ACS Catalysis

Research Article

pubs.acs.org/acscatalysis

Deriving Chemically Essential Interactions Based on Active Site Alignments and Quantum Chemical Calculations: A Case Study on Glycoside Hydrolases

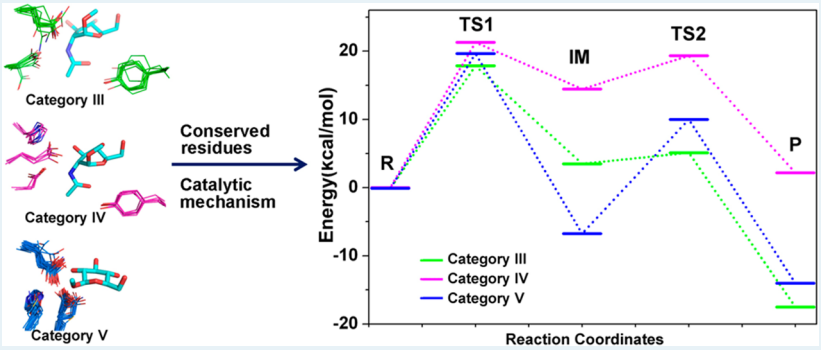
Yinliang Zhang,[†] Zheng Zhao,*[,§] and Haiyan Liu*[,†,‡,§]

[†]School of Life Sciences, University of Science and Technology of China, 443 Huangshan Road, Hefei, Anhui 230027, China

[‡]Hefei National Laboratory for Physical Sciences at the Microscales, Hefei, Anhui 230027, China

[§]Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, Anhui 230031, China

Supporting Information



ABSTRACT: We here use an approach of active site alignment and clustering of many evolutionarily distant enzymes catalyzing alike reactions to identify conserved residues/interactions that may play key chemical roles in catalysis. Then density functional theory (DFT) calculations on cluster models are used to investigate the chemical essentialness of such residues/interactions and its mechanistic basis. We apply this approach to 130 glycoside hydrolases (GHs) of the $(\beta\alpha)_8$-barrel fold. These enzymes adopt either a classical retaining mechanism or a substrate-assisted intramolecular nucleophilic attack mechanism, both in need of a general acid/general base residue for catalysis. On the basis of the multiple active site alignments, the enzyme active sites can be clustered into six categories. The conserved or convergently evolved hydrogen bond/salt bridge involving the general acid/general base in different categories suggests the importance of this interaction. DFT calculations indicate that its presence may reduce the energetic barrier by as large as 17−20 kcal mol^{-1}. The mechanistic explanation for this large effect is that a proton transfer from the general acid to the leaving group takes place before the nucleophile attacks the transition state. The large energetic effect suggests that this interaction should be considered as chemically essential, although it is realized with varied residue types in different GH categories. In addition, for the substrate-assisted mechanism, an interaction between the substrate nucleophile group and a tyrosine is found to have been convergently evolved in enzymes of two different categories. This interaction does not seem to have favorable effects on the energetic barrier. Instead, it might contribute to reducing the activation entropy. In summary, active site alignment of distant enzymes combined with quantum mechanical calculation may comprise a powerful approach to obtain new insights into enzyme catalysis.

KEYWORDS: *active site alignments, quantum chemistry calculations, glycoside hydrolase, conserved interactions, general acid/general base catalysis*

1. INTRODUCTION

Through chemical intuitions combined with extensive biochemical studies, especially structural analysis and site-directed mutagenesis of representative enzymes, general mechanisms of many enzymes catalyzed reactions have been depicted.[1,2] A general mechanism usually includes the basic chemical steps as well as some core catalytic residues that participated in these steps, such as the residues forming chemical bonds with the substrate in covalent catalysis, or the residues providing/accepting protons to/from the substrate in general acid/general base catalysis.[1,3,4] Beyond such a general mechanism and besides the core catalytic residues it considers, it is often still interesting to ask to what extent the basic chemical steps are assisted by the remaining environment of an enzyme's active site, and how. If theoretical analysis or other evidence indicates high activation barriers associated with one or more of the basic chemical steps without extra interactions other than those proposed in the general mechanism,[5,6] one may even

Received: October 31, 2014
Revised: March 12, 2015
Published: March 13, 2015

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

raise the question whether there are additional residues and/or chemical interactions that contribute critically to catalysis, and if yes, what are these residues/interactions? To answer such questions is important not only for deeper mechanistic insights but also for inhibitor design[7] and enzyme design.[8]

Questions regarding key catalytic residues in an enzyme were usually addressed by human inspection of the three-dimensional organizations of the active sites of individual enzymes.[3,4] Candidate residues were suggested and then verified using experiments such as site-directed mutagenesis.[2] Although this approach has yielded extensive knowledge about enzyme catalysis, the suggestion of key residues based on structure inspection is somewhat subjective. Some chemically critical residues/interactions might be overlooked. Sequence alignment of homologous enzymes has for long played an important role in elucidating enzymatic mechanisms by telling the evolutionary variability of different residues.[3] However, sequence alignments are more reliable for enzymes that are evolutionarily close to each other. Such alignments usually yield many highly conserved residues in the active site, not all of them playing important chemical roles. In addition, important interactions conserved in structure but not in sequence[9] may not be recognized.

Enzymes with different overall sequences and structures may share similar catalytic mechanisms. Therefore, focusing the comparisons between different enzymes on the active sites instead of their overall sequences or structures could be more efficient and accurate. As evolutionarily distant proteins are compared, conserved residues and interactions are more likely to be catalytically important. In this work, we use a structure-based sequence-order-independent active site alignment approach (SMAP) to identify chemically essential residues/interactions in enzyme catalysis.[10−12] In this approach, the active sites of a relatively large number of evolutionarily distant enzymes catalyzing alike reactions are compared and aligned. The SMAP method is used to perform the pairwise alignments. Then a multiple active site alignment is built from all pairwise alignments. A minimum model of the reaction center can be proposed on the basis of the aligned active sites. Such a model may include additional residues/interactions besides those indicated in the respective general mechanism. The necessity to include them can be verified using quantum mechanical (QM) calculations.

We applied this approach to glycoside hydrolases (GHs)[13−16] of the $(\beta/\alpha)_8$-barrel fold or the triosephosphate isomerase (TIM)-fold.[17,18] GHs are a widely distributed group of enzymes. They cleave glycosidic bonds in glycosides, glycans, and glycol conjugates, and they are important candidates in the development of biofuels and in disease research.[19,20] The TIM-fold is one of the most ancient protein folds[17] and the most commonly adopted fold in GHs.[18] To date, GHs of the TIM-fold have been observed to use either a classical mechanism that retains the anomeric configuration (Figure 1A) or a substrate-assisted mechanism (Figure 1B).[15,16,18] The classical mechanism was first proposed by Koshland in 1953.[21] It involves two core catalytic residues, both being either glutamate or aspartate, to provide two carboxyl groups, one acting as a nucleophile to attack on the carbon center of the scissor glycosidic bond and the other as a general acid to protonate the leaving glycosidic oxygen. The second carboxylate also acts as a general base to deprotonate the nucleophilic water in the subsequent hydrolysis step (Figure 1A). The substrate-assisted mechanism works with *N*-acetyl-glucosamine-containing substrates.[15,16] It also uses two core catalytic carboxylates. One plays the same roles of the general acid/general base as in the classical mechanism, whereas the other stabilizes the charge-redistributed

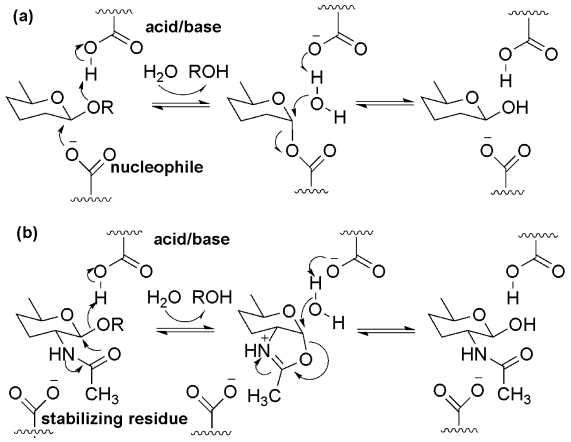


Figure 1. General mechanisms of TIM-fold GHs. (a) The classical two-step retaining mechanism. (b) The substrate-assisted intramolecular nucleophilic attack mechanism.

intermediate (Figure 1B). Comprehensive classifications based on sequence and structure information on a large number of TIM-barrel GHs are available in the database of carbohydrate-active enzymes (CAZy).[22] The compiled sequence and structure data provide solid starting ground for applying the active site comparison approach to these enzymes. In addition, existing knowledge about the catalytic mechanisms provide good references for the judgment of results for validity and for new insights.

We use the SMAP program developed by Bourne and co-workers to carry out the pairwise comparisons of the active sites of TIM-fold GHs.[10−12] The program uses a graph algorithm to perform a structure-based sequence-order independent alignment between the two sets of amino acid residues forming the active sites, respectively, with the alignment score defined based on the chemical similarity between aligned residues. The significance of an alignment is defined as the probability for two unrelated active sites to produce a better-scoring alignment than the alignment in consideration. The method has been applied mainly to predict target proteins for a given ligand based on the known binding sites for the target in some template proteins.[23,24] Here we apply it to identify structurally conserved residues at the active sites of TIM-fold GHs. For this purpose, we developed a tool to merge the pairwise alignments into a single multiple alignment including all considered GH active sites. This allows us to group the GHs into different categories, each has a characteristic set of structurally and chemically conserved active site residues. Within each category, some specific interactions involving the chemically active groups of the substrate or of the core catalytic residues emerge as being highly conserved, suggesting their potential importance for the proposed chemical steps in respective general mechanisms.

To look at the essentialness of these residues/interactions for catalysis, QM calculations are carried out to investigate their effects on the activation barriers of the chemical steps proposed for the classical retaining mechanism or substrate-assisted mechanism. Here we have used gas phase cluster models,[6] which include only the chemically active parts of the active site and the substrate. More sophisticated approaches including combined quantum mechanical/molecular mechanical (QM/MM) models have been widely used to simulate enzymatic reactions (for reviews, see refs 25,26). Such methods allow the effects of the entire enzymatic and solution environment on the chemical pathways and the transition barriers to be modeled, with contributions of thermodynamics

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

Table 1. PDB Chains and GH Families Covered by Different Active Site Categories

category	PDB entries	GH families
I	1GJW_A, 1HX0_A, 1IV8_A, 1JDC_A, 1JG9_A, 1LWJ_A, 1MXG_A, 1UA7_A, 1UH4_A, 2BHZ_A, 2D2O_A, 2D3N_A, 2FHF_A, 2GDV_A, 2GVY_A, 2QPU_A, 2VR5_A, 2YA1_A, 2YA2_A, 2Z1K_A, 3AXI_A, 3BC9_A, 3BMW_A, 3EDF_A, 3FAX_A, 3K8M_A, 3VGF_A, 3VU2_A, 3WDJ_A, 3ZOA_A, 3ZT5_A, 4E2O_A, 4J3V_A, 3AIB_A, 3KLL_A, 1ESW_A	GH13, GH70, GH77
II	1T0O_A, 1UAS_A, 3A23_A, 3LRL_A, 4DO4_A, 2F2H_A, 2G3N_A, 2XVK_A, 3L4Y_A, 3MKK_A, 3W37_A, 4AMW_A, 4BA0_A, 4KWU_A, 2YFO_A	GH27, GH31, GH36
III	1EOM_A, 1LLO_A, 1UR9_A, 2A3E_A, 3A4X_A, 3ALG_A, 3ARQ_A, 3CO4_A, 3N17_A, 3WL1_A, 4B16_A, 4MB4_A, 4PTM_A	GH18
IV	1C7S_A, 1NOW_A, 3OZP_A, 3SUV_A, 4AZ6_A, 4H04_A	GH20
V	1UWS_A, 1V03_A, 3AIR_A, 3QOM_A, 3VIG_A, 4PBG_A, 1JZ8_A, 2VJX_A, 2VZS_A, 3HN3_A, 3OB8_A, 1CEN_A, 1ECE_A, 1QNR_A, 1UZ4_A, 2CKR_A, 2JEQ_A, 2OSX_A, 2WHL_A, 3AOF_A, 3N9K_A, 3NDZ_A, 3PZI_A, 3QHO_A, 3ZMR_A, 4HU0_A, 8A3H_A, 4GZJ_A, 1ODZ_A, 2BVT_A, 2V3G_A, 3VPL_A, 3WDR_A, 4CD5_A, 2NT0_A, 2Y24_A, 3OGV_A, 3THC_A, 3W5G_A, 4E8C_A, 1UHV_A, 4KH2_A, 1KWK_A, 3TTY_A, 4BQ4_A, 1QW9_A, 2VRQ_A, 3UG4_A, 2GFT_A, 4CCD_A, 2W62_A, 3VNZ_A, 4AW7_A, 4CD8_A	GH1, GH2, GH5, GH17, GH26, GH30, GH35, GH39, GH42, GH50, GH51, GH53, GH59, GH72, GH79, GH86, GH113
VI	1E5N_A, 1US2_A, 1V0L_A, 2D22_A, 2W5F_A, 3W25_A	GH10

fluctuations to free energy barriers taken into account.[27] In the current context, however, small QM models that are limited to include only the interactions common to different enzymes are preferred over QM/MM models. With a QM/MM model considering an enzyme-specific environment, it would be difficult to separate the effects of enzyme-specific interactions from those of the common interactions. In addition, the effects we are looking at are so significant that they can already be captured largely by relatively small gas phase cluster models that can be treated purely quantum mechanically.[6]

We note that there have been a number of previous QM/MM studies on several GHs of the TIM-fold, including the studies of Jitonnom et al.[28,29] and Greig et al.[30] on different GH enzymes adopting the substrate-assisted mechanism, and the study of Badieyan et al.[31] on an enzyme adopting the classical retaining mechanism. Compared with these studies, the shift of focus from specific/individual enzymes to common features of different enzymes allows us to obtain new insights into the roles of key residues shared by different enzymes. More details will be given in discussion of results.

2. MATERIALS AND METHODS

2.1. Data Set of TIM-Fold GHs. Glycoside hydrolase (GH) families of $(\beta/\alpha)_8$ fold were retrieved from the CAZy database,[22] as described below. There are a total of 1468 PDB entries.[32] From them, we found 747 structures that contained at least one carbohydrate ligand. Then we selected only one enzyme to represent a group of enzymes having higher than 40% sequence identity with the representative enzyme. This led to a final data set of 130 single protein chains representing 26 GH families according to the CAZy classification (see Table 1).

2.2. Pair-Wise Comparisons of Binding Sites. The carbohydrate binding sites of each pair of GH were compared and aligned using the SMAP software[10−12] version 2.0. An all-against-all scheme was used to compare the 130 enzyme structures in a pair-wise manner. As the method including the reported significance of alignments is not symmetric with respect to changing the order of the two GHs in a pair, each GH pair was compared twice with the order of the two enzymes swapped. The alignment with a larger significance (i.e., a smaller SMAP P-value[11]) was retained. We have also used another active site alignment program, ProBiS[33,34] to validate the pairwise binding site alignments generated by SMAP. SMAP and ProBiS indeed give quite similar results: pairs that give significant SMAP P-values are always associated with large ProBiS Z-scores (Figure S1). For these pairs, the corresponding alignments are also equivalent (results not shown).

2.3. Build Multiple Structure Alignments. The pairwise alignments were merged into one multiple active site alignment that included all considered enzymes. Not all the pairwise alignments are consistent with each other (for example, supposing residue A in enzyme 1 is aligned to residue B in enzyme 2 and to residue C in enzyme 3 in respective pairwise alignments, residues B and C may not necessarily be aligned to each other in the pairwise alignment of enzyme 2 versus enzyme 3). To resolve such conflicts during the building of a multiple active site alignment, we developed a heuristic procedure called P2M. It comprises the following steps.

1. A graph G is constructed to summarize all the pairwise alignment results. Each vertex (V) of G represents one residue in one of the enzymes to be aligned. An edge between two vertices indicates that the two residues represented by them are aligned to each other in the pairwise alignment of their containing enzymes.
2. A column of the multiple active site alignment can be represented by a connected subgraph C of graph G. Note that no two vertices of C can be associated with the same enzyme for C to represent an alignment column. Such a subgraph C is extracted in the following manner. First, a degree of connection between any two vertices associated with different enzymes is calculated as the total number of paths connecting them in G. Then the pair of vertices with the highest degree of connection is taken as initial seeds to build the subgraph C by iterations described in step 3.
3. Check each vertex that is not in C but connected to C, to see if it is directly connected to more than a threshold fraction (here 50%) of vertices in C. If yes, the vertex is added to C. If two vertices in C would be associated with the same enzyme, the one with a smaller number of connections within C is removed from C. Then each previous vertex in C is rechecked to see if it remains to be sufficiently connected within the enlarged C. If not, it is also removed from C. This process is repeated until C no longer increases. An aligned column is formed by the residues corresponding to the vertices in C. It contains residues that are densely aligned to each other in the pairwise alignments (all each residue is aligned to more than 50% of the remaining residues in the column), and no two residues are from the same enzyme.
4. The vertices and edges in the subgraph C are removed from G, and step 2 is reentered.
5. Steps 2−4 are repeated until G is empty. The aligned columns are put together to produce the overall multiple enzyme alignment.

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

2.4. Cluster the Enzymes and the Aligned Positions. The enzymes and aligned positions (columns) are clustered based on the multiple active-site alignment. The aligned columns are ranked in descending order based on the number of enzymes contributing to each of them. The first 304 columns, each containing at least four aligned residues, have been considered in subsequent analysis. The alignment is converted into a matrix of elements 0 and 1. Each row of the matrix corresponds to an enzyme. Each column corresponds to an aligned position. An element of 0 indicates the absence of an aligned position from an enzyme, while an element of 1 indicates the opposite. Similarity between any two enzymes (aligned positions) is defined as the Pearson correlation between the corresponding two row (column) vectors of the 0−1 matrix. With this definition, hierarchical clustering of the enzymes as well as of the aligned positions is performed using the Cluster 3.0 software.[35] Based on the clustering, the 130 enzymes are found to fall into six categories. For each category, a set of conserved active site residues emerge from the multiple active site alignment.

2.5. QM Calculations. Guided by the general mechanisms in Figure 1, small sets of residues/interactions that are potentially very important for the chemical steps could be recognized for respective enzyme categories. For an enzyme category, this set may include residues and/or interactions beyond those depicted in the respective general chemical mechanisms. Various cluster models are defined to investigate whether the extra residues/interactions are truly essential for the chemical steps to be feasible by first principle DFT calculations. A cluster model is usually extracted from a representative PDB structure, containing several catalytic side chains surrounding the substrate. The side chains are connected to methyl groups fixed at positions of corresponding Cα atoms in the PDB structure. A reaction coordinate is chosen based on the possible chemical mechanism. The energy profile along the reaction coordinate is obtained by geometry optimization at the B3LYP/6-31G* level[36−38] of theory with the reaction coordinate restrained to change from the reactant to the product (see also captions of Supporting Information Figures S2−S4). Specifically, a harmonic restraining potential of the form $V_{\text{restrain}} = 1/2\ k_{\text{restrain}}\ (R_c - R_c^{\text{ref}})^2$ is applied to the system. The reaction coordinate R_c is defined as a combination of interatomic distances associated with forming/breaking covalent bonds in the concerned mechanism, namely, $R_c = \sum_i c_i d_i$, in which d_i are the distances, with $c_i = 1$ if the bond is to be broken or $c_i = -1$ if the bond is to be formed. The restraining force constant k_{restrain} is 1000 kcal mol^{-1} Å^{-2}. The reference reaction coordinate value R_c^{ref} starts from that of starting reactant and is changed by 0.1 Å after one point has been optimized to optimize the next point. The actual reaction coordinate values after restrained optimization are always within 0.01 Å from the respective reference values.

For results of this reaction coordinate driving approach to be meaningful, it is important to make sure that the system closely follows a single minimum energy path throughout the energy profile. This can be determined from continuity of the curves of individual distances and of the energy changing with the reaction coordinate. Thus, geometries and energies are monitored to ensure smooth and continuous reaction paths. When not certain, the reaction coordinate driving calculation has been repeated in the backward direction using the final optimized structure of the forward calculation as starting point. Small hysteresis indicates that both the forward and backward profiles closely follow the same single minimum energy path. Composition of the various cluster models and the respective reaction coordinates are given in Results and in captions of Supporting Information (Figures S2−S4). QM calculations have been performed using the Gaussian03 program using the program's default convergence criteria for geometry optimization (maximum force below 0.00045 and root-mean-square force below 0.000300 in atomic unit).[39] All calculations have been carried out on a DAWNING-A840 computer.

3. RESULTS

3.1. Alignment and Clustering of Active Sites. The complete multiple alignment merged from the all-against-all pairwise alignments between the 130 enzyme active sites contained 304 columns (see Supporting Information Table S1). Figure 2 depicts the 0−1 alignment matrix, with the orders of the alignment columns as well as of the enzymes rearranged to have elements belonging to same hierarchical clusters next to each other. The pattern of rectangular blocks suggests that the 130 enzymes can be partitioned into six main categories (Table 1). The different categories obtained here through three-dimensional structural alignment of only the active sites are in general consistent with the respective CAZy classifications which have been derived from overall sequence comparisons, indicating validity of the SMAP alignments. At the level of GH Family of CAZy, different categories cover varied numbers of GH families (Table 1). At the level of GH Clan of CAZy, the GH-families clustered into category I form Clan-H in the CAZy classification of GHs.[22] The GH families clustered into category II belong to Clan-D. The GHs clustered into categories III and IV both belong to Clan-K, while GH families clustered into Categories V and VI both belong to Clan-A.

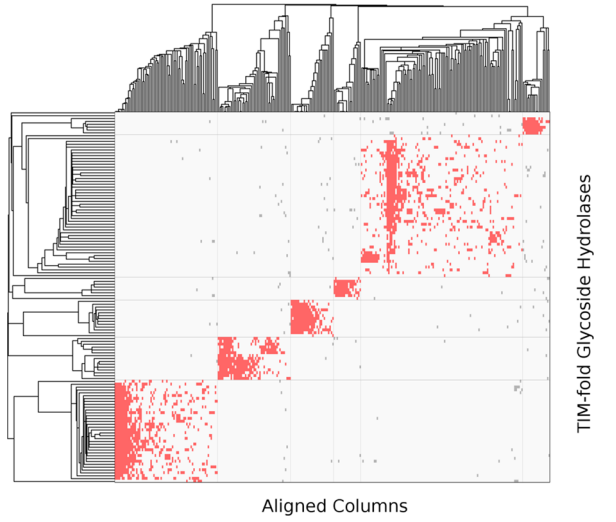


Figure 2. Multiple alignment matrix showing the effects of clustering. Each column corresponds to an aligned column of amino acid residues, and each row corresponds to a GH protein. A filled area represents the presence of an aligned residue in the corresponding GH protein at the corresponding aligned column, while an empty area represents a gap. To show the effects of clustering, the columns and proteins have been ordered so that similar elements are next to each other to form a continuous block.

3.2. Conserved Residues and Interactions at the Active Sites of Different Categories. In the following results, a conserved residue in a category of GH active sites will be labeled by on which one of the $(\beta\alpha)_8$ structure unit the residue is located, followed by the conserved residue type (for example, "4-D" stands for an aspartate on the fourth $(\beta\alpha)$ unit of the TIM-barrel fold). If circular permutations have taken place in some of the member GH families (for example, GH-70 compared with GH-13 or GH-77 in category I), the numbering of the $(\beta\alpha)$ unit will be just based on

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

one circular permutation form. The correspondences of the conserved positions to their actual sequence locations in a representative GH (indicated by PDB ID and chain ID in a PDB structure) belonging to the category are given under the sequence conservation logos for different categories in Figures 3a−8a.

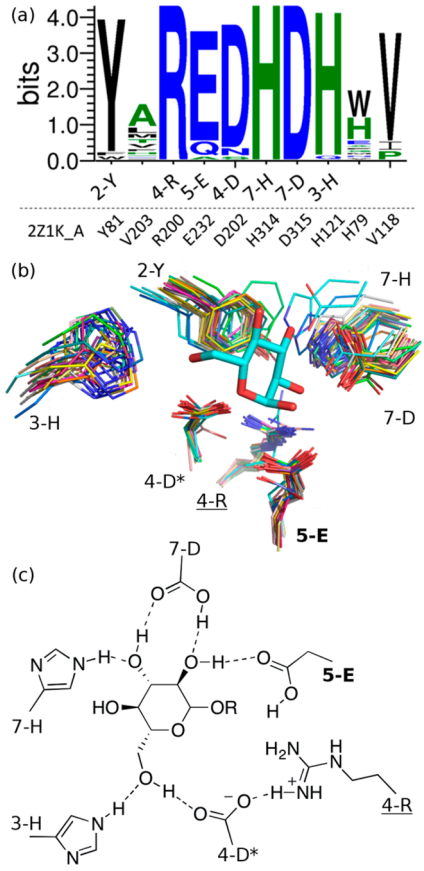


Figure 3. Conserved residues of active sites in category I. (a) Sequence conservation logo of the most conserved aligned positions. The amino acid types observed at a position are shown by the stacked one-letter codes. The size of a letter is proportional to the frequency of the corresponding amino acid type. (b) Superimposed conserved residues surrounding a partial substrate. (c) Schematic representation of the interactions between these residues. The number in a residue label represents the ($\beta\alpha$) unit on which the residue is located, while the letter indicates the conserved residue type. The general acid/general base is indicated with a bold label. The nucleophile is indicated by a star. The residue forming specific interactions with the core catalytic residues is underlined. Sequence conservation logos in this and in subsequent figures have been generated using WebLogo.[85]

Category I. GHs in this category contained seven active site residues that have been aligned. The conserved residue types can be visualized from the sequence profile for these positions shown as a sequence logo in Figure 3a. Figure 3b shows superposed conserved residues from different GHs surrounding a glycosidic partial substrate. The conserved interactions involving these residues are schematically drawn in Figure 3c. Residue 4-D is the catalytic nucleophile in the classical retaining mechanism, while residue 5-E is the general acid/general base. Most of the remaining conserved residues interact with the chemically inactive parts of the substrate, except for residue 4-R, whose guanidine group forms a salt bridge with the nucleophilic residue 4-D.

Structural or functional roles of the conserved residues in Figure 3 have been discussed in a number of previous studies on various GHs. For examples, residues 4-D and 5-E have been mutated to investigate their roles in catalysis.[40−45] Residues 2-Y, 3-H, and 7-H have also been mutated.[46,47] Residue 7-D has been suggested to play an important role in stabilizing the transition state.[48,49] It together with the residues 4-D and 5-E has been named a catalytic triad.[50] From the aligned active sites here, the main role of residue 7-D seems to be not a chemical one but to hold the sugar ring in position through hydrogen bonding with the 2-hydroxyl. Residue 4-R was mostly conserved, but was replaced by Lys in some GH-77 proteins.[51,52] Mutations of this residue in human pancreatic α-Amylase resulted in a 20−450-fold decrease in the activity of the enzyme toward starch and shifted the pH optimum to a more basic pH.[53] No significant structural changes of the catalytic nucleophile and only a minor reorientation of the carbonyl group of acid/base catalyst have been observed in the mutant. Thus, its contribution to catalysis had been attributed mainly to electrostatic effects on transition state stabilization.[53]

Category II. Active sites of this category contained characteristic residues whose sequence locations, three-dimensional arrangements, and specific interactions are shown in Figure 4. The core catalytic residues in the classical retaining mechanisms correspond to residues 4-D (nucleophile) and 6-D (general acid/general base). In some GHs of this active site category (members of family GH-27), there is also a conserved arginine (residue 6-R) that forms a salt bridge with residue 6-D. However, the residue aligned to this

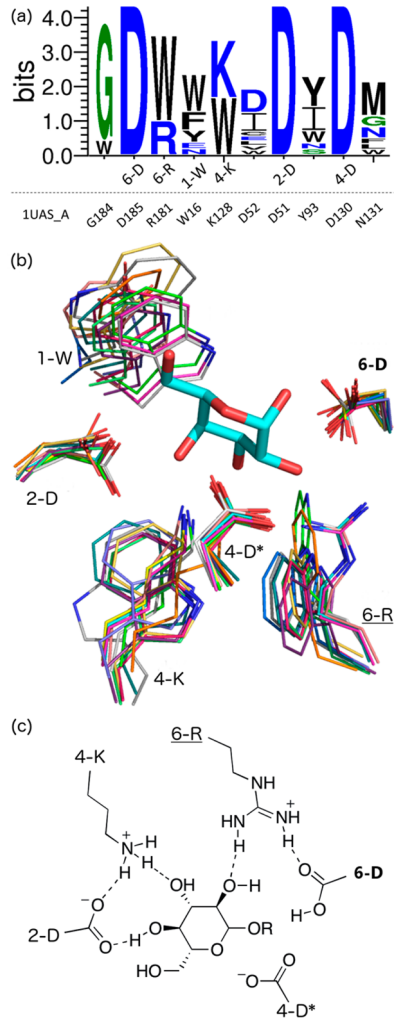


Figure 4. Same as Figure 3, but for active sites of category II.

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

position is a tryptophan in other GHs (members of families GH-31 and GH-36).

GHs in this category may fall into two subcategories, one subcategory comprising the GH-27 and GH-36 family members, the other the GH-31 family members. Within the GH-31 family, there are additional conserved residues and interactions (see Supporting Information Figure S5). Among them is an arginine residue 5-R interacting with the catalytic general acid/general base.

Category III. The conserved sequences, structures, and interactions are shown in Figure 5. GHs in this category adopt

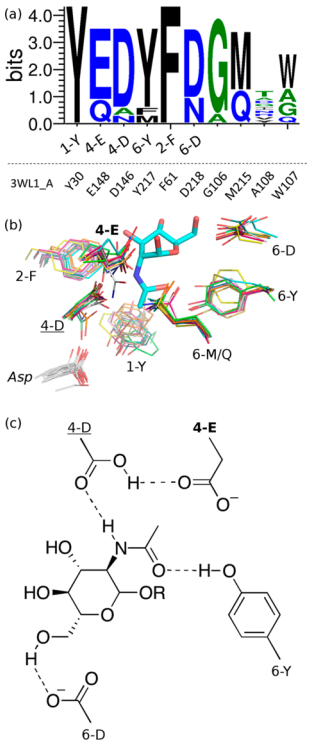


Figure 5. Same as Figure 3, but for active sites of category III.

the substrate-assisted mechanism. The conserved residue 4-E provides the general acid/general base. Another conserved residue 4-D stabilizes the intermediate. In some PDB structures, Ala, Asn, or Gln occur in these positions because structures of mutants were reported.[44,54−57]

The aligned structures show two conserved interactions that involve the chemically active groups suggested by the general mechanism but not included in the description of the mechanism. One is the hydrogen bond interaction between the two core catalytic residues 4-D and 4-E (Figure 5c). The other is the hydrogen bond interaction between the phenol of the conserved residue 6-Y and the substrate *N*-acetyl carbonyl, which is the suggested nucleophile in the substrate-assisted mechanism. In addition, 1-Y is hydrogen bonded with another conserved Asp but far from the substrate (shown in white in Figure 5b). The residue interacts with the *N*-acetyl group from the other side can also be well aligned and is relatively conserved (residue 6-M/Q).

Previous analysis of the conserved residues in this category includes the mutation of 6-Y, which was found to reduce specific activity by 2 orders of magnitude.[58] An interesting QM/MM study by Jitonnom et al. specifically discussed the roles of 4-D and 6-Y in *Serratia marcescens* Chitinase B.[29] They suggested that a neutral 4-D is preferred to provide electrostatic stabilization of the oxazolinium ion intermediate formed in the reaction and that 6-Y plays a critical role in the deglycosylation step of the reaction. The residue 6-D has been suggested to promote distortion of the sugar ring and to increase the pK_a of the catalytic acid.[58] The role of another conserved Asp (shown in white in Figure 5b) has also been suggested to form a catalytic DxDxE motif with 4-D and 4-E.[59]

Category IV. The conserved sequences, structures, and interactions are shown in Figure 6. GHs in this category also

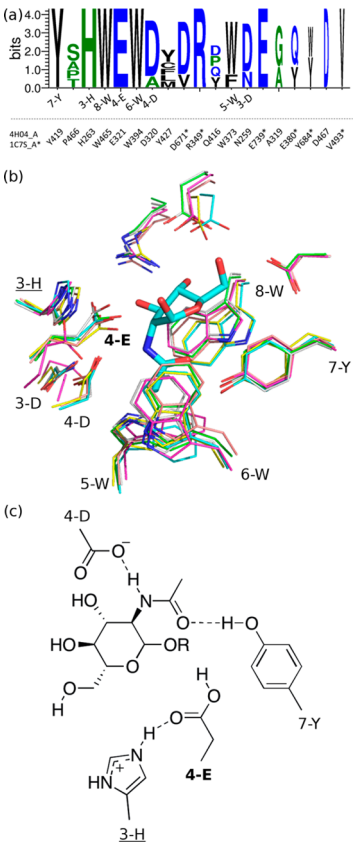


Figure 6. Same as Figure 3, but for active sites of category IV.

adopt the substrate-assisted mechanism, with residue 4-E as the general acid/general base and residue 4-D to stabilize the intermediate. Unlike category III, there is no direct hydrogen bonding interaction between the two residues. Instead, the side chain of a conserved histidine (residue 3-H) forms a hydrogen bond with residue 4-E. As in category III GHs, the side chain of a conserved tyrosine (residue 7-Y) is also observed to interact with the *N*-acetyl carbonyl. This tyrosine, however, is located on a different β strand as compared with the residue 6-Y in category III GHs. In addition, the *N*-acetyl group of the substrate is wrapped by the side chains of three conserved tryptophans 5-W, 6-W, and 8-W (Figure 6 B).

Previously, mutation of 7-Y to Phe in an enzyme of this category was found to reduce both the K_m and k_{cat} significantly.[60] The mutation of 3-H to Phe also showed reduced k_{cat}.[60]

Category V. Conserved sequences, structures, and interactions for the GHs in this category are shown in Figure 7. These GHs adopt the classical retaining mechanism, with the conserved residue 7-E as the nucleophile and residue 4-E as the general acid/general base. In most GHs of this category, there is a conserved hydrogen bond between the chemically active 4-E and the side chain of a histidine (residue 6-H).

Previously, a QM/MM simulation has indicated significant contributions of 4-N and 6-Y to catalysis, and some minor effects of

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

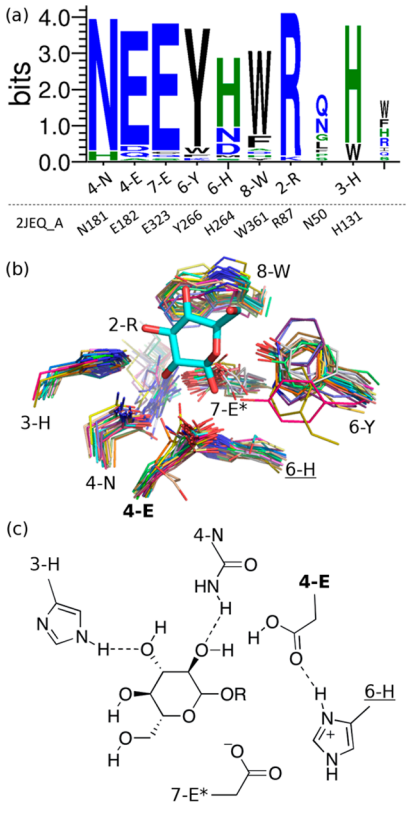


Figure 7. Same as Figure 3, but for active sites of category V.

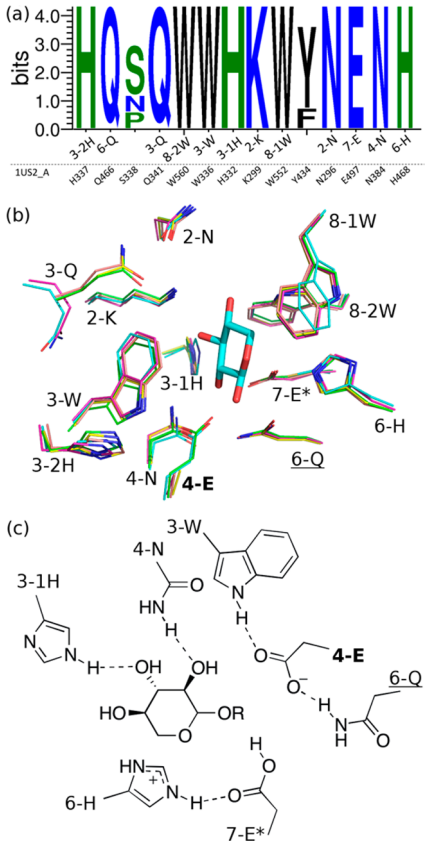


Figure 8. Same as Figure 3, but for active sites of category VI.

3-H.[31] In β-mannosidase, 4-N is replaced by Asp. Mutations and inhibitor analyses suggested that this position contributes to substrate recognition and transition-state stabilization.[61] Mutation of 6-Y to Phe reduces the hydrolytic activity and results in a slight conformational change of the general acid/general base.[62] Residue 6-H has been suggested to form a distinctive catalytic module Glu−His−Glu with the two catalytic glutamates.[63] Mutation of residue 2-R suggested that it is involved in the structural organization of the protein.[64] It was also suggested to play a role in the activation of the nucleophile.[65] Mutation of 2-R may also be related to some diseases.[66] Residue 8-W has been discussed for its stacking with substrates.[67]

Category VI. Conserved sequences, structures, and interactions for the GHs in this category are shown in Figure 8. As GHs in category V, these GHs also adopt the classical retaining mechanism, with the conserved residue 7-E as the nucleophile and residue 4-E as the general acid/general base (Residue 4-E is absent from the sequence logo in Figure 8a because it has been mutated in four out of six PDB structures in this category). Characteristic interactions involving the core catalytic residues include a hydrogen bond between 4-E and the side chain of a glutamine (residue 6-Q), and another hydrogen bond (or salt bridge) between 7-E and the side chain of a histidine (residue 6-H).

Previously, residues 4-N and 4-E have been recognized as a conserved NE pair. Suzuki et al. provided snapshots of the components of the entire reaction using double mutant of the conserved NE pair, including the E-S complex, the covalent intermediate, breakdown of the intermediate and the enzyme−product (E-P) complex.[68] Residue 6-H was suggested to stabilize the nucleophile residue 7-E through electrostatic interactions.[69,70]

3.3. Comparing the Active Sites of Different Categories. Although the active sites in different categories usually do not share aligned columns in the multiple alignment (Figure 2), reasonably significant pairwise alignment scores were observed occasionally, suggesting that the organizations of the active sites in different categories could still be related. Figure 9 shows for each category the locations of the core catalytic residues and some of their interacting residues in the overall $(\alpha\beta)_8$ fold. Like in many other enzymes of this fold, these residues are located at the C-terminal end of the β strands or the N-terminal end of the $\beta\alpha$ loops. Categories I, II, III, and IV all have a core catalytic Asp (residue 4-D) at a similar location, although this Asp has been suggested to play different chemical roles: although it acts as the nucleophile of the classical retaining mechanism in categories I and II, it is the key intermediate-stabilizing residue in the substrate-assisted mechanism in categories III and IV. Beside residue 4-D, remaining catalytic residues in categories I and II do have similar locations. These two categories may be evolutionarily related, with the other catalytic residues evolved independently despite their similar chemical roles. Categories III and IV may also be evolutionarily related. However, the catalytically similar residues 4-D and 4-E are separated by one residue in category III but are next to each other in category IV. The arrangement of the two residues in category III makes it possible for residue 4-D to form a direct hydrogen bond with the residue 4-E, making it a better general acid/general base in the chemical steps. Although this interaction is missing in category IV, a conserved histidine not found in category III is hydrogen bonded to 4-E. In addition, the conserved Tyr that has similar interactions with the substrate carbonyl in categories III and IV does not locate on the same β strand in different categories, suggesting it to be a result of independent but convergent evolution. Categories V and VI and categories III and IV are similar in the location of their catalytically conserved Glu (residue 4-E) in the overall fold. This Glu plays a similar role as general acid/general base. Categories V and VI are further similar to each other in their

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

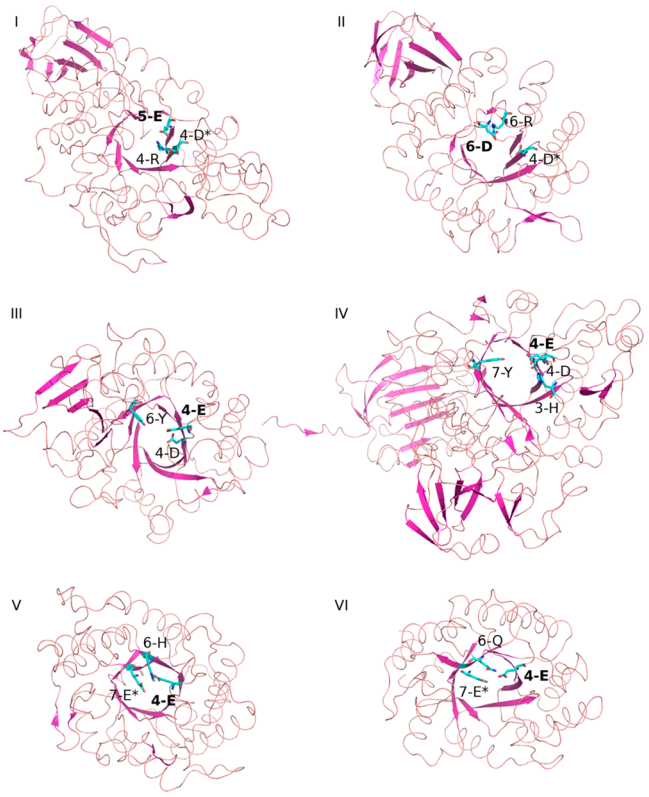


Figure 9. Locations of core catalytic residues and some of their interacting partners in the overall $(\alpha/\beta)_8$ fold. The labels I–VI represent categories I–VI, respectively. Residues are labeled by the index of the α/β unit and one-letter code residue type. Bold labels indicate the general acid. Stars indicate the nucleophile.

location of the nucleophilic Glu residue (residue 7-E) and the residue specifically interacting with the general acid/general base residue 4-E (residue 6-H in category V and residue 6-Q in category VI).

From the above comparisons, several observations emerged regarding interactions that are conserved across evolutionarily distant enzymes and potentially have significant effects on the chemical steps. These interactions have not been emphasized in the general mechanisms, but they directly involve the chemically active groups proposed in the general mechanism. A most ubiquitous interaction is the hydrogen bonding interaction or even charged interaction involving the general acid/general base residue. Another interaction that may be essential for the substrate-assisted mechanism is the interaction between a tyrosine and the substrate carbonyl, because different evolutionary processes have probably led to the same interaction.

3.4. Effects of Conserved Interactions Determined by Quantum Mechanical Calculations. To verify the essentiality of the above specific interactions and to investigate the roles of them in the chemical steps, we extracted models of small clusters from the active sites of representative enzymes, and we investigated energetics of respective chemical processes using first principle density functional theory calculations. Three enzymes have been selected on the basis of their suggested chemical mechanisms and available structural data. The selected enzymes, *Ostrinia furnacalis* Group I Chitinase, β-hexosaminidase from *Paenibacillus*, and family 5 xyloglucanase from *Paenibacillus*, belongs to categories III, IV, and V, respectively. The respective PDB IDs are 3WL1, 3SUV, and 2JEQ.

In 3WL1, which adopts the substrate-assisted mechanism, the general acid/general base E148 interacts directly with another core catalytic residue D146. In 3SUV and 2JEQ, the respective general acid/general base residues (E322 in 3SUV and E182 in 2JEQ) both interact with a histidine (H258 in 3SUV and H264 in 2JEQ). 3SUV adopts the substrate-assisted mechanism, whereas 2JEQ adopts the classical retaining mechanism. The effects of the conserved substrate carbonyl-binding tyrosine in the category III and category IV enzymes are also investigated by including respective tyrosines (Y217 of 3WL1 and Y395 of 3SUV) in some of the cluster models. Although the QM model representing the classical retaining mechanism has been extracted from one member of Category V among the four Categories (Categories I, II, V, and VI) that adopt such a mechanism, the model comprised key residues/interactions common to all these Categories. Except for the starting structure and the boundary atom positions, the QM model is not specific for the chosen enzyme, so repetitive calculations with QM models extracted from other categories of enzymes are unnecessary. Compositions of the different cluster models, including possible protonation states and reaction schemes, are summarized in Figure 10. In this figure and in the following discussions, the label of a model includes the enzyme category and the one-letter codes of residues included. For residues with alternative protonation states, the bare one letter amino acid code represents a deprotonated state, whereas an "H" in parentheses following the one letter code indicates the protonated state (The general acid is always considered to be protonated in the reactant state, so its protonation state was not labeled for clarity). For example, CM_III_YED represents a cluster model for enzyme category III containing side chains of a Tyr, the protonated Glu as the general acid and an Asp in the deprotonated state, while CM_III_YED (H) represents the same model except that the last Asp is in the protonated state.

Energetics of the CM_III Models. Based on the changes of relevant interatomic distances along the reaction pathways (Supporting Information Figure S2) obtained from the reaction coordinate driven calculations (See caption of Supporting Information Figure S2), the chemical events should take place in the following order: proton transfer from the general acid E148 to the substrate leaving group, intramolecular nucleophilic attack leading to transition state TS1, dissociation of the leaving group from TS1 to form the intermediate. The hydrolysis of the intermediate takes place in a reversed order. Figure 11a shows the calculated energetics of the different CM_III models. From model CM_III_YED to CM_III_YED(H), there is a large decrease (ca. 17 kcal mol^{-1}) in the energy barrier associated with TS1, suggesting that the protonation of D146 is of critical importance. Only a protonated D146 can form a hydrogen bond with the E148. This is consistent with the fact that proton transfer from E148 to the leaving group is an early event in the chemical process. Then at TS1 the D146 is fully deprotonated and much more strongly stabilized by hydrogen bonding interactions than at the reactant state. From CM_III_YED(H) to CM_III_ED(H), there is a small decrease in the TS1 barrier, suggesting that Y217 does not contribute in an energetically favorable way to the nucleophilic substitution step. We also have investigated the hydrolysis step using the cluster CM_III_YED(H). It turns out that the hydrolysis step has a much lower transition barrier than the nucleophilic substitution step (Figure 11a). So the overall energetic effect of Y217 is not favorable (actually may be slightly unfavorable based on the energy profiles in Figure 11a) on the chemical steps, even though it lowers the barrier of the hydrolysis step. This unfavorable energetic effects on the nucleophilic attack step is understandable

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559–2572

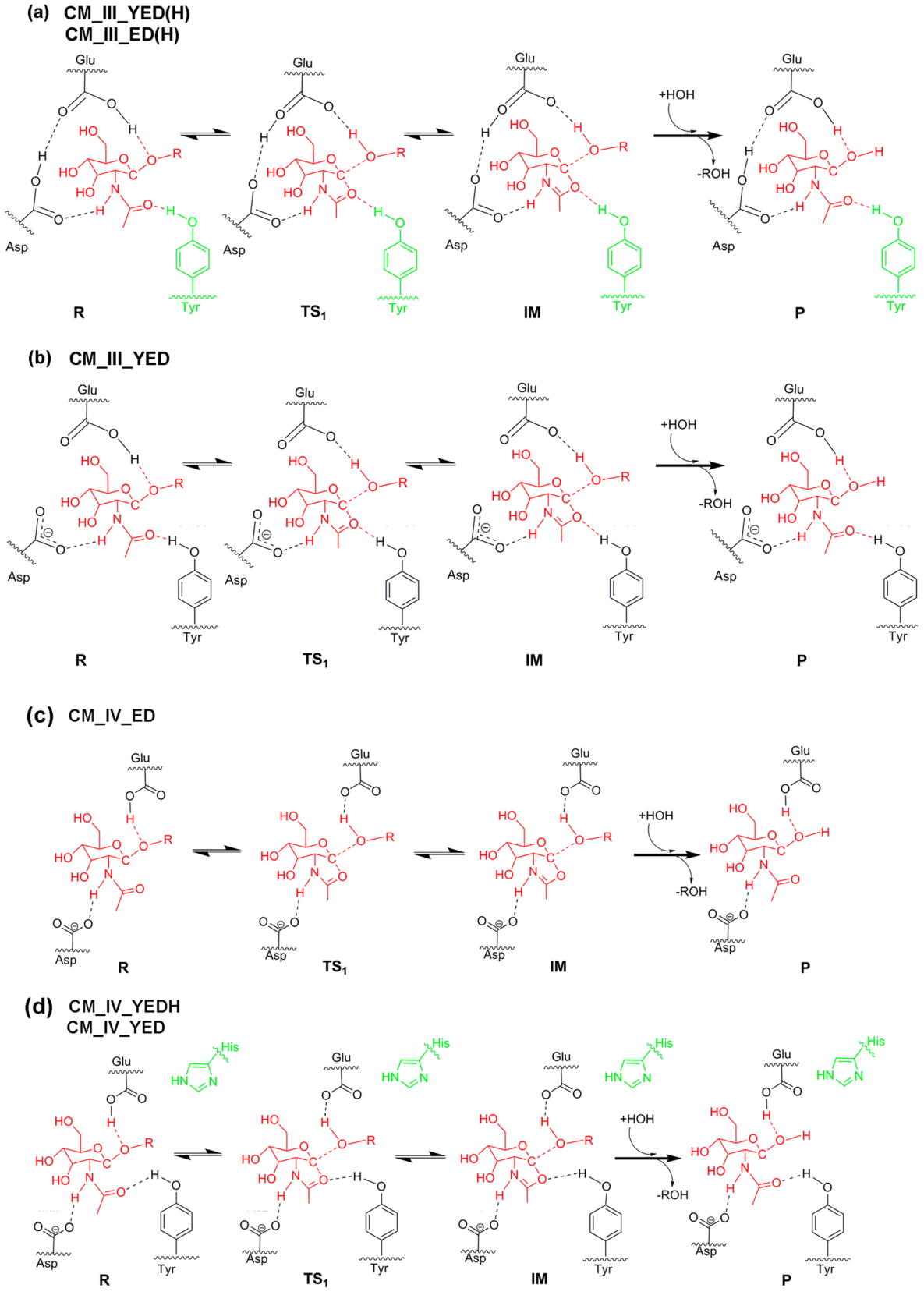


Figure 10. continued

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

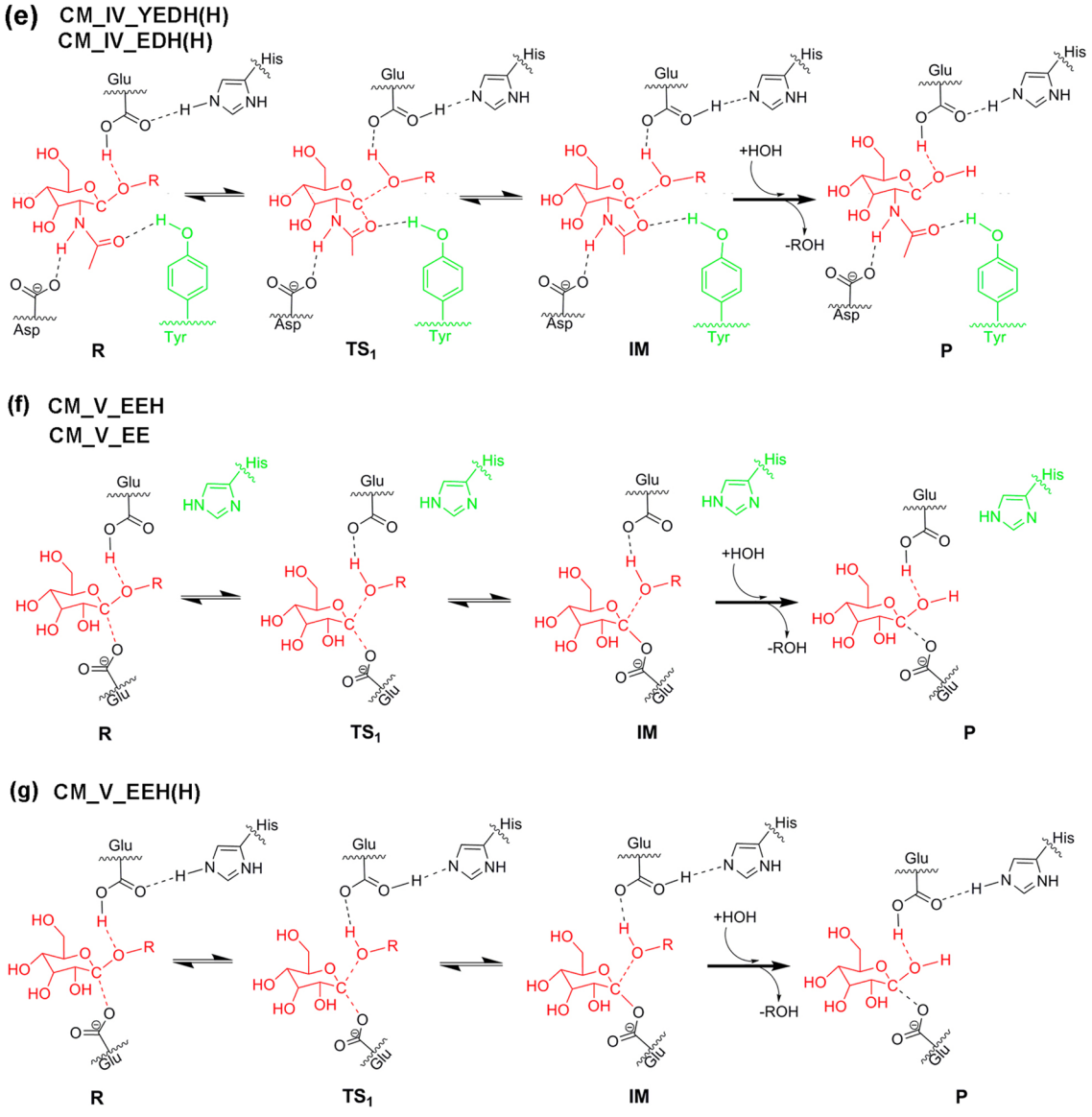


Figure 10. Proposed reaction schemes from reactant (R) to product (P) for the largest cluster models representing active sites of different categories. In the smaller models, only some of the side chains are included (which residues are included are indicated by complete model labels as described in the main text). The transition state 1 (TS1) occurs between the reactant and the intermediate (IM). (a)−(g) show different models with varied protonation schemes. The residues that may be absent in smaller models are colored in green. (a) CM_III_YED(H) or CM_III_ED(H); (b) CM_III_YED; (c) CM_IV_ED; (d) CM_IV_YEDH or CM_IV_YED; (e) CM_IV_YEDH(H) or CM_IV_EDH(H); (f) CM_V_EEH or CM_V_EE; (g) CM_V_EEH(H).

from the expected charge redistribution during this step: negative charge density around the nucleophilic carbonyl is expected to decrease, thus the hydrogen bond with Y217 would be weakened at the transition state relative to the reactant.

Energetics of the CM_IV Models. The various distance changes along the optimized reaction pathways (Supporting Information Figure S3) suggested a similar order of chemical events as in the CM_III models. Again, from model CM_IV_ED to CM_IV_EDH(H), and then to CM_IV_YEDH(H), the presence of a hydrogen bond with the general acid/general base and the increase of its strength have very large effects on the energetic barrier associated with TS1 (Figure 11b). Also in agreement with the results on the CM_III models, Y395 has small unfavorable energetic effects on the nucleophilic attack (Figure 11b). The hydrolysis step is also associated with lower barriers than the nucleophilic substitution step in CM_IV_YEDH(H).

Energetics of the CM_V Models. The order of chemical steps according to the distance changes along the reaction pathways (Supporting Information Figure S4) are the following: proton transfer to the leaving group from the general acid/base E182; then E182 is stabilized by the proton transferred from H264; nulceophilic attack that leads to the transition state TS1; and subsequent dissociation of the leaving group. As shown in Figure 11c, a protonated H264 on the reactant is important. The energy barrier of 19.64 kcal mol^{-1} of the CM_V_EEH(H) model with a protonated H264 is much lower than the energy barrier of 38.22 kcal mol^{-1} of the CM_V_EEH model with a neutral H264. Without H264 residue in the CM_V_EE model, the energy barrier is even higher (39.24 kcal mol^{-1}).

On the basis of the QM results (Figure 11a−c), we can conclude that the hydrogen bonding or salt bridge interactions with the general acid/general base are indeed essential for GH catalysis. Its energetic effects in gas phase can be as large as 17−20 kcal mol^{-1},

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

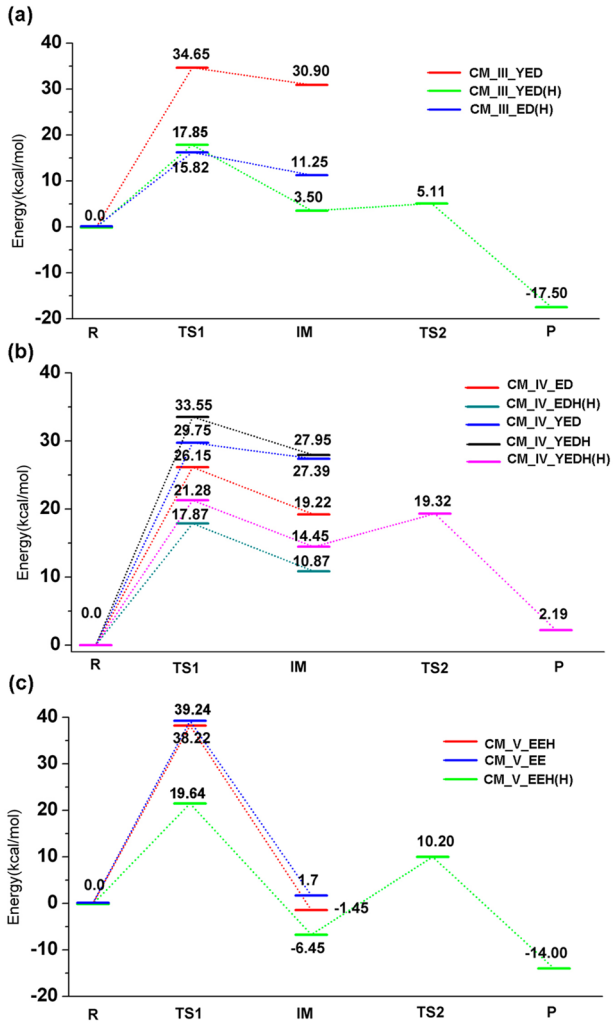


Figure 11. Calculated energy profiles for different cluster models. (a)−(c) correspond to results for the series of CM_III, CM_IV, and CM_V models, respectively. The TS energies have been taken as maxima on respective potential energy curves, not exactly the first order saddle points. The hydrolysis part of the calculated energy profile has been simply shifted to give the same energy at the IM state as the nucleophilic substitution part.

bringing the potential energy barriers from above 30 kcal mol^{-1} for the bare models without such interactions to around 17−20 kcal mol^{-1}. These may be considered as reasonably low gas phase barrier heights for viable enzyme mechanisms.

We think that such large effects can be largely understood based on the order of chemical events that lead to the transition state. From the distance change curves in Supporting Information Figures S2−4, we can see that the proton transfer from the general acid is always predicted to complete before the nucleophilic attack transition state. These curves have been obtained by reaction coordinate driving calculations in which the order of events have not been assumed; that is., the nucleophilic attack coordinate (d_2−d_1, see Figure S2−S4) and the proton transfer coordinates (d_4−d_3 and d_6−d_5, see also Figure S2−S4) have been treated in the same way in the overall driving coordinate. This conclusion should be robust because calculations on the three QM models extracted from different enzymes gave the same order of events. In addition, reaction coordinate driving calculations using only the nucleophilic attack coordinate d_2−d_1 as the driving coordinate have been attempted for several models (results not shown). Then the proton transfers take place automatically and abruptly (shown as sudden jumps of the respective interatomic distances) after a certain point along the calculated paths. The associated energy barriers are always much higher. Proton transfer from the general acid can significantly lower the energy of the transition state for nucleophilic attack by stabilizing the leaving group. Then the strong hydrogen bond/charged interaction with the general acid can significantly stabilize the transition state by stabilizing the deprotonated general acid. From the calculations, we also see that the bare models without including this interaction are all associated with so high transition barriers that they should not be considered as feasible at room temperature by themselves. Thus, we think this interaction should qualify as an essential component of the chemical mechanisms.

In fact, such a strong hydrogen bond interaction involving the general acid has been noted in GHs that are not of the TIM-fold. An example is the family 7 GHs, which are of the β-jelly roll fold. An example structure of this GH family is the Cellobiohydrolase from a Fungus *Heterobasidion irregulare* (PDB entry 2XSP).[71] In this enzyme, the suggested general acid E219 and a conserved D216 showed a relative arrangement for hydrogen bonding similar to that between E148 and D146 in 3WLl of Category III discussed here. The interaction there has been suggested to be important for catalysis.[72]

The above suggested mechanism requires the residues interacting with the general acid to be protonated in the reactant states. We used ProPKA3.1[73−77] (https://github.com/jensengroup/propka-3.1) to estimate the pK_a of these residues. Ligand has been included as part of the system for these estimations. The results are 12.8 for D146 in 3WL1, 15.31 for H258 in 3SUV, and 7.12 for H264 in 2JEQ. Such pK_a values are consistent with the assumption that these residues are protonated in the reactant state. On the other hand, the predicted pK_a of the proposed general acid residues are acidic (6.50 for E148 in 3WL1,

Table 2. Energy Barriers (ΔE) Calculated by Continuum Solvation with Different Dielectric Constants[a]

	models	ΔE (ε = 0.0)	ΔE (ε = 4.9)	ΔE (ε = 20.70)	ΔE (ε = 78.39)
CM_III	CM_III_YED(H)	17.85	28.46	29.24	30.18
	CM_III_YED	34.65	33.02	32.45	30.6
	$\Delta\Delta E$	−16.8	−14.56	−3.21	−0.42
CM_IV	CM_IV_YEDH(H)	21.28	22.03	22.97	23.18
	CM_IV_YEDH	33.55	31.02	29.1	27.28
	$\Delta\Delta E$	−12.27	−8.99	−6.13	−4.1
CM_V	CM_V_EEH(H)	19.64	25.29	26.42	27.36
	CM_V_EEH	38.22	34.99	33.92	31.32
	$\Delta\Delta E$	−18.58	−9.7	−7.5	−3.94

[a] $\Delta\Delta E$s are the differences between the energy barriers associated with models in different protonation states. The values are in kcal mol^{-1}.

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

3.69 for E322 in 3SUV, and 3.02 for E182 in 2JEQ). It is possible that the substrate imposes a pK_a shift on the catalytic residues: as these latter group of residues should be protonated in the reactant, otherwise even the general chemical mechanism would not be possible.

The above results have been obtained without considering possible dielectric screening effects. To look at if dielectric screening may substantially affect the results, we recalculated the energetic barriers for clusters CM_III_YED and CM_III_YED-(H), CM_IV_YEDH, CM_IV_YEDH(H), CM_V_EEH and CM_V_EEH(H) in different dielectric continuums by using the polarizable continuum model.[78] Single-point calculations on the gas phase-optimized geometries have been used. The results are summarized in Table 2. In general, increasing dielectric constant lead to increased TS1 barriers. This can be understood on the basis of the overall increased charge delocalization upon the nucleophilic attack. The effects of the hydrogen bond/salt bridge with the general acid−general base are reduced in high dielectric mediums (see the large $\Delta\Delta E$ values in Table 2) because of dielectric screening. However, inside the enzyme active sites, the effective dielectric constants will be small, so the effects would most probably remain quite large.

The calculated energetic effects of Y217 and Y395 in respective CM models may seem to conflict with their independent appearance in evolutionarily optimized substrate-assisted GH enzymes. Although a role of such a Tyr to lower the barrier of the hydrolysis step instead of the nucleophilic substitution step has been suggested,[29] this explanation is somewhat unsatisfactory as the hydrolysis step is associated with lower barriers than the nucleophilic substitution step. Here we speculate that the role of this Tyr may be related to activation entropy, which has not been included in the energetic barriers. Compared with the linearly branched flexible reactant, the transition state of the intramolecular nucleophilic attack step leads to a rigid five-membered ring (Figure 1B). This change would be expected to be associated with highly unfavorable activation entropies. We have seen that in the aligned active sites, the substrate *N*-acetyl group is sandwiched by conserved residues with bulky side chains. This may serve to reduce the flexibility of the substrate and thus reduce the activation entropy. But such sandwich interactions might still be not enough. The presence of the conserved Tyr might further reduce the entropy of the reactant state (at the transition state, the substrate structure is by itself very rigid and Tyr should not change the entropy much). To just qualitatively look at the possibility of this effect, we used vibrational analysis to estimate the activation entropies in the CM_III_ED(H) and CM_III_YED(H) models, the respective results of around −3 cal mol^{-1} K^{-1} for the former and ca. 6 cal mol^{-1} K^{-1} for the latter indicate that Tyr might indeed increase the activation entropy, even though the vibrational analysis should have significantly underestimated the flexibility of the substrate N-acetyl in reactant state in the CM_III_ED(H) cluster. Activation entropy from vibrational analysis for CM_IV_EDH(H) (ca. 7 cal mol^{-1} K^{-1}) and CM_IV_YEDH(H) (ca. 10 cal mol^{-1} K^{-1}) show the same trend. We must, however, emphasize that the discussions regarding the possible role of this Tyr to reduce activation entropy, although reasonable, are highly speculative. The entropy estimation from vibrational analysis is presented just to show that the sign of entropy change estimated in this way does not contradict this speculation. To address this issue more reliably may require free energy profile calculations inside an enzyme environment using QM/MM models and may be subject of future investigations.

QM modeling in the current work has focused on verifying and rationalizing chemically essential interactions deduced from common active site features of different enzymes. This focus has led to insights different from a number of previous QM/MM studies focusing on individual GH enzymes. Jitonnom et al. have reported QM/MM studies on a member of the GH family 18 which belongs to category III in our active site-based clustering.[28,29] They investigated the effects of different protonation states of an Asp142 (corresponding to residue 4-D in Figure 5) in the reactant state. However, the focus had been on the role of this Asp in stabilizing the oxazolinium ion intermediate, for which a protonated Asp142 would actually be less effective. The possibility of a protonated Asp142 to assist the catalytic general acid/general base group had not been fully explored or discussed in their studies. Another QM/MM study by Greig et al. investigated an enzyme that belongs to Category IV.[30] Again, the focus there had been on the pK_a of the oxazolinium ion intermediate inside an enzymatic environment relative to solution. The interaction between the conserved 3-H and the general acid 4-E (see Figure 6), the main point of interest in our study, was not their focus. A QM/MM study on a member of the Category V enzymes has also been reported by Badieyan et al.[31] That study has focused on active site residues interacting with the glucose moiety. The conserved interaction between 6-H and the general acid 4-E (see Figure 7) was also not their focus. Actually, the barrier for the glycosylation step calculated in that work was relatively high, being above 29 kcal mol^{-1}. The barriers for the gas phase cluster models of Category V enzymes without a charged 6-H to assist the general acid were also high (Figure 11c).

We would like to point out that the hydrogen bond or salt bridge interaction involving the general acid have been widely noticed in experimental structural analyses of different GHs (for examples, see references[63,79]), and its potential roles have been discussed along with other specific interactions observed for respective enzyme. For some enzymes, site-directed mutagenesis have been used to probe the effects of this interaction.[63,79] Relative to these works, the current work highlighted the general existence of this interaction in TIM-fold GHs across wide families, suggested it to be essential for the chemical process, and provided insights to rationalize this essentiality.

4. CONCLUSIONS

We have presented an approach of combining active site alignment with quantum mechanical calculations to reveal chemically essential interactions in enzymes that use the same or similar general mechanisms to catalyze alike reactions. Applying this approach to TIM-fold GHs suggested the chemical essentialness of a hydrogen bond/salt bridge interaction with the general acid/general base, a feature examined in a number of previous studies of specific enzymes but not emphasized in the general mechanisms. It is suggested that the essentialness of this interaction is due to the early proton transfer from the general acid to the leaving group before the transition state. In those TIM-fold GHs adopting the substrate-assisted mechanism, a tyrosine conserved in three-dimensional position but not in sequence position is hydrogen bonded to the nucleophilic carbonyl. This Tyr was not found to contribute favorably to lower the energetic barrier from the QM calculations. We speculated that the role of this interaction could be to lower the activation barrier through entropic effects.

The active site alignment approach allows for common active site structures and interactions in evolutionarily distant enzymes to be extracted in a systematic manner. In addition, protein structure models predicted through reliable comparative modeling (e.g.,

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

model structures built based on templates of >40% sequence identity) may also be considered as valid inputs for active site alignment, thus extending applicability of the approach. On the other hand, we note that for active site comparisons, other methods besides SMAP have been proposed, such as ProBiS,[33,34] CMASA[80] or the method of Steinkellner et al.[81] For our purpose, these methods could have been used as well to achieve the same goal. Active site comparisons based on overall structure alignment methods such as MUSTANG have also been reported.[82−84] Alignment of overall structures usually, though not always, leads to the most sensible alignment of active sites. In addition, residues of similar spatial but different sequence locations (for example, the Tyr interacting with the substrate carbonyl in the Category III and Category IV enzymes) may only be aligned by the active site alignment approach. In principle, approaches like SMAP or ProBiS may also be applied to find similarities across different folds. In practice, however, the similarity between the active sites of GHs of different folds seems to be below the threshold of being detectable by the SMAP or ProBiS approach in their general form. Earlier we have mentioned the similarity in the arrangement of two chemically important residues of the β-jelly roll fold family 7 GH enzymes to the TIM fold GHs considered here. Automatic alignments through SMAP or ProBiS are not yet able to align the respective active sites with significant *P*-values or *Z*-scores. That is probably because all residues surrounding the binding site have been treated as of equal importance for significant alignments, while the conservation across fold families can be limited to only a few key residues. The main reason that the SMAP alignment approach was not able to detect the active sites of non TIM-barrel GHs is that all residues surrounding the binding site have been considered to be of equal importance, while the conservation across different folds seems to be limited to only the few key catalytic residues. It may be possible to circumvent this limitation by restricting the alignment to include only these key residues. However, for this approach to be practical, one also needs to find an accompanying strategy to avoid yielding a large number of false positive alignments.

Compared with models constructed by inspecting unaligned individual enzymes, quantum mechanical models based on examining the aligned active sites may better capture features that are truly essential for catalysis. Our results on the TIM-fold GHs indicate that given a well-known general enzymatic mechanism, the basic chemical steps may still be unfeasible in lack of some chemically essential interactions. The proposed approach of combining active-site-oriented bioinformatics analysis with QM modeling provides a useful way to gain insights into such aspects of enzyme catalysis.

■ ASSOCIATED CONTENT

Supporting Information

The following files are available free of charge on the ACS Publications website at DOI: 10.1021/cs501709d.

Validation of binding site alignments; subcategory of Catergory II; energy profiles and distance profiles for the CM_III, CM_IV and CM_V series of cluster models (PDF)

Multiple active site alignment (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: hyliu@ustc.edu.cn.

*E-mail: zhengzh@mail.ustc.edu.cn.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Bourne for providing the SMAP program. This work has been supported by Chinese Natural Science Foundation (Grant Nos. 31370755 and 21173203) and the Ministry of Science and Technology (Grant No. 2012AA02A704).

■ REFERENCES

(1) Fersht, A. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*; W.H. Freeman: New York, 1999.
(2) Peracchi, A. *Trends Biochem. Sci.* **2001**, *26*, 497−503.
(3) Zvelebil, M. J.; Sternberg, M. J. *Protein Eng., Des. Sel.* **1988**, *2*, 127−138.
(4) Bartlett, G. J.; Porter, C. T.; Borkakoti, N.; Thornton, J. M. *J. Mol. Biol.* **2002**, *324*, 105−121.
(5) Tantillo, D. J.; Chen, J.; Houk, K. N. *Curr. Opin. Chem. Biol.* **1998**, *2*, 743−750.
(6) Siegbahn, P. E.; Himo, F. *JBIC, J. Biol. Inorg. Chem.* **2009**, *14*, 643−651.
(7) Silverman, R. B. *Methods Enzymol.* **1995**, *249*, 240−283.
(8) Kiss, G.; Celebi-Olcum, N.; Moretti, R.; Baker, D.; Houk, K. N. *Angew. Chem., Int. Ed.* **2013**, *52*, 5700−5725.
(9) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. *Protein Sci.* **1996**, *5*, 1001−1013.
(10) Xie, L.; Bourne, P. E. *BMC Bioinf.* **2007**, *8* (Suppl4), S9.
(11) Xie, L.; Bourne, P. E. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 5441−5446.
(12) Xie, L.; Xie, L.; Bourne, P. E. *Bioinformatics* **2009**, *25*, i305−312.
(13) Davies, G.; Henrissat, B. *Structure* **1995**, *3*, 853−859.
(14) Henrissat, B.; Callebaut, I.; Fabrega, S.; Lehn, P.; Mornon, J. P.; Davies, G. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 7090−7094.
(15) Vocadlo, D. J.; Davies, G. J. *Curr. Opin. Chem. Biol.* **2008**, *12*, 539−555.
(16) Vuong, T. V.; Wilson, D. B. *Biotechnol. Bioeng.* **2010**, *107*, 195−205.
(17) Nagano, N.; Orengo, C. A.; Thornton, J. M. *J. Mol. Biol.* **2002**, *321*, 741−765.
(18) Rigden, D. J.; Jedrzejas, M. J.; de Mello, L. V. *FEBS Lett.* **2003**, *544*, 103−111.
(19) Ogawa, S.; Yuasa, H. *Curr. Top. Med. Chem.* **2009**, *9*, 1.
(20) Song, L.; Laguerre, S.; Dumon, C.; Bozonnet, S.; O'Donohue, M. J. *Bioresour. Technol.* **2010**, *101*, 8237−8243.
(21) Koshland, D. E. *Biol. Rev. Camb. Philos. Soc.* **1953**, *28*, 416−436.
(22) Lombard, V.; Golaconda Ramulu, H.; Drula, E.; Coutinho, P. M.; Henrissat, B. *Nucleic Acids Res.* **2014**, *42*, D490−495.
(23) Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. *PLoS Comput. Biol.* **2009**, *5*, e1000423.
(24) Xie, L.; Li, J.; Xie, L.; Bourne, P. E. *PLoS Comput. Biol.* **2009**, *5*, e1000387.
(25) Warshel, A. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425−443.
(26) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198−1229.
(27) Zhang, Y. K.; Liu, H. Y.; Yang, W. T. *J. Chem. Phys.* **2000**, *112*, 3483−3492.
(28) Jitonnom, J.; Lee, V. S.; Nimmanpipug, P.; Rowlands, H. A.; Mulholland, A. J. *Biochemistry* **2011**, *50*, 4697−4711.
(29) Jitonnom, J.; Limb, M. A.; Mulholland, A. J. *J. Phys. Chem. B* **2014**, *118*, 4771−4783.
(30) Greig, I. R.; Zahariev, F.; Withers, S. G. *J. Am. Chem. Soc.* **2008**, *130*, 17620−17628.
(31) Badieyan, S.; Bevan, D. R.; Zhang, C. *Biochemistry* **2012**, *51*, 8907−8918.
(32) Rose, P. W.; Bi, C.; Bluhm, W. F.; Christie, C. H.; Dimitropoulos, D.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Prlic, A.; Quesada, M.; Quinn, G. B.; Ramos, A. G.; Westbrook, J. D.; Young, J.; Zardecki, C.; Berman, H. M.; Bourne, P. E. *Nucleic Acids Res.* **2013**, *41*, D475−482.
(33) Konc, J.; Cesnik, T.; Konc, J. T.; Penca, M.; Janezic, D. *J. Chem. Inf. Model.* **2012**, *52*, 604−612.

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572

(34) Konc, J.; Janezic, D. *Bioinformatics* **2010**, *26*, 1160−1168.
(35) de Hoon, M. J.; Imoto, S.; Nolan, J.; Miyano, S. *Bioinformatics* **2004**, *20*, 1453−1454.
(36) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200−1211.
(37) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(38) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623−11627.
(39) 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.; Bakken, V.; 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, 2003.
(40) Przylas, I.; Terada, Y.; Fujii, K.; Takaha, T.; Saenger, W.; Strater, N. *Eur. J. Biochem.* **2000**, *267*, 6903−6913.
(41) Roujeinikova, A.; Raasch, C.; Sedelnikova, S.; Liebl, W.; Rice, D. W. *J. Mol. Biol.* **2002**, *321*, 149−162.
(42) Linden, A.; Mayans, O.; Meyer-Klaucke, W.; Antranikian, G.; Wilmanns, M. *J. Biol. Chem.* **2003**, *278*, 9875−9884.
(43) Kramhoft, B.; Bak-Jensen, K. S.; Mori, H.; Juge, N.; Nohr, J.; Svensson, B. *Biochemistry* **2005**, *44*, 1824−1832.
(44) Hsieh, Y. C.; Wu, Y. J.; Chiang, T. Y.; Kuo, C. Y.; Shrestha, K. L.; Chao, C. F.; Huang, Y. C.; Chuankhayan, P.; Wu, W. G.; Li, Y. K.; Chen, C. J. *J. Biol. Chem.* **2010**, *285*, 31603−31615.
(45) Ito, K.; Ito, S.; Shimamura, T.; Weyand, S.; Kawarasaki, Y.; Misaka, T.; Abe, K.; Kobayashi, T.; Cameron, A. D.; Iwata, S. *J. Mol. Biol.* **2011**, *408*, 177−186.
(46) Matsui, I.; Yoneda, S.; Ishikawa, K.; Miyairi, S.; Fukui, S.; Umeyama, H.; Honda, K. *Biochemistry* **1994**, *33*, 451−458.
(47) MacGregor, E. A.; Janecek, S.; Svensson, B. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.* **2001**, *1546*, 1−20.
(48) Rydberg, E. H.; Li, C.; Maurus, R.; Overall, C. M.; Brayer, G. D.; Withers, S. G. *Biochemistry* **2002**, *41*, 4492−4502.
(49) Kaper, T.; Leemhuis, H.; Uitdehaag, J. C.; van der Veen, B. A.; Dijkstra, B. W.; van der Maarel, M. J.; Dijkhuizen, L. *Biochemistry* **2007**, *46*, 5261−5269.
(50) Machovic, M.; Janecek, S. *Biologia* **2003**, *58*, 1127−1132.
(51) Godany, A.; Vidova, B.; Janecek, S. *FEMS Microbiol. Lett.* **2008**, *284*, 84−91.
(52) Janecek, S.; Svensson, B.; MacGregor, E. A. *Cell. Mol. Life Sci.* **2014**, *71*, 1149−1170.
(53) Numao, S.; Maurus, R.; Sidhu, G.; Wang, Y.; Overall, C. M.; Brayer, G. D.; Withers, S. G. *Biochemistry* **2002**, *41*, 215−225.
(54) Vaaje-Kolstad, G.; Vasella, A.; Peter, M. G.; Netter, C.; Houston, D. R.; Westereng, B.; Synstad, B.; Eijsink, V. G.; van Aalten, D. M. *J. Biol. Chem.* **2004**, *279*, 3612−3619.
(55) Tsuji, H.; Nishimura, S.; Inui, T.; Kado, Y.; Ishikawa, K.; Nakamura, T.; Uegaki, K. *FEBS J.* **2010**, *277*, 2683−2695.
(56) Ohnuma, T.; Numata, T.; Osawa, T.; Mizuhara, M.; Varum, K. M.; Fukamizo, T. *Plant Mol. Biol.* **2011**, *75*, 291−304.
(57) Malecki, P. H.; Vorgias, C. E.; Petoukhov, M. V.; Svergun, D. I.; Rypniewski, W. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2014**, *70*, 676−684.
(58) Synstad, B.; Gaseidnes, S.; Van Aalten, D. M.; Vriend, G.; Nielsen, J. E.; Eijsink, V. G. *Eur. J. Biochem.* **2004**, *271*, 253−262.
(59) van Aalten, D. M.; Komander, D.; Synstad, B.; Gaseidnes, S.; Peter, M. G.; Eijsink, V. G. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 8979−8984.
(60) Ito, T.; Katayama, T.; Hattie, M.; Sakurama, H.; Wada, J.; Suzuki, R.; Ashida, H.; Wakagi, T.; Yamamoto, K.; Stubbs, K. A.; Fushinobu, S. *J. Biol. Chem.* **2013**, *288*, 11795−11806.
(61) Kaper, T.; van Heusden, H. H.; van Loo, B.; Vasella, A.; van der Oost, J.; de Vos, W. M. *Biochemistry* **2002**, *41*, 4147−4155.
(62) Kim, H. W.; Ishikawa, K. *Biochem. J.* **2011**, *437*, 223−230.
(63) Zheng, B.; Yang, W.; Zhao, X.; Wang, Y.; Lou, Z.; Rao, Z.; Feng, Y. *J. Biol. Chem.* **2012**, *287*, 8336−8346.
(64) Belaich, A.; Fierobe, H. P.; Baty, D.; Busetta, B.; Bagnara-Tardif, C.; Gaudin, C.; Belaich, J. P. *J. Bacteriol.* **1992**, *174*, 4677−4682.
(65) Durand, P.; Lehn, P.; Callebaut, I.; Fabrega, S.; Henrissat, B.; Mornon, J. P. *Glycobiology* **1997**, *7*, 277−284.
(66) Tomatsu, S.; Fukuda, S.; Sukegawa, K.; Ikedo, Y.; Yamada, S.; Yamada, Y.; Sasaki, T.; Okamoto, H.; Kuwahara, T.; Yamaguchi, S.; et al. *Am. J. Hum. Genet.* **1991**, *48*, 89−96.
(67) Gloster, T. M.; Ibatullin, F. M.; Macauley, K.; Eklof, J. M.; Roberts, S.; Turkenburg, J. P.; Bjornvad, M. E.; Jorgensen, P. L.; Danielsen, S.; Johansen, K. S.; Borchert, T. V.; Wilson, K. S.; Brumer, H.; Davies, G. J. *J. Biol. Chem.* **2007**, *282*, 19177−19189.
(68) Suzuki, R.; Fujimoto, Z.; Ito, S.; Kawahara, S.; Kaneko, S.; Taira, K.; Hasegawa, T.; Kuno, A. *J. Biochem.* **2009**, *146*, 61−70.
(69) Notenboom, V.; Birsan, C.; Nitz, M.; Rose, D. R.; Warren, R. A.; Withers, S. G. *Nat. Struct. Biol.* **1998**, *5*, 812−818.
(70) Liu, J.; Zhang, C.; Xu, D. *J. Mol. Graphics Modell.* **2012**, *37*, 67−76.
(71) Momeni, M. H.; Payne, C. M.; Hansson, H.; Mikkelsen, N. E.; Svedberg, J.; Engstrom, A.; Sandgren, M.; Beckham, G. T.; Stahlberg, J. *J. Biol. Chem.* **2013**, *288*, 5861−5872.
(72) Stahlberg, J.; Divne, C.; Koivula, A.; Piens, K.; Claeyssens, M.; Teeri, T. T.; Jones, T. A. *J. Mol. Biol.* **1996**, *264*, 337−349.
(73) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins: Struct., Funct., Genet.* **2005**, *61*, 704−721.
(74) Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins: Struct., Funct., Genet.* **2008**, *73*, 765−783.
(75) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.
(76) Rostkowski, M.; Olsson, M. H.; Sondergaard, C. R.; Jensen, J. H. *BMC Struct. Biol.* **2011**, *11*, 6.
(77) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011**, *7*, 2284−2295.
(78) Scalmani, G.; Frisch, M. J. *J. Chem. Phys.* **2010**, *132*, 114110.
(79) Oyama, T.; Schmitz, G. E.; Dodd, D.; Han, Y.; Burnett, A.; Nagasawa, N.; Mackie, R. I.; Nakamura, H.; Morikawa, K.; Cann, I. *PLoS One* **2013**, *8*, e80448.
(80) Li, G. H.; Huang, J. F. *BMC Bioinf.* **2010**, *11*, 439.
(81) Steinkellner, G.; Gruber, C. C.; Pavkov-Keller, T.; Binter, A.; Steiner, K.; Winkler, C.; Lyskowski, A.; Schwamberger, O.; Oberer, M.; Schwab, H.; Faber, K.; Macheroux, P.; Gruber, K. *Nat. Commun.* **2014**, *5*, Articleno. 4150.
(82) Wei, Y.; Ringe, D.; Wilson, M. A.; Ondrechen, M. J. *PLoS Comput. Biol.* **2007**, *3*, e10.
(83) Suplatov, D. A.; Arzhanik, V. K.; Svedas, V. K. *ActaNaturae* **2011**, *3*, 93−98.
(84) Konagurthu, A. S.; Whisstock, J. C.; Stuckey, P. J.; Lesk, A. M. *Proteins: Struct., Funct., Genet.* **2006**, *64*, 559−574.
(85) Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. *Genome Res.* **2004**, *14*, 1188−1190.

NOTE ADDED AFTER ASAP PUBLICATION

After this paper was published ASAP March 23, 2015, corrections were made to Figure 10. The corrected version was reposted March 24, 2015.

DOI: 10.1021/cs501709d
ACS Catal. 2015, 5, 2559−2572