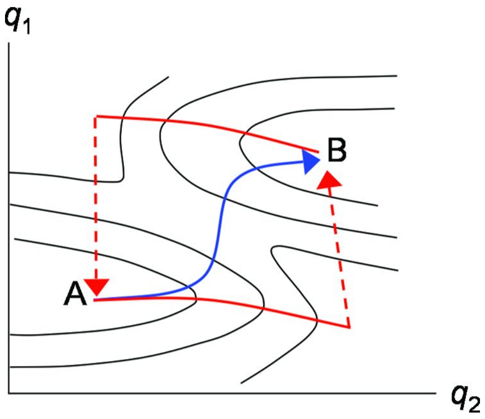


Figure 2. Free energy surface with two stable basins, A and B. Because q_1 incurs a large change between the reactant (A) and the product state (B), one might assume q_1 is the reaction coordinate. However, free energy calculations along either of the coordinates q_1 or q_2 would show hysteresis effects if the barrier is large. (see color insert)

The Central Idea of Transition Path Sampling

Conventional molecular dynamics (MD) and Monte Carlo (MC) algorithms are typically used to study properties of stable or metastable states. However, in studying the transitions between two metastable states, an MD trajectory initiated in basin A on Figure 1 may take a very long time to escape to basin B if the barrier is high. Even if a trajectory initiated in basin A does eventually cross the barrier to basin B, the fraction of time spent at the top of the barrier, the region of interest for understanding the mechanism, will be miniscule relative to the overall simulation time. Thus using conventional MD or MC methods to simulate barrier crossings

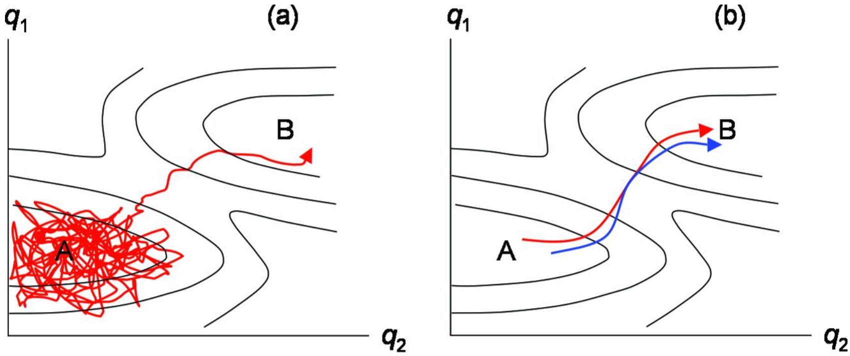


Figure 3. Trajectories initiated on a free energy surface. (a) A trajectory initiated in basin A spends the majority of the simulation time in basin A. If the barrier is low enough and the simulation long enough, there may be a possibility of observing a rare event in which the system overcomes the barrier to reach basin B. However, in the case of high barriers and (or) large systems, this is unlikely and incredibly inefficient. (b) A trajectory initiated from an intermediate point along a reactive trajectory is able to quickly move to basin A or basin B, depending on the initial configuration and momenta of the system. (see color insert)

“wastes” the overwhelming majority of the simulation time, as illustrated in Figure 3(a).

To circumvent the inability to efficiently sample regions of high free energy in real systems, Chandler and co-workers pioneered the technique of transition path sampling (TPS) (*13–18*), which utilizes a Monte Carlo type algorithm in trajectory space. TPS generates a sequence of reactive trajectories by modifying each trajectory and then accepting or rejecting the new trajectory based on its statistical weight in the “path space” of reactive trajectories. Each trajectory is generated according to the natural dynamics of the system, so that the trajectories do not contain artifacts from being forced along a pre-chosen coordinate like the trajectories from steered-MD and similar methods. Thus, the central idea of TPS is to generate the true unbiased ensemble of barrier crossing trajectories without simulating the long time that a standard simulation would waste between barrier crossing events.

Each trajectory has a probability in "path space" proportional to the probability to start at the initial point in phase space (the Boltzmann distribution) multiplied by a series of transition probabilities for following the trajectory from one timeslice to the next starting from time 0 to the total trajectory duration T. The TPS algorithm accepts trajectories according to their probability in trajectory space with the constraint that accepted trajectories must connect basins A and B. To generate new trajectories, TPS employs shooting and shifting moves. Shooting moves create a new trajectory from an old trajectory by changing the momentum slightly at a random timeslice t along the previous trajectory. Then the dynamics are propagated forward and backward in time to times 0 and T. In a shifting move, a length of time Δt is cut from one end of the previous trajectory and then the dynamics are propagated from the other end for an additional time Δt to regain a trajectory of duration T. Most formulations of TPS use dynamics that conserve the ensemble in which the reaction is being studied. In this way the acceptance rule for new trajectories becomes simple: trajectories are accepted if they connect states A and B.

To date, path sampling has been applied to a wide range of systems, such as conformational changes in model peptides (*29*), the folding of small proteins (*30*), conformational transitions in biomolecules (*31*), micelle formation (*32*), hydrophobic polymer collapse (*33*), nanoparticle assembly (*34*), ion association in solution (*19*), autoionization of water (*35*), order-disorder transitions in glass forming model systems (*36*), the nucleation of hexagonal ice (*37*), structural changes in inorganic nanocrystals (*38*), solid-solid polymorph transformations in organic crystals (*39*, *40*), nucleation of sodium chloride from solution (*41*), chemical reactions (*42*), etc. From both a physical and computational standpoint, many of these problems are directly analogous to problems encountered in understanding molecular-level events in fundamental research problems in biomass construction and conversion, which typically involve solvated systems and diffusive processes. Therefore we propose that path sampling approaches will be an integral part of the computational toolkit for understanding the mechanisms in the synthesis and degradation of lignocellulosic material, both in the natural world and for use in the biofuels industry.

TPS has been extensively reviewed elsewhere by the original developers (*13*, *18*). To avoid unnecessary overlap, we instead focus on introducing the latest developments to computational groups studying problems in the construction and conversion of biomass, a research community that is beginning to incorporate simulation as a tool to study molecular-level details of systems and processes of interest (*43–50*). In Section 2, we discuss the importance of quantitative basin definitions, highlight methods for finding initial pathways, briefly describe the algorithm of the original TPS method, and discuss at length a new version of TPS with many advantages over the original algorithm, Aimless Shooting (*51*, *52*). In Section 3, we discuss three methods used to extract the reaction coordinate from path sampling methods: the p_B-histogram test (*19*, *53*), the Genetic Neural Network approach (*54*), and likelihood maximization (*51*, *52*). We also discuss the utility of the p_B histogram test to verify the reaction coordinate and quantify the error associated with the reaction coordinate from the resulting histogram. Section 4 describes a new method to calculate free energy with a path sampling approach, equilibrium path sampling (*55*), which is based on the BOLAS algorithm developed by Radhakrishnan and Schlick (*56*) and from the hybrid MD/MC approach as used by Auer and Frenkel (*57*). In Section 5, we review methods to calculate rate constants in diffusive systems, and we conclude in Section 6 with our perspective on path sampling approaches in biomass construction and conversion problems and future methodology development.

2. Path Sampling Algorithms

This section discusses some practical considerations in collecting a transition path ensemble with path sampling. We stress the importance of accurate basin definitions, outline methods for finding initial pathways, summarize the original

TPS algorithm, and describe a new path sampling algorithm, aimless shooting, in detail.

Basin Definitions

TPS is designed to function without knowledge of the reaction coordinate. However, the Monte Carlo-like path action, described in the next sub-section, requires quantitative descriptions of the reactant (A) and product (B) basins. The path action is described as a function of population functions, denoted $h_A(x)$ and $h_B(x)$, where $h_A(x) = 1$ if x is in basin A and 0 otherwise, with the corresponding values for $h_B(x)$ if x is in basin B or not in basin B. Defining the basins is an *ad hoc* process for each new research problem, but there are three rules to guide the development of practical basin definitions:

1. The basin definitions must include typical equilibrium fluctuations within each basin.
2. Basins A and B must not overlap during the collection of the transition path ensemble, or TPS will find a path from A to B that never leaves A .
3. To optimize the reaction coordinate accurately, the basin definitions should leave as much configuration space as possible assigned to the “no mans land” between the two basins.

Typically, basin definitions are constructed by running long MD simulations in the reactant and product basins. From the equilibrium simulations, the fluctuations in various OPs can be monitored and quantitative basin definitions can be established by selecting windows along several of the OPs whose distributions in A and B do not overlap. For instance, in the case of a conformational change in a protein, possible basin definitions include φ-ψ angles, distances of particular residues, native contacts, radius of gyration, or hydrogen-bonding between residues (*58*). It is essential that basin definitions be carefully developed prior to conducting path sampling.

Finding Initial Pathways

Path sampling generates a sequence of reactive trajectories by modifying the previous trajectory in the sequence. Thus path sampling methods require an initial reactive trajectory. This initial pathway does not necessarily have to be an unbiased dynamical pathway. Although there is no general formula for obtaining an initial pathway, several methods used in the literature to harvest initial trajectories include:

1. Long, unbiased trajectories (*40*)
2. Minimum energy path or minimum free energy path methods (*3*, *7*, *28*, *59*)
3. Umbrella sampling along an assumed reaction coordinate (*40*)
4. Targeted or steered MD (*60*)
5. High-temperature sampling (*58*, *61–63*)

6. Alteration of the Hamiltonian (*41*)
7. Bias annealing over an assumed reaction coordinate (*64*)

Running long, unbiased trajectories at the conditions of interest is the most straightforward method to obtain an initial pathway (*40*); however if the free energy barrier is significantly greater than *kT* and the system is large, this may take an overly excessive amount of computational time. Methods such as umbrella sampling (*65*), in which a series of harmonic restraints are placed on the system along the assumed reaction coordinate in overlapping windows can yield a pathway more efficiently than running unbiased trajectories (*39*, *40*, *66*). However, this approach requires judicious selection of an assumed reaction coordinate that will yield a physically reasonable initial pathway. In a similar vein, targeted MD (*60*) can be used to obtain an initial pathway as in (*39*). In cases where system stability as a function of temperature is not significantly variable, running unbiased simulations at high temperature can accelerate transitions (*61*). For sodium chloride nucleation from solution, Zahn cleverly adjusted the van der Waals radius of the solute and solvent ions to promote nucleation in a computationally accessible simulation (*41*). When path sampling was started from this initial pathway, the potential was changed back to the original parameter set.

In the case of biasing the system to follow a particular pathway, or assumed reaction coordinate to obtain an initial pathway, it is essential to equilibrate the system in trajectory space once path sampling is initiated. For instance, if MD umbrella sampling is used to harvest an initial pathway for a conformational change in a protein or a nucleation event, the first subset of trajectories collected with a path sampling algorithm should be discarded because the system will anneal to the true free energy landscape at the conditions of interest, rather than the assumed free energy landscape.

A particular method of note to obtain an initial pathway that systematically minimizes the need for trajectory space equilibration was developed by Hu, Ma, and Dinner (*31*). Their method uses a bias annealing approach in which a trajectory is first harvested with steered MD with a large force constant. From this initial, biased pathway, successive pathways are generated iteratively by reducing the force constant and firing trajectories from random points along the previous pathway in both directions. This approach has been successfully applied to harvest an initial pathway in nucleotide flipping by a DNA repair protein (*31*, *64*). We anticipate that this method will prove to be useful for many problems in which conformational transitions in biological or macromolecular systems are the processes of interest.

Once an initial pathway is generated, it is essential to know approximately where the transition state region is located along the initial pathway as input into a path sampling algorithm, as will be discussed. The transition state region along the initial trajectory can be found by shooting multiple, randomly seeded trajectories from different configurations along the initial trajectory and noting the final configuration. If trajectories from a given configuration sometimes end in A and sometimes end in B, this suggests that the selected configuration is located near the transition state region. An approximate p_B=1/2 point along the trajectory can be identified efficiently from a bisection algorithm.

Transition Path Sampling

As mentioned previously, several excellent reviews of TPS are available from the original developers (*13*, *18*). Reference (*18*) reviews many of the working aspects of TPS applied to real systems. Many of the ideas originally discussed therein apply to other path sampling approaches. Here we briefly highlight the basics of the original TPS algorithm.

To generate new trajectories, the original versions of TPS employed shooting and shifting moves. Shooting moves create a new trajectory from an old trajectory by changing the momentum slightly at a random timeslice t along the previous trajectory. Then the dynamics are propagated forward and backward in time to times 0 and T. In a shifting move, a length of time Δt is cut from one end of the previous trajectory and then the dynamics are propagated from the other end for an additional time Δt to regain a trajectory of duration T. Most formulations of TPS use dynamics for the shooting and shifting moves that conserve the ensemble in which the reaction is being studied. With this choice of dynamics, the acceptance rule for trajectories from shooting and shifting moves becomes simple: new trajectories are accepted if they connect states A and B and rejected otherwise.

Shooting moves are described here so that the reader can understand the differences between the original TPS algorithm and aimless shooting. For the "shooting" move, the algorithm is as follows:

1. Select a timeslice t along the trajectory randomly, denoted by $\mathbf{x}_t$.
2. Perturb the momenta by a vector δp from the original momenta.
3. Propagate the dynamical equations of motion forwards and backwards in time by some length $T/2$ in each direction.
4. Accept the new trajectory if it joins the reactant and product states, and reject the new trajectory if it does not connect the reactant and product basins.

A shooting move is illustrated in Figure 4, similar to that shown in reference (*18*). The size of the momentum perturbation in the shooting moves, δp, is adjusted to obtain reasonable acceptance rates of approximately 40-50% as in conventional Monte Carlo simulation.

For a "shifting" move, the trajectory is translated in time by some adjustable parameter, δt. This move is analogous to polymer reptation. Shifting moves help TPS explore trajectory space by enabling a rapid mechanism for relaxation of those reactive trajectories that barely have time to reach one basin after spending the vast majority of the trajectory duration in the other basin.

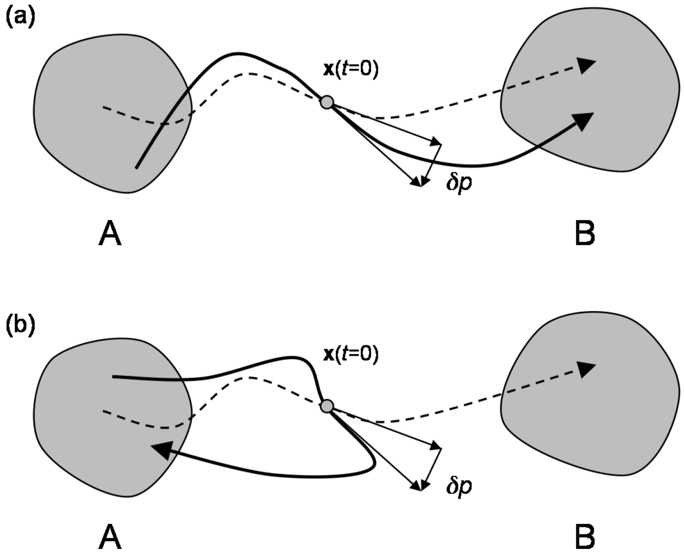


Figure 4. Examples of shooting moves for TPS. In both panes, the original trajectory is shown as a dotted line. The new trajectory is shown as a solid line. (a) Example of an "accepted" trajectory in which the trajectory connects basin A and basin B. (b) Example of a "rejected" trajectory in which the trajectory connects basin A to itself.

Applications of the original TPS algorithm to diffusive systems such as crystal nucleation (*67*) and protein folding (*58*, *62*) exhibited very low acceptance rates. The commitment time for these types of rare events is usually on the nanosecond timescale (or greater) implying a rough free energy landscape from the top of the free energy barrier to a given basin. This problem has lead to new developments in the path sampling community to circumvent this problem including the addition of a weak stochastic element through use of an Andersen thermostat applied to half trajectories (*68*), the transition interface sampling technique (*67*, *69*, *70*), shooting moves with submachine precision (*71*), and specialized move sets with double ended constraints in the path interior (*72*). Each of these methods to improve acceptance in diffusive dynamics uses a strategy that generates successive trajectories that are highly similar. Since each trajectory is costly, especially for highly diffusive systems, it is preferable to generate trajectories that rapidly diverge from each other while maintaining reasonable acceptance rates. A new method that accomplishes these goals simultaneously, aimless shooting, is discussed in the next section.

Aimless Shooting

Peters and Trout recently developed a new approach to path sampling, dubbed aimless shooting, which replaces the shooting and shifting moves from the original TPS algorithm with a single new type of shooting move. There are several distinct advantages of aimless shooting over TPS that make it a significant development in the field of path sampling. In aimless shooting, the momenta are drawn from the Boltzmann distribution for each new trajectory; hence the trajectories de-correlate more quickly in aimless shooting than the conventional TPS shooting moves. In addition, aimless shooting automatically keeps the system in the transition state region (more accurately, where p_B is near ½, which will be discussed in Section 3), thus maintaining high acceptance rates. Another advantage is that there is only one adjustable parameter in the algorithm, Δt. The total length of an aimless shooting trajectory is $(T + \Delta t)$ and each trajectory has three segments:

1. a "backward" trajectory from $\mathbf{x}(t{=}0)$ to $\mathbf{x}(t{=}{-}T/2)$
2. a "connector" trajectory from $\mathbf{x}(t{=}0)$ to $\mathbf{x}(t{=}\Delta t)$
3. a "forward" trajectory from $\mathbf{x}(t{=}0)$ to $\mathbf{x}(t{=}\Delta t + T/2)$

These segments are illustrated in Figure 5. To initiate aimless shooting, a small time interval Δt is chosen such that $\Delta t << T$ along the initial pathway of length T. Typically, we have found that $\Delta t = 0.01{\cdot}T$ yields favorable acceptance rates. Along the initial trajectory, a configuration is selected that is close to the transition state region. This initial configuration is typically found by firing several randomly seeded trajectories from points along the initial pathway, as discussed earlier. From here, the aimless shooting algorithm, illustrated in Figure 5, works as follows:

1. From the previous trajectory, select $\mathbf{x}(t{=}0)$ or $\mathbf{x}(t{=}\Delta t)$ as the shooting point with 50% probability for the two choices. Save the shooting point as $\mathbf{x}_{\text{new}}(t{=}0)$.
2. Draw new velocities from the Boltzmann distribution at the chosen shooting point.
3. Propagate the dynamics backwards for $-T/2$, i.e. reverse the momenta and run a forward trajectory for time $T/2$.
4. Propagate the dynamics forwards in time by Δt from the shooting point, and save the configuration, $\mathbf{x}_{\text{new}}(t{=}\Delta t)$.
5. Continue the forward trajectory for $+T/2$.
6. Accept the new trajectory if it joins the reactant and product states, and reject the new trajectory if it does not just as in the original TPS Monte Carlo-like path action.

An "inconclusive" trajectory is defined as a trajectory in which one or both ends do not reach a basin. Just as in the original TPS algorithm, frequent inconclusive trajectories indicate that the trajectory length is insufficient or that the initial pathway is of poor quality. After the initial iteration of aimless shooting, shooting points are selected between $\mathbf{x}(t{=}0)$ and $\mathbf{x}(t{=}\Delta t)$. In this way, aimless shooting ensures that one point of two from which to shoot will yield reactive trajectories.

Aimless shooting has some very useful properties. First, each shooting move generates an independent realization of the committor probability because the momenta are chosen fresh from the Boltzmann distribution. Second, Peters and

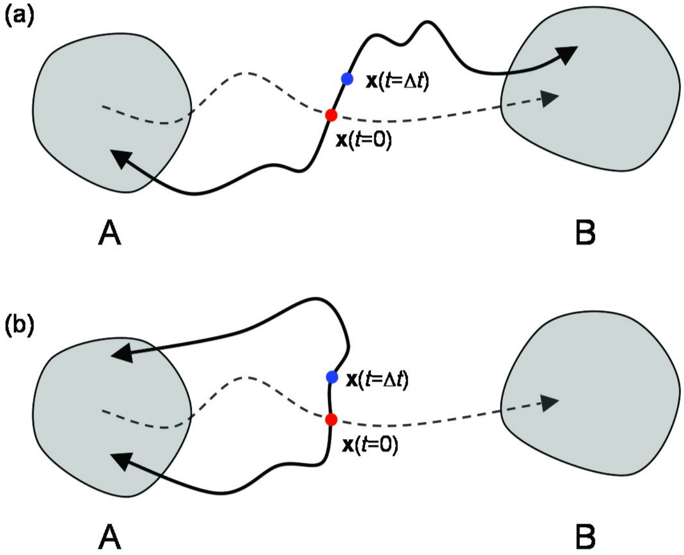


*Figure 5. The aimless shooting algorithm. In both panes, the original trajectory is shown as a dotted line; the new trajectory is shown as a solid line; the point chosen along the previous trajectory at **x**(t=0) is shown as a red dot; and the point chosen at **x**(t=Δt) is shown as a blue dot. (a) Example of an "accepted" trajectory. For the next iteration of aimless shooting, the shooting point is selected randomly between configurations on the new trajectory at **x**(t=0) and **x**(t=Δt). (b) Example of a rejected trajectory. For the next iteration of aimless shooting, the shooting point is selected randomly between configurations on the old trajectory at **x**(t=0) and **x**(t=Δt), the latter of which is not shown. (see color insert)*

Trout showed that the shooting points themselves are approximately distributed according to:

$$\rho(\mathbf{x}_{\mathrm{sp}}) = \mathrm{p}(\mathrm{TP} \mid \mathbf{x}_{\mathrm{sp}})\ \mathrm{p}(\mathbf{x}_{\mathrm{sp}} \mid \mathrm{TP}) \qquad (1)$$

where $\mathbf{x_{sp}}$ is configuration of the system at a shooting point (**sp**), $\rho(\mathbf{x_{sp}})$ is the distribution of shooting point configurations, p(TP|$\mathbf{x_{sp}}$) is the probability of being on a transition path given a shooting point configuration, and p($\mathbf{x_{sp}}$|TP) is the probability of being at a shooting point configuration given being on a transition path. This distribution of shooting points is confined to the transition pathway by the factor p(**x**|TP). The distribution of shooting points is also confined to the zone along the transition pathway where the factor p(TP|**x**) is nonzero. Because of this property, aimless shooting automatically maintains a high acceptance probability, even without using small momentum perturbations as in the original TPS algorithm. Furthermore, the zone where p(TP|**x**) is large also corresponds to the transition state region (where p_B(**x**) is near ½ as will be discussed in Section 3). Having many shooting points in the transition state region helps aimless shooting generate data that will help identify transition states. Figure 6 depicts the difference between aimless shooting and shooting-shifting strategies for TPS.

Aimless shooting also has advantages for sampling transition paths in diffusive systems. Aimless shooting has been applied to understand highly diffusive processes like the nucleation of polymorph transitions in terephthalic acid (*39*, *40*) and the nucleation of Lennard-Jonesium crystals from the melt (Beckham and Peters, in preparation). As discussed in the previous sub-section, several investigators have modified the shooting-shifting algorithm to address diffusive systems including sub-machine precision momentum perturbations (*71*), shooting half-trajectories with a stochastic element (*68*), and incorporating specialized moves with double-ended constraints in the interior of the path (*72*). These modified path sampling schemes all generate a new trajectory that is very similar to the old trajectory. By comparison, successive aimless shooting trajectories rapidly diverge from one another at the shooting point while still maintaining a high acceptance rate. In summary, aimless shooting is a TPS method with all of the features of the original method. However, the new aimless shooting version has these advantages:

1. Each shooting point represents an independent realization of p_B, which is a useful feature for reaction coordinate identification as will be discussed in Section 3.
2. Shooting points are automatically distributed with $0 < p_B < 1$ with a distribution that is peaked near ½. This feature leads to high acceptance, even for highly diffusive barrier crossing dynamics.
3. Aimless shooting trajectories diverge from each other more rapidly than those from previous shooting/shifting schemes.
4. Aimless shooting is easily implemented with existing molecular dynamics packages and has just one adjustable parameter, Δt.

Scripts for the aimless shooting algorithm with extensive comments are available for download at http://www.engineering.ucsb.edu/~baronp/codes.html. The aimless shooting scripts are written for implementation on a standard Linux cluster. A "master" shell script contains the loop over the successive aimless shooting paths, calls the MD code as a parallel executable, then calls a Fortran executable to determine the values of the basin definition OPs, performs the Monte Carlo path action, and saves the appropriate files for subsequent analysis. Two sets of MD code for the simulation of a terephthalic acid nucleation problem (*39*, *40*) are available for the shooting moves, written currently for CHARMM (*73*) and NAMD (*74*). A sample Fortran code, also for the terephthalic acid problem (*39*, *40*) is available, which reads in the output from the MD code. The scripts are easily altered for other systems of interest and other MD codes.

Comparison of Aimless Shooting to Other Sampling Methods

Several other sampling methods have been proposed for highly diffusive problems. These include forward flux sampling (*75–78*), the string method in collective variables (*59*), milestoning (*79*, *80*), and transition interface sampling (*69*, *70*, *81*, *82*). The efficiency of transition interface sampling and forward flux sampling decreases when performed with inaccurate coordinates.

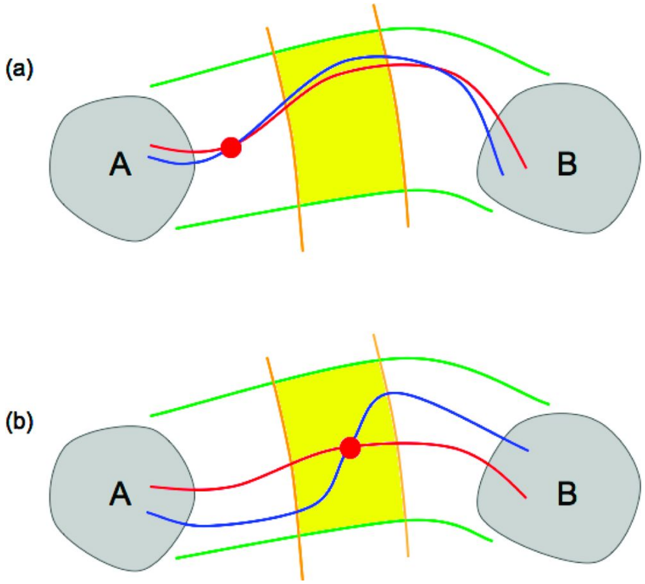


Figure 6. Differences between TPS and AS move sets. The transition state region is shown in yellow shading. (a) TPS generates shooting points outside of the transition state region (filled in yellow shading), thus for highly diffusive processes, the acceptance rates can be very low. (b) Aimless shooting distributes shooting points confined to the transition state region (filled in yellow shading) according to the factor $p(\mathbf{x}|TP)$, thus demonstrating good acceptance rates, even for highly diffusive processes. (see color insert)

Sampling procedures involved in the string method and milestoning depends on pre-identified variables. If the subspace of predetermined variables does not include the reaction coordinate, the string method and milestoning approaches may give incorrect results.

3. Finding the Reaction Coordinate from Path Sampling Data

The result from path sampling is an ensemble of trajectories connecting the reactant and product basins, which is commonly referred to as the transition path ensemble. In the development of TPS, one of the major research challenges was to develop an automated way to extract the reaction coordinate from the transition path ensemble. This section describes three methods, in the chronological order of development, to find the reaction coordinate from TPS and aimless shooting data, namely the histogram test, the Genetic Neural Network (GNN) method, and likelihood maximization. The histogram test, often called a committor analysis, represents the first method to identify a reaction coordinate, whereas the GNN and likelihood maximization each made significant advances in efficiency to identify reaction coordinates.

p_B and the Histogram Test

The committor probability, denoted as $p_B(\mathbf{x})$, is the probability that a given configuration **x** will commit to the product basin "B" if initiated with a velocity from the Boltzmann distribution. $p_B(\mathbf{x})$ takes the value of 1 for configurations in the product basin, 0 for configurations in the reactant basin, and ½ for transition states. Isosurfaces of an accurate physical reaction coordinate, $r(\mathbf{x})$=constant, should closely approximate the isocommittor surfaces, $p_B(\mathbf{x})$=constant (*18*), as depicted in Figure 7.

The committor probability has long been interpreted as a reaction coordinate (*18*, *19*, *53*, *54*, *83*, *84*). The p_B histogram test, also called a committor analysis, begins with an assumed reaction coordinate q to test and usually with a free energy calculation to identify the putative transition state location q^* along the assumed coordinate. A Boltzmann distributed sample of atomistic configurations **x** with $q(\mathbf{x})=q^*$ is harvested by constrained sampling. At each harvested configuration **x**, $p_B(\mathbf{x})$ is estimated by initiating trajectories from **x** with Boltzmann distributed velocities. The fraction of these trajectories that commit to the product basin is an estimate of p_B at that configuration. The p_B-estimates from each configuration are then combined into a histogram of estimated committor probabilities. A putative transition state surface from an accurate reaction coordinate will give a histogram that is closely centered on p_B=½.

Figure 8 depicts the process involved in the histogram test as it is commonly performed in the literature (*18*, *19*, *53*). A putative reaction coordinate and a corresponding transition state region is selected. In Figure 8(a), the reaction coordinate chosen is q_2 and a particular value q_2^* is chosen as the transition state along q_2. Boltzmann distributed points are generated, shown as black circles, along with the putative transition state surface and multiple trajectories are fired from each point. As the reaction coordinate chosen in Figure 8(a) is not representative of the true reaction coordinate, a collection of points will result in a number of trajectories that reach basin A only (shown in red) and a number of trajectories that only reach basin B (shown in blue). When plotted as the fraction of points that reach basin B from any given point, this will result in a histogram with peaks near p_B=0 and p_B=1. A histogram of this nature denotes a poor reaction coordinate choice and that other collective variables are important components of the reaction coordinate. Alternatively, if the reaction coordinate is chosen properly as shown in Figure 8(b) where the reaction coordinate is a function of q_1 and q_2, trajectories fired from the transition state surface will result in an approximately equal probability of reaching basin A or basin B from any given point, thus resulting in a histogram peaked near p_B = ½.

As commonly performed in the literature, the histogram test is extremely expensive because it is an iterative, trial-and-error process. Moreover, the histogram is only a qualitative measure of reaction coordinate error because protocol dependent sampling errors add noise to the p_B-estimates. Peters deconvoluted the sampling error in the histogram to quantify the actual error in the continuous distribution of p_B-values on the dividing surface (*85*). These distributions are depicted in Figure 9.

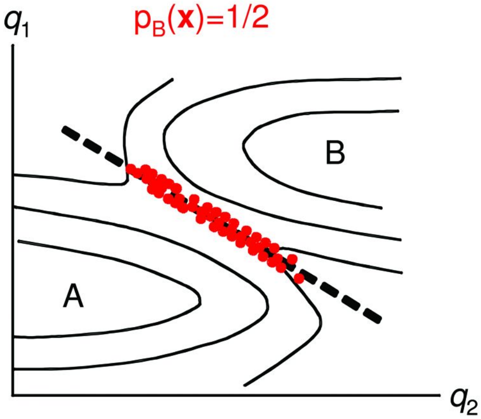


*Figure 7. Isosurfaces of an accurate physical reaction coordinate r(**x**) should coincide with the location of specific values of the committor probability. In particular, structures on the dividing surface r(**x**)=r* should all be transition states as defined by p_B(**x**)=½. (see color insert)*

The mean and variance of the actual committor distribution can be obtained from the mean and variance of the histogram:

$$\mu_\mathrm{p} = \mu_\mathrm{H} \tag{2}$$

$$\sigma_\mathrm{p}^2 = \frac{N\sigma_\mathrm{H}^2}{N-1} - \frac{\mu_\mathrm{H}(1-\mu_\mathrm{H})}{N-1} \tag{3}$$

where μ_H is the histogram mean, σ_H is the histogram variance, μ_p is the intrinsic mean, σ_p is the intrinsic variance, and N is the number of shooting points per histogram. These formulas provide a quantitative range of committor probability values on a predicted transition state surface:

$$\mathrm{p_B} = \mu_\mathrm{p} \pm \sigma_\mathrm{p} \tag{4}$$

Peters showed that μ_P and σ_P can be obtained by a relatively inexpensive calculation. For reaction coordinates with errors of $\sigma P > 0.15$, μ_P and σ_P can be estimated to within 10% of their values with 2000 trajectories partitioned among 200 $\mathrm{p_B}$-estimates. Thus the quantitative measure of reaction coordinate error is often more than a factor of ten less expensive than the original committor analysis procedure.

The Genetic Neural Network Method

The $\mathrm{p_B}$ histogram test is computationally expensive, especially when the trial and error approach for finding reaction coordinates requires many iterations of the histogram test with different putative reaction coordinates (*63*). The first method to find the reaction coordinate in a more systematic manner was developed by Ma and Dinner. Their seminal advance in this area was the Genetic Neural network (GNN) method for optimizing the reaction coordinate (*54*). The GNN method

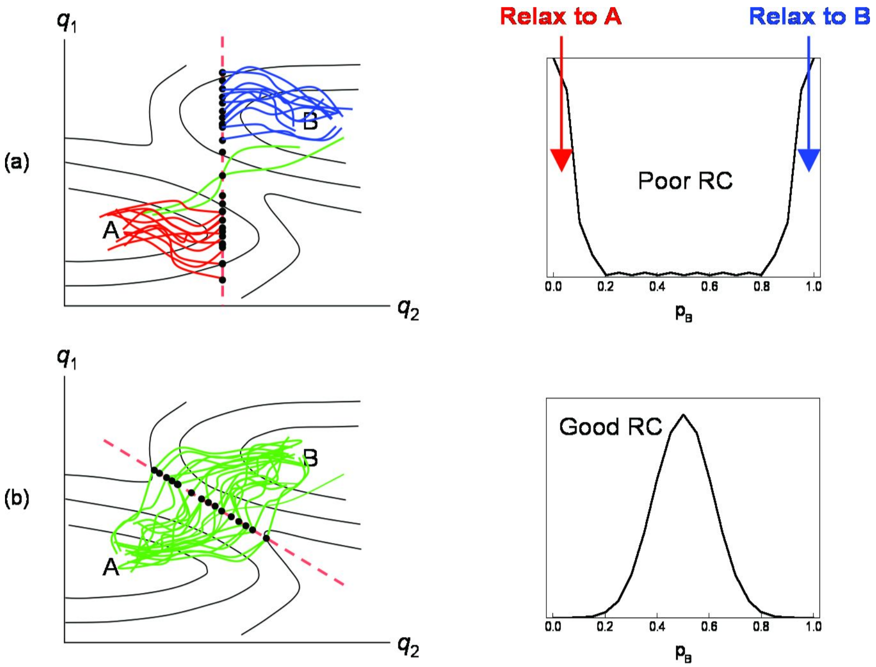


Figure 8. The p_B histogram test. A putative reaction coordinate and corresponding transition state surface (shown as the dotted red line) are proposed. Boltzmann distributed points along the putative transition surface are generated as shown in black circles. Many trajectories are fired from each point and the fraction of trajectories from each point that reach B are plotted as a histogram. (a) Example of a poor choice of a reaction coordinate. The points closer to A result in trajectories (shown in red) that all relax to A. The points closer to B result in trajectories (shown in blue) that all relax to basin B. The resulting histogram is therefore peaked close to $p_B = 0$ and $p_B = 1$, denoting a poor reaction coordinate choice. (b) Example of a judicious choice of reaction coordinate. The points approximate the transition state surface well so the histogram is peaked at p_B=½, which is indicative of the transition state surface. (see color insert)

begins with an ensemble of trajectories from TPS. $p_B(\mathbf{x})$ estimates are computed at points along the transition paths and a training set of p_B estimates is selected so that the estimates are evenly distributed with p_B values between zero and one. One and two level perceptron models of the reaction coordinate are constructed from a database of possible components of the reaction coordinate. A genetic algorithm sorts through the perceptron models and finds the best model reaction coordinate by comparing their square error scores. Each score is determined by minimizing the square error between the computed and perceptron-predicted p_B estimates in the training set, as given by Equation 5 and illustrated in Figure 10.

$$S = \sum_{k=1}^{training\ set} (\hat{p}_B(\mathbf{x_k}) - p_M(\mathbf{x_k}))^2 \qquad (5)$$

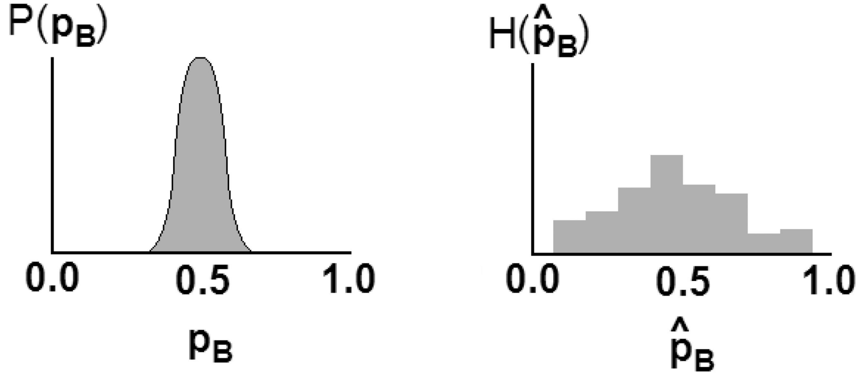


Figure 9. The intrinsic committor distribution $P(p_B)$ corresponds to the distribution of infinitely accurate committor probability estimates. Variance in the distribution $P(p_B)$ is entirely due to reaction coordinate error. The distribution $H(p_B)$ corresponds to the distribution of p_B-estimates in a histogram. The histogram has a larger variance because of binomial sampling error in the p_B-estimation process.

The GNN approach of Ma and Dinner provided the first good reaction coordinate for isomerization of the alanine dipeptide represented by a united atom model in explicit solvent. GNN's success for the alanine dipeptide was remarkable for several reasons. This challenging problem had thwarted all previous attempts to find a good reaction coordinate using Ramachandran angles and other internal coordinates. Second, the reaction coordinate identified by GNN confirmed the suspected importance of solvent dynamics in the mechanism of this reaction. Specifically, GNN found that the local orientation of solvent dipoles must preorganize to enable the conformational change. Because the solvent preorganization occurs on a longer timescale than fluctuations in the Ramachandran angles, the solvent orientations are an integral part of the reaction coordinate.

In the original work of Ma and Dinner on the isomerization of the alanine dipeptide, a training set constructed from 194,700 trajectories gave a coordinate from GNN that was not extremely accurate. Later work suggests that this is primarily due to missing coordinates in the set of candidate order parameters (*31*). However our subsequent analysis demonstrates that p_B-estimates and square error scoring as used by GNN make inherently less efficient use of trajectory data than likelihood maximization (*51*).

Likelihood Maximization

Peters and Trout introduced the method of likelihood maximization to obtain reaction coordinates from path sampling data (*52*). Likelihood maximization is an information theory-based approach in which a large set of putative reaction coordinates are exhaustively tested to fit to the reaction coordinate, given by the committor function, $p_B(r(\mathbf{x}))$. It is different from the previous methods for

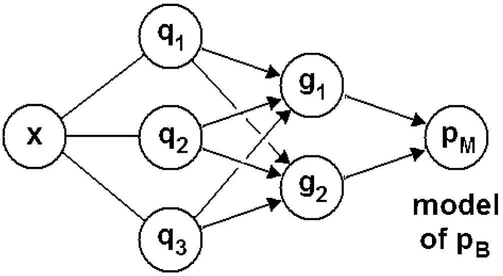


Figure 10. The Genetic Neural Network method uses one and two level perceptron models (g_1 and g_2) based on possible components of the reaction coordinate ($q_1…q_N$) to fit a model for the reaction coordinate (p_M) to p_B estimates from the transition path ensemble.

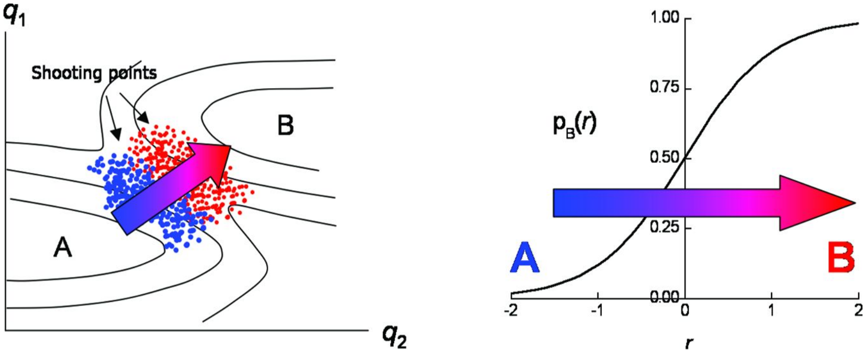


Figure 11. Likelihood maximization is designed to extract the best approximation to the reaction coordinate from the transition path ensemble collected by aimless shooting. (a) Forward-trajectory outcomes for points on a free energy surface between A and B, colored by outcome. Blue points relax to basin A and red points relax to basin B. The transition state region is located at the "interface" where the shooting point outcomes switch from relaxing to basin A to relaxing to basin B. (b) The committor probability function, $p_B(r)$, is a function of the model reaction coordinate, r, that fits the shooting point data. All model reaction coordinates can be shifted so that the value $r = 0$ marks the transition state location. (see color insert)

identifying reaction coordinates, which rely on committor probability estimates. Instead, the likelihood maximization approach is the first to use a rigorous probabilistic framework based on *realizations* of a committor probability. We stress that p_B realizations are different from p_B estimates. To that end, likelihood maximization makes efficient use of the information collected during aimless shooting, wherein each trajectory is an independent realization of the committor probability. The concept of the technique is illustrated in Figure 11.

Table 1. Scheme for likelihood maximization. The shooting points are collected with the outcomes in the forward direction and candidate OP values at each point

Shooting point	*Forward outcome*	*Candidate order parameter values*
$\mathbf{x}_1$	B	$q_1(\mathbf{x}_1)\ q_2(\mathbf{x}_1)\ q_3(\mathbf{x}_1)\ \ldots\ q_{\mathrm{N}}(\mathbf{x}_1)$
$\mathbf{x}_2$	A	$q_1(\mathbf{x}_1)\ q_2(\mathbf{x}_1)\ q_3(\mathbf{x}_1)\ \ldots\ q_{\mathrm{N}}(\mathbf{x}_1)$
$\mathbf{x}_3$	A	$q_1(\mathbf{x}_1)\ q_2(\mathbf{x}_1)\ q_3(\mathbf{x}_1)\ \ldots\ q_{\mathrm{N}}(\mathbf{x}_1)$
…	…	…
$\mathbf{x}_n$	B	$q_1(\mathbf{x}_1)\ q_2(\mathbf{x}_1)\ q_3(\mathbf{x}_1)\ \ldots\ q_{\mathrm{N}}(\mathbf{x}_1)$

Likelihood maximization is designed to infer the reaction coordinate from shooting points that relax to A and those that relax to B (or the "length" of the arrow in Figure 11). In addition, the reaction coordinate can be a function of multiple OPs, and likelihood maximization will extract the best approximation from a candidate list of variables (the "direction" of the vector in Figure 11).

In the first iteration of the likelihood maximization method, Peters and Trout recognized that for a good reaction coordinate, denoted here as $r(\mathbf{x})$, the probability of being on a transition path, $\mathrm{p}(\mathrm{TP}|\mathbf{x})$ depends only on the reaction coordinate. Therefore, the method introduced in the first study describing likelihood maximization is based on finding the function $\mathrm{p}(\mathrm{TP}|r(\mathbf{x}))$ that can best explain $\mathrm{p}(\mathrm{TP}|\mathbf{x})$ (*52*). It was later realized that fitting the reaction coordinate to $\mathrm{p}(\mathrm{TP}|\mathbf{x})$ shifts the reaction coordinate to fit the tails of $\mathrm{p}(\mathrm{TP}|\mathbf{x})$ at the expense of accuracy in the transition state region (*51*). In a subsequent study, the likelihood maximization method was improved by fitting the reaction coordinate from the shooting point outcomes to the probability of reaching basin B from a given configuration, or $\mathrm{p}_{\mathrm{B}}(\mathbf{x})$, the committor probability function described above (*51*). The likelihood maximization method based on p_{B} only uses the forward trajectory outcomes because when the dynamics are not completely diffusive, the forward and backward outcomes at shooting points may be (negatively) correlated. Likelihood maximization as presented in this review and the paper by Peters, Beckham, and Trout (*51*) are appropriate for both ballistic and diffusive dynamics.

Likelihood maximization works as follows: a transition path ensemble is collected with aimless shooting. During aimless shooting the basin outcome of each forward shooting point is saved (regardless of whether the shooting move is accepted or rejected), and the shooting point configurations are also saved. A list of order parameters that may be part of the reaction coordinate is generated based on the configurations at the shooting points, given by $q_1 \ldots q_{\mathrm{N}}$. This results in a data file with one row for each of the n shooting points as shown in Table 1.

The data are used to optimize models of the committor probability. The one constraint is that the committor probability should approach zero at small values of the reaction coordinate and it should approach one for large values of the reaction coordinate. We use a single-level perceptron:

$$\mathrm{p_B}(r) = \frac{1}{2}\left(1 + \tanh(r)\right) \tag{6}$$

More flexible models like the multilevel neurons of Ma and Dinner may be needed in some cases. In the single level perceptron models, the argument of the tanh function should be some monotonic function of the candidate OPs. In the simplest case, $r(\mathbf{q})$ may be a linear combination of M OPs as follows:

$$r(\mathbf{q}) = \alpha_0 + \sum_{k=1}^{k=M} \alpha_k q_k \tag{7}$$

The linear combination model is actually quite flexible because the OPs themselves may be nonlinear combinations of other OPs. If the model reaction coordinate given by Equations 6 and 7 is correct, then the likelihood of observing the data in the table is:

$$L = \prod_{x_k \to \mathrm{B}} p_{\mathrm{B}}(r(\mathbf{x}_k)) \prod_{x_k \to \mathrm{A}} (1 - p_{\mathrm{B}}(r(\mathbf{x}_k))) \tag{8}$$

The notation in Equation 8, x_k→B denotes a product over all shooting points (x_k) that result in trajectories that reach basin B and vice versa for x_k→A. The coefficients in the model reaction coordinate of Equation 7 are varied to maximize the log likelihood score in Equation 8. This maximization is performed for each combination of proposed OPs up to models of the reaction coordinate with M component OPs. The resulting reaction coordinate models given by $r(\mathbf{q})$ are ranked in terms of their respective log likelihood scores. For all models with a specific number of component variables, M in Expression 7, the reaction coordinate model with the maximum likelihood score is the best model. When comparing models with different numbers of component variables, it should be remembered that the models also have different numbers of fitting coefficients. The best model with M+1 component variables will always have a higher likelihood score than the best model with M component variables. However, the improvement may not be physically significant. Peters and Trout proposed using the Bayesian Information Criterion (BIC) to decide whether an improvement was significant. The BIC is given by ½ln(N) where N is the number of trajectories collected with aimless shooting. If the difference in two likelihood scores for two reaction coordinate models is much greater than the BIC, the model with an additional component variable is a physically significant improvement. The BIC test is thus useful to determine the number of component OPs (M in Expression 7) that are necessary to describe the reaction coordinate.

Following likelihood maximization, the quality of the data as fit to the committor probability function given in Equation 6 can be assessed by plotting the committor function with the data, as illustrated in Figure 12. For the "data" in Figure 12, the reaction coordinate is separated into reasonably sized bins, and the ratio of trajectories that reach B from a given bin is divided by the total number of trajectories from that bin. The "model" curve is simply the function given by Equation 6. The error bars are on the "model" and are from the binomial standard

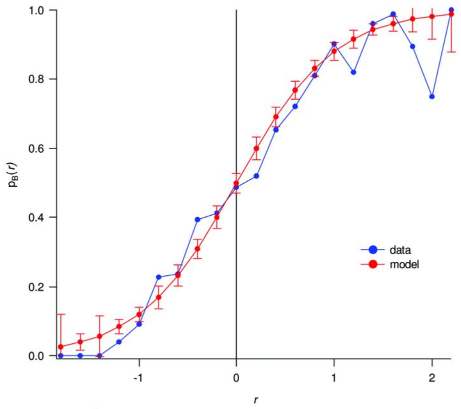


Figure 12. The committor probability function given by Equation 6 ("model", shown in red) and the data from the transition path ensemble collected with aimless shooting ("data", shown in blue). This curve is used to determine the quality of the reaction coordinate approximation from likelihood maximization. (see color insert)

deviations based on the mean in each bin: $[\mu(1\text{-}\mu)/n]^{1/2}$ as described in Peters *et al.* (*51*).

Agreement between the model and the aimless shooting data as shown in Figure 12 is a necessary, but not sufficient, test for successful identification of an accurate coordinate. To ensure that likelihood maximization has identified an appropriate reaction coordinate, the predicted reaction coordinate should still be subjected to a histogram test. This is necessary because the aimless shooting data are harvested from the distribution of equation (5), i.e. from a p(TP|**x**)-weighted transition path ensemble and not from the equilibrium ensemble. By contrast, the committor probability distribution on a reaction coordinate isosurface is the distribution of committor probabilities from an *equilibrium distribution of states on the reaction coordinate isosurface*. Thus, aimless shooting and likelihood maximization may identify a coordinate that accurately describes dynamics along the reaction channel, but fails to separate this motion from excursions into an off pathway branch of the stable basin A. Problems like this have not been encountered in any of the previous applications of likelihood maximization (*30*, *39*, *40*, *51*, *52*), but Figure 13 provides a schematic illustration of this possibility.

The example in Figure 13 illustrates why the final reaction coordinate should still be rigorously tested for accuracy. The quantitative version of committor analysis developed by Peters (*85*) can reduce the computational demands of this final step in identifying an accurate reaction coordinate.

A C code written for the likelihood maximization step is available for download at http://www.engineering.ucsb.edu/~baronp/codes.html. The script is written to take a text file input with shooting point outcomes and values of the candidate OPs at each shooting point. Instructions on how to use the code are included in the distribution. The applications for the likelihood maximization code are in references (*39*, *40*).

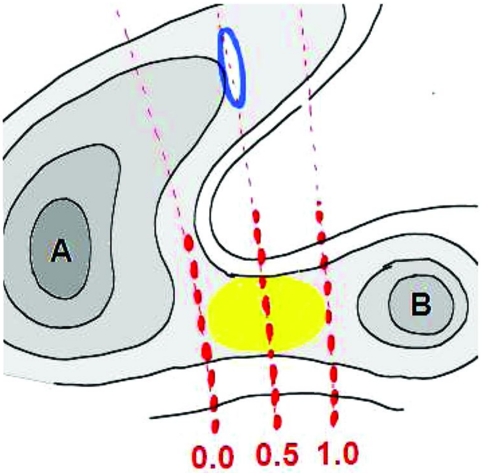


Figure 13. For a free energy landscape such as this, aimless shooting points will be located in the zone that is colored yellow. Near the yellow zone, the reaction coordinate isosurfaces from likelihood maximization are a good approximation to the true isocommittor surfaces. However, the reaction coordinate may be inaccurate in regions of the reactant basin that were not visited during aimless shooting. In this case, a constrained equilibrium simulation on the predicted p_B=½ surface would visit many points that are actually committed to the A basin (the blue region). Even though the reaction coordinate appears accurate according to the test in Figure 12, the histogram for this coordinate would be peaked at p_B=0 instead of p_B=½. (see color insert)

Other Methods To Find Reaction Coordinates

Two other methods have been reported for identifying reaction coordinates from path sampling data. Antoniou and Schwartz (*86*) and also Best and Hummer (*87*, *88*) have proposed methods to optimize the parameterization of the separatrix (p_B=1/2). We note that an accurate description of the p_B=1/2 surface is a necessary but not sufficient condition to ensure an accurate reaction coordinate. Coordinates having one isosurface that parameterizes the separatrix may still poorly describe earlier and later stages of the reaction progress. Errors at later and earlier stages can lead to hysteresis problems in free energy calculations and inaccurate dynamical models for motion along the reaction pathway. Thus, the methods of Antoniou and Schwartz (*86*) and Best and Hummer (*87*, *88*) are less robust and general than GNN and Likelihood Maximization approaches.

4. Calculating the Free Energy from a Known Reaction Coordinate

As described in the previous sections, careful identification of the reaction coordinate is essential to obtain accurate free energy barriers and meaningful kinetics of a rare event. Once the reaction coordinate is verified, there are many

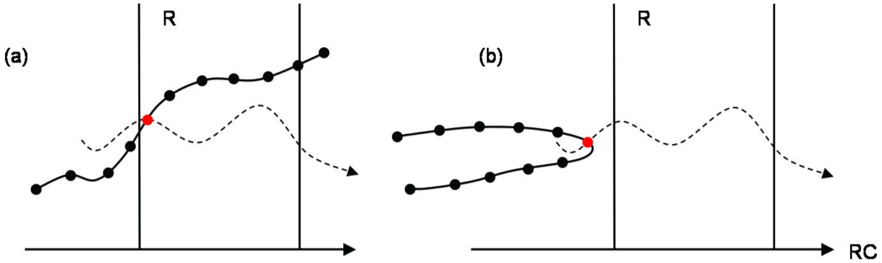


Figure 14. Illustration of the equilibrium path sampling method. The panes show a window region, R, in which EPS is being conducted. The previous trajectory is shown in both panes as a dotted line. The starting point for both trajectories is a random frame along the previous trajectory, labeled in red. (a) Example of an accepted trajectory. Several points along the trajectory are in the window. (b) Example of a rejected trajectory that never visited the region R. (see color insert)

methods available to calculate the free energy barrier with a known reaction coordinate including umbrella sampling (*65*, *89*), blue moon sampling (*90*), lambda dynamics (*91*), metadynamics (*92*), the finite-temperature string method (*59*), adaptive umbrella sampling (*93*), hyperdynamics (*94*), single-point hybrid MC-MD (*57*), BOLAS (*56*), Wang-Landau sampling (*95*), and nonequilibrium methods (*96*).

Of particular note to calculate free energy barriers is the transition path theory (TPT) of Vanden-Eijnden and co-workers (*3*, *97*, *98*). The TPT theory provides a framework for computing both isocommittor surfaces and the current field in a reaction pathway. However, TPT does require the component variables of the reaction coordinate as an input to the theory. Thus, TPT is not an alternative procedure to identify reaction coordinates, except in cases where the component variables in the reaction coordinate are obvious. Instead, TPT should be viewed as a rigorous formalism for studying reaction dynamics once aimless shooting and likelihood maximization or the GNN method has identified the important collective variables.

Here we describe a variation on the BOLAS method for computing free energies by path sampling (*56*), originally developed by Radhakrishnan and Schlick. As described in Peters *et al.* (*55*), the name equilibrium path sampling (EPS) has been proposed to clarify some confusion over the distribution of paths that the BOLAS algorithm generates. BOLAS and EPS combine the features of Monte Carlo umbrella sampling with features from path sampling. Both methods are useful for computing the potential of mean force (PMF) along a reaction coordinate that is not easily differentiable. CHARMM, NAMD, and other packages provide many methods for computing a PMF along simple coordinates like distances, angles, dihedrals, and functions of these quantities, but not for computing a PMF along a non-differentiable coordinate. Aimless shooting and likelihood maximization frequently identify complex collective coordinates that are not easily differentiated. Thus BOLAS and EPS provide a way to use even the most complex reaction coordinates that are identified by aimless shooting and likelihood maximization.

The following discussion of EPS assumes some familiarity with the MD umbrella sampling technique, which is briefly discussed here. In umbrella sampling, multiple windows along a reaction coordinate are specified and a harmonic restraint is placed on a reaction coordinate value at the center of the window. MD simulations are run in each window along the reaction coordinate with the restraint active. The potential of mean force (the reversible work) is calculated from the negative natural log of the resulting probability distribution of the system in each window. The weighted histogram analysis method (WHAM) is used to match up the overlapping portions of the windows (*99*). For more details, see references (*65*, *89*), and (*99*).

The algorithm for EPS is as follows: from a known reaction coordinate *r*, windows are specified, denoted by R. The configurations are denoted by **x**.

1. Select one of (k + 1)-timeslices on the previous trajectory: $x^{(o)}(0\Delta t)$, $x^{(o)}(1\Delta t)$… $x^{(o)}(k\Delta t)$.
2. Select a random integer *j* between 0 and *k*. Let the point selected in step (1) be timeslice $j\Delta t$ on a new trajectory: $\mathbf{x}^{(n)}(j\Delta t)$
3. Draw momenta **p** from the Boltzmann distribution and propagate the dynamical equations forward in time from $x^{(n)}(j\Delta t)$ to $x^{(n)}(k\Delta t)$. Also reverse the initial momenta **p** and propagate the dynamical equations of motion back in time to $x^{(n)}(0\Delta t)$.
4. Accept the new trajectory $x^{(n)}(0\Delta t)$, $x^{(n)}(1\Delta t)$… $x^{(n)}(k\Delta t)$ if any timeslice is in R. Reject the trajectory if all of the timeslices are outside of R.

This methodology is illustrated in Figure 14. In Figure 14(a), an example of an accepted trajectory is shown (as a solid line) that is started from a random frame of the previous trajectory (labeled in red). Figure 14(b) shows a trajectory started outside of the window region R, which does not enter the window. This trajectory would be rejected and the previous trajectory would be counted again in the probability distribution along the appropriate reaction coordinate.

As discussed in reference (*55*), many points will extend beyond the window R. However, these points are not distributed according to the equilibrium distribution. To remedy this, as in umbrella sampling, window regions should be selected such that windows overlap. From there, WHAM (*99*) can be used to connect the resulting free energy distributions between windows. In addition, the length of trajectories used in previous applications of EPS is quite short. As with umbrella sampling, the trajectory length and the window width should be adjusted to achieve good convergence and sampling efficiency. For our previous applications, we estimate the diffusion time of the system along the verified reaction coordinate in a given window size, then use this characteristic diffusion time as our trajectory length. Specifically for windows along methane hopping from different hydrate clathrate cages, the window length was 0.1 ps (*55*) and 0.5 ps for a small processive enzyme diffusing on a hydrophobic surface (*100*).

A C code written for EPS as conducted in reference (*55*) is available for download at http://www.engineering.ucsb.edu/~baronp/codes.html. Additional scripts written to resemble the aimless shooting code for EPS are also available for implementation in CHARMM for nucleation in the Lennard-Jones system (Beckham and Peters, in preparation).

5. From Accurate Reaction Coordinate to the Rate Constant

The advantages of identifying a reaction coordinate for computing rate constants are outlined in a paper by Hillier *et al.* (*101*). They compared the TPS procedure for computing a reaction rate without any knowledge of the reaction coordinate (*16*, *17*) to methods for computing the reaction rate from a PMF along the reaction coordinate. Using a reaction coordinate was 27.5 times faster than the approach outlined in references (*16*, *17*). By providing a coordinate that accurately projects the high dimensional barrier crossing dynamics onto a single coordinate, aimless shooting and likelihood maximization enable simpler and more accurate calculations of reaction rates.

The Transition State Theory Rate

Many prevalent rate theories are formulated as corrections to the transition state theory (TST) rate constant. For elementary reaction steps that break and make strong bonds, TST rate constants are often quite accurate. The lengths of the bonds being broken and formed are often sufficient components of the reaction coordinate for these reactions (*102*, *103*). Only a few atoms move significantly in crossing the barrier, and these move over distances of just a few Angstroms. Thus, friction between motion along the reaction coordinate and the bath modes tends to be sufficiently weak that the initial barrier crossing occurs successfully. Then, for a high barrier corresponding to strong bond breaking, even weak friction can dissipate enough of the large amount of energy in the reaction coordinate to prevent an energy-diffusion limit type barrier recrossing. Thus TST for the breaking of strong bonds, even in enzymes, tends to be a good approximation (*104*, *105*). To our knowledge the energy diffusion limit has not been encountered for reactions in biological systems.

Enzymatic bond-breaking and bond-making reactions are extremely important for biomass conversion, but excellent reviews have been given elsewhere (*105*–*107*). Instead, we focus on methods for reactions with greater friction for motion along the reaction coordinate. Examples include conformational and allosteric transitions of biomolecules, competitive binding steps that expel water from a hydrophobic surface or from an enzyme active site, and reactions that involve desolvation like sugar recombination reactions in solution and cellulase processivity initiation.

In the intermediate and high friction regimes TST can be formulated in terms of the square root of $\omega_A{}^2$, the force constant for (undamped) motion along the reaction coordinate in the reactant basin and in terms of the free energy difference between the transition state and the reactant minimum

$$k_{TST} = \frac{\omega}{2\pi}\exp\left[-\beta F_{\neq}\right] \qquad (9)$$

$$k_{TST} = \frac{1}{2x_{\mathrm{A}}} \frac{\langle \delta[q-q^*] |\dot{q}| \rangle}{\langle \delta[q-q^*] \rangle} \qquad (10)$$

where x_{A} is the fraction of time the system spends in the reactant state when the reactant and product are inter-converting at equilibrium.

Reaction Rates with Intermediate Friction

For intermediate values of friction along the reaction coordinate, one could make direct use of Kramers theory to estimate a rate constant. However, the Bennett-Chandler method (*108*, *109*) is a practical strategy that incorporates the real dynamical sources of friction that are being modeled by the "friction" in the Kramers' picture. The Bennett-Chandler method correlates the flux through the dividing surface at the initial moment of crossing with the probability that the initial trajectories remain in the product state some time later. The reactive flux approaches the TST flux at the initial time *t=0*, and decays to a plateau as the velocity along the reaction coordinate relaxes from the initial condition. When normalized by the TST flux, the reactive flux correlation function decays from a value of 1 at *t=0* to a value of κ at the plateau.

$$\kappa(t) = \frac{\langle \dot{q}(0)\delta[q(0)-q^*]h[q(t)-q^*] \rangle}{\langle |\dot{q}(0)|\delta[q(0)-q^*]/2 \rangle} \qquad (11)$$

For extremely long times, *i.e.* the timescale between reaction events, the reactive flux correlation function decays to zero. The plateau value is the transmission coefficient, which is a dynamical correction to the rate constant:

$$k = \kappa(\tau_{\mathrm{plateau}})\, \mathrm{k_{TST}} \qquad (12)$$

The transmission coefficient also serves as a correction to the rate constant for reaction coordinate error, but as the reaction coordinate error becomes large, the transmission coefficient often becomes too small to compute (*110*). Because the efficiency of a transmission coefficient calculation is directly influenced by the accuracy of the reaction coordinate, aimless shooting and likelihood maximization can facilitate rate constant calculations for reactions with intermediate friction.

Figure 15 shows the free energy surface for methane hopping between a donor and an acceptor cage in a methane hydrate, as described in detail in reference (*55*). The dividing surface is shown as a heavy black curve on the free energy surface. The dividing surface location was chosen because of a broad shallow intermediate between the acceptor and donor states. Note the dividing surface in this case is not at p_B=½. In some cases, the optimal dividing surface is not the p_B=½ surface, but another isosurface of the committor probability. In this case the p_B=½ surface is within the shallow intermediate basin. A dividing surface at the plane of symmetry (which must be the p_B=½ surface) would result in a very low transmission coefficient. The mistaken view that optimal dividing surfaces must rigidly be identified with the p_B=½ surface is a common source of confusion. We emphasize this example to show that a good reaction coordinate may indeed give a p_B=½ surface that is not an optimal dividing surface for application of transition

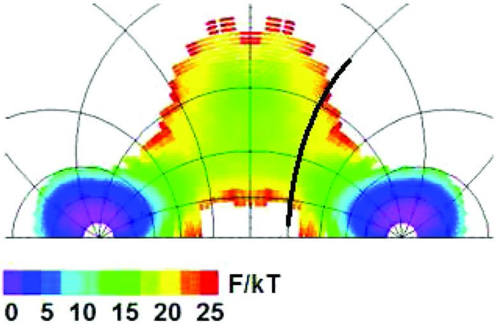


Figure 15. Free energy surface for methane molecule hopping between 2 cages of a methane hydrate. The black line denotes the chosen dividing surface used to calculate the rate constant, which is not at p_B=½ because of the broad shallow basin around p_B=½, which would yield a very low transmission coefficient, κ. Results taken from reference (55). (see color insert)

state theory or for computing a transmission coefficient. In this case the dividing surface we have chosen corresponds to p_B=0.65.

Figure 16 shows the reactive flux correlation function with a plateau that gives a transmission coefficient of *κ=0.3*. The surrounding water molecules exert friction on the methane molecule leading to the small transmission coefficient, *κ*. Figure 17 shows the evolution of a swarm of trajectories initiated from the dividing surface at *t*=0. The dynamics of the swarm exactly corresponds to the free energy surface, behavior that can only be expected for accurate reaction coordinates.

Reaction Rates within the High Friction Limit

For biological systems involving conformational changes of solvated macromolecules, the dynamics along a reaction coordinate are likely to fall in the intermediate to high friction Kramers regime (*30*, *68*, *111*, *112*). The discussion below will show that Kramers' theory in the high friction limit is remarkably simple when a one-dimensional reaction coordinate can be identified. Now with methods like GNN and likelihood maximization this first step is feasible.

As the friction becomes very high, the transmission coefficient decreases, and the reactive flux correlation function becomes difficult to calculate accurately (*110*). Thus for the very high friction regime, a better strategy is to compute a diffusivity for motion along the reaction coordinate. The diffusivity is related to the friction by the Einstein relation, $D = kT/\gamma$, but the friction itself is difficult to obtain. A more practical strategy is to compute the mean squared displacement along the reaction coordinate from an initial condition.

$$\left\langle (r(t)-\langle r(t)\rangle)^2 \right\rangle = 2Dt \tag{13}$$

The duration of trajectories launched from the top of the barrier to calculate *D* must be carefully chosen. In particular, the duration should be longer than the short time Ornstein-Uhlenbeck ballistic behavior of the mean squared displacement, but

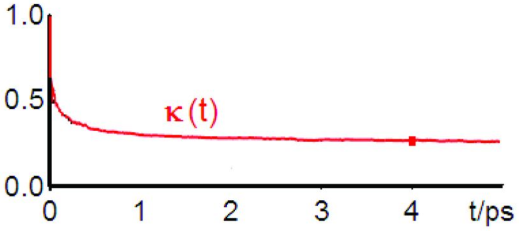


Figure 16. The reactive flux correlation function exhibits a plateau corresponding to the correction factor to the TST rate constant given in Equation 12. Results taken from reference (55). (see color insert)

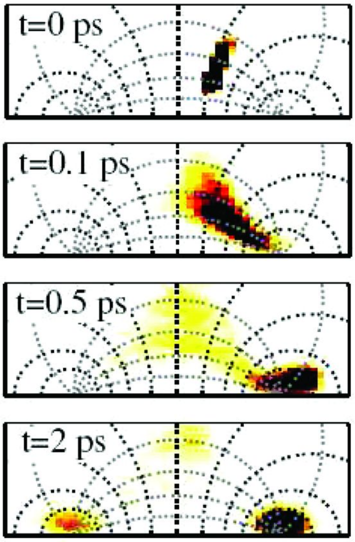


Figure 17. Swarms of trajectories initiated from the dividing surface shown in Figure 15. The resulting endpoints at t=2 ps correspond to the results from the free energy surface above. For more details, see reference (55). (see color insert)

shorter than the time for trajectories to fall significantly from the barrier top. When used within these bounds, Equation 13 provides a simple alternative to methods for computing small transmission coefficients. For very high friction where velocity correlations decay long before the trajectory has left the top of the barrier, the rate constant is:

$$k_{\gamma \gg \omega_{\neq}} = \left[\int_{\cup} \exp[-\beta F(x)]dx \int_{\cap} \frac{\exp[\beta F(x)]}{D(x)} dx \right]^{-1} \qquad (14)$$

Equations 13 and 14 clearly show how accurate 1-D reaction coordinates can help relate simulations of activated processes to experimental kinetics. Once an accurate 1-D reaction coordinate is identified, computing rate constants is reduced to computing the free energy profile, estimating a single diffusion constant along the reaction coordinate, and straightforward quadrature.

6. Summary and Outlook

One of the longstanding promises of molecular simulation is to enable rational design of more effective catalysts and inhibitors by providing molecular level, mechanistic insight into rare events. Determining the reaction coordinate accurately without prior assumptions that may lead to erroneous rate predictions or erroneous insights about the mechanism is an essential first step in using simulations to understand kinetics.

To that end, this review outlines advances in methods to find reaction coordinates from path sampling approaches. Specifically we detailed the aimless shooting approach, which improves on the original transition path sampling algorithm for applications in highly diffusive systems because diffusive dynamics are probable characteristics of many biological and solvated systems. We also discussed the likelihood maximization approach for identifying accurate reaction coordinates from many candidate variables. Equilibrium path sampling and BOLAS free energy methods were discussed as methods to compute the potential of mean force even for coordinates that are not easily differentiated. The path sampling based free energy calculations enable us to make use of highly complex order parameters that may be identified in searching for an accurate reaction coordinate. In kinetic studies of rare events, we expect that these methods and further improvements on these methods will enable determination of accurate reaction coordinates and more accurate rate calculations. In the field of biomass conversion, we expect that the methods described here will form an integral part of the computational toolkit to characterize the molecular-level mechanisms occurring in the construction and degradation of the plant cell wall in both natural and engineered systems.

Acknowledgments

We especially thank Bernhardt Trout for the opportunity to work in his group at MIT, where several methods described in this review were developed. We benefited greatly from his guidance and advice. We have also benefited from many helpful discussions with the developers of path sampling methods including Peter Bolhuis, Christoph Dellago, Ravi Radhakrishnan, David Chandler, Aaron Dinner, and Michael Hagan.

References

1. Baker, J. *J. Comput. Chem.* **1986**, *7*, 385–395.
2. Cerjan, C. J.; Miller, W. H. *J. Chem. Phys.* **1981**, *75*, 2800–2806.
3. Weinan, E.; Ren, W.; Vanden-Eijnden, E. *Phys. Rev. B* **2002**, *66*, 1.
4. Faires, J. D.; Burden, R. L. *Numerical Methods*; PWS-Kent: Boston, 1993.
5. Fischer, S.; Karplus, M. *Chem. Phys. Lett.* **1992**, *194*, 252–261.
6. Halgren, T. A.; Lipscomb, W. N. *Chem. Phys. Lett.* **1977**, *49*, 225–232.
7. Henkelman, G.; Jonsson, H. *J. Chem. Phys.* **1999**, *111*, 7010–7022.
8. Jensen, F. *Introduction to Computational Chemistry*; Wiley: New York, 1999.

9. Mills, G.; Jonsson, H. *Phys. Rev. Lett.* **1994**, *72*, 1124–1127.
10. Peters, B.; Heyden, A.; Bell, A. T.; Chakrabory, A. K. *J. Chem. Phys.* **2004**, *120*, 7877.
11. Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514–1527.
12. Pratt, L. R. *J. Chem. Phys.* **1986**, *85*, 5045–5048.
13. Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.
14. Bolhuis, P. G.; Dellago, C.; Chandler, D. *Faraday Discuss.* **1998**, *110*, 421–436.
15. Dellago, C.; Bolhuis, P. G.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 9236–9245.
16. Dellago, C.; Bolhuis, P. G.; Chandler, D. *J. Chem. Phys.* **1999**, *110*, 6617–6625.
17. Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964–1977.
18. Dellago, C.; Bolhuis, P. G.; Geissler, P. G. *Adv. Chem. Phys.* **2001**, *123*, 1–86.
19. Geissler, P. G.; Dellago, C.; Chandler, D. *J. Phys. Chem. B* **1999**, *103*, 3706–3710.
20. McCormick, T. A.; Chandler, D. *J. Phys. Chem. B* **2003**, *107*, 2796–2801.
21. Pettitt, B. M.; Karplus, M. *Chem. Phys. Lett.* **1985**, *121*, 194–201.
22. Apostolakis, J.; Ferrara, P.; Caflisch, A. *J. Chem. Phys.* **1999**, *110*, 2099–2108.
23. Scarsi, M.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. B* **1998**, *102*, 3637–3641.
24. Marrone, T. J.; Gilson, M. K.; McCammon, J. A. *J. Phys. Chem.* **1996**, *100*, 1439–1441.
25. Bartels, C.; Karplus, M. *J. Comput. Chem.* **1997**, *18*, 1450–1462.
26. Tobias, D. J.; Brooks, C. L. *J. Phys. Chem.* **1992**, *96*, 3864–3870.
27. Lazaridis, T.; Tobias, D. J.; Brooks, C. L.; Paulaitis, M. E. *J. Chem. Phys.* **1991**, *95*, 7612–7625.
28. Chu, J. W.; Brooks, B. R.; Trout, B. L. *J. Am. Chem. Soc.* **2004**, *126*, 16601–16607.
29. Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877–5882.
30. Juraszek, J.; Bolhuis, P. G. *Biophys. J.* **2008**, *95*, 4246–4257.
31. Hu, J.; Ma, A.; Dinner, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4615–4620.
32. Pool, R.; Bolhuis, P. G. *J. Chem. Phys.* **2007**, *126*, 244703.
33. ten Wolde, P. R.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6539–6543.
34. Willard, A. P.; Chandler, D. *J. Phys. Chem. B* **2008**, *112*, 6187–6192.
35. Geissler, P. G.; Dellago, C.; Chandler, D.; Hutter, J.; Parrinello, M. *Science* **2001**, *291*, 2121–2124.
36. Hedges, L. O.; Jack, R. L.; Garrahan, J. P.; Chandler, D. *Science* **2009**, *323*, 1309–1313.
37. Radhakrishnan, R.; Trout, B. L. *J. Am. Chem. Soc.* **2003**, *125*, 7743–7747.

38. Grünwald, M.; Rabini, E.; Dellago, C. *Phys. Rev. Lett.* **2006**, *96*, 255701.
39. Beckham, G. T.; Peters, B.; Trout, B. L. *J. Phys. Chem. B* **2008**, *112*, 7460–7466.
40. Beckham, G. T.; Peters, B.; Variankaval, N.; Starbuck, C.; Trout, B. L. *J. Am. Chem. Soc.* **2007**, *129*, 4714–4724.
41. Zahn, D. *Phys. Rev. Lett.* **2004**, *92*, 040801.
42. Rowley, C. N.; Woo, T. K. *J. Am. Chem. Soc.* **2008**, *130*, 7218–7219.
43. Petridis, L.; Smith, J. C. *J. Comput. Chem.* **2008**, *30*, 457–467.
44. Matthews, J. F.; Skopec, C. E.; Mason, P. E.; Zuccato, P.; Torget, R. W.; Sugiyama, J.; Himmel, M. E.; Brady, J. W. *Carbohydr. Res.* **2006**, *341*, 138–152.
45. Nimlos, M. R.; Matthews, J. F.; Crowley, M. F.; Walker, R. C.; Chukkapalli, G.; Brady, J. W.; Adney, W. S.; Cleary, J. M.; Zhong, L.; Himmel, M. E. *Protein Eng., Des. Sel.* **2007**, *20*, 179–187.
46. Koivula, A.; Ruohonen, L.; Wohlfahrt, G.; Reinikainen, T.; Teeri, T. T.; Piens, K.; Claeyssens, M.; Weber, M.; Vasella, A.; Becker, D.; et al. *J. Am. Chem. Soc.* **2002**, *124*, 10015–10024.
47. Petersen, L.; Ardevol, A.; Rovira, C.; Reilly, P. J. *J. Phys. Chem. B.* **2009**, *113*, 7331–7339.
48. Zhao, X.; Rignall, T. R.; McCabe, C.; Adney, W. S.; Himmel, M. E. *Chem. Phys. Lett.* **2008**, *460*, 284–288.
49. Fushinobu, S.; Mertz, B.; Hill, A. D.; Hidaka, M.; Kitaoka, M.; Reilly, P. J. *Carbohydr. Res.* **2008**, *343*, 1023–1033.
50. Mulakala, C.; Reilly, P. J. *Proteins: Struct., Funct., Bioinform.* **2005**, *60*, 598–605.
51. Peters, B.; Beckham, G. T.; Trout, B. L. *J. Chem. Phys.* **2007**, *127*, 034109.
52. Peters, B.; Trout, B. L. *J. Chem. Phys.* **2006**, *125*, 054108.
53. Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334–350.
54. Ma, A.; Dinner, A. R. *J. Phys. Chem. B* **2005**, *109*, 6769–6779.
55. Peters, B.; Zimmermann, N. E. R.; Beckham, G. T.; Tester, J. W.; Trout, B. L. *J. Am. Chem. Soc.* **2008**, *130*, 17342–17350.
56. Radhakrishnan, R.; Schlick, T. *J. Chem. Phys.* **2004**, *121*, 2436–2444.
57. Auer, S.; Frenkel, D. *Annu. Rev. Phys. Chem.* **2004**, *55*, 333–361.
58. Bolhuis, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12129–12134.
59. Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. *J. Chem. Phys.* **2006**, *125*, 024106.
60. Schlitter, J.; Engels, M.; Kruger, P.; Jacoby, E.; Wollmer, A. *Mol. Sim.* **1993**, *10*, 291–308.
61. Rowley, C. N.; Woo, T. K. *J. Chem. Phys.* **2007**, *126*, 024110.
62. Bolhuis, P. G. *Biophys. J.* **2005**, *88*, 50–61.
63. Hagan, M. F.; Dinner, A. R.; Chandler, D.; Chakrabory, A. K. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13922–13927.
64. Hu, J.; Ma, A.; Dinner, A. R. *J. Chem. Phys.* **2006**, *125*, 114101.
65. Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.
66. ten Wolde, P. R.; Ruiz-Montero, M. J.; Frenkel, D. *J. Chem. Phys.* **1996**, *104*, 9932–9947.

67. Moroni, D.; ten Wolde, P. R.; Bolhuis, P. G. *Phys. Rev. Lett.* **2005**, *94*, 235703.
68. Bolhuis, P. G. *J. Phys.: Condens. Matter* **2003**, *15*, S113–S120.
69. Moroni, D.; Bolhuis, P. G.; van Erp, T. S. *J. Chem. Phys.* **2004**, *120*, 4055–4065.
70. van Erp, T. S.; Bolhuis, P. G. *J. Comput. Phys.* **2005**, *205*, 157–181.
71. Grünwald, M.; Dellago, C.; Geissler, P. L. *J. Chem. Phys.* **2008**, *129*, 194101.
72. Miller, T. F.; Predescu, C. *J. Chem. Phys.* **2007**, *126*, 144102.
73. MacKerell, A. D.; Wiórkiewicz-Kuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946–11975.
74. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
75. Allen, R. J.; Frenkel, D.; ten Wolde, P. R. *J. Chem. Phys.* **2006**, *124*, 194111.
76. Allen, R. J.; Frenkel, D.; ten Wolde, P. R. *J. Chem. Phys.* **2006**, *124*, 024102.
77. Borrero, E. E.; Escobedo, F. A. *J. Chem. Phys.* **2007**, *127*, 164101.
78. Borrero, E. E.; Escobedo, F. A. *J. Chem. Phys.* **2008**, *129*, 024115.
79. Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880–10889.
80. West, A. M. A.; Elber, R.; Shalloway, D. *J. Chem. Phys.* **2007**, *126*, 145104.
81. Moroni, D.; van Erp, T. S.; Bolhuis, P. G. *Physica A* **2004**, *340*, 395–401.
82. van Erp, T. S.; Moroni, D.; Bolhuis, P. G. *J. Chem. Phys.* **2003**, *118*, 7762–7774.
83. Onsager, L. *Phys. Rev.* **1938**, *54*, 554–557.
84. Weinan, E.; Ren, W. Q.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2005**, *413*, 242–247.
85. Peters, B. *J. Chem. Phys.* **2006**, *125*, 241101.
86. Antoniou, D.; Schwartz, S. D. *J. Chem. Phys.* **2009**, *130*, 151103.
87. Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6732–6737.
88. Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci.* **2010**, *107*, 1088–1093.
89. Kottalam, J.; Case, D. A. *J. Am. Chem. Soc.* **1988**, *110*, 7690–7697.
90. Carter, E. A.; Ciccotti, G.; Hynes, J. T.; Kapral, R. *Chem. Phys. Lett.* **1989**, *156*, 472.
91. Kong, X.; Brooks, C. L. *J. Chem. Phys.* **1996**, *106*, 2414–2423.
92. Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.
93. Maragakis, P.; van der Vaart, A.; Karplus, M. *J. Phys. Chem. B* **2009**, *113*, 4664–4673.
94. Voter, A. F. *Phys. Rev. Lett.* **1997**, *78*, 3908–3911.
95. Wang, F.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050.
96. Park, S.; Schulten, K. *J. Chem. Phys.* **2004**, *120*, 5946–5961.
97. Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130*, 194101.
98. Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130*, 194103.
99. Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

100. Bu, L.; Beckham, G. T.; Crowley, M. F.; Chang, C. H.; Matthews, J. F.; Bomble, Y. J.; Adney, W. S.; Himmel, M. E.; Nimlos, M. R. *J. Phys. Chem. B* **2009**, *113*, 10994–11002.

101. Dimelow, R. J.; Bryce, R. A.; Masters, A. J.; Hillier, I. H.; Burton, N. A. *J. Chem. Phys.* **2006**, *124*, 114113.

102. Chu, J. W.; Brooks, B. R.; Trout, B. L. *J. Am. Chem. Soc.* **2004**, *126*, 16601–16607.

103. Lo, C.; Giurumescu, C. A.; Radhakrishnan, R.; Trout, B. L. *Mol. Phys.* **2004**, *102*, 281–288.

104. Ju, L. P.; Han, K. L.; Zhang, J. Z. H. *J. Comput. Chem.* **2009**, *30*, 305–316.

105. Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. *Chem. Rev.* **2006**, *106*, 3210–3235.

106. Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, *303*, 186–195.

107. Gao, J.; Ma, S.; Major, D. T.; Nam, K.; Pu, J.; Truhlar, D. G. *Chem. Rev.* **2006**, *106*, 3188–3209.

108. Chandler, D. *J. Chem. Phys.* **1978**, *68*, 2959–2970.

109. Straub, J. E.; Berne, B. J. *J. Chem. Phys.* **1985**, *83*, 1138–1139.

110. Frenkel D.; Smit B. *Understanding Molecular Simulations: From Algorithms to Applications*, 2002.

111. Schulten, K.; Schulten, Z.; Szabo, A. *J. Chem. Phys.* **1981**, *74*, 4426–4432.

112. Jacob, M.; Geeves, M.; Holtermann, G.; Schmid, F. X. *Nat. Struct. Biol.* **1999**, *6*, 923–926.

Subject Index

A

Aimless shooting, path sampling algorithms, 308, 310*f*, 311, 312*f*, 317*f*, 320*f*, 321*f*
Alcohols and H abstraction reactions, 208, 209, 211*f*, 212, 213*f*, 214, 216*f*
All-atom and coarse-grained representations
  CbhA, 85*f*
  Cel48A, 84*f*
  Cel5B, 85*f*
  CipA from *C. thermocellum*, 83*f*
  cohesin from CipA, 84*f*
Allomorphs
  cellulose, man-made, 20
    II, 21*f*
    III_I, 21*f*
  cellulose, natural, 18
    chain polarity, 19*f*
    Iα and Iβ crystal packing, 20*f*
Anisole
  atoms, 60*t*
  dihedral, 63*f*
  model compounds, 65*f*
AS. *See* Aimless shooting, path sampling algorithms
Atomistic model of CBM1, 106*f*
Atomistic simulation
  cellulosomal protein complexes, 55
  lignocellulosic biomass, 55, 57

B

Bacillus agaradhaerens Cel5A
  deglycosylation, 149
    free energy profile, 151*f*
  glycosyl-enzyme intermediate
    active site, 149*f*
    deglycosylation, 149*f*
    structure, 145*f*
BDE. *See* Bond dissociation energies, water elimination from alcohol
Beta-glucose
  chair conformation, 141*f*
  Stoddart's diagram, 141*f*
Biomass
  composition, 253*f*
  gas velocity vectors, 252*f*
  gasification and pyrolysis
    CFD modeling, 273, 275
    CFD process design and optimization, 281
  lignocellulosic, simulation, 55, 57
  mass distribution, 252*f*
  pyrolysis, 249*f*, 251*t*
  thermochemical conversion
    catalytic surface, 261
    compound wavelet matrix method, 257, 258, 259
    computational savings, 267*f*
    configuration and flow parameters, 251*t*
    continuum level computational fluid dynamics, 248
    coupling methods, 257
    kinetic models, 201
    kinetic Monte Carlo method, 256, 265*f*, 266*f*
    lattice Boltzmann method, 253
    motivation, 246
    multiphysics components, 247, 265*f*
    multiscale/multiphysics modeling, 245
    physical and chemical processes, 279
    processes and modeling, 277
    scales encountered, 246
    time parallel compound wavelet matrix method, 262, 263, 266*f*, 267
    time parallel method, 263*f*
B3LYP/6-31G(d) method, water elimination from alcohols, 227*f*, 228*f*
Bond dissociation energies, water elimination from alcohol, 222*t*

C

Carbohydrate-binding module
  atomistic model, 106*f*
  CBH I, 108*f*
  CBH II, 108*f*
  cellulose surface, translation, 99, 107
  coarse-grained model, 106, 106*f*
Carbohydrates, coarse-grained model, 102
CBD. *See* Cellulose binding domain
CBH. *See* Cellobiohydrolase
CbhA
  all-atom and coarse-grained representations, 84*f*

construct, 80*f*
CBH I. *See* Cellobiohydrolase I
CBM. *See* Carbohydrate-binding module
CBM1. *See* Family 1 carbohydrate-binding module
CBS-4M, 236*f*
homolysis, 238*f*
levels of theory, 227*f*, 228*f*
CBS-QB3, 236*f*
calculations, 222*t*
data, 228*f*
homolysis, 238*f*
Cel5A
B. agaradhaerens, 145*f*
deglycosylation, 149, 151*f*
glycosyl-enzyme intermediate, 145*f*, 149, 149*f*
QM/MM simulations, 146
Cel6A, 143*f*
Cel9A
catalytic and immunoglobulin-like modules, 81*f*
construct, 80*f*
Ig-Gh9
cross-correlation map, 92*f*
modules, 88*f*, 89*f*
normal mode analysis, 89
simulation, 95*f*
Cel12A
H. grisea, 145*f*
glycosylation, 146, 147*f*, 148*f*
Michaelis complex, 147*f*
QM/MM simulations, 146
Cel48A, 84*f*
Cel5B, 85*f*
Cellobiohydrolase, 101, 120, 136, 137, 137*t*
Cellobiohydrolase 1
linker
cellulose surface, 128*f*, 129*f*
polypeptide, 126*f*
Cellobiohydrolase I, 122*f*
amino acid sequence, 122*f*
CBM1, 108*f*, 109, 110*f*, 112*f*, 113*f*
cellulase systems, 121
glycosylation, 122*f*
linker polypeptide, 121, 122*f*
Cellobiohydrolase II, 109
CBM1, 108*f*, 110*f*
Cellobiose, 136*f*
adiabatic potential energy maps, 24*f*
conformational energy mapping, 22
crystal, 29, 29*t*
CSFF force field, 24*f*, 26*f*, 28*f*, 28*t*
GLYCAM06 force field, 24*f*, 26*f*, 27*f*, 28*f*, 28*t*
hydroxymethyl combinations, 26*f*
unit cell parameters, 29*t*
Cellotetraose
9 glycosyl hydrolase, 80*f*
H. grisea Cel12A, 145*f*
Cellulase
active sites, 135, 137
linker peptides, 119
energy storage, 119
simulation, 126
QM/MM analysis, 135
systems, 120
cellobiohydrolase I, 121
linker polypeptide, 121
Cellulose, 136*f*
allomorphs, 18, 19*f*, 20*f*
background, 18
binding domain, 122*f*
bond length distribution, 105*f*
coarse-grained beads, 103*f*
coarse-grained bond B1-B4, 105*f*
coarse-grained model, 104
fibrils
hydrogen bonding, 40, 42*f*, 43*f*
twist, 40*f*, 41*f*, 46*f*
hydrolysis, enzymes, 137*t*
man-made
allomorphs, 20
II, 21*f*
III_I, 21*f*
natural, 18
plant cell wall components, 67*f*
simulations, 17
structure, 17
surface
carbohydrate-binding module, 99
CBH 1 linker polypeptide, 126*f*
CBM1 of CBH I, 113*f*
uses, 18
Cellulose binding domain, 122*f*
Cellulose Iβ crystal, 30*f*
Cellulosomal enzymes, 84
scaffoldin molecule, 87*f*
simulation box, 87*f*
Cellulosome
assembly, coarse-grained modeling, 80
bonded terms, 81
cellulosomal enzymes, 84
cohesin and dockerin, 83
functional form and parameters, 81
non-bonded terms, 81
scaffold subunit, 82
cohesin-dockerin complex, 79*f*
computational methods, 79
concept and architecture, 77
C. thermocellum, 79*f*

  - enzymes in simulation box, 88*t*
  - modeling, 64, 75
  - modules, function, 78
  - molecular dynamics simulations, 92
  - protein complexes
    - architecture, 78*t*
    - atomistic simulation, 55
- Cellulosome-integrating protein
  - *C. thermocellum*, 77*f*
    - all-atom representation, 83*f*
    - coarse-grained representation, 83*f*
  - cohesin, 84*f*
- CelS, structure, 140*f*
- CFD. *See* Computational fluid dynamics
- Ciccotti's method, 124
- CipA. *See* Cellulosome-integrating protein
- *Clostridium thermocellum*
  - CbhA, 80*f*
  - Cel9A, 80*f*
  - cellulosome
    - cohesin-dockerin complex, 79*f*
  - CipA
    - all-atom representation, 77*f*, 83*f*
- Coarse-grained representation
  - atomistic cellulose residue model, 103*f*
  - CBM, 106*f*
  - CBM1, 111
  - cellulose, 104
  - partially populated scaffold, 87*f*
- Cohesin
  - CipA, 84*f*
  - dockerin complex, 64
    - crystal structure, 67*f*
    - *C. thermocellum*, cellulosome, 79*f*
    - free energy landscape, 68, 69*f*
- Compound wavelet matrix
  - multiscale/multiphysics components, coupling, 257, 258, 259
  - temporal, 263*f*, 266*f*
- Computational fluid dynamics
  - biomass gasification and pyrolysis, 273
    - computational resources and modeling challenges, 289
    - descriptive to predictive, 289
    - models, 284, 287, 287*t*
    - numerical experiments, 282
    - process design and optimization, 281
    - reactor scale-up, 283
    - realistic configurations, 285
- Crystal structure reorganization
  - CSFF, one ns simulation, 31
    - hydrogen bonds, 34*f*
    - structures, 33*f*
  - diagonal crystal and DP 40 crystals, ten ns simulations
    - CSFF, 35*f*, 36*f*
    - GLYCAM06, 36*f*
    - heating and equilibration period, 35*f*
    - Iβ crystals, 36*f*
  - hydrogen bond patterns, 32*f*
- CSFF
  - crystal structure reorganization, 35*f*
    - one ns simulation, 31
    - ten ns simulations, 35*f*
  - DP 40, 36-chain crystal, 37*f*, 42*f*
  - force field, 24*f*, 26*f*, 41*f*
    - cellobiose, 28*f*, 28*t*
    - cellobiose hydroxymethyl, 26*f*
  - hydrogen bonds, 43*f*
  - methanol dimer hydrogen, 45*f*
- CWM. *See* Compound wavelet matrix
- Cyclohexene, 234*f*
- Cystine, 110*f*

D

- Deglycosylation
  - *B. agaradhaerens* Cel5A, 149, 149*f*, 151*f*
- β-D-glucopyranose-gg, 187*f*
- β-D-glucose, 22*f*
- Dihedrals
  - anisole, 63*f*
  - β-O-4' linkage, 66*t*
- Discrete particle model, simulation, 280*f*
- DP 14, 36-chain crystal
  - CSFF force field, 47*f*
  - twist, 46*f*
- DP 40, 36-chain crystal
  - CSFF, 42*f*, 43*f*
  - GLYCAM06, 42*f*, 43*f*
  - glycosidic linkages, 37*f*, 39*f*
  - twist, 46*f*
- DP 40, 16-chains crystal
  - CSFF force field, 47*f*
  - twist, 46*f*
- DPM. *See* Discrete particle model

E

- Endoglucanase, 56, 76, 78, 84, 120, 136, 137*t*
- Enzymes
  - cellulose, hydrolysis, 137*t*
  - reaction mechanism, 136, 139*f*
  - substrates, action, 135
- EPS. *See* Equilibrium path sampling
- Equilibrium path sampling, 322*f*
- Ethylene glycol, 207*f*
- Evans-Polanyi plot, 221*f*

Exoglucanase, 56, 78, 102, 120, 136

F

Family 1 carbohydrate-binding module
  atomistic model, 106*f*
  CBH I, 108*f*, 109, 110*f*, 113*f*
  CBH II, 109, 110*f*
  cellulose surface, translation, 107
  coarse-grained model, 111
  molecular dynamics simulation, 108*f*
Family 48 enzyme, 141*f*
Fluidized-bed reactor, 249*f*, 276*f*, 278*f*
Free energy
  B. agaradhaerens Cel5A, 151*f*
  basins, 302*f*, 303*f*
  calculations, 123
  CBH1 linker and cellulose surface, 128*f*, 129*f*, 130*f*
  Ciccotti's method, 124
  cohesin-dockerin dissociation, 68, 69*f*
  deglycosylation, 151*f*
  landscape, 68
  methane hydrate, 326*f*, 327*f*
  reaction coordinate, 300, 302*f*, 321
  umbrella sampling, 125

G

Gasification and pyrolysis, 273
Gas-phase elimination of water from alcohols
  bond dissociation energies, 222*t*
  Evans-Polanyi plot, 221*f*
  rate rules, 216, 229
    development, 219
    lower levels of theory, performance, 226
    transferability, 222
  unimolecular rate constants, 220*f*
Gas velocity vectors
  biomass mass distribution, 252*f*
  void fraction, 252*f*
Genetic Neural Network method, 314, 317*f*
GH9. *See* 9 Glycosyl hydrolase and cellotetraose
GLYCAM06
  crystal structure reorganization, 36*f*
  force field, 24*f*, 26*f*, 27*f*, 28*f*, 28*t*
  simulations, 46*f*
  twist of cellulose fibrils, 46*f*
Glycoside hydrolases, 142
Glycosidic bond hydrolysis
  inverting mechanism, 139*f*
  retaining mechanism, 139*f*
Glycosidic linkages
  DP 40, 36-chain crystals, 37*f*
    CSFF, 39*f*
    GLYCAM06, 39*f*
Glycosylation
  CBH I, 122*f*
  H. grisea Cel12A, 147*f*, 148*f*
9 Glycosyl hydrolase and cellotetraose, 80*f*
GNN method. *See* Genetic neural network method
Guaiacyl, 59*f*

H

H abstraction reactions
  alcohol, 211*f*
    C-H bond, 209, 213*f*, 214, 216*f*, 220*f*
    CH_3 radicals, 208, 217*f*, 218*f*, 219*f*, 220*f*
    H atoms, 208, 209, 211*f*, 219*f*
    hydroxyl group, 209
    methyl radicals, 214
  rate rules, 208
Heated biomass particle, structural changes, 279*f*
Heterogeneous chemical reactor, 247*f*
Homolysis
  CBS-4M, 238*f*
  CBS-QB3, 238*f*
  1,3-H shift, 238*f*
  1,5-H shift, 238*f*
  rate constants, 238*f*
Humicola grisea Cel12A
  cellotetraose, 145*f*
  glycosylation, 147*f*, 148*f*
  Michaelis complex, 147*f*
  QM/MM simulations, 146
Hydrogen bond
  cellulose fibrils, 40*f*
  crystal structure reorganization, 34*f*
  methanol dimer, interaction energy, 44*f*
  patterns, 32*f*
Hydroxymethyl rotamers, 22*f*

I

Iβ crystals
  CSFF 10 ns simulations, 36*f*
  GLYCAM06 10 ns simulations, 36*f*
Ig-Gh9, 89
  atomic fluctuations, 91*f*, 93*f*, 95*f*

Cel9A, 88*f*, 89*f*
cross-correlation map, 92*f*
molecular dynamics simulation, 94*f*
modules, 88*f*
molecular dynamics simulation, 93*f*, 95*f*
Ig-like protein, 80*f*
Interaction energy
water:anisole, 62*t*
water:lignin-dimer, 62*t*
Inverting mechanism, glycosidic bond hydrolysis, 139*f*

K

Kinetic models, thermal conversion of biomass, 201
Kinetic Monte Carlo method, 256, 259*f*, 260, 265*f*
KMC. *See* Kinetic Monte Carlo method

L

Lattice Boltzmann method, 253
kinetic Monte Carlo, coupling, 259*f*, 260
LBM. *See* Lattice Boltzmann method
Lignin
decomposition, 235
molecular mechanics force field
bond and angle vibrations, 61
dihedral parameters, 61
force field validation, 62
parameterization strategy, 58
partial atomic charges, optimization, 59
plant cell wall components, 67*f*
Lignocellulosic biomass
atomistic simulation, 55, 57
models, 63
Likelihood maximization, 313, 316, 317*f*, 318*t*, 321*f*
Linker polypeptide in CBH I, 121

M

MC simulations. *See* Monte Carlo simulations
MD simulations. *See* Molecular dynamics simulations
Meso-scale modeling, polysaccharides in plant cell walls, 99
Methane hydrate and free energy surface, 326*f*
Methanol dimer
hydrogen bond
CSFF, 45*f*
GLYCAM06, 45*f*
interaction energy, 45*f*
interaction energy, 44*f*
Modeling cellulosomes, 64
multiscale methods, 75
Molecular dynamics simulations
CBH 1 linker polypeptide, 126*f*
CBM1, 108*f*, 112*f*
cellulosome, 92
Ig-Gh9, 93*f*, 95*f*
unit cell properties, 66*t*
Molecular model
CBH I
amino acid sequence, 122*f*
linker peptide, 122*f*
Molecular simulation methods
accuracy
force field accuracy, 162
polarizable force fields, 163
QM/MM, 164
dynamics and conformational sampling, 155
interaction potential, 156
molecular dynamics simulations, 92
choice of time step, 158
integrator, 158
simulations and convergence, interpretation, 160
statistical ensembles, 158
Monte Carlo sampling, 161
speed
biased sampling, 167
coarse-graining, 173
multi-scale methods, 170
replica exchange simulations, 169
sampling techniques, enhanced, 166
solvent environments, implicit description, 171
time scales, 165
Monolignols
guaiacyl, 59*f*
p-coumaryl, 59*f*
sinapyl, 59*f*
Monte Carlo simulations, 155, 161
Multi-scale methods, 170, 171*f*
Multiscale/multiphysics modeling, biomass thermochemical processes, 245

N

- NMA. *See* Normal mode analysis
- Normal mode analysis
  - Cel9A, 89
  - Ig-Gh9, 91*f*

P

- Parameterization strategy, 60*f*
- Particle-laden flows
  - biomass thermochemical conversion, 277
  - gas phase coupling, 280
  - hydrodynamics, 277
  - reactive particles, 281
  - solid phase description, 277
- Path sampling
  - accurate reaction coordinate, 299
  - algorithms, 304
    - aimless shooting methods, 308, 310*f*, 311
    - basin definitions, 305
    - initial pathways, 305
    - transition path sampling, 307
  - reaction coordinate, 321
    - Genetic neural network method, 314, 317*f*
    - likelihood maximization, 316, 317*f*, 318*t*, 320*f*, 321*f*
    - pB histogram test, 313, 315*f*, 316*f*
  - reaction coordinates, 299
  - transition, 302
- pB histogram test, 313, 315*f*, 316*f*
- *p*-Coumaryl, 59*f*
- PES. *See* Potential energy surfaces
- Phenethyl phenyl ether
  - Arrhenius rate expressions, 239*t*
  - decomposition, 234*f*
- PHP. *See p*-Hydroxyphenyl
- *p*-Hydroxyphenyl, 59*f*, 65*f*
- Plant cell wall
  - cellulose, 66*t*
  - lignin, 66*t*
  - polysaccharides, 99
- Polysaccharides, 99
- Potential energy surfaces
  - phenyl vinylethyl ether, 236*f*
  - vinyl vinylethyl ether, 236*f*
- PPE. *See* Phenethyl phenyl ether
- Product gas mass fraction, 255*f*
- Pyrolysis, 273

Q

- QM/MM. *See* Quantum mechanical/ molecular mechanical
- Quantum mechanical methodology, 180
  - sugar thermochemistry
    - balanced equations, 185
    - composite methods, 181
    - homodesmotic reactions, 185
- Quantum mechanical/molecular mechanical
  - analysis
    - cellulase active sites, 135
    - enzymes on substrates, actions, 135
    - model setups, 144
  - free energy simulation, 151*f*
  - molecular simulation methods, 135
  - simulation, 148*f*
    - Cel5A, 146
    - glycosylation, 146
    - *H. grisea* Cell12A, 146

R

- Rate rules
  - gas-phase elimination of water from alcohols, 216
  - H abstraction reactions, 216
- RDA. *See* Retro-Diels-Alder
- Reaction coordinate
  - free energy, 300, 302*f*, 303*f*, 314*f*, 321
  - isosurfaces, 314*f*
  - path sampling, 299, 312
  - rate constant
    - high friction limit, 326
    - intermediate friction, 325
    - methane hydrate, 326*f*, 327*f*
    - transition state theory, 324, 327*f*
- Retaining mechanism, 139*f*
- Retro-Diels-Alder
  - cyclohexene, 234*f*
  - rate constants, 234*f*
  - reactions, 230

S

- Scaffoldin molecule and cellulosomal enzymes, 87*f*
- Serine, 122*f*
- Simulations
  - atomic

associated cellulosomal protein complexes, 55
lignocellulosic biomass, 55
Cel9A, 95*f*
cellulose, structure, 17
molecular methods, 155
Sinapyl, 59*f*
Stoddart's diagram and beta-glucose, 141*f*
Sugar
extension, 187, 189*f*
ring
distortion, 140
subsite -1, conformational changes, 143*f*
thermochemistry
derived thermodynamic data, 195*t*
quantum mechanical modeling, 179

T

Tar gas mass fraction, 255*f*
Thermal conversion of biomass
calculation methodology
electronic structure calculations, 204
rate expressions, 206
thermodynamic properties, 205
first principle methods, 201
kinetic models, 201
rate estimation rules, 201
See also Biomass
Thermodynamic data
computation and literature, 195*t*
derived, 195*t*
Theronine, 122*f*
Time-parallel compound wavelet matrix method, biomass thermochemical conversion, 262, 263, 266*f*, 267
Time parallel method, 263*f*
tpCWM. *See* Time-parallel compound wavelet matrix method
TP method. *See* Time parallel method
TPS. *See* Transition path sampling
Transition path sampling, 302, 308*f*, 312*f*
Transition state theory rate, 324, 327*f*
TST rate. *See* Transition state theory rate
Twist
cellulose fibrils
CSFF diagonal crystal, 40*f*
GLYCAM06 simulations, 46*f*
Tyrosine, 110*f*

U

Unit cell
cellulose Iβ crystal, 30*f*
small-molecule-dimer, 66*t*

V

Vinyl vinylethyl ether, 236*f*
Void fraction with gas velocity vectors, 252*f*

W

Water:anisole-dimer, 62*t*
Water elimination from alcohols
B3LYP/6-31G, 230*f*
B3LYP/6-31G(d), 227*f*, 228*f*
bond dissociation energies, 222*t*
CBS-4M, 227*f*, 228*f*, 230*f*
CBS-QB3
data, 228*f*
level, 230*f*
rate constants, 227*f*, 228*f*, 232*f*
B3LYP/6-31G, 227*f*
CBS-4M levels of theory, 227*f*
TST, 227*f*
Water:lignin-dimer, 62*t*