ACS SYMPOSIUM SERIES 1076

Library Design, Search Methods, and Applications of Fragment-Based Drug Design

EDITED BY

Rachelle J. Bienstock

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.fw001

Library Design, Search Methods, and Applications of Fragment-Based Drug Design

In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.fw001

ACS SYMPOSIUM SERIES **1076**

Library Design, Search Methods, and Applications of Fragment-Based Drug Design

Rachelle J. Bienstock, Editor
National Institute of Environmental Health Sciences
National Institutes of Health
Research Triangle Park, North Carolina

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.fw001

Sponsored by the
ACS Division of Chemical Information
ACS Division of Computers in Chemistry

American Chemical Society, Washington, DC

Distributed in print by Oxford University Press, Inc.

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.fw001

Library of Congress Cataloging-in-Publication Data

Library of Congress Cataloging-in-Publication Data

Library design, search methods, and applications of fragment-based drug design / Rachelle J. Bienstock, editor ; sponsored by the ACS Division of Chemical Information and ACS Division of Computers in Chemistry.
p. ; cm. -- (ACS symposium series ; 1076)
Includes bibliographical references and index.
ISBN 978-0-8412-2492-6 (alk. paper)
I. Bienstock, Rachelle J. II. American Chemical Society. Division of Chemical Information. III. American Chemical Society. Division of Computers in Chemistry. IV. Series: ACS symposium series ; 1076. 0097-6156
[DNLM: 1. Combinatorial Chemistry Techniques. 2. Drug Design. 3. Drug Discovery--methods. 4. Small Molecule Libraries. QV 744]

615.1′9--dc23
2011035971

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.

Copyright © 2011 American Chemical Society

Distributed in print by Oxford University Press, Inc.

All Rights Reserved. Reprographic copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Act is allowed for internal use only, provided that a per-chapter fee of $40.25 plus $0.75 per page is paid to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. Republication or reproduction for sale of pages in this book is permitted only under license from ACS. Direct these and other permission requests to ACS Copyright Office, Publications Division, 1155 16th Street, N.W., Washington, DC 20036.

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

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.fw001

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

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.pr001

Preface

This volume is the result of Division of Chemical Information (CINF) symposia organized around the popular topic of fragment-based ligand design conducted during two Spring American Chemical Society National Meetings. The first symposium, *Library Design, Search Methods and Applications of Fragment-Based Drug Design,* was held during the 2009 Spring National ACS meeting in Salt Lake City, Utah. The second *Fragment-Based Drug Design: Success Stories Due to Novel Computational Methods Applications,* followed one year later at the 2010 Spring National meeting in San Francisco. All presenters at both meetings were invited to submit a written chapter summary of their oral presentations for this volume. The sessions at both of these meetings were extremely popular and the talks were presented to standing room only audiences. This reflects the strong interest in the technique and application of fragment-based drug discovery computational associated methods.

Computational, ‘rational’, or structure-based drug design methods have been promulgated since the early 1980s evolving over time. From statistical methods like QSAR, to shape based methods such as pharmacophore analysis, combinatorial chemistry and high throughput virtual screening, different computational methods have become part of the pharmaceutical company drug discovery arsenal. As pharmaceutical research and development time and costs have increased, there is constant commercial pressure to develop new and better methods for the identification of novel chemical entities. The success of the human genome project followed by protein structure initiatives, and molecular biology pathway analysis, have brought forth a plethora of new ideas for understanding disease pathways and new targets ripe for drug discovery. Into this arena has stepped the methodology of fragment-based drug or ligand design. This volume covers computational methods in fragment-based ligand design and their application to fragment library design, library screening and fragment docking methods and computational methodologies for fragment linking, merging and growing. It touches on some success stories in the development of potential leads using these fragment-based computational methods as well.

Readers who are new to the field of fragment-based ligand design as well as veterans of computational drug discovery methods and pharmaceutical and medical chemistry will be interested in the contents of this volume. In addition to chemists, mathematicians and statisticians, may be interested in this volume as well, in that they may see where novel algorithms, mathematical and statistical techniques, can be applied to the difficulties of library searching, and for the development of better fragment docking and scoring methods. Fragment-based drug discovery has rapidly risen as evidenced by both the increasing number of

In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

industrial and academic groups with interest and publications in this area as well as documentation of at least 13 different companies (Abbott, Astex Therapeutics, Aventis, Burnham Institue, Novartis, Plexxikon , Roche, SGX Pharmaceuticals, Schering-Plough, Sunesis, Triad, Vernalis, Vertex) with successful leads as a result of fragment-based screening (Hajduk PJ and Greer, J Nature Reviews Drug Discovery 6, 2007, 211-219.) The development of twenty two clinical candidates currently in Phase I, or II clinical trials are attributed to the application of fragment-based drug discovery methods (Law, R J Comput Aided Mol Design 2009, 23:459-473).

I would like to thank all those who agreed to write chapters for this volume as well as all those who participated in the two ACS symposia on this topic. Both symposia were filled with interesting presentations and discussions. I would also like to thank, Mr. Tim Marney, my editor at ACS publications for his assistance with all the intricate technical details involved in producing this volume, as well as Mr. Bob Hauserman, Senior Acquisitions Editor at ACS for noting the topical significance of the symposium and its merit for book publication. Of course, I always owe a debt of gratitude to my family, my husband Scott Snyder and daughters, Julia and Shira, for the support which they provide throughout all life's endeavors.

Dr. Rachelle J. Bienstock

Contract Senior Research Scientist
National Institute of Environmental Health Sciences
National Institutes of Health
P.O. Box 12233, MD F0-011
Research Triangle Park, North Carolina 27709
(919)541-3397 (telephone)
biensto1@niehs.nih.gov (e-mail)

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.pr001

Chapter 1

Overview: Fragment-Based Drug Design

Rachelle J. Bienstock

National Institute of Environmental Health Sciences, P.O. Box 12233,
MD F0-011, Research Triangle Park, North Carolina 27709

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

Fragment-based drug design has recently risen to great prominence as a new methodology for novel lead identification. This chapter is a general overview of computational methods for all three phases of fragment-based ligand design: (1) Designing and Searching Fragment Libraries, (2) Computational Screening: Docking, and (3) Leads from Fragments: Fragment Growing and Linking. Appendix 1 at the end of this chapter summarizes a large number of computational methods and software programs available with associated web sites and references.

Introduction

Fragment-based drug discovery has emerged as a new and promising computational methodology for efficiently increasing diversity space leading to novel leads and therefore NCEs (new chemical entities). For many researchers in the drug discovery area, high throughput screening (HTS) has not lived up to its original “great expectations”. Many HTS screens failed to deliver good starting molecules for drug discovery, as often the databases searched were composed of nondrug like molecules or compounds which, as products of combinatorial approaches, were difficult to transform into lead compounds. The preference for fragment-based design over other methods, such as high throughput screening, rests largely with the enhanced screening of a more impressive conformational space with a smaller starting number of compounds. A large HTS screen library contains 10^5-10^6 compounds, which still samples only a fraction of the chemical space of small molecules (on the order of 10^{60}- 10^{100} compounds), but world require huge amounts of testing. However, combinatorial combination of 3 different fragments of a small 100 fragment database would yield 10^6 different compounds and require less experimental assaying or virtual screening (*1*).

© 2011 American Chemical Society
In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Significantly higher hit rates have been reported with fragments than with HTS. HTS has on the order of a .1% or less successful hit rate while fragment-based design is estimated to have typically a 3-5% hit rate (depending on the target, hit rates are typically 8-10% for kinase targets and 2-3% for protein-protein interactions), and because the fragments are small when docked into the binding site, the binding site can be explored in new ways not accessible to large molecules (*2*). With fragments there is higher probability that a designed compound can be synthesized because it will be based on a small simple scaffold. In short, fragment-based design marries elements of rational design with the diversity of random virtual screening which makes it particularly attractive.

What are fragments? Fragments are low molecular weight (MW< 250Da), small organic molecules, that actually have low affinity (100 μM−10 mM) for binding to the target. These fragments are then embellished, grown and linked to create high affinity lead compounds with high selectivity. However, because these smaller fragments are weaker binders than leads and hits identified through HTS, the experimental methods for confirming binding must be more sensitive. Hits from fragment screening methods usually would not be identified as potential leads in HTS screens, and therefore provide for novel templates.

How do you design a good, diverse fragment library or database? Usually 2D fingerprint methods (Tanimoto and/or Daylight are among some popular methods) or statistical models based on fingerprint connectivity or pharmacophore models are used to analyze the library database for diversity (*3*). A popular set of rules, referred to as the "rule of three" has been proposed for the fragments comprising a library : MW less than 300 Da; less than or equal to 3 H bond acceptors or donors; ClogP less than or equal to 3; 3 or less rotatable bonds and a polar surface area less than 60 Å^2. "Drug-likeness" is an important characteristic for a fragment library and frequently fragment databases are compared to a database of known drug compounds, or often known drugs are dissected into fragments to develop a fragment database. High ligand efficiency (LE> or = 3.0), defined as the binding energy per heavy atom of the structure, is required for good leads.

Once a fragment library has been developed, a screening method must be implemented where fragments are screened against the pharmaceutical target of interest. Screening can be performed virtually *in silico* using computational methods, or experimentally using x-ray, NMR(SAR by NMR) (*4*), surface plasmon resonance (SPR, Biacore) , mass spectrometry, isothermal titration calorimetry or protein thermal unfolding experimental methods. Soaking crystals with small fragments can be used as a method to identify fragments as well, the CrystaLEAD method of x-ray based fragment library screening (*5*).

Computational screening, using docking methodologies for fragments, is particularly challenging due to some of the same concerns as general docking. Problems with docking fragments include identification of the interaction or binding site for the fragment; binding cavities can be much larger than smaller fragments so there can be difficulties in predicting binding modes, and scoring functions are not optimized for fragments as they are designed for larger drug-like molecules. For fragments a metric other than RMSD must be used when docking, such as a structural fingerprint score, feature scoring, volume clustering, or fragment pharmacophore feature identification (*6*). In a review by Marcou and

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

Rognan of FlexX, Glide, Gold and Surflex on 42 protein –fragment complexes compared to x-ray data these docking algorithms could predict correct binding of fragments 40, 70, 70 and 60% of the time respectively (*7*).

Difficult targets for which fragment-based screening is particularly beneficial are those that are large and open to solvent and for targets which do not have well developed compound libraries. Structural information, such as from x-ray or NMR studies of the target must be available for both computational docking of fragments and successful linking and growing of fragments to fill the target binding pocket. As a result, many of the first successful applications of fragment-based ligand design have been exclusively in the oncology therapeutic area with many focused solely on kinase targets (*8*). Fragments are in general more rigid than molecules and easier to dock computationally due to fewer degrees of freedom. Fragment drug design is based on two fundamental assumptions- molecular recognition by receptor occurs due to the presence of a fundamental core structure or fragment, and that the properties of active fragments are additive and can be combined .

Fragment approaches can be complimentary to other lead generation methods and are often used in conjunction with other methods. Sometimes a set of fragments is assayed at high concentration (1mM) simultaneously with typical fragment screening concentrations (100 μM). Also sequential methodologies are employed following fragment-based design with a more sensitive screening technique. Often different hit finding methods work better with different types of targets. For example for the BACE-1 Alzheimer's target, a high throughput μM screening of 200,000 compounds at Evotec did not yield any leads, however a screen of a 20,000 compound fragment library led to two dozen confirmed good hits tested as 1mM binders using SPR (surface plasmon resonance) testing. In fragment-based drug design frequently the idea of a 'privileged structure" is employed where certain types of substructures are effective with particular types of targets, i.e. hydrosamates with matrix metalloproteases, benzamidines with serine proteases, aminopyrididines with kinases and ATP containing proteins (*9*).

Fragment-based discovery has claimed early successes with 50 perspective small molecule leads with good ligand efficiency (*10*). Several companies are developing drugs based on fragment screening which currently are in the clinic, including Abbott, Astex, SGX Pharm, and Plexxikon. Recently, on the practical fragments blog (http://practicalfragments.blogspot.com/) edited by Drs. Dan Erlandson and Teddy Zartler a list of 44 companies were compiled with research efforts in the area of fragment-based ligand design and drug discovery methods. These are examples of some targets where fragment-based ligand design has been successfully applied: BACE-1 (beta-secretase) for Alzheimers disease (*11–13*), Urokinase (*14*), Phosphodiesterase 4 (*15*), Bcl-XL for cancer,(a protein-protein interaction target) (*16*): Thrombin (*17*) ,Aurora kinase inhibitor (*18*) HSP90 inhibitor (*19*).

This volume covers the development of computational methods and their application in the area of fragment-based drug discovery. There are three basic steps involved in fragment-based computational drug design. The initial step is the design of a good fragment library, the second step is computational docking, ranking or screening of the fragments within the library and the third step is computational methods for growing, linking or combining of the fragments to

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

yield lead compounds. In the application of computational *in silico* screening methods for fragments, many of the same issues and concerns apply as do for computational docking of compounds in general. The targets used in computational screening, usually structures solved by NMR or x-ray methods, are rigid without flexibility in computational docking, and must be properly prepared adding hydrogens, and side chains assigned proper protonation states. There is a need for improved computational methodologies to score, rank and categorize fragments docked to targets as well as for methodologies for growing or linking the fragments to form complete molecules.

All three computational aspects were discussed within these ACS symposiums and I will briefly outline the discussions and presentations in these areas. The chapters within this volume discuss each of the methods presented, described by their developers. Successful applications to kinase, GPCR, CNS targets, HSP 90 and drugs targeting protein-protein interfaces were presented at the meeting. As we are focusing solely on computational methods in this volume, experimental methodologies will not be discussed.

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

I. Designing and Searching Fragment Libraries

What comprises a good fragment library? The properties of a good fragment library are : diversity of physicochemical properties, molecular diversity, aqueous solubility, drug- like molecules, MW < 300 Da, good ligand efficiency (LE) (free energy of binding a ligand averaged over each non hydrogen atom) and involve taking lipophilicity into consideration. Several novel computational approaches to the development of fragment libraries were reported at this ACS symposium. This is a rapidly developing area involving both chemoinformatics and modeling tools.

One approach to the design of fragment libraries is the use of methods which perform computational deconstruction of known drugs (*20*) for example the DAIM method developed by Peter Kolb and Amedeo Caflish (Decomposition and Identification of Molecules) (*21*) or retrosynthesis and combinatorial analysis, such as the RECAP method (*22*). CoLibri is a commercial program (BioSolveIT) that can be used for "shredding" compounds for the development of fragment libraries for virtual screening.

The FTrees-FS software (*23*) performs fragment space similarity searching and fragment assembly. (Searching web interface freely available: http://public.zbh.uni-hamburg.de/ftrees/query.py; software available commercially from BioSolveIT). Dr. Carsten Detering (BioSolveIT) reported on the development and application of the FTrees-FS methodology and its advantages in using synthetically accessible compounds from a giant virtual chemistry fragment space library. FTrees-FS represents molecules in reduced graph representation (feature trees) for easier and faster searching and extracts cores and identifies link atoms ,with a reagent list, to attach link atoms (*24*). Information concerning chirality and 3D properties of molecules is not included, only topology information (*25*)(*26*)(*27*)(*28*). Dr. Atipat Rojnuckarin, reported on work conducted at ArQule, implementing a novel targeting strategy to design type IV kinase inhibitors. This involved application of the FTrees-FS software to construct searchable fragment space based on the ArQule kinase inhibitor compound library to identify novel type IV kinase inhibitors with improved ease in synthesis.

Dr. J. Robert Fischer, Zentrum für Bioinformatik, Universität Hamburg, described LoFT, a library optimizer using Feature Trees, which is a tool for focused combinatorial library design (*29*). LoFT uses a reduced topological graph descriptor to match feature tree nodes to compare and search fragment space. Using LoFT, a fragment space can be searched and cores and reagents selected according to selected physicochemical properties. LoFT will then optimize and compile a complete fragment sub library with the described properties based on filters and descriptors in the scoring function. LoFT searches a fragment space consisting of combinatorial libraries with a unique scaffold and uses feature tree descriptors representing the molecules as unrooted nodes on a tree with topology, connectivity and physicochemical properties conserved. The difference between FTrees-FS and LoFT, is that LoFT results in sub libraries which focus on a single core placement for a more focused library design. For validation, LoFT was applied to several drug design scenarios. Starting with known drug molecules, focused libraries were generated with desired property ranges.

Dr. Christof Wegscheid-Gerlach, Bayer-Schering Pharma, reported on applications of BRICS (Breaking into Restrosynthetically interesting Chemical Substructures), a modification of RECAP to fragment molecules according to 11 default bond rules including comprehensive modeling of ring substitution and cleavage of sulfur groups. (BRICS fragment spaces are publicly available- http://ww.zbh.uni-hamburg.de/BRICS) (*30*) BRICS is a compilation of a new rule set for breaking up interesting chemical structures into fragments. Dr. Wegscheid-Gerlach discussed applications of BRICS to shred the WDI and Zinc databases and compared BRICS enhanced performance compared to RECAP. He then presented several successful applications demonstrating novel scaffold hops yielding Sorafenib, Fasudil and Erlotinib.

BROOD (Commerical software from Openeye http://www.eyesopen.com/brood) searches databases of chemical fragments to identify and select fragments with similarities to the query fragment and can perform bioiosteric replacements to develop new leads. BROOD also generates analogs to leads by assembling and replacing different fragments based on shape, electrostatics and molecular properties with graphical tools for fragment editing. BROOD is accompanied by CHOMP which serves as a molecular fragmentor and MERGE for fragment merging. Additionally, if a crystal structure for the target protein is known, BROOD can use information from the protein structure to eliminate fragments which will not fit in the binding pocket and will verify protein-ligand close contacts (*31*).

Dr. Ijen Chen, Vernalis, reported on the use of SeeDs (Structural Exploitation of Experimental Drug Startpoints) a method which uses pharmacophore fingerprints to facilitate fragment library design. The SeeDs library enumeration creates diversity through the use of pharmacophore triangles to create a fragment library. Fingerprints of 3 point pharmacophore triangles (acceptor, donor and hydrophobid) are used and these pharmacophore fingerprints can screen drug like chemical space and pick out the known preferred substructures (*32*, *33*). The first SeeDs library was based on MW and desired chemical features judged

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

by medicinal chemistry, solutibility and tractability. This was used for in house docking followed by a merging of fragments to create novel leads for PDK1 and Hsp90, ATPases and kinases. Average screen hit rates of 5.6% were reported by Dr. Chen for the SeeDs library screen on a dozen diverse targets.

Dr. Francois Delfaud (MEDIT- SA http://medit-pharma.com/index.php?page=MEDIT-SA-Products-Drug-Design-Softwares) reported on Med Sumo, a database searching method to identify similar binding surfaces of macromolecules based on specific similar chemical features and properties. Med Sumo can be used in conjunction with Med-Portions to identify small ligand fragment structures binding to these surfaces. Med-Portions is a protein-ligand database which includes a binding site database and protein-fragment database. Once the Protein-Ligand has a solved structure in the PDB, it can be converted to a protein-fragment pattern (MED-Portions). The Med-Portion database can be mined with a library of small molecules for detection of protein pocket similarities and alignments to identify new binding fragments. It generates fragments sharing some surface interaction features with the query protein-ligand solved structure surface (taken from PDB). A chapter in this volume describes this work in detail.

Dr. Valerie J. Gillet and her colleagues in the Chemoinformatics Research Group, University of Sheffield, reported on a *de novo* design method using reaction vectors and its application to fragment library design. In fragment library *de novo* design, one of the concerns is the design of molecules that can be synthesized easily and cost effectively. Dr. Gillet's method for library design involves a knowledge-based approach using reaction vectors that describe structural changes at the reaction center within a reaction database. Dr. Gillet has written a chapter for this volume which describes her method.

Dr. Qiong Yuan and her group at Chemical Abstracts Service described SubScape for SciFinder (www.cas.org/products/scifindr/subscpe) and its use for substructure searching to facilitate FBDD (fragment-based drug design). Fragments can be analyzed and sorted and organized with associated experimental and predicted properties with each substructure including bioactivity. Chemical space around fragments hits can be optimized using 2D Tanimoto similarity searches (www.cas.org/products/scifindr/subscpe).

Dr. John Badger , DeltaG Technologies in conjunction with Zenobia Therapeutics, reported on the design and application of fragment libraries for crystallography studies. Zenobia applied this strategy to a CNS target (Parkinson's disease), LRRK2, using rule-based filtering software to generate appropriate fragments for crystallographic screening methods. Compounds were searched for drug like core substructures (SDSearch) and then the rule of three was implemented and additional special useful filters (i.e. blood-brain barrier permeability) were used (*34*). SD search is an in house search tool which is comprised of Zenobia's small fragment library. The Target LRRK2- leucine rich repeat kinase was use for "pseudo docking" the lead fragment. In this way key binding interactions could be indentified and a first scaffold screen with a focused library was performed followed by a second round focused around the best hits. LeadModel3D was used for docking. The best hits were selected for an experimental screen. Dr. Badger describes this work in detail in a chapter in this volume.

Drs. Ammar Abdo, and Naomie Salim, Universiti Teknologi Malaysia, Faculty of Computer Science & Information Systems, introduced a novel similarity-based virtual screening approach based on a Bayesian interference network. Their network permits a combination of multiple queries and molecular representations and weighing schemes. They feel that their method surpasses the Tanimoto similarity approach and offers a reasonable method to assess 2D similarity between structures. A chapter in this volume discusses this work.

Dr. Tobias Lippert, Zentrum für Bioinformatik, Universität Hamburg , presented ,Qsearch: a pharmacophore-based search in fragment space. QSearch is an iterative search method using molecular evolution to allow a search of fragment space for molecules that can fulfill the criteria of a three dimensional pharmacophore. The search method uses an evolutionary approach where partial solutions evolve to fit the posed query by adding, deleting or replacing fragments. The fitness of a partial solution is calculated by its ability to obey the constraints of the pharmacophore. An example was presented using a thrombin query (PDB structure 1c4v) and focused fragment space as input with known thrombin inhibitors cleaved with BRICS rules which gave 800 fragments.

Dr. J. D. MacCuish and colleagues at Mesa Analytics & Computing, Inc., reported on a method for shape clustering of fragment databases using both 3D shape fingerprints (generated via Quasi-Monte Carlo integration) and 2D structure fingerprints. Individual clusters are then analyzed with 3D shape fingerprints incorporating substructure information, akin to substructure commonality programs with 2D fingerprints, such as Stigmata and ChemTattoo.

Dr. V. V. Poroikov, Department for Bioinformatics, Institute of Biomedical Chemistry of Rus. Acad. Med. Sci., reported on the PASS method that predicts more than 3000 biological activities of a database of chemical compounds (http://www.ibmc.msk.ru/PASS) (*35*). Prediction is based on SAR (structure activity) analysis of the training set containing over 200,000 biologically active compounds collected from different sources. PASS calculates the impact of each atom in a molecule into a certain activity using MNA descriptors for each particular atom and its immediate neighbors. These estimations could be used for identification of fragments responsible for binding chemical compounds with a specific target, and for further computer-aided design or generation of new "candidates" with the required biological activity.

Drs. Y. Xu, H. Jansen, and E. Martin of the Novartis Institutes for Biomedical Research, proposed a method when binding site information is known, modifications can be proposed using a “cut and fit" and "fit and cut “ method. The "cut and fit" approach fragments a compound database replacing part of a lead molecule with fragments; the fit and cut starts with a complete molecule from a compound database, and determines whether this molecule fits. The fragments selected are separated and merged with relevant parts of the lead molecule. Finally, the fitting of the new modifications are confirmed with docking method. This method has produced interesting ideas in multiple kinase projects. As a validation of the method, a case study with the P38 ATP pocket and the MDDR database was described.

ALTA (Anchor-Based Library Tailoring) a focused chemical library obtained by prioritizing molecular fragments according to their docking energy is a

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

technique developed by Dr. Peter Kolb and Dr. Caflish (*36*). An example was presented in the talk by Dr. Kolb of the small molecule ABT-737 mapping bound to Bcl-XL and this method is discussed further in a chapter by Dr. Kolb in this volume.

II. Computational Screening: Docking

After designing fragment libraries and screening these libraries for hits, the next step in the process is to computationally dock these fragments into the targets or receptors to determine energetically favorable binding site positions for fragments and functional groups. Once a good fragment hit is found, it is developed into a lead by linking, growing or merging. Fragments can be positioned in the binding cleft of protein targets and then grown, attached or linked to fill the binding pockets optimizing steric, electrostatic, van der Waals and hydrogen bonds. Some of the commonly used software methods available for fragment positioning in the past are the methods GRID, MCSS, SPROUT, MUSIC, LUDI, Sklegen (De Novo Pharmaceuticals) and Superstar (CCDC). At the ACS symposiums presentations were given on the methods whose descriptions follow.

The MCSS method (Multiple Copy Simultaneous Search), and the Miranker and Karplus paper (1991), are considered by many to be the one of the first examples of fragment-based ligand docking. The MCSS method (currently implemented as a commercial product in Accelrys Discovery Studio software suite) uses simultaneous molecular mechanics minimization of fragments in the active site using CHARMm force field and the fragments are ranked by their MCSS energy score. The MCSS method can be used to determine fragment binding modes. Dr. Jürgen Koska and colleagues from Accelrys and Pfizer presented a paper at this meeting docking small fragments using MCSS minimization. Accelrys has taken the original MCSS method and incorporated it into a fully automated Pipeline Pilot workflow and demonstrated performance with scoring and placing of fragments with correct poses in several protein-fragment complexes.

Drs. Dima Kozakov, and Sandor Vajda, Boston University, gave a presentation on FTMAP, a method also based on MCSS, to find druggable sites at protein-protein interfaces using computational fragment mapping. Developing drugs which target protein-protein interfaces has been a recent area of significant interest. The FTMAP method uses small molecules, fragments or groups on the surface as probes and finds and clusters their most energetically favorable positions. In this way, the hot spots for drug binding are identified. When proteins interact as binding partners, although there might be a large binding surface area, there are specific essential “hot spots” where they interact. Using a small organic molecule probe, and an efficient FT algorithm (FTMAP) for sampling the surface area in a grid like manner, clusters of probe consensus binding sites can be identified. The largest consensus cluster is the most significant “hotspot” for binding. This methodology grew out of a experimental techniques for solving protein x-ray structures called Multiple solvent crystal structures (MSCS)

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

(*37*), where each solved structure is solved with multiple organic solvents . When these structures are compared, the organic solvent molecules cluster in consensus sites that are significant functional hot spots. The idea was to develop a computational method which could identify these functional hot spots to use in place of experimental x-ray crystallography. The output from FTMAP is a PDB file and the 6 lowest energy cluster representations for reach probe, the number of nonbonded interactions between the probes and residues, and the number of H bonds between probes and each residue. These docked organic fragments can then serve as a starting point for fragment-based drug design. This group has applied this method to several protein-protein interaction systems, including interleukin-2, Bcl-xL, MDM2, HPV-11 E2, ZipA, TNF-alpha, and NEMO. The FT map server is freely available for users http://ftmap.bu.edu (*38*).

Dr. Peter Kolb, (Professor Amedeo Caflish's group, University of Zurich), gave a presentation on the application of several of the methods for fragment design developed by this group: DAIM, SEED, FFLD and GANDI. The Caflisch group has designed several publicly available programs (http://www.biochem-caflisch.uzh.ch/download/) which work together for fragment ligand design. DAIM for decomposition of molecules into a fragment library: SEED (Solvation Energy for Exhaustive Docking) for docking fragments and FFLD for molecule docking based on docked fragment locations found using SEED. After automatically decomposing molecules in a library into fragments (DAIM) they can be docked and the docked fragments ranked (SEED). SEED docks the fragments and the favorable poses of anchor fragments are used for FFLD (Fast Flexible Ligand Docking) which docks the molecules designed, so the two programs SEED and FFLD are designed to work together. The docking program SEED docks and orients fragments into a binding pocket (*39*, *40*),and the binding energy estimated. The free energy of binding can be calculated for multiple poses (LIECE (linear interaction energy with continuum electrostatic method).

GANDI is a Genetic Algorithm-based *de novo* design of inhibitors and is a program for fragment-based *de novo* ligand design (*41*). GANDI performs automatic design of molecules within known binding site structures. It includes a novel simultaneous energy minimization and a term forcing 3D similarity to known inhibitors or ligands through 3D overlap. GANDI's fragment method joins predocked fragments with linkers, which are evaluated with a search algorithm. Dr. Kolb has written a chapter describing applications of these methods in detail.

Dr. Zsolt Zsoldos, SimBioSys Inc, (http://www.simbiosys.ca/) gave a presentation on the application of the fragment-based docking and linking engine of eHiTS . Any docking method used for molecules can be used for fragments as well. However, many of the conventional docking methods have problems with fragments largely due to the fact that the cavities are large and the fragments small and therefore the fragments are not sufficiently constrained for docking within the cavity. Since eHITs works by breaking down larger ligands into small fragments and docking them independently and then reconnecting the fragment poses it has resulted in about a 0.5 Å RMSD small fragment pose prediction and is capable of linking the fragments without loss of information (*42–44*). Dr. Zsoldos has a chapter in this volume describing applications of eHiTS for fragment docking.

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

III. Leads from Fragments: Fragment Growing and Linking

The time required to develop good leads from fragment-based screens can be longer than other methods since fragments involve additional development work to evolve into leads. Good methods for growing and linking fragments can aid this process. There are several general methodologies for creating a lead compound from fragments: (1) linking two or more fragments that bind to different parts of the target pocket to create a lead with a chemical bridge compound or linker (2) fragment self binding- where two or more fragments bind or connect to each other due to chemistry without a linker ("click chemistry" is an example) (3) simply optimizing the fragment itself alone and essentially using the fragment itself as a lead (4) "fragment evolution" where functional groups which bind to the target binding partner are added to increase the fragments affinity . Commonly used computational methods for fragment linking in the past include: Caveat, HOOK, Recore, Allegrow, Confirm, MED-SuMo LEGEND (*45*) LUDI (*46*), GROWMOL (*47*) LigBuilder (*48*) SkelGen (*49*, *50*) SMoG (*51*) LUDI, HOOK (*52*),, PRO_LIGAND (*53*), LigBuilder, SPLICE/RACHEL (*54*), CAVEAT (*55*, *56*), CLIX and LUDI GROW, LEGEND, LORE, GEMINI. GRID and many similar methods place fragments on grid points in the active site and determine favorable interactions. HSITE maps hydrogen bonding regions of the enzyme active site. Other methods for *in situ* fragment linking involve dynamic combinatorial chemistry, "click chemistry", and fragment tethering- disulfide bond formation between cys residues in the protein and the fragment. .

GroupBuild (*57*) was one of the first design linking methods, described as a "fragment-based method". GroupBuild uses a library of organic templates and a force field describing nonbond interactions between the ligand and enzyme to build drug candidates that have steric and electrostatic properties that fit with the enzyme binding site.

Dr. A. Peter Johnson and colleagues from the University of Leeds, reported on SPROUT (*58*), a computational tool for growing fragments (now available commercially through SimBiosys Inc. http://www.simbiosys.com/sprout/index.html). SPROUT was an older program originally developed for *de novo* ligand design, however it is useful for fragment linking as well. When information is known about a fragment and its binding pose, SPROUT with two or more fragments, is able to link them together, redocking to maintain the original poses, also permitting some movement , limited by user selected tolerances. SPROUT is also capable of fragment growth (evolution). SPROUT locates binding pockets and identifies potential interaction sites (H bonding, hydrophobic, covalent, metal, user defined) and docks molecular fragments to target sites, and generates novel chemical structures from templates and clusters. Using SynSPROUT, the fragments are linked through virtual synthetic chemistry. LeadOpt has reaction information and starting material information to predict virtual reaction and synthesis of the ligands which are known to bind and works together with SynSPROUT.

FlexNovo is useful for large fragment fragment space database docking (*59*), when the active site structure is known. FlexNovo is based on the FlexX flexible docking algorithm with structural information for the target active site and pharmacophore type constraints for the fragments. The fragments in the library are combined and linked and docked in the active site with a calculated score (*60*). Recore, (http://www.zbh.uni-hamburg.de/en/research/computational-molecular-design/projects.html), (developed by Mattias Rarey ,ZBH Hamburg, in conjunction with Hoffmann LaRoche AG Basel Switzerland, commercial product from BioSolveIt) (*61*) is a new fragment replacement tool for fast searching of conformations (similar to the earlier CAVEAT) however combined with a large search domain focusing on druglike structures and including pharmacophore type searching. The result is that Recore provides for effective scaffold hopping, 3D core replacement and fragment linking and merging. The fragment database is created from 3D structures with cleavage rules defined by SMARTS patterns. While ReCore uses indexed searching, it additionally does core replacement, and fragment linking and growing and merging.

AlleGrow, is a program which can be used in the design of cyclic scaffolds which connect and incorporate fragments. It is an update of the earlier GrowMol (*62*)(RS Bohacek). Cyclic scaffolds are fairly common in bioactive molecules. CONFIRM (Connecting Fragments Found in Receptor Molecules) is a linking approach with a search library for bridges for fragments and automation of linking and docking to a target (*63*). A prepared library, of bridges and links, is used as linkers to fragments and docked. CONFRIM retrieves molecular fragments based on distances and atom types and searches a database of bridges using a substructure pattern search. These bridges are linked to the fragments in combinatorial ways and then proposed molecules are docked computationally.

Dr. Jacob Durant (Dr. McCammon's group, University of San Diego) gave a presentation on AutoGrow a fragment growing method using an initial core and randomly adding fragments to the core scaffold which are then dynamically docked into the protein . It is a genetic algorithm so the compounds that dock the best become the scaffolds for the next generation in an iterative fashion (*64*). It is freely available for download: http://autogrow.ucsd.edu/ . AutoGrow and AutoClick are based on an evolutionary algorithm which starts with an initial scaffold mutation operator and replaces it with a molecular fragment and docks it into the target structure.

BREED (licensed from Vertex and implemented as scripts fragment_join.py and fragment_link.py in Glide XP for fragment docking within the Schrodinger software suite http://www.schrodinger.com/scriptcenter/#Fragments, and also available as part of the Chemical Computing Group MOE software suite http://www.chemcomp.com/software-sbd.htm) (*65*), is an automated computational method to create new inhibitors by joining fragments from ligands whose structures bound to the target are already known. Scaffold structures of known ligands are superimposed so that similar bonds match and can then be split and recombined in different combinations to generate new ligands. This method is dependent on having structural information. It is based on and similar to the earlier SPLICE method (*66*). BREED has been used successfully by Vertex to design HIV protease inhibitors and kinase inhibitors . Methods like BREED, that replace groups hanging off a scaffold belong to the group of methods referred to as "scaffold hoping". BREED detects bonds from different ligands in close proximity in spatial alignment when the ligands are superimposed. Fragments

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

can be recombined to form chimeric compounds. This is referred to as ‘fragment shuffling” (BREED, RECORE, FLUX and MED-Hybridise are all methods in this category). In addition to BREED, Chemical Computing Group also has tools within the MOE software suite for scaffold replacement, fragment linking and growing , and medchem transformations including pharmacophore features (see http://www.chemcomp.com/journal/newscaffold.htm).

Dr. Peter Kutchukian (Harvard) gave a presentation on FOG, (Fragment Optimized Growth Algorithm), a statistically biased growth of fragments to produce compounds with certain features that appear with high occurrence in the training database (*67*). It uses a Markov Chain approach with branching treating each new fragment as a new transition probability and training on a database of bioactive compounds. Fragments must have shape and energetic complementary to the binding pockets, and synthetic feasibility so that certain types of molecular connections are favored in growing fragments and others forbidden. The result is the generation of molecules that have “druglikeness” properties (i.e. occupying drug like chemical space). Dr. Kutchukian has written a chapter in this volume describing FOG.

Dr. Dan Erlanson, (Carmot Therapeutics, Inc.), developed a technology, Chemotype Evolution (*68*), which uses rapid *in-situ* chemistry to expand a fragment into a diverse range of hits. Chemotype Evolution begins with a “Bait fragment” which is modified with a group or fragment and then used for screening so that a new “chemotype” is generated. This leads to custom library generation in an iterative fashion. The chemotype evolution method is directed at finding a “good fragment”, and is based on elaborating fragments found using any method: a starting “bait” fragment can be a “privileged” pharmacophore derived from a known inhibitor, substrate, or cofactor, or a fragment identified through a previous screen. Through iterative application of Chemotype Evolution, the starting fragment can be transformed into novel, varied “chemotype”, while desired properties can be enhanced by incorporating counter screens. This techinique was applied to the challenging design of Aurora A fragments in an adaptive kinase pocket with DFG loop movement (*69*).

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

IV. Examples of Successful Applications of Fragment Design Process

Dr. Valerio Berdini, Astex Therapeutics, gave a presentation on the discovery of AT7519, a novel CDK inhibitor (which at the time of the meeting was in clinical trials) and AT9283, using fragment-based drug design methods. Fragments used for the CDK project were used to develop novel Aurora kinase inhibitors. This work led to the identification of AT9283 which is also was in clinical trials at the time of the meeting. AT7519 inhibited CDK2 with an IC_{50}=0.047 mircomolar and LE =0.42 (*70*). AT9283 (Pyrazil-4-yl Urea) is an Aurora kinase inhibitor with IC_{50}=0.91 mircomolar and LE =0.59 (*71*).

Dr. Miles Congreve, Heptares Therapeutics, discussed fragment-based screening of stabilized G protein-coupled receptors. GPCRs are difficult targets due to their conformational flexibility, heterogeneity and instability outside the cell membrane. Heptares is establishing a GPCR targeted fragment library by using a uniquely stabilized receptor. Heptares "STAR technology" involves iteratively introducing small mutations which stabilize and trap GPCR conformations. A GPCR targeted fragment library is underoing development using this technique in conjunction with biophysical mapping using Surface Plasmon Resonance (Biacore) and TINS, target immobilized NMR screening.

Dr. Richard J. Law, Evotec, discussed a novel histamine GPCR family antagonist designed by fragment screening and molecular modeling. Applying a small fragment collection to the screening of three histamine receptors, the goal was to identify subtype specific antagonists. This resulted in fragment hits by building H3 and H4 receptor models based on similarity to known GPCR crystal structures and optimizing them using a series of molecular dynamics procedures. These models were used for docking procedures to reveal the bioactive conformation of the bound ligands, with a view to structure-guided fragment-to-lead expansion. A subsequent shape-based analogue search provided a short list of hits from which novel submicromolar and lead-like H3 and H4 antagonists were obtained. Evotec's substructure search and Gold docking was based on virtual screening (using OpenEye ROCS) and NOE docking and QM calculations.

Dr. Francois Delfaud, MEDIT SA, presented mitotic kinesin Eg5 inhibitors generation by MEDIT's computational MED-Portion based drug design. Eg5, a mitotic kinesin is exclusively involved in the formation and function of the mitotic spindle and has attracted interest as an anticancer drug target. Eg5 is co-crystallized with several inhibitors bound to its allosteric binding pocket. Each of these occupies a pocket formed by loop5/helix α2. Recently designed inhibitors additionally occupy a hydrophobic pocket of this site. The goal of the present study was to identify new fragments which fill this hydrophobic pocket and might be interesting chemical moieties to design new inhibitors. Dr. F Delfaud presented the application of the MEDIT SA software Med Fragmentor, and Med-Sumo application to this problem which is dicussed in detail in a chapter which follows.

Dr. Vicki L. Nienaber, Zenobia Therapeutics, discussed the application of fragment-based design methods to particular issues involved with CNS drug design-mainly dealing with the challenge presented by compounds that cross the blood brain barrier. Their target enzyme is a LRRK2 kinase (Parkinson Disease target) using a chemical property filter and a structural computational fragment screen. Dr. Nienaber discussed several positive outcomes from their screens which will be very good potential leads for new drug development in this area. A chapter on this work in detail is included in this volume (*72*).

The chapters which follow in this volume discuss these specific examples of fragment-based drug design successful application and novel computational method development in detail. A table accompanying this chapter summarizes computational methods available for fragment-based drug design along with reference information. I hope this volume conveys the excitement generated by the development and promise of fragment-based ligand design for drug discovery and development. Hopefully the years which follow will show the fruition of this promise with drugs in clinical development.

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

Appendix 1.

Software Used to Facilitate Fragment Based Drug Discovery

Name of program/algorithm	Organization or Company	Website	reference	comments
I. Designing Fragment Libraries				
BRICS model and fragment spaces	Jörg Degen, Andrea Zaliani , Matthias Rarey, Center for Bioinformatics University of Hamburg; Christof Wegscheid-Gerlach , Bayer Schering Pharma AG	http://www.zbh.uni-hamburg.de/en/research/computational-molecular-design/software.html	Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M, *ChemMedChem* **2008**, 10, 1503-7.	academic
RECAP fragmentor (retrosynthesis and combinatorial analysis)	Part of the fragmentor in The Chemaxon software suite (original development: GlaxoWelcome UK)	http://www.chemaxon.com/jchem/doc/user/fragment_recap.html	Lewell XQ, Judd DB, Watson SP, Hann MM.*J Chem Inf Comput Sci*, **1998**, 38, 511-522.	commercial
DAIM (decomposition and identification of molecules)	Peter Kolb and Amedeo Caflish, Department of Biochemistry, University of Zurich	http://www.biochem-caflisch.uzh.ch/download/	Kolb P, Caflisch A.*J Med Chem.* **2006,** 49, 7384-92.	academic
CoLibri (fragmentor and assembler of fragment libraries)	BioSolveIt GmbH	http://www.biosolveit.de/CoLibri/index.html?ct=1	Lessel, U., Wellenzohn, B. Lilienthal, M. Claussen, H., *J. Chem. Inf. Model.*, **2009**, 49, 270–279. Boehm, M. Wu, T., Claussen, H., Lemmen, C., *J. Med. Chem.*, **2008**, 51, 2468–2480	commercial
FTrees	Matthias Rarey	http://public.zbh.uni-	Rarey,	WebServer

	Center for Bioinformatics, University of Hamburg	hamburg.de/ftrees/query.py http://www.biosolveit.de/FTrees-FS/	M.,Dixon, J.S. *J. Comp Aided Mol Des*, **1998**, 12, 471-490.	(academic) for installable software BioSolveIT (commercial)
LOFT	J. Robert Fischer and Mattias Rarey, Center for Bioinformatics, University of Hamburg	http://www.zbh.uni-hamburg.de/home.html	J Fischer, JR.,Lessel U.,Rarey,M *J Chem, Inf. Model* **2010,** 50, 1-21.	
SeeDs (Structural Exploitation of Experimental Startpoints)	Ijen Chen and Roderick E Hubbard , University of York and Vernalis,	http://www.york.ac.uk/chemistry/staff/academic/h-n/rhubbard/	Hubbard RE, Davis B, Chen I, Drysdale MJ., *Curr Topics in Medicinal Chemistry* **2007**, 7, 1568-1581.	
MED-Portions	MEDIT	http://medit-pharma.com/index.php?page=MEDIT-SA-Products-Drug-Design-Softwares	F Moriaud, O Doppelt-Azeroual, L Martin, K Oguievetskaia, K Kosch, A vorotyntsev, SA Adcock, F Delfaud, **2009**, *JCIM*, 49, 280-94.	commercial
SubScape for SciFinder	CAS	ww.cas.org/roducts/scifindr/subscpe		commercial
BROOD (CHOMP)	OpenEye	http://www.eyesopen.com/brood	Chen, X. and Wang, W., *Annual Reports in Medicinal Chemistry*, #38, Elsevier, Inc., **2003.**	commercial
fragment_molecule.py (fragmentor script: Schrödinger Glide	Schrödinger	http://www.schrodinger.com/scriptcenter/#Fragments http://www.schrodinger.c	Loving K, Salam NK, Sherman W. *J Comput Aided*	commercial

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

fragment library)		om/productpage/14/5/73/	*Mol Des*. **2009,** 23, 541-554.	
QSearch	Tobias Lippert and Mattias Rarey, Center for Bioinformatics University of Hamburg	http://www.zbh.uni-hamburg.de/en/research/computational-molecular-design/software.html		
II. Computational Screening : Docking				
FTMAP (MSCS)	Dima Kozakov and Sandor Vajda , Boston University and Acpharis	http://ftmap.bu.edu/ http://www.acpharis.com/	Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, Mattos C, Vajda S.,Bioinformatics. **2009** , 25,621-7.	academic and commercial
SEED (Solvation Energy for Exhaustive Docking) ALTA (Anchor Based Library Tailoring)	Amedeo Caflish's group , Department of Biochemistry, University of Zurich,	http://www.biochem-caflisch.uzh.ch/download/ http://www.biochem-caflisch.uzh.ch/publication/103/structure-based-tailoring-of-compound-libraries-for-high-throughput-screening-discovery-of-novel-ephb4-kinase-inhibitors.html	Majeux, N., Scarsi, M., Caflisch, A.,*Proteins*. **2001,** 42, 256-68. Kolb P, Kipouros CB, Huang D, Caflisch A. *Proteins***, 2008**, 73,11-18.	academic
GRID	Molecular Discovery Ltd.	http://www.moldiscovery.com/soft_grid.php	Goodford, P., *JMedChem,* **1985**, 28, 849-857.	commercial
FlexNovo	BiosolveIt GmbH	http://www.biosolveit.de/software/libraries.html/	Lessel, U: Wellenzohn B; Lilientahl M; Claussen H *J Chem Inf Comput Sci* **2009**, 49, 270-279. Boehm M; Wu TY; Claussen, H; Lemmen C *J Med.Chem*,**200**	commercial

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

			8, 51, 2468-80.	
MCSS	Accelrys (Introduced Discovery Studio 2.5)	http://blog.accelrys.com/tag/mcss/	A. Miranker and M. Karplus, *Proteins*, **1991**, 11, 29-34. Eisen, Karplus and Hubbard, *Proteins*, **1994**, 19, 199-221.	commercial
BROOD	OpenEye	http://www.eyesopen.com/brood	Chen, X. and Wang, W., *Annual Reports in Medicinal Chemistry*, #38, Elsevier, Inc., **2003.**	commercial
eHITS Electronic High Throughput Screening	SimBioSys Inc.	http://www.simbiosys.ca/ehits/index.html	Zsoldos, Z., Reid, D. Simon, A., S.B. Sadjad, S.b., A.P. Johnson, A.P., *J.Mol.Graph.Modeling*. **2007**, 26, 198-212.	commercial
fragment_selector.py (ScriptCenter) scores and filters docked fragment poses from Glide based on ligand efficiency	Schrödinger	(Glide product page *http://www.schrodinger.com* *http://www.schrodinger.com/scriptcenter/#Fragments* (ScriptCenter)	Loving K, Salam NK, Sherman W. *J Comput Aided Mol Des*. **2009,** 23, 541-554.	commercial
DLD (dynamic ligand design)	Andrew Miranker and Martin Karplus Harvard		Miranker and Karplus , Proteins, 23, 472-490; Stultz and Karplus, Proteins, 2000, 40, 258-89	academic
PhDock Pharmacophore based docking The method is implemented in DOCK 4.0	Diane Joseph-McCarthy,Bert E. Thomas IV, Michael Belmarsh, Demetri Moustakas, and Juan C. Alvarez *Wyeth Research*	DOCK is freely available http://dock.compbio.ucsf.edu/	Joseph-McCarthy, D., Thomas , BE., IV, Belmarsh, M., Moustakas, D., Alvarez , JC., *Proteins*, **2003**, 51,172–	academic

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

			188 .	
SILCS = Site Identification by Ligand Competitive Saturation	alex@ outerbanks.umaryland.edu		Raman EP, Yu W, Guvench O, Mackerell AD.*J Chem Inf Model.* **2011** , 51,877-96.	academic
MUSIC-(part of BOSS)	Developed by Dr Heather Carlson (Michigan) in conjunction with Drs. William Jorgensen (Yale) and Dr. Andrew McCammon (UCSD) and others	http://sitemaker.umich.edu/carlsonlab/home.html http://mccammon.ucsd.edu/pubs/abstracts00.html	Carlson, H.A., Masukawa, K.M., Rubins, K.,Bushman,F. D., Jorgensen, W.L., Lins, R.D.,Briggs,|J. M., McCammon, J.A., *J.Med.Chem.* **,** **2000***,* 43,2100-14.	academic
Superstar-	CCDC	http://www.ccdc.cam.ac.uk/products/life_sciences/superstar/	Taylor, R.D., Verdonk, M.L., *J.Mol.Bio.***,** **1999** 289, 1093-1108. Verdonk , M.L., Cole, J.C., Watson, P., Gillet, V., Willett , P., *J.Mol.Bio* , **2001**, 307, 841-859.	commercial
III. Fragment Growing and Linking (Scaffold Hopping)				
BREED-linking breed.py Schrödinger script	Schrödinger (licensed from Vertex)	http://www.schrodinger.com/scriptcenter/#Fragments	AC Pierce , G. Rao, GW Bernis, *J Med Chem*, **2004,** 47, 2768-75.	commercial
CombiGlide (core hopping)	Schrödinger	http://www.schrodinger.com/products/14/2/		commercial
combine_fragments.py (2009 ScriptCenter) direct joining and linking of	Schrödinger	http://www.schrodinger.com/scriptcenter/#Fragments		commercial

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

fragments				
Fragment based pharmacophore; Initial glide XP docking-phase (pharmacophore hypothesis from top fragments followed by Phase DB search and scoring)	Schrödinger	http://www.schrodinger.com/	Loving K, Salam NK, Sherman W. *J Comput Aided Mol Des*. **2009,** 23, 541-554.	commercial
CONFIRM (Connecting Fragments Found in Receptor Molecules) (implanted as an automated protocol in Accelrys Pipeline Pilot)	efeyfant@wyeth.com		DC Thompson, RA Denny, R Nilakantan and C Humblet and D Joseph-McCarthy and E Feyfant . *J Comput Aided Mol Des*. **2008**, 22,761-772.	
AutoGrow-	Jacob D. Durrant, Rommie E. Amaro and J. Andrew McCammon, University of California San Diego	freely available for download: http://autogrow.ucsd.edu/	JD Durrant, RE Amaro and J Andrew McCammon, *Chem Biol Drug Des* **2009**, 73, 168-178.	academic
FOG	Peter Kutchukian And Shakhnovich (Harvard)	http://www-shakh.harvard.edu/research/index.html#drug	Kutchukian, PS., Lou, D., Shakhnovich, EI., *J Chem Inf* Model. **2009** Jul;49(7):1630-42.	academic
CCG-MOE procedure	Scaffold Replacement in MOE	http://www.chemcomp.com/software-sbd.htm http://www.chemcomp.com/journal/scaffold.htm	Deschênes, A., Sourial, E.*J. Chem. Comp. Group* (**2007**).	commercial
CAVEAT (linking fragments	Paul Bartlett (UC Berkeley) Licenses for CAVEAT and the TRIAD and ILIAD databases are available from the	http://www.cchem.berkeley.edu/pabgrp/Data/caveat.html	Lauri G, Bartlett PA. *J Comput Aided Mol Des*. **1994**, 8, 51-66. PA Bartlett, et	academic

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

	University of California Office of Technology Licensing		al, in Molecular Recognition: Chemical and Biological Problems: SM Roberts, ed Royal Society of chemistry, 182-196, 1989.	
Recore	BioSolveIt Scaffold replacement	http://www.biosolveit.de/recore/	Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M; *J Chem Inf Model.* **2007** 47,390-9.	commercial
SMoG and CombiSmoG	Shakhanovich's group, Harvard University	http://www-shakh.harvard.edu/~smog/	DeWitte, RS., Shakhnovich, EI**.,** *JACS*, **1996,** 118, 11733-44.	academic
MED-SuMo	MEDIT	http://medit-pharma.com/index.php?page=MEDIT-SA-Products-Drug-Design-Softwares	F Moriaud, O Doppelt-Azeroual, L Martin, K Oguievetskaia, K Kosch, A vorotyntsev, SA Adcock, F Delfaud, **2009**, *J Chem Inf Model.*, 49, 280-94.	commercial
VFL: Virtual Fragment Linking model performance classification according to rules implemented in Pipeline Pilot	Center for Proteomic Chemistry, Novartis Institutes for BioMedical Research	meir.glick@novartis.com	Crisman TJ, Bender A, Milik M, Jenkins JL, Scheiber J, Sukuru SC, Fejzo J, Hommel U, Davies JW, Glick M. ,*J Med Chem.* **2008**, 24,2481-91.	
FragFCA-	LIMES Program Unit Chemical	http://www.limes.uni-bonn.de/forschung/abteil	Lounkine, E., Auer, J., and	academic

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

	Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-UniVersita¨t, Bonn, Germany	ungen/Bajorath/labwebsite/research	Bajorath, J., *J. Med Chem* **2008**, 5342-48.	
Skelgen Fragment based design; scaffold hopping	DeNovo Pharmaceuticals	http://www.denovopharma.com/page2.asp?PageID=484	Todorov , N., Dean, PM., . *J Comput Aided Mol Des*, **1997**, 11, 175-192	commercial
LigBuilder (2) growing and linking-	jfpei@pku.edu.cn (J.F.P.); Jianfeng Pei lhlai@pku.edu.cn (L.H.L.). Luhua Lai Peking University, Beijing, China	The Cavity 1.0 program can be downloaded at http://mdl.ipc.pku.edu.cn/. Academic users can acquire the LigBuilder 2.0 package by contacting the authors.	Wang, R., Gao Y., Lai, L., *J Mol Model,* **2000**, 6, 498-516. Yuan, Y., Pei, J., Lai, L., *J. Chem. Inf. Model.*, **2011**, 51 , 1083–1091.	academic
TOPAS (DOGS - Design Of Genuine Structures)	The Computer Assisted Drug Design Group, ETH Zurich	http://www.pharma.ethz.ch/institute_groups/computer_drug_design/research/index	Schneider, P., Tanrikulu, Y. and Schneider, G., *Curr. Med. Chem*., **2009,** 16, 258-266.	academic
BROOD	OpenEye	http://www.eyesopen.com/brood	Blomberg, N., Cosgrove, DA., Kenny,PW., Kolmodin, K., *J. Comp.Aided Mol.Design,* **2009***, 23, 513-525.*	commercial
SPROUT	Originally developed by V. Gillet, P. Johnson et al., The University of Leeds (Now commercial Simbiosys)	http://www.simbiosys.ca/sprout/index.html	Gillet V, Johnson AP, Mata P, Sike S, Williams P.*J Comput Aided Mol Des.* **1993**, 7,127-53.Gillet VJ, Newell W, Mata P, Myatt	commercial

			G, Sike S, Zsoldos Z, Johnson AP.*J Chem Inf Comput Sci.* **1994** 34,207-17.	
LUDI	H Bohm (Accelrys) Part of the Accelrys commercial software tools	www.accelrys.com	*Bohm, HJ., J.Comput Aided Mol Des* **1992**, 6, 61-78. Bohm, HJ, *J.Comput Aided Mol Des*, **1992**, 6, 593-606.	commercial
SYNOPSIS= Synthetize and Optimize System in Silico	MolMo Services BVBA Retie Belgium	http://www.molmo.be/synopsis.html	Vinkers HM, de Jonge MR, Daeyaert FF, Heeres J, Koymans LM, van Lenthe JH, Lewi PJ, Timmerman H, Van Aken K, Janssen PA. *J Med Chem.* **2003 ,** 46,2765-73.	commercial
GANDI- linking	Dr. A Caflisch Universitat Zurich	http://www.biochem-caflisch.uzh.ch/download/	F Dey and A Caflisch, *J Chem Inf Model*, **2008**, 48, 679-90.	academic
AlleGrow (GrowMol)	Boston De Novo Design Regine Bohacek	http://www.bostondenovo.com/Allegrow.htm	RS Bohacek and C McMartin, *JACS*, **1994**, 116, 5560-5571	commercial
LEA3D-	Centre de Biochimie Structurale CNRS, , France, and Unite´ de Chimie Organique and Laboratoire de Chimie Structurale	http://chemoinfo.ipmc.cnrs.fr/lea.html	Douguet,D., Munier-Lehmann, H., Labesse, G and Pochet|, S., *J Med Chem*, **2005**, 48, 2457-68.	academic
	des Macromole´cules Institut Pasteur,			
GRID	Molecular Discovery Ltd. (commercial)	http://www.moldiscovery.com/soft_grid.php	Goodford, PJ, *JMedChem,* **1985,** 28, 849-857.	commercial
Automated Fragment shuffling workflow	Bayer AG, Corporate Development - Innovation, ulrich.rester.ur@bayer-ag.de		B Nisius and U Rester, , *J Chem Inf Model*, **2009,** 49, 1211-11.	

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

References

1. Zoete, V.; Grosdidier, A.; Michielin, O. *J. Cell. Mol. Med.* **2009**, *13*, 238–48.
2. Schuffenhauer, A.; Ruedisser, S.; Marzinzik, A. L.; Jahnke, W.; Blommers, M.; Selzer, P.; Jacoby, E. *Curr. Top. Med. Chem.* **2005**, *5*, 751–762.
3. Crisman, T. J.; Bender, A.; Milik, M.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C.; Fejzo, J.; Hommel, U.; Davies, J. W.; Glick, M. *J. Med. Chem.* **2008**, *51*, 2481–91.
4. Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. *Science* **1996**, *274*, 1531–4.
5. Nienaber, V. L.; Richardson, P. L.; Klighofer, V.; Bouska, J. J.; Giranda, V. L.; Greer *Nat Biotechnol* **2000**, *18*, 1105–8.
6. Loving, K.; Salam, N. K.; Sherman, W. *J. Comput.- Aided Mol. Des.* **2009**, *3*, 541–554.
7. Marcou, G.; Rognan, D. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.
8. Sun, C.; Petro A. M.; Hajudk, P. J. *J. Comput.-Aided Mol. Des.* **2011**, accessed online July 6.
9. Sutherland, J. J.; Higgs, R. E.; Watson, I.; Vieth, M. *J. Med. Chem.* **2008**, *51*, 2689–2700.
10. Nisius, B.; Rester, U. *J. Chem. Inf. Model.* **2009**, *49* (5), 1211–22.
11. Geschwindner, S.; Olsson, L. L.; Albert, J. S.; Deinum, J.; Edwards, P. D.; de Beer, T.; Folmer, R. H. *J. Med. Chem.* **2007**, *50*, 5903–11.
12. Edwards, P. D.; Albert, J. S.; Sylvester, M.; Aharony, D.; Andisik, D.; Callaghan, O.; Campbell, J. B.; Carr, R. A.; Chessari, G.; Congreve, M.; Frederickson, M.; Folmer, R. H.; Geschwindner, S.; Koether, G.; Kolmodin, K.; Krumrine, J.; Mauger, R. C.; Murray, C. W.; Olsson, L. L.; Patel, S.; Spear, N.; Tian, G. *J. Med. Chem.* **2007**, *50*, 5912–25.
13. Murray, C. W.; Callaghan, O.; Chessari, G.; Cleasby, A.; Congreve, M.; Frederickson, M.; Hartshorn, M. J.; McMenamin, R.; Patel, S.; Wallis, N. *J. Med. Chem.* **2007**, *50*, 1116–23.
14. Frederickson, M.; Callaghan, O.; Chessari, G.; Congreve, M.; Cowan, S. R.; Matthews, J. E.; McMenamin, R.; Smith, D. M.; Vinković, M.; Wallis, N. G. M *J. Med. Chem.* **2008**, *51*, 183–6.
15. Card, G. L.; Blasdel, L.; England, B. P.; Zhang, C.; Suzuki, Y.; Gillette, S.; Fong, D.; Ibrahim, P. N.; Artis, D. R.; Bollag, G.; Milburn, M. V.; Kim, S. H.; Schlessinger, J.; Zhang, K. Y. *Nat. Biotechnol.* **2005**, *23*, 201–7.
16. Oltersdorf, T.; Elmore, S. W.; Shoemaker, A. R.; Armstrong, R. C.; Augeri, D. J.; Belli, B. A.; Bruncko, M.; Deckwerth, T. L.; Dinges, J.; Hajduk, P. J.; Joseph, M. K.; Kitada, S.; Korsmeyer, S. J.; Kunzer, A. R.; Letai, A.; Li, C.; Mitten, M. J.; Nettesheim, D. G.; Ng, S.; Nimmer, P. M.; O'Connor, J. M.; Oleksijew, A.; Petros, A. M.; Reed, J. C.; Shen, W.; Tahir, S. K.; Thompson, C. B.; Tomaselli, K. J.; Wang, B.; Wendt, M. D.; Zhang, H.; Fesik, S. W.; Rosenberg, S. H. *Nature* **2005**, *435*, 677–81.
17. Howard, N.; Abell, C.; Blakemore, W.; Chessari, G.; Congreve, M.; Howard, S.; Jhoti, H.; Murray, C. W.; Seavers, L. C.; van Montfort, R. L. *J. Med. Chem.* **2006**Feb23, *49* (4), 1346–55.

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

18. Warner, S. L.; Bashyam, S.; Vankayalapati, H.; Bearss, D. J.; Han, H.; Mahadevan, D.; Von Hoff, D. D.; Hurley, L. H. *Mol. Cancer. Ther.* **2006**, *5*, 1764–73.
19. Huth, J. R.; Park, C.; Petros, A. M.; Kunzer, A. R.; Wendt, M. D.; Wang, X.; Lynch, C. L.; Mack, J. C.; Swift, K. M.; Judge, R. A.; Chen, J.; Richardson, P. L.; Jin, S.; Tahir, S. K.; Matayoshi, E. D.; Dorwin, S. A.; Ladror, U. S.; Severin, J. M.; Walter, K. A.; Bartley, D. M.; Fesik, S. W.; Elmore, S. W.; Hajduk, P. J. *Chem. Biol. Drug Des.* **2007**, *70*, 1–12.
20. Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. *J. Med. Chem.* **2004**, *47*, 224–232.
21. Kolb, P.; Caflisch, A. *J. Med. Chem.* **2006**, *49*, 7384–92.
22. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
23. Rarey, M.; Dixon, J. S. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.
24. Rarey, M.; Stahl, M. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 479–520.
25. Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. *J. Chem. Inf. Comput. Sci.* **2009**, *49*, 270–279.
26. Rarey, M.; Dixon, J. S. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.
27. Boehm, M.; Wu, T. Y.; Claussen, H.; Lemmen, C *J. Med.Chem.* **2008**, *51*, 2468–80.
28. Lessel, U.; Wellenzohn, B.; Lilientahl, M.; Claussen, H. *J. Chem. Inf. Comput. Sci.* **2009**, *49*, 270–279.
29. J Fischer, J. R.; Lessel, U.; Rarey, M. *J. Chem. Inf. Model.* **2010**, *50*, 1–21.
30. Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. *ChemMedChem* **2008**, *10*, 1503–7.
31. Chen, X.; Wang, W. *Annual Reports in Medicinal Chemistry*; Elsevier, Inc.: New York, 2003; volume 38.
32. Baurin, N.; Aboul-Ela, F.; Barril, X.; Davis, B.; Drysdale, M.; Dymock, B.; Finch, H.; Fromont, C.; Richardson, C.; Simmonite, H.; Hubbard, R. E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2157–66.
33. Hubbard, R. E.; Davis, B.; Chen, I.; Drysdale, M. J. *Curr. Top. Med. Chem.* **2007**, *7*, 1568–1581.
34. Hartshorn, M. J.; Murray, C. W.; Cleasby, A.; Frederickson, M.; Tickle, I. J.; Jhoti, H. *J. Med. Chem.* **2005**, *48*, 403–13.
35. Filimonov, D. A.; Poroikov, V. V. In *Chemoinformatics Approaches to Virtual Screening*; Varnek, A., Tropsha, A., Eds.; RSC Publishing: London, 2008; pp 182−216.
36. Kolb, P.; Kipouros, C. B.; Huang, D.; Caflisch, A. *Proteins* **2008**, *73*, 11–18.
37. Mattos, C.; Bellamacina, C. R.; Peisach, E.; Pereira, A.; Vitkup, D.; Petsko, G. A.; Ringe, D. *J. Mol. Biol.* **2006**Apr14, *357* (5), 1471–82.
38. Brenke, R.; Kozakov, D.; Chuang, G. Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. *Bioinformatics* **2009**, *25*, 621–7.
39. Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. *Proteins* **1999**, *37*, 88–105.
40. Majeux, N.; Scarsi, M.; Caflisch, A. *Proteins* **2001**, *42*, 256–68.
41. Dey, F.; Caflisch, A. *J. Chem. Inf. Model.* **2008**, *48*, 679–690.

42. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. *J. Med. Chem.* **2007**, *50*, 726–41.
43. Congreve, M.; Aharony, D.; Albert, J.; Callaghan, O.; Campbell, J.; Carr, R. A.; Chessari, G.; Cowan, S.; Edwards, P. D.; Frederickson, M.; McMenamin, R.; Murray, C. W.; Patel, S.; Wallis, N. *J. Med. Chem.* **2007**, *50*, 1124.
44. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. *J. Mol. Graphics Modell.* **2007**, *26*, 198–212.
45. Nishibata, Y.; Itai, A. *Tetrahedron* **1991**, *47*, 8985–90.
46. Bohm, H. J. *J. Mol. Recognit.* **1993**, *6*, 131–137.
47. Bohacek, R. S.; McMartin, C. *J. Am. Chem. Soc.* **1994**, *116*, 5560–5571.
48. Wang, R.; Gao, Y.; Lai, L. *J. Mol. Model.* **2000**, *6*, 498–516.
49. Todorow, N. P.; Dean, P. M. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 175–192.
50. Todorow, N. P.; Dean, P. M *J. Comput.-Aided Mol. Des.* **1998**, *12*, 335–349.
51. Ishchenko, A. V.; Shakhnovich, E. I. *J. Med. Chem.* **2002**, *45*, 2770–80.
52. Miranker, A.; Karplus, M. *Proteins* **1995**, *23*, 472–90.
53. Clark, D. D.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowyca, B.; Westhead, D. R. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 13–32.
54. Ho, C. M. W; Marshall, G. R. *J Comput.-Aided Mol. Des.* **1993**, *7*, 623–647.
55. Lauri, G.; Bartlett, P. A. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 51–66.
56. Yang, Y. L.; Nesterenko, D. V.; Trump, R. P.; Yamaguchi, K.; Bartlett, P. A.; Drueckhammer, D. G. *J. Chem. Inf. Model.* **2005**, *45*, 1820–2005.
57. Rotstein, S. H.; Murko, M. A. *J. Med. Chem.* **1993**, *36*, 1700–1710.
58. Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 127–153.
59. Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. *J. Chem. Inf. Comput. Sci.* **2007**, *47*, 390–9.
60. Degen, J.; Rarey, M. *ChemMedChem* **2006**, *8*, 854–68.
61. Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. *J. Chem. Inf. Model.* **2007**, *47* (2), 390–399.
62. Ripka, A. S.; Satyshur, K. A.; Bohacek, R. S.; Rich, D. H. *Org. Lett.* **2001**, *3* (15), 2309–12.
63. Thompson, D. C.; Denny, R. A.; Nilakantan, R.; Humblet, C.; McCarthy, D. J.; Feyfant, E. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 761–772.
64. Durrant, J. D.; Amaro, R. E.; McCammon, J. A. *Chem. Biol. Drug Des.* **2009**, *73*, 168–178.
65. Pierce, A. C.; Rao, G.; Bemis, G. W. *J. Med. Chem.* **2004**, *47*, 2768–2775.
66. Ho, C. M. W; Marshall, G. R. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 623–647.
67. Erlanson, D. A.; Lam, J. W.; Wiesmann, C.; Luong, T. N.; Simmons, R. L.; DeLano, W. L.; Choong, I. C.; Burdett, M. T.; Flanagan, W. M.; Lee, D.; Gordon, E. M.; O'Brien, T. *Nat. Biotechnol.* **2003**, *21*, 308–314.
68. Kutchukain, P.; Lou, D.; Shakhnovick, E. I. *J. Chem. Inf. Model.* **2009**, *49*, 1630–1642.
69. Cancilla, M. T.; He, M. M.; Viswanathan, N.; Simmons, R. L.; Taylor, M.; Fung, A. D.; Cao, K.; Erlanson, D. A. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3978.

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

70. Wyatt, P. G.; Woodhead, A. J.; Berdini, V.; Boulstridge, J. A.; Carr, M. G.; Cross, D. M.; Davis, D. J.; Devine, L. A.; Early, T. R.; Feltell, R. E.; Lewis, E. J.; McMenamin, R. L.; Navarro, E. F.; O'Brien, M. A.; O'Reilly, M.; Reule, M.; Saxty, G.; Seavers, L. C.; Smith, D. M.; Squires, M. S.; Trewartha, G.; Walker, M. T.; Woolford, A. J. *J. Med. Chem.* **2008**, *51*, 4986–99.
71. Howard, S.; Berdini, V.; Boulstridge, J. A.; Carr, M. G.; Cross, D. M.; Curry, J.; Devine, L. A.; Early, T. R.; Fazal, L.; Gill, A. L.; Heathcote, M.; Maman, S.; Matthews, J. E.; McMenamin, R. L.; Navarro, E. F.; O'Brien, M. A.; O'Reilly, M.; Rees, D. C.; Reule, M.; Tisi, D.; Williams, G.; Vinković, M.; Wyatt, P. G. *J. Med. Chem.* **2009**, *52*, 379–388.
72. Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. *Proteins* **2003**, *52*, 137–145.

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch001

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

Chapter 2

Validation of Reaction Vectors for *de Novo* Design

Dimitar Hristozov,*,[1] Michael Bodkin,[1] Beining Chen,[2] Hina Patel,[3] and Valerie J. Gillet[3]

[1]Eli Lilly UK, Erl Wood Manor, Windlesham, Surrey GU20 6PH
[2]Department of Chemistry, University of Sheffield, Western Bank, Sheffield S10 2TN
[3]Department of Information Studies, Regent Court, 211 Portobello St., University of Sheffield, Western Bank, Sheffield S1 4DP
***E-mail: hristozov_dimitar_nonlilly@lilly.com**

A detailed validation of a new *de novo* design algorithm for the *in silico* generation of synthetically accessible compounds is presented. The algorithm is based on reaction vectors which describe the changes that take place at a reaction centre and which have been extracted from a knowledge-base of reactions. In the *de novo* design context, novel chemical compounds are generated by applying the reaction vectors to new starting materials. Here the algorithm is validated by attempting to reproduce a large number of diverse chemical reactions. On average, 90% of the reactions investigated (ranging from functional group interconversions to complex rearrangements) were successfully reproduced, thus showing the general applicability of the proposed algorithm.

Introduction

Virtual screening has become commonplace in drug discovery with computational techniques such as similarity searching and protein-ligand docking routinely used to predict the bioactive properties of molecules. However, these techniques are usually applied to databases of compounds which have already been synthesised which therefore limits the novelty that can be accessed. *De novo* design, on the other hand, refers to the design of previously unknown

© 2011 American Chemical Society
In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

compounds to fit a set of constraints, for example, to fit into the binding site of a target protein or to fit to a pharmacophore derived from known active compounds (*1*). *De novo* design is appealing since it provides a way of discovering novel compounds of therapeutic potential. However, chemical space is enormous (it has been estimated that up to 10^{60} compounds could exist with <30 atoms using the common elements C,N,O,S (*2*)) and the number of compounds that has already been synthesised represents a tiny fraction of this space (for example, there are around 48 million compounds in the CAS registry file (*3*)). Thus, there are vast areas of chemistry space that are currently uncharted (*4*). This clearly presents opportunities for the design of novel compounds, however, it also presents *de novo* design programs with significant challenges in how to navigate through this space to find useful compounds. The primary design constraints, such as fit to a protein binding site, clearly provide one way of focussing the search and, nowadays it is recognised that any compounds suggested for synthesis should fit multiple design objectives thus further restricting the search space (*5*), for example, predicted binding to a protein and acceptable ADMET properties (Adsorption, Distribution, Metabolism, Excretion and Toxicity). However, while multiple design constraints can be effective in directing the search towards compounds with promising characteristics, it is still the case that many of these theoretical compounds will not actually be synthetically accessible.

Although programs for *de novo* design first appeared around twenty years ago (*6*, *7*), a common failing of these early attempts was the lack of synthetic knowledge that was built into the methods, so that although compounds could be designed to fit the specified constraints, typically they were unappealing to chemists due to a lack of synthetic tractability. Thus, a recent focus in *de novo* design has been the incorporation of synthetic accessibility into the design process and several different approaches have been developed. These include the use of scoring methods to predict synthetic accessibility post structure generation (*8–10*), the use of fragment connection probabilities obtained from databases of molecules (*11*), and the use of chemical reaction transforms to restrict the structures that are generated (*12–14*). Many of the latter approaches are restricted to a small number of reactions which thus severely limits the structures that can be generated. Furthermore, they often require the use of atom-mapping techniques which involve mapping the atoms in the product to those in the reactant in order to identify the reaction centre and which are therefore computationally expensive to operate.

Reaction transforms have also been used retrosynthetically in Computer-Assisted Synthesis Design (CASD) systems where a synthetic route to a target molecule is suggested by applying retrosynthetic structural transformations (*15*). The early approaches were based on manual encoding of transformations by experts and took account of the conditions necessary for each reaction to occur (*16*), however, more recent approaches have attempted to automate the rule generation process in order to exploit reaction databases. In the automated approaches, the core of the reaction is usually identified using atom-mapping techniques and the core can then be extended to include the environment of the reaction, either based on distance to the reaction centre (*17*) or through the application of rules to identify relevant neighbouring atoms, as in the recent Route Designer method (*18*).

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

We have developed a reaction transform approach to *de novo* design that is based on the automatic extraction of the reaction centre and its environment into what are known as reaction vectors (*19*). The reaction vectors are then applied in the forwards synthetic direction to suggest novel molecules that could be made from a given starting material. The reaction vectors are based on atom-pair descriptors and are derived through a simple subtraction of the reactant descriptors from the product descriptors. Thus, they encode the atoms and bonds that are removed from the reactant(s) together with the new atoms and bonds required to form the product(s). The use of atom-pair descriptors also allows the environment of the reaction to be encoded based on distance to the reaction centre. Reaction vectors are very rapid to calculate since atom mapping information is not required and they can be calculated from virtually any database of organic reactions to form a knowledge-base for use in *de novo* design. Given such a knowledge-base of reaction vectors and a starting material, then our method selects one or more reaction vectors and applies them algorithmically to transform the starting material into potential new product molecules. Since the suggested transformations are based on known reactions a degree of confidence is provided on the synthetic accessibility of the virtual products.

Our method aims to make use of the vast number of organic reactions stored in different data sources – both commercial databases of reactions and in-house reaction data such as that encoded in electronic laboratory notebooks. The *de novo* design method has been developed using a modular approach to enable the knowledge-base to be easily extended to include new reactions and to allow it to be tailored to specific design scenarios, for example, to explore potential products that could be made from a given starting material and a given set of reagents using a specific reaction.

Here we focus on a detailed validation of the approach that is based on reproducing known reactions that cover a wide range of different reaction types in a wide variety of different environments. We first describe the reaction vectors themselves and give a brief overview of the structure generation algorithm in which a molecular transformation is applied to generate a product molecule: full details of the algorithms have already been provided in (*19*).

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

Reaction Vectors Overview

A reaction vector is generated automatically from a reaction and encodes the difference between the product(s) and the reactant(s) in vector form. Our work is based on reaction vectors as described by Broughton et al. (*20*) who developed them for assessing the similarity between reactions. Similar approaches to the representation of reactions were first suggested nearly 40 years ago (*21*, *22*) and recently reaction vectors have been used to describe the relationships between pairs of molecules with the aim of finding local QSAR models (*23*). Here we use the reaction vectors to encode molecular transformations present in reaction databases so that they can be applied to previously unseen molecules in order to generate novel product molecules.

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

The reaction vectors used here are a combination of atom-pairs at one and two bonds separation, respectively. We use the atom-pair notation introduced by Carhart et al. (*24*) (atom type-separation-atom type) in which separation indicates the number of atoms in the shortest bond-by-bond path that contains both atoms 1 and 2, so that atom-pair 2 (AP2) indicates two atoms which are bonded, and atom-pair 3 (AP3) indicates atoms at two bonds separation. The definitions of AP2 and AP3 are shown below where each atom is represented by element type *(X)*, number of non-hydrogen connections *(h)*, number of π electrons *(p)*, and number of ring memberships *(r)*. The bond order (1 = single bond, 2 = double bond, 3 = triple bond and 4=aromatic bond) is also included for AP2 descriptors. The combination of AP2 and AP3 was found to provide an effective balance between generalising the reactions to allow novel molecules to be generated, while including sufficient of the environment to maintain specificity.

$$\text{AP2: } X1(h,p,r)\text{-}2(BO)\text{-}X2(h,p,r) \qquad (1)$$

$$\text{AP3: } X1(h,p,r)\text{-}3\text{-}X2(h,p,r) \qquad (2)$$

A reactant is represented by an atom-pair vector in which the number of occurrences of each atom-pair is recorded; where there is more than one reactant the reactant vectors are summed. The same process is applied to the products to generate a product vector. A reaction vector, *RxnV*, is then generated by subtracting the reactant vector, *RctV*, from the product vector, *PrdV*:

$$RxnV = PrdV - RctV \qquad (3)$$

Atom-pairs that are unchanged by the reaction, that is, that occur with the same frequency in the reactant and product vectors, do not appear in the reaction vector. Atom-pairs with negative counts indicate atom-pairs that are removed from the reactant(s) and those with positive counts represent atom-pairs that are added to the reactant(s) to form the product(s). The AP2 descriptors in the reaction vector describe the bonds that are directly involved in the reaction (and the atoms incident on the bonds) and the AP3 descriptors extend the environment of the reaction to include atoms and bonds that are one bond away from the reaction centre. The reaction from which a reaction vector is derived is known as the parent reaction and there is a many-to-one relationship between parent reactions and reaction vectors, that is, several parent reactions may be represented by the same reaction vector. The reaction vector for a Beckmann rearrangement is shown in Figure 1. The AP2s are shown as shaded and the AP3s are unshaded. The AP2 descriptors indicate that there are three bonds broken (“lost”) and three bonds made (“gained”) in the course of the reaction. The AP3 descriptors indicate the environments in which the bonds occur.

Structure Generation Algorithm

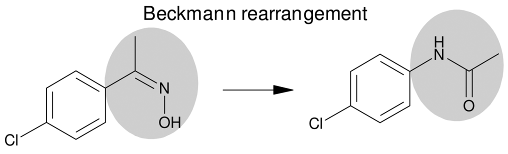
A structure generation algorithm has been developed which is able to generate a virtual product molecule(s) from a starting material and an appropriate reaction vector (i.e. one which describes bonds that are present in the starting material).

The algorithm is described in detail in (*19*) and is summarised in Figures 2 and 3 using the same Beckmann rearrangement reaction shown in Figure 1.

The first step, Figure 2, involves removing bonds from the starting material according to the negative AP2 descriptors in the reaction vector to form a starting fragment (the AP3 descriptors are used to ensure that the bonds removed occur in the correct environment). The positive AP2 descriptors in the reaction vector are then used to add bonds to the fragment(s) to generate a product molecule, Figure 3. The structure generation proceeds via a breadth-first search in which all possible bonds are added in all possible ways. The AP3 descriptors are used to prune the search tree by defining the wider environment of the newly created bonds.

The example shown in Figures 2 and 3 demonstrates how a reaction vector can be applied to the reactant of the reaction from which it was derived. In this work, we provide a comprehensive validation of the algorithm by attempting to reproduce the known product(s) for a large set of reaction vectors and their parent reactants. This is considered a necessary validation: if the algorithm is unable to reproduce the reactions used to construct its knowledge base then its utility in *de novo* design would be questionable. However, it should be born in mind that the real aim of our reaction vector method is to generate novel molecules by selecting and applying reaction vectors to previously unknown starting materials.

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002



Negative APs		Positive APs	
1-	C(3,2,1)-2(1)-C(3,1,0)	1+	C(3,2,1)-2(1)-N(2,0,0)
2-	C(3,1,0)-2(2)-N(2,1,0)	2+	C(3,1,0)-2(1)-N(2,0,0)
3-	N(2,1,0)-2(1)-O(1,0,0)	3+	C(3,1,0)-2(2)-O(1,1,0)
a-	C(3,2,1)-3-N(2,1,0)	a+	C(3,2,1)-3-C(3,1,0)
b-	C(3,2,1)-3-C(1,0,0)	b+	C(2,2,1)-3-N(2,0,0)
c-	C(3,1,0)-3-C(2,2,1)	c+	C(2,2,1)-3-N(2,0,0)
d-	C(3,1,0)-3-C(2,2,1)	d+	N(2,0,0)-3-C(1,1,0)
e-	C(3,1,0)-3-O(1,0,0)	e+	N(2,0,0)-3-O(1,1,0)
f-	N(2,1,0)-3-C(1,0,0)	f+	O(1,1,0)-3-C(1,0,0)

Figure 1. The reaction vector generated for a Beckmann rearrangement reaction.

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

Negative APs	
1-	C(3,2,1)-2(1)-C(3,1,0)
2-	C(3,1,0)-2(2)-N(2,1,0)
3-	N(2,1,0)-2(1)-O(1,0,0)
a-	C(3,2,1)-3-N(2,1,0)
b-	C(3,2,1)-3-C(1,0,0)
c-	C(3,1,0)-3-C(2,2,1)
d-	C(3,1,0)-3-C(2,2,1)
e-	C(3,1,0)-3-O(1,0,0)
f-	N(2,1,0)-3-C(1,0,0)

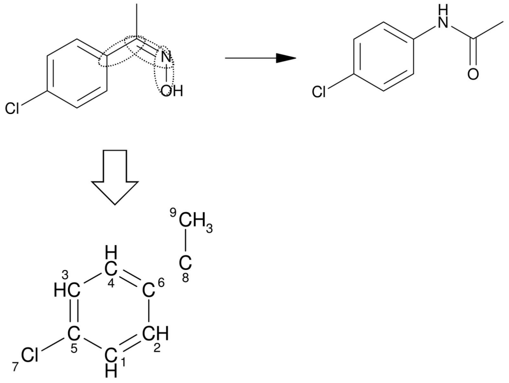


Figure 2. The first step in the reaction generation algorithm: removing bonds from the reactant.

Positive APs	
1+	C(3,2,1)-2(1)-N(2,0,0)
2+	C(3,1,0)-2(1)-N(2,0,0)
3+	C(3,1,0)-2(2)-O(1,1,0)
a+	C(3,2,1)-3-C(3,1,0)
b+	C(2,2,1)-3-N(2,0,0)
c+	C(2,2,1)-3-N(2,0,0)
d+	N(2,0,0)-3-C(1,1,0)
e+	N(2,0,0)-3-O(1,1,0)
f+	O(1,1,0)-3-C(1,0,0)

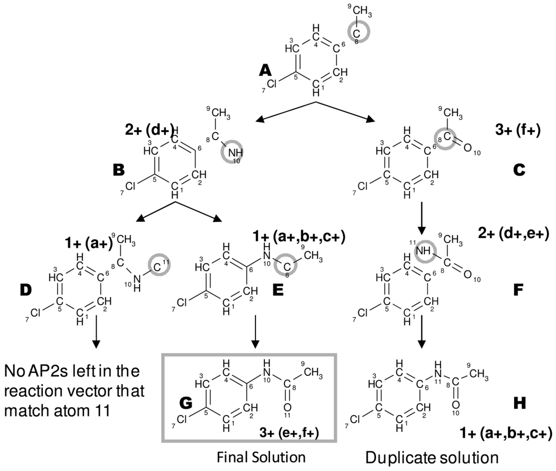


Figure 3. The second step in the reaction generation algorithm: Adding bonds to generate a product molecule.

Reproducing Reactions from the Knowledge Base

The reaction vector approach has been validated by attempting to reproduce the known product(s) from a reactant(s) and a reaction vector for a wide range of different reaction types. A set of 5,695 reactions covering the 28 different reaction types shown in Table I was extracted from the Lilly collection of commercially available databases. Many of the reactions (44%) were initially found to be incomplete, for example, there were missing fragments in the reactants or products; reagents were present which were not part of the reaction itself; they were not stoichiometrically balanced; or there was more than one product due to the presence of structural isomers. The reactions were therefore processed by a reaction cleaning algorithm to ensure that the same number of carbon atoms appeared on each side of the reaction (*19*). A small number of reactions were rejected including those which could not be cleaned with our algorithm (2%) and those consisting of more than two reactants or two products. The number of reactions remaining for each reaction type is shown in Table I. Reaction vectors were calculated for each reaction and stored. The validation procedure then consisted of extracting each reaction in turn, retrieving the corresponding reaction vector and applying it to generate a product molecule. The generated product was then compared with the known product of the parent reaction, according to the scheme shown in Figure 4. A time-out of 30 seconds was applied and reactions that exceeded the limit were reported as failed.

In 75% of cases, the structure generation algorithm took less than 0.05 seconds to run, with the median run time 0.015 seconds per reaction. The results are summarised in Table I as the number and percentage of reactions in each class that were successfully reproduced. Figure 5 shows the success rates for different reaction types. For 11 of the reaction classes the success rates were 100% (see Figure 6 for examples of these reaction types) and for 20 of the classes the success rates were higher than 90% (Figure 7 shows examples from the nine reaction classes with success rate >90% and <100%). Taken together, these represent a range of different reaction types in which the reactions occur in a wide variety of environments. They vary from straightforward functional group interconversions through to more complex rearrangements. They include reactions consisting of the conversion of a single reactant to a single product through to reactions consisting of two reactants and two products in which both products are successfully reproduced.

For seven of the reaction classes, the success rates were less than 90% but higher than 60%. The worst case was encountered when trying to reproduce a number of Fischer indole synthesis reactions. However, 40% of the reactions could still be reproduced successfully. Sample reactions which were successfully reproduced from these eight classes are shown in Figure 8.

The majority of the failures (470 reactions which represent ~8.3% of the complete set) were due to reaching the time-out limit of 30 seconds. In many cases, these are reactions which involve the formation of complex ring systems. The algorithm reaches the time-out due to there being too many possibilities for both the removal of bonds from the starting material and the addition of bonds to the resulting fragment. An example of some of these cases is shown in Figure 9. The effect of increasing the time-out was investigated by doubling the allowed time from 30 to 60 seconds. Solutions were then generated for approximately 4.5% (21 of 470) of the previously failed reactions. However, the increase in success rate was relatively low.

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

Table I. Success rates in reproducing different types of organic reactions

Reaction Type	*Number of Reactions*	*Correctly Reproduced*	
		Number	*Percent*
Epoxide reduction	450	449	99.8
Epoxide formation	450	444	98.7
Ester to amide	172	172	100.0
Alcohol dehydration	171	169	98.8
Claisen rearrangement	61	54	88.5
Beckmann rearrangement	123	123	100.0
Friedyl Crafts acylation	113	113	100.0
Olefin metathesis	9	7	77.8
Dieckmann condensation	98	91	92.9
Nitro reduction	231	230	99.6
Alkene oxidation	272	272	100.0
Cope rearrangement	453	306	67.5
Aldol condensation	134	134	100.0
Alcohol amination	97	97	100.0
Amide reduction	51	51	100.0
Diels-Alder hetero	441	320	72.6
Ether halogenation	58	58	100.0
Ozonolysis	132	125	94.7
Claisen condensation	98	77	78.6
Carboxylic acids to aldehydes	194	194	100.0
Nitrile reduction	102	102	100.0
Diels-Alder cycloaddition	106	65	61.3
Fischer indole	230	94	40.9
Alkene halogenation	310	281	90.6
Nitrile hyrdrolysis	460	460	100.0
Olefination	455	427	93.8
Wittig-Horner	211	190	90.0

Continued on next page.

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

Table I. (Continued). Success rates in reproducing different types of organic reactions

Reaction Type	*Number of Reactions*	*Correctly Reproduced*	
		Number	*Percent*
Robinson annulation	13	10	76.9
Total	5,695	5,115	89.8

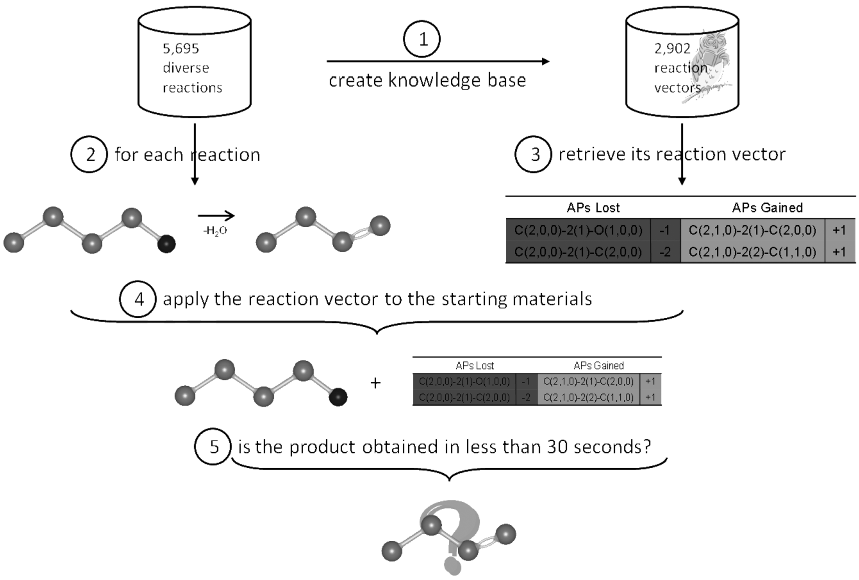

5,695 diverse reactions
① create knowledge base
2,902 reaction vectors
② for each reaction
-H₂O
③ retrieve its reaction vector
APs Lost
C(2,0,0)-2(1)-O(1,0,0) -1
C(2,0,0)-2(1)-C(2,0,0) -2
APs Gained
C(2,1,0)-2(1)-C(2,0,0) +1
C(2,1,0)-2(2)-C(1,1,0) +1
④ apply the reaction vector to the starting materials
+
⑤ is the product obtained in less than 30 seconds?

Figure 4. The validation procedure.

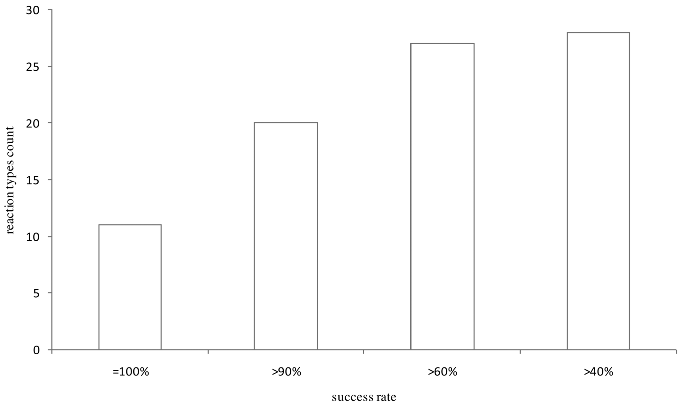


Figure 5. Cumulative reaction types count per success rate.

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

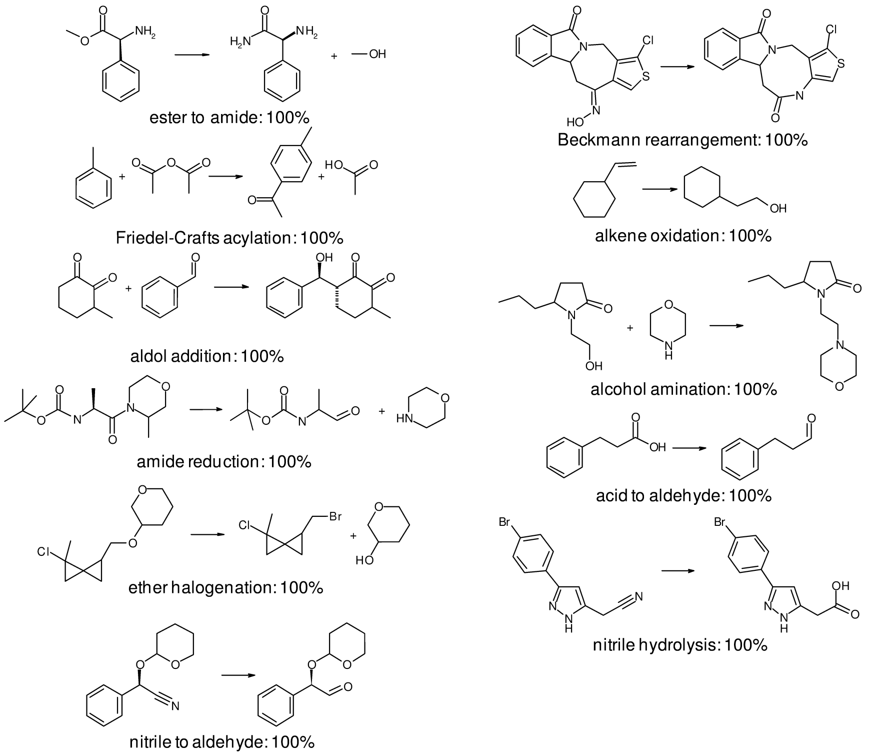

ester to amide: 100%
Friedel-Crafts acylation: 100%
aldol addition: 100%
amide reduction: 100%
ether halogenation: 100%
nitrile to aldehyde: 100%
Beckmann rearrangement: 100%
alkene oxidation: 100%
alcohol amination: 100%
acid to aldehyde: 100%
nitrile hydrolysis: 100%

Figure 6. Example reactions from the 11 classes with 100% success rate.

In a small number of cases (110 reactions, or ~2%, at the 30s time-out), the algorithm terminates before reaching the time-out without generating a solution. There are two main reasons why this may happen. First, the reconstruction tree (illustrated in Figure 3) can often lead to duplicate states through the addition of bonds in different orders. To speed up the algorithm an attempt is made to avoid exploring the same state two or more times. A "state" in this context refers not only to the structural fragment being explored, but also to the atom-pairs left for addition and the attachment points, i.e. atoms, in the fragment. A hash function is used to compare states in an efficient manner and transforms each state into a single integer number, called a hash code. The hash function is designed to assign different hash codes to different states, however, there is a small possibility of collisions, that is, two different states being assigned the same hash code. Since the hash codes are used to declare two states identical, only one of these states will be examined and it could be that the correct branch in the reconstruction tree is pruned. While the probability of hash code collisions is low, the second reason occurs more frequently and underlines one of the limitations of the reaction vectors approach. Remember that a reaction vector is given by the difference between the reactant vector and the product vector (eq 3). In the majority of cases, eq 3 results in a vector of atom-pair counts which reflects the bond breaking and making in the underlying reaction, as desired. However, in some reactions, such as the one exemplified in Figure 10, it is possible that the new bonds that are made in the product are represented by the same atom-pairs that occur at the reaction centre in the starting material. In such cases, these pairs cancel each other out and are not represented in the reaction vector. This then leads to either incorrect fragmentation of the starting material or insufficient atom-pairs for reconstruction of the product. This situation has been seen in complex rearrangement reactions and often indicates that the parent chemical reaction is either a tandem reaction (as in the example shown in Figure 10 which involves the formation of a new cycle in addition to epoxide reduction) or is a reaction in which several steps have been condensed into a single reaction. Failed reactions of such complexity do not cause undue concern for *de novo* design since it is likely that they would not be high on a chemists' list of preferred reactions. Nevertheless, we are currently investigating extensions to the basic algorithm that would permit these reactions to be applied *in silico* so that they could be included in a knowledge-base of reactions if so desired.

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

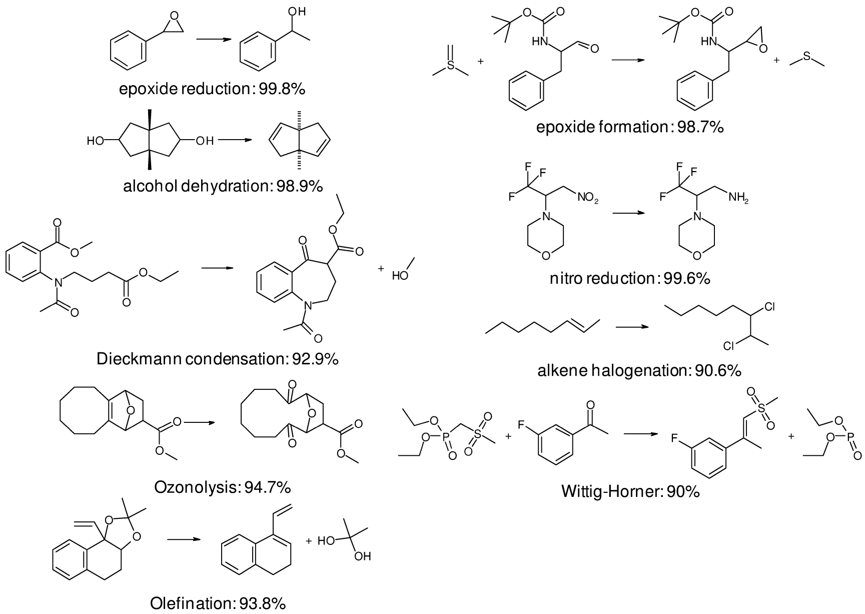


Figure 7. Example reactions from classes where success rate is >90% and <100%.

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

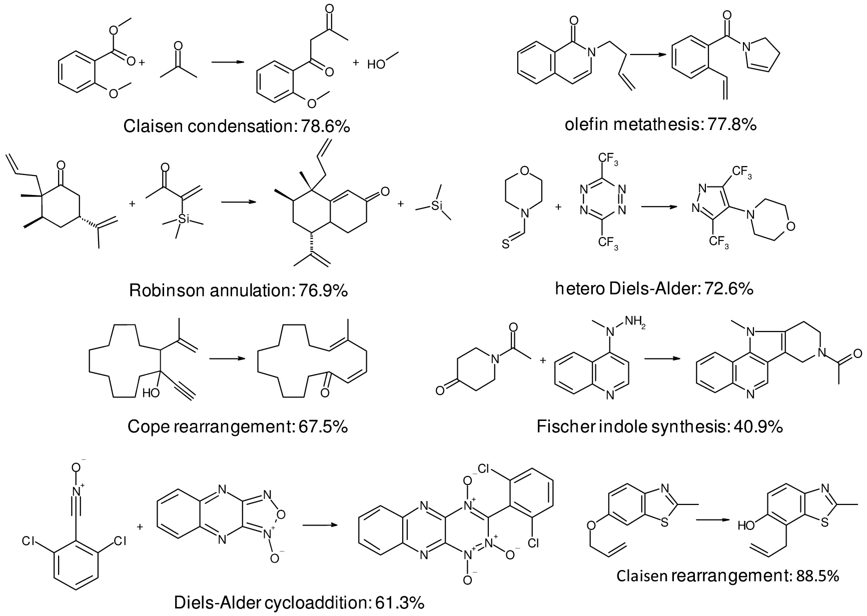

Claisen condensation: 78.6%
olefin metathesis: 77.8%
Robinson annulation: 76.9%
hetero Diels-Alder: 72.6%
Cope rearrangement: 67.5%
Fischer indole synthesis: 40.9%
Claisen rearrangement: 88.5%
Diels-Alder cycloaddition: 61.3%

Figure 8. Examples of reactions from reaction classes with <90% success.

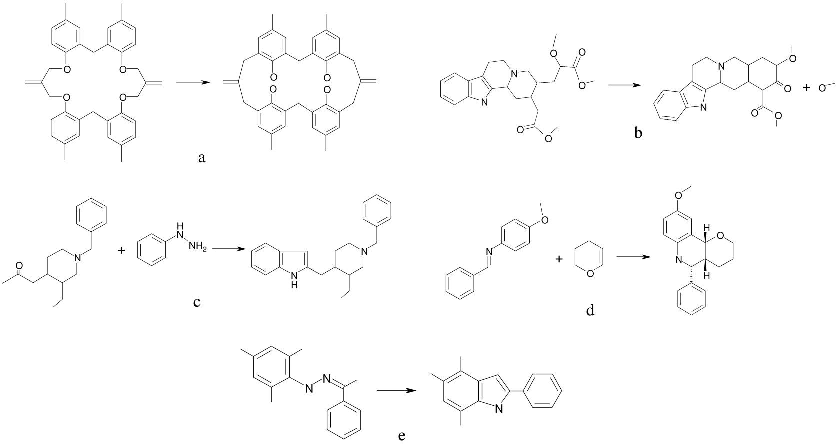


Figure 9. Examples of reactions which were not reproduced within the 30 seconds timeout. a) Claisen rearrangement; b) Dieckmann condensation; c) Cope rearrangement; d) hetero Diels-Alder; e) Fischer indole synthesis.

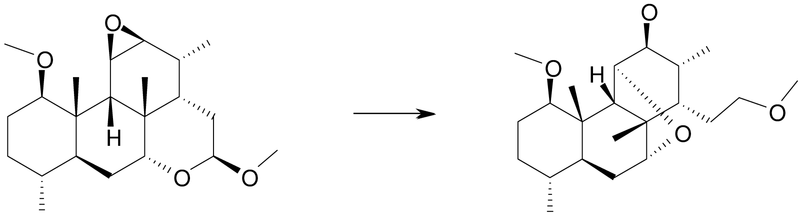


Figure 10. A tandem reaction in which the new bonds formed in the product generate atom pairs already present in the starting material.

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

Conclusions

The reaction vector approach has been validated by reproducing a large number of reactions in a variety of different environments. The average success rate was around 90% and for many of the reaction types the success rate was in excess of 90%. An analysis of the failed reactions has revealed some deficiencies in the algorithm especially for reactions involving complex ring systems and we are currently investigating extensions to the approach to increase the variety of reactions that can be handled.

While the validation reported here has been based on reproducing known reactions, the intended use of the algorithm is to apply molecular transformations to previously unknown starting materials in order to generate novel products *in silico*. We have previously introduced a desktop tool which achieves this goal and which can be run in a variety of different modes (*19*): for example, to explore possible compounds that could be made from a given lead compound in a lead optimisation experiment. The next step in the development of the method will be the incorporation of the structure generation algorithm into an iterative *de novo* design tool that is capable of suggesting multi-step syntheses and which will be driven by multiple design constraints.

Acknowledgments

We gratefully acknowledge Howard Broughton for many helpful discussions throughout this work. We thank the Engineering and Physical Sciences Research Council and Eli Lilly for funding, and the Royal Society and the Wolfson Foundation for laboratory support. Programmes were written using JOELib (University of Tübingen, Germany), Marvin Beans (ChemAxon) and KNIME (University of Konstanz, Germany) and we gratefully acknowledge the provision of software by these organisations and the support provided.

References

1. Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649–663.
2. Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.
3. CAS Registry File, Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210. http://www.cas.org (accessed July 28, 2009).
4. Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R. Heteroaromatic rings of the future. *J. Med. Chem.* **2009**, *52*, 2952–2963.
5. Ekins, S.; Boulanger, B.; Swaan, P. W.; Hupcey, M. A. Z. Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 381–401.
6. Lewis, R. A.; Leach, A. R. Current methods for site-directed structure generation. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 467–475.
7. Gillet, V. J.; Johnson, A. P. Structure Generation for De Novo Design. In *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*; Martin, Y. C., Willett, P., Eds.; American Chemical Society: Washington, 1998; pp 149−174.
8. Boda, K.; Johnson, A. P. Molecular complexity analysis of de novo designed ligands. *J. Med. Chem.* **2006**, *49*, 5869–5879.
9. Boda, K.; Seidel, T.; Gasteiger, J. Structure and reaction based evaluation of synthetic accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 311–325.
10. Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.
11. Kutchukian, P. S.; Lou, D.; Shakhnovich, E. I. FOG: Fragment optimized growth algorithm for the de novo generation of molecules occupying druglike chemical space. *J. Chem. Inf. Model.* **2009**, *49*, 1630–1642.
12. Fechner, U.; Schneider, G. Flux (1): A virtual synthesis scheme for fragment-based de novo design. *J. Chem. Inf. Model.* **2006**, *46*, 699–707.
13. Schürer, S. C.; Tyagi, P.; Muskal, S. A. Prospective exploration of synthetically feasible, medicinally relevant chemical space. *J. Chem. Inf. Model.* **2005**, *45*, 239–248.
14. Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. J. SYNOPSIS: SYNthesize and OPtimize system in silico. *J. Med. Chem.* **2003**, *46*, 2765–2773.
15. Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247–266.
16. Corey, E. J. *The Logic of Chemical Synthesis*; John Wiley & Sons. Inc: New York, 1995.
17. Satoh, K.; Funatsu, K. A novel approach to retrosynthetic analysis using knowledge bases derived from reaction databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 316–325.

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

18. Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route designer: A retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Model.* **2009**, *49*, 593–602.
19. Patel, H.; Bodkin, M. J.; Chen, B. N.; Gillet, V. J. Knowledge-based approach to de novo design using reaction vectors. *J. Chem. Inf. Model.* **2009**, *49*, 1163–1184.
20. Broughton, H. B.; Hunt, P. A.; MacKey, M. D. Methods for Classifying and Searching Chemical Reactions. U.S. Patent Application 2003/0182094 A1, 2003.
21. Harrison, J. M.; Lynch, M. F. Computer analysis of chemical reactions for storage and retrieval. *J. Chem. Soc. C* **1970**, 2082–2087.
22. Ugi, I.; Gillespi, P. Chemistry and logical structures. 3. Representation of chemical systems and interconversions by BE matrices and their transformation properties. *Angew. Chem., Int. Ed.* **1971**, *10*, 914–915.
23. Sheridan, R. P.; Hunt, P.; Culberson, J. C. Molecular transformations as a way of finding and exploiting consistent local QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 180–192.
24. Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular-features in structure activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

Downloaded by UNIV OF GUELPH LIBRARY on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch002

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

Chapter 3

Design and Application of Fragment Libraries for Protein Crystallography

Computational Approaches to Compound Selection

John Badger*

DeltaG Technologies, 4360 Benhurst Ave., San Diego, CA 92122, USA
***E-mail: info1.dgtech@gmail.com**

The x-ray diffraction analysis of protein crystals soaked in libraries of fragment compounds is able to identify those small compounds that specifically bind to critical sites on the protein. This crystal structure data may be used in the subsequent design of focused scaffold libraries for early lead discovery. By applying simple computational tools to search through the several million off-the-shelf screening compounds currently available it is possible to implement fragment screening methodologies in academic and small biotechnology laboratory environments.

Background

Fragment-based screening for lead discovery (FBLD) is now an established methodology with a proven record of success. Hajduk and Greer have reviewed a decade of positive results from early leads to the clinic, listing 47 compounds at significant development stages, including four compounds in clinical trials and two compounds that were developed within two years of project inception (*1*). Influential theoretical work that has encouraged protein screening with low molecular weight compounds emphasizes the *ligand efficiency*, the binding energy per non-hydrogen atom, rather than the total binding affinity for deciding which compound hits from the initial screen are most likely to evolve into clinical candidates (*2*). Specifically, relatively low affinity compounds may be considered useful for lead development if the ligand efficiency exceeds 0.3 and this factor favors lower molecular weight compounds for a given binding affinity.

© 2011 American Chemical Society
In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

In a separate development, feature analysis modeling of the probability of a compound binding to a specific site on a protein indicates an exponentially falling probability with increasing compound complexity (*3*). This result suggests that screening libraries containing relatively small and simple compounds will yield higher hit rates than libraries containing larger, more complex molecules although the reduced number of interaction points will mean that the binding affinities will often be relatively weak. As a practical matter, it is considered simpler to manage a compound's chemical properties through the lead development process by growing it from a small starting compound than by attempting to modify a compound which is already approaching its maximum tolerable size for use as an orally deliverable drug.

The development of FBLD may have a greater practical impact than just a re-interpretation of existing conventional high throughput screening technologies (HTS) towards the inclusion of lower molecular weight compounds in screening libraries. Screening with small fragment libraries that contain 100-1000's of compounds rather than the traditionally large HTS libraries containing ~1,000,000 compounds drastically reduces the cost and data management complexity of the initial screen; fragment screening is a practical methodology that can be applied in small pilot programs within academic and small biotechnology environments.

Some specific therapeutic areas may be particularly well suited to the application of FBLD. Fragment screening appears to be an appropriate methodology for CNS drug discovery and development because the chemical properties that typify fragment compounds are comparable to the compound properties required for passive transport across the blood-brain barrier (BBB). CNS compounds capable of passive BBB transport are small molecules with low molecular weights (<450Da and a mean value of ~350Da), that contain limited numbers of hydrogen bonds (8-10 may be an upper limit), a relatively low polar surface area (usually <60Å^2) and an optimal solubility (logP) of ~2 (*4*). It has been argued that an expansion of use of HTS screening methodologies in the pharmaceutical industry, in which the screening libraries typically contain relatively large compounds on the edge of acceptability for BBB permeability, has tended to decelerate the development of for drugs for CNS diseases (*4*).

Other rationales for performing fragment-based screening include the identification of compound binding motifs on novel target proteins for which there is little prior knowledge and to search for novel compound on well-established and heavily studied targets. In some cases fragment screening has been applied as a 'last shot' for the discovery of lead compounds after other approaches have failed.

Concepts

Crystallographic Screening and Fragment Library Parameters

At first glance, x-ray crystallography would seem to be an unappealing approach for initial target screening since it requires the availability of many stable well-diffracting protein crystals and the data collection requirements are both technology and resource intensive. The crystallographic screening

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

experiment involves soaking previously grown protein crystals in solutions of fragment compounds, mounting and freezing the soaked crystals and then performing data collection on the individual crystals at a suitable x-ray source. Despite considerable advancements in robotics for automated crystal handling these operations unavoidably require significant quantities of pure, well-folded protein (to grow the crystals) and careful manual manipulation to transfer and mount individual crystals prior to data collection.

However, in the larger picture of developing a fragment-based lead discovery project with a successful outcome, some of these issues are not as specifically limiting to the selection of crystallography as the screening technique as they appear. Regardless of the technique used for the initial screen, the pace and chances of success of a FBLD project are greatly enhanced by the production of high quality protein samples and the availability of accurate three-dimensional structure data on protein:compound complexes. It could be argued that many of the pre-requisites for crystallographic screening simply front-load otherwise desirable experimental factors as technical necessities for launching a project.

X-ray crystallographic data collection no-longer requires an investment in in-house data collection equipment but can be carried out on an *ad hoc* basis at synchrotron radiation facilities. Many biotechnology companies have abandoned support for in-house equipment and now conduct data collection operations through outsourcing or use of their own staff at synchrotron facilities. The methodology requires expertise (a protein crystallographer) and laboratory resources for protein and crystal preparation but only at the ‘small science’ level already found in many academic laboratories and biotechnology companies.

The number of compounds that may be screened crystallographically is limited by available resources for protein and crystal production and many laboratories might consider that the preparation and data collection on ~100 crystals to be a reasonable level of effort for the first screening attempt on a protein. At a third generation synchrotron source equipped with instrumentation including robotic crystal handling apparatus, the collection of 100 complete data sets on 100 reasonably well-diffracting crystals might require ~50 real-time hours of synchrotron beam time.

In order to increase the number of compounds that can be screened in a single experiment it is usual to soak protein crystals in cocktails that containing 3-10 different compounds each. The compounds in each cocktail are typically chosen so as to appear shape diverse in x-rays in order to give the best chance for an unambiguous identification of the bound compound in the resulting electron density map. In assigning compounds to shape diverse groups, the important fact is that the scattering of x-rays from an atom is proportional to atomic number. For this reason the scattering from the hydrogen atoms is usually too weak to distinguish them above the noise level in the final image of the structure, the scattering from all of most common elements in biological samples (C, O, N) is rather similar, which makes them difficult to distinguish from each other, but any heavier atoms can provide significant and distinctive markers in the compounds that contain them. Consequently, the shape diverse mixtures used in these experiments typically consist of differently sized cyclic compounds and mix small and large side groups. After making use of compound sampling in mixtures, the

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

feasible size for a fragment library for x-ray crystallographic screening is usually in the range of 300-2000 individual compounds.

Obviously, a crystallographic fragment screening library is restricted to contain far fewer compounds than is typically tested using other screening methods. Compensation for this limited practical library size is obtained by exploiting the sensitivity of crystallographic screening. Recalling the discussion of molecular complexity (*3*) it is possible to survey a great deal of chemical space with only ~1000 compounds provided that the screening compounds are sufficiently small. Since crystallographic screening allows for the detection of weakly binding compounds with binding affinities in the low millimolar regime, the technique is applicable to the detection of very low molecular weight fragments and hit rates from crystallographic screening campaigns with small fragments are typically in the 1-5% range. Some published examples of fragment libraries that have been specifically designed and successfully used for protein crystallographic studies include the 'Drug Fragment Set' described in pioneering work from the Astex group (327 compounds, mainly single cyclic groups with most molecular weights 100-250Da, ref (*5*)), a library used by a group at the University of Washington for inhibitor design studies on the *Trypanosoma brucei* Nucleoside 2-Deoxyribosyltransferase (304 small compounds, in which all hits are single core cyclic compounds, ref (*5*)) and the 'Fragments of Life' library from deCODE Biostructures (1329 compounds with mean molecular weight 182.5Da, ref (*7*)). In contrast, many of the fragment collections and subsets available from commercial vendors contain far larger numbers of compounds and most of the compounds they contain have molecular weights above 200Da. These fragment collections require further filtering in order to create appropriate libraries for crystallographic screening experiments. Exceptions are the fragment libraries from Zenobia Therapeutics (352 compounds with mean molecular weight of 155Da) and Maybridge (1000 compounds with a mean molecular weight of 178Da)

Exploiting Information from Crystallographic Fragment Screening Data

The analysis of crystal diffraction data from a fragment screening experiment typically involves placement of a previously solved model of the protein structure in the crystal cell followed by automated refinement of this model against the data and output of an electron density map for inspection. Electron density features in the map that are not accounted for by the protein or bound water molecules may indicate a bound fragment compound. The fragment is subsequently fitted to match the shape of the extra density and refined to provide an accurate model of the fragment position and conformation on the protein. Technical details for the crystallographic structure solution processes have been given in individual publications (for example, ref (*6*)) but the essential point is that this structure information indicates that the compound is usefully bound (*i.e.* in a critical functional site) and shows which atoms interact with the protein. Simple visual inspection of the structure data for the bound fragment in the context of the protein shows which atomic sites are available for substitution and indicates the appropriate sizes and chemical types for substitution. This structure data also invites the application of more sophisticated computational docking approaches to help design follow-up screening libraries.

The initial fragment structure data may also be used to design follow-on libraries which need not contain only those cores identified from the hits in the initial screen but may also contain other compounds that contain the same binding motifs (*i.e.* the very small groups of ~3-5 atoms responsible for specific interactions with the protein). The crystallographic data encourages computational approaches to 'scaffold hopping' at the very earliest stage of analysis.

Although the three-dimensional structure information allows exploitation of structure constraints in the design of efficient follow-up libraries, information on binding constants must be obtained using another biophysical technique, usually on subsequent generations of more potent compounds. One key point in the development of lead compounds, emphasized by the Astex group, is that the binding affinity of follow-up compounds should be carefully monitored and matched to the compound size so that a relatively high ligand efficiency is maintained. Although structure modeling methods (see below) may be usefully employed in the selection of follow-up compounds, a robust development process also involves evaluating their true binding modes through the determination of representative cocrystal structures.

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

Crystallographic Fragment Screening without Crystals

The experimental data from a crystallographic fragment screening experiments may be interpreted as determining both specific substructures that bind to the protein and their associated binding motifs. From this perspective other structural data, including known protein structures that contain larger natural and synthetic ligands, might also be examined in order to identify critical binding interactions and substructures. This type of compound dissection may be used to create 'hypothetical fragment hits' that inspire the design of focused fragment or scaffold libraries that incorporate these binding interactions or substructures. While not as reliable as true fragment hits, this approach is distinct from virtual screening by computational docking. These hypothetical fragments are sterically allowable and do make experimentally demonstrated interactions with the protein.

Fragment compounds are relatively unconstrained and are able to make relatively optimal interactions with the protein. Although the position of a fragment that is a substructure of a larger compound may be somewhat compromised, a well chosen subgroup should still provide a significant binding energy. Obviously, the compound dissection approach limits the novelty of suggested screening compounds to some extent but by focusing on deriving the key interaction motifs these libraries may still contain novel chemotypes, outside the initial compound data.

In some cases experimental fragment screening has tended to recapitulate and confirm information that was embedded in structure data involving larger ligands. For example, several series of hsp90 inhibitors contain resorcinol or adenine substructures and these compounds may be found as fragment hits in crystallographic fragment screens (*8*) as well as within previous structures containing larger ligands.

The Rule-of-Three (Ro3) and Crystallographic Screening Libraries

Fragment libraries are typically designed so that the compounds they contain lie within an appropriate range of chemical properties *and* are appropriate chemical types for lead development. Computational filtering, and analysis tools may be used to select compounds for these libraries from much larger collections. Further 'expert' input on a final fragment library design might be provided by assessing the resulting compounds for toxicities, instabilities and synthetic utility.

General libraries of screening compounds are typically constrained so that their chemical properties are 'drug-like' and Lipinski's rule-of-five (Ro5) is often invoked as a selection criteria. In a self-conscious echo of the Ro5, a rule-of-three (Ro3) that is appropriate guiding the selection of smaller molecules used in fragment screening has been proposed (*9*). The limits on chemical properties expressed by the Ro3 are that: (i) the compound molecular weights should be < 300Da, (ii) the number of hydrogen bond donors should be ≤ 3, (iii) the number of hydrogen bond acceptors should be ≤3. Additionally, (iv) the number of rotatable bonds should be ≤3 and, (vi) the polar surface area should be ≤60Å^2 . The Ro3 is intended to impose some useful, sensible bounds on compounds within a fragment library and, like the Ro5, builds in potential for drug development while still at the screening phase of a project. Somewhat less clear is the extent to which Ro3 is related to the hit rate since a recent study (*8*), with a library of molecules selected from chemical types suitable for chemical development, shows very little dependence of hit rate on the values of individual compound properties. Commercial vendors of fragment screening libraries usually bolster the acceptability of their collections by stating compliance with Ro3.

From the specific perspective of a fragment library design for crystallographic screening an upper bound of molecular weight of 300Da is excessively permissive because it allows many complex, multi-core compounds to enter the screening collection. The arguments linking molecular complexity to library size imply that most compounds in a small crystallographic screening library should be of the single core type (simple decorated ring systems), with a limited number of biaryl compounds, and this requirement implies that typical molecular weights for appropriate compounds will be <200Da.

A subtle issue with rigid adherence to the Ro3 in fragment library designs is that some useful drug-like cores are likely to be excluded or under-represented. The chemical diversity of a library might be enriched by allowing some limited exceptions. Molecules that include ring systems containing several nitrogen atoms (for example, purines and pteridines) are potentially excluded on the grounds that they contain excessive numbers of hydrogen bond acceptors and/or too large a polar surface area. (A further complication, particularly in defining some nitrogen atom types, is that different computational tools provide potentially differing results when calculating the compound properties). A specific example illustrating the dangers in an over-strict adherence to specific filters is the fragment-based ligand design for hsp90, where crystallographic fragment hits

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

may be obtained with resorcinol and adenine and lead series have been based on compounds including these substructures (*8*). Resorcinol might be considered by many to be too small for inclusion in a library (MW=110Da) and, depending on the property calculation algorithm, adenine might be considered to contain too many hydrogen bonds and contain an excessive polar surface area.

One additional consideration for libraries that are intended for use in crystallographic screening is that quoted solubilities are calculated properties and a relatively high solubility is required in the experimental setting. After allowing for a dilution factor that arises from screening in cocktails and assuming that the crystals are stable in up to ~10% DMSO it is necessary that the individual compounds are soluble at concentrations of ~200mM in DMSO in order to achieve concentrations of ~5mM of each compound within the protein crystal. *i.e.* so that compounds with affinities that bind with affinities in the low mM range may be detected. A solubility of ~200mM in DMSO is an order of magnitude higher than the 10-20mM solutions often provided for screening by chemical assay. In this author's experience ~90% of compounds selected to comply with the Ro3 prove to be sufficiently soluble for crystallographic work.

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

Implementation

Design of Project-Purposed Screening Libraries

Appropriate software tools may be used to develop screening libraries tailored to the scientific and logistical needs of specific projects. Practical work on fragment and scaffold library design has been carried out by this author using the *SDsearch* software (with some calculations performed via *OpenBabel2*) but analogous calculations could be carried out using chemical database software that provides tools for filtering on chemical properties or by using commercially available filtering software.

SDsearch is able to apply standard chemical property filters of the types embodied in the Ro3 using vendor annotations or by calculation. In addition, *SDsearch* is able to filter compounds according to substructures defined by SMILES strings and according to small motifs designated by three-dimensional distance restraints. A convenient analysis capability is that *SDsearch* may accept a target list of SMILES strings and classify the compounds in a potential library according to the matching of substructures in this list. By placing more complex and specific definitions at the top of the list, and making the search order dependent so that assigned compounds are not reused, problems of multiple matches, where, for example, a naphthalene core type would also be classed amongst benzene types are avoided. The software outputs a file of unclassified molecules so that these can be examined and those substructures added to the list of SMILES definitions in order to develop a complete coverage over possible cyclic structures within the libraries.

Tabulations of compounds into clusters representing the available cyclic substructures provide convenient information for designing focused libraries, aimed at a specific protein target and these libraries will usually be quite small. However, when designing a general screening library with compounds from multiple chemical vendors a practical issue is how best to reduce the number of available compounds to a usable value while maintaining chemical diversity. A possible approach is to maintain the full diversity of chemotypes but reduce the number of compounds in the more heavily populated core groups. It seems to this author that methods involving the elimination of the closely related similar compounds, leaving just representative examples, are predicated on the assumption that chemically related compounds will all produce hits, albeit with some variations in affinity. This assumption is less likely to be true for screening with small fragment compounds, where shifting a single side group atom has a relatively high chance of impacting a critical binding surface. Simply refiltering overpopulated core groups to preserve the compounds with the best chemical properties for subsequent exploitation (for example, using a more restrictive range of solubilities) might provide a more pragmatic solution.

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

Available Screening Compounds

Quite extensive collections of compounds for protein screening are currently available from commercial vendors. The SD files supplied by just four of the larger vendors (Chembridge, Maybridge, Enamine, Life Chemicals) list ~2 million available compounds from which to design fragment libraries. However, the number of available compounds sharply decreases at the low molecular weight end of the range and is relatively limited for MW < 200Da. At the time of writing (Summer, 2009) the screening collections from these four vendors are able to provide 1230, 1364, 1119 and 558 compounds respectively that meet the Ro3 property criteria and have molecular weights restricted to <200Da. Almost all of these compounds contain cyclic core structures and the sets of individual fragment compounds in these collections are almost completely disjoint; for no pair of vendors is there more than a 6% overlap of individual compounds. From this perspective there are over 4000 unique and available low molecular weight compounds from which to design a fragment library for crystallographic work.

The fragments in the the fragment screening concept are usually taken to be fragments of known drugs or chemical substructures that might be considered suitable components of drugs (*5*). A suitable fragment library should contain a diverse and balanced collection of appropriate fragment cores. A direct approach to analyzing and controlling the available chemical diversity in a fragment library is to simply enumerate cyclic core types within the putative fragment compound subsets drawn from these collections. This analysis shows where vendor collections differ in available chemotypes and where it might be useful to merge sets of fragment compounds obtained from multiple vendors. These searches identify missing or under populated core types, including types which could be expanded if the Ro3 was relaxed.

In total, there appear to be ~240 different cyclic cores covered by the four vendors, Chembridge, Maybridge, Enamine, Life Chemicals for Ro3 compliant subsets with MW <200Da. A cyclic core is defined here as a single or fully conjugated ring system that is unique in terms of chemical composition and atom hybridization state. In particular, nitrogen atoms within cyclic systems are differentiated as to whether they lack any substituents, they are hydrogenated or they are connected to a larger R-group. The greatest numbers of different cyclic cores are found in the Chembridge and Maybridge collections with each collection covering ~2/3 of the total number of core classes and, when combined, covering almost all cores classes. In most cases the numbers of compounds assigned to each cyclic core type is quite small and is variably distributed across vendors - there are many singlet classes and over 3/4 of classes contain fewer than ten compounds. Benzene cores represent an outstandingly large class of compounds that contains between ~1/4 and ~1/3 of all fragment compounds, depending on vendor.

Structure-Based Exploitation of Fragment hits

A simple and rapid targeted docking approach that creates small enriched scaffold libraries as follow-up to hits from the initial fragment screen may prove useful. The compounds in these enriched scaffold libraries incorporate the substructures identified from crystallographic fragment screens (or computationally, as hypothetical fragments) provided that the compound may achieve a docked binding mode that is compatible with the steric restraints of the protein structure.

The input for the docking screening process is a set of compounds that has been prefiltered to appropriate chemical properties and which contain a specific binding substructure. The Ro3 constraints with a size limit of 275Da in order to allow predominance of biaryl compound types has been used in some experiments at Zenobia Therapeutics.

A pipeline process takes each compound in turn from the prefiltered compound set and (i) generates a 3D structure, (ii) explodes the 3D structure in torsion space to span all available conformations, (iii) superimposes each pose into the target site by matching atoms from the binding motif onto the equivalent set of atoms in the experimentally determined protein-ligand structure, (iv) evaluates each pose as to whether the compound atoms clash with the surrounding protein and scores it according to a simple contact potential to obtain the most probable binding mode.

If all poses of a compound clash with the protein then the compound is considered not feasible and may be rejected from the library. The modeling approximations within this procedure are justified by the context in which it is applied and a deliberate laxity in judging a steric clash in the pose evaluation. Specifically, the buildup of modeling errors in restricting poses to those defined by ideal torsion angles is mitigated by using the application on small compounds with relatively few rotatable bonds (≤ 3 according to Ro3). Similarly, the rigid docking approximation, which does not allow relaxation of small steric overlaps, is mitigated by only rejecting compounds that make significant, multiple van der Waals overlap with the protein.

Overall, this computationally cheap procedure aims to rapidly eliminate from experimental consideration those compounds for which there appears to be no feasible binding mode, facilitating a rapid experimental follow-through with the library design.

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

Example: Application to Kinase CNS Targets

A pilot study carried out at Zenobia Therapeutics on the protein kinase GSK3ß illustrates the type of results that might be obtained using structure-based screening with low molecular weight compounds. GSK3ß has been implicated in several disease areas, including bipolar disorder and Alzheimer's disease but only one clinical trial (for Alzheimer's disease and involving a non-ATP competitive compound) is currently underway and most of the published high affinity compounds may be too large to penetrate the blood-brain barrier.

Scaffold compounds that contain characteristic kinase hinge-binding motifs observed in previously known GSK3ß protein-ligand structures (*10*) were selected from a commercial screening collection containing ~450,000 compounds. Kinases are generally favorable targets for this kind of study because several types of compound that form two hydrogen bonds to the protein backbone in the hinge-binding region (*11*) are known to derive a significant binding energy from these interactions. The chemical properties of the selected compounds were restricted to meet the Ro3 in a mass range chosen to select mainly biaryl compounds (mean molecular weight was 253Da). These compounds were additionally filtered using a predictive equation for BBB passage (*12*) that incorporates chemical parameters from the Ro3 as factors. The 303 compounds obtained after applying these filters covered multiple chemotypes; some compounds contain fragments of known ligands but other compounds only resemble them in the sense of containing the small atomic motifs characteristic of kinase hinge-binders. Passing these compounds through the docking procedure described above reduced the library to 155 compounds that appeared feasible in the sense of fitting within the binding site on GSK3ß. This reduced set was then visually filtered and rebalanced across types of binding substructures to create exactly one plate of 82 compounds for chemical activity assay (*i.e.* a 96 well plate with two columns left empty for control compounds). Factors involved in this final selection included compound availability, overt similarities and docking contact potential score.

In a single-point (duplicated) assay with compound concentrations set at 20mM, 30 compounds gave >80% inhibition of GSK3ß activity. Follow-up determinations of IC50 for four of the most potent compounds yielded one compound with IC50 at ~1uM and three other compounds in the 6-70uM range. While these binding affinities are relatively low when compared to hits typically resulting from HTS screens the compounds are all small (MW 222-291Da, corresponding to 14-15 non-hydrogen atoms), have high ligand efficiencies (0.41-0.55) and have chemical properties well within a range applicable for orally available CNS drugs capable of crossing the blood-brain barrier.

Comparisons of this hit rate with parallel work on other CNS kinase targets for which the protein structures are not yet available suggest that the focused scaffold library is enriched by a factor of 5-10 by the addition of the structure-specific docking analysis.

Our work on other CNS kinase targets has shown that one further round of library designs with commercially available compounds, extending to slightly larger molecular weights, may yield potencies in the 0.1-1μM range, which may

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

be sufficient to initiate cell assays. As the compound design process becomes more informed and ideas for compound design become increasingly detailed and specific, there are limits to extent to which commercially available compounds meet the needs of a lead development project. At that point it may be necessary to perform specific synthetic chemistry to further develop promising molecules.

Conclusion

Fragment screening is flexible approach for the initial discovery phase of inhibitor design on a variety of protein targets. The approach is enabled as a low cost screening approach by the accessibility of suitable compounds that may be cherry-picked from commercial screening collections using simple computational approaches.

Acknowledgments

Vicki Nienaber (Zenobia Therapeutics) posed the question of how best to manage the size and composition of a fragment library when an excess of feasible compounds is available. Ruo Steensma (Steensma Consulting) suggested some approaches to reducing library sizes and provided advice on the definition of cyclic cores. The work on GSK3ß and other kinases implicated in CNS diseases was carried out at Zenobia therapeutics. The *SDsearch* software is available from the author (JB).

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

References

1. Hajduk, P. J.; Greer, J. *Nature Reviews: Drug Discovery* **2007**, *6*, 211–219.
2. Kuntz, I. D.; Chen, K.; Sharp, P. A.; Kollman, P. A. *PNAS* **1999**, *96*, 9997–10002.
3. Hann, M. M.; Leach, A. R.; Harper, G. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–864.
4. Pardridge, W. M. *J. Am. Soc. Exp. NeuroTherapeut.* **2005**, *2*, 3–14.
5. Hartshorn, M. J.; Murray, C. W.; Cleasby, A.; Frederickson, M.; Tickle, I. J.; Jhoti, H. *J.Med.Chem.* **2005**, *48*, 403–413.
6. Bosch, J.; Robien, M. A.; Mehlin, C.; Boni, E.; Riechers, A.; Buckner, F. S.; Van Vooris, W. C.; Myler, P. J.; Worthey, E. A.; DeTitta, G.; Luft, J. R.; Lauricella, A.; Gulde, S.; Anderson, L. A.; Kalyuzhniy, O; Neely, H. M.; Ross, J.; Earnest, T. N.; Soltis, M.; Schoenfeld, L.; Zucker, F.; Merritt, E. A.; Fan, E.; Verlinde, C. L. M. J.; Hol, W. G. J. *J.Med.Chem.* **2006**, *49*, 5439–5946.
7. Davies, D. R.; Mamat, B.; Magnusson, O. T; Christensen, J.; Haraldsson, M. H.; Rama Mishra, R.; Pease, B.; Hansen, E.; Singh, J.; Zembower, D.; Kim, H.; Kiselyov, A. S.; Burgin, A. B.; Gurney, M. E.; Stewart, L. J. *J.Med.Chem.* **2009**, *52*, 4694–4715.
8. Hubbard, R. E.; Davis, B.; Chen, I.; Drysdale, M. J. *Curr. Topics Med. Chem.* **2007**, *7*, 1–14.

9. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. *Drug Discovery Today* **2003**, *8*, 876–877.
10. Meijer, L.; Flajot, M.; Greengard, P. *Trends Pharmacol. Sci.* **2004**, *25*, 471–480.
11. Nobel, M. E. M; Endicott, J. A.; Johnson, L. N. *Science* **2004**, *303*, 1800–1805.
12. Feher, M.; Sourial, E.; Schmidt, J. M. *Int. J. Pharm.* **2000**, *201*, 239–247.

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch003

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

Chapter 4

Ligand-Based Virtual Screening Using Bayesian Inference Network

Ammar Abdo* and Naomie Salim

Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, 81310 Skudai, Malaysia
***E-mail: ammar.abdo@gmail.com**

The concept of molecular similarity has been widely used in rational drug design, where functionally similar molecules are sought by searching molecular databases for structurally similar molecules. In conventional 2D similarity methods, uncertainty in each stage of the similarity process is not considered and molecular features that do not relate to a particular biological activity carry the same weight as the important ones. In addition, since different methods have been found to retrieve different subsets of actives from the database, it is advisable to use several search methods where possible. A novel similarity searching approach using a Bayesian inference network (BIN) is introduced, where a database is ranked in order of decreasing probability of bioactivity. Our experiments on the MDDR database demonstrate that the BIN provided an interesting alternative to existing tools for ligand-based virtual screening, especially when the actives molecules being sought have a high degree of structural homogeneity. In such cases, the BIN substantially outperformed the conventional Tanimoto-based similarity searching system.

Introduction

Virtual screening is the name given to a range of computational tools for searching chemical databases to filter out the unwanted compounds or to assess the probability that each molecule will exhibit the same activity against a specific biological target. These tools can be used to reduce drug discovery cost by

© 2011 American Chemical Society

removing undesired compounds as early as possible and providing only those compounds that have the largest a priori probabilities of activity for conventional biological screening.

Many virtual screening approaches have been implemented for searching chemical databases, such as, substructure searching, similarity, docking and QSAR. These approaches can be categorized as structure-based approaches (e.g. ligand-protein docking) which can be used when the 3D structure of a biological target is available, and ligand-based approaches, which are applicable in the case of the absence of such structural information. Similarity searching is an example of a ligand-based approach. However, selecting the appropriate virtual screening approach depends on the amount and type of data that are available before a meaningful query can be formed. In addition, there is a substantial difference in the computational expense of different types of virtual screening approaches; the most notable difference is that the 3D versions require the generation of reasonable conformers of the molecules in the database.

Similarity searching is the simplest and one of the most widely used tools for ligand-based virtual screening. That is because this technique requires just a single known bioactive molecule as the starting-point for a database search. Similarity searching methods can be categorised according to the dimensionality of molecular structure used for determining the similarity, namely 2D similarity methods and 3D similarity methods.

Over the years, many ways of measuring the structural similarity of molecules have been introduced (*1–5*) with similarity measure based on the number of substructural fragments common to a pair of structures and a simple association coefficient (e.g. Tanimoto, Cosine) being the most common (*1*, *6*). There are, however, many other similarity methods in which the structural similarity between molecules is computed. The effectiveness of any similarity method has been found to vary greatly from one biological activity to another in a way that is difficult to predict (*2*). In addition, different methods have been found to retrieve different subsets of actives from the database, so it is advisable to use several search methods where possible.

Many studies in information retrieval have proved that retrieval models based on inference networks give significant improvements in retrieval performance compared to conventional models (*7–10*). In the chemoinformatics field, many techniques originate from the information retrieval field, where many similarities have been identified between them (*11*). There are several analogies between textual information retrieval and chemoinformatics (*11*), and these have led to recent work by Abdo and Salim (*12*, *13*) developing a ligand-based virtual screening method that uses BIN and 2D fingerprints. Experiments with a subset of the MDL Drug Data Report (MDDR) (*14*) database demonstrated that the BIN provided an interesting alternative to existing similarity search approaches. Similar results were obtained by Chen *et al.* (*15*), who used a BIN to search the MDDR and World Of Molecular Bioactivity (WOMBAT) databases. In this work we report the use of BIN for ligand-based virtual screening when a single and multiple reference structures are used. In addition, the BIN has been evaluated using an additional database rather than our previous work (*12*, *13*).

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

Similarity Inference Network Model

The basic model for similarity inference network, shown in Figure 1, consists of two component networks: a compound network and a query network. The compound network represents the compound collection. The compound network is built once for a given collection and its structure does not change during query processing. The query network consists of a single node which represents the user's activity-need and one or more query node representations. A query network is built for each query and is modified during query processing as the query is refined or additional representations are added in an attempt to better characterize the activity-need. The compound and query networks are connected though links between their feature nodes. Each node is binary-valued and takes on one of two values from the set {*true, false*}.

Compound Network

The compound network shown in Figure 1 is a simple direct acyclic graph (DAG) consisting of compound nodes (c_j) as roots, and feature nodes (f_i) as leaves. If we let C be the set of compounds and F be the set of features where the cardinality of these sets is n_c and n_f respectively, then the event space represented by the compound network is $E_c=C\times F$. Since all propositions are binary valued, the size of the event space is $2^{n_c} \times 2^{n_f}$.

Each compound node represents an actual compound in the collection. A compound node corresponds to the event that a specific compound has been observed. Each compound node has one or more feature nodes as children. Each feature node has one or more compound node as parents. The feature nodes can be divided into several subsets, each corresponding to a single molecular descriptor type that has been applied to the compound. For example, descriptors that represent properties of whole molecules such as log*P* and molar reactivity, descriptors that can be calculated from 2D graph representations of structures such as topological indices and 2D fingerprints, and descriptors such as pharmacophore keys that require 3D representations of structures. For simplicity, we only consider a weighted 2D fingerprint (integer fingerprint) in which each feature is being weighted by the frequency of its occurrence in the molecule. Therefore, the number of the feature nodes corresponds to the length of the molecular descriptor used to characterize the compound.

We represent the assignment of a specific feature to a compound by drawing a directed arc to the feature node from the compound node. In this case, the presence or absence of a link corresponds to the binary assignment of features to compounds. Each compound node has a prior probability associated with it that describes the probability of observing that compound. This prior probability will generally be set to *1/(collection size)* and this probability will be small for real collections. Each feature node contains a specification of the conditional probability associated with the node given its set of parent compound nodes. This specification incorporates the effect of any weighting associated with the feature node.

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

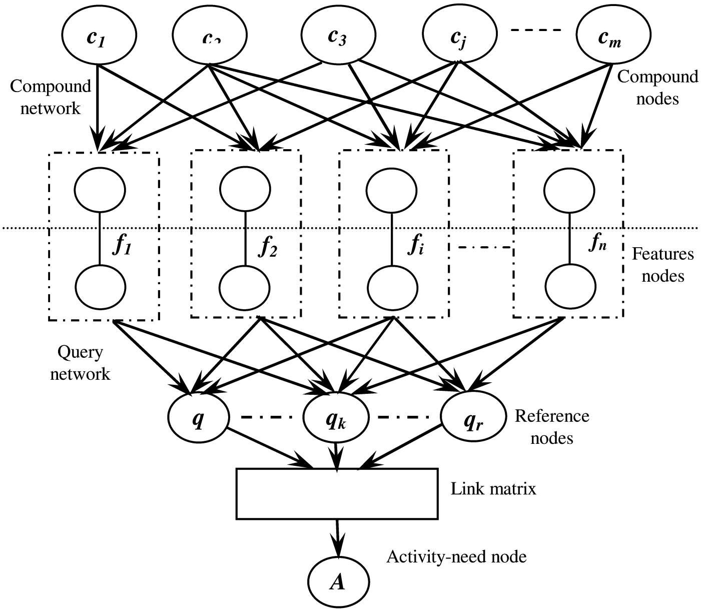


Figure 1. Molecular inference network model.

Query Network

The query network is an "inverted" DAG with a single leaf that corresponds to the event that an activity-need is met and multiple roots that correspond to the features that express the activity-need. A set of intermediate query nodes may also be used when multiple queries are used to express the activity-need. The roots of the query network are query features. A single query feature node has a single compound feature node as parent. A query feature node contains a specification of its dependence on a single parent compound feature node. The query feature nodes define the mapping between the features used to represent the compound collection and the features used to describe the query. In our model, the relation between query and compound feature nodes is 1:1 and completely dependent because the same descriptor is used to describe compounds and query.

However, the attachment of the query features nodes to the compound network has no effect on the basic structure of the compound network. Therefore, none of the existing links need change and none of the conditional probability specification stored in the nodes are modified.

Weighting Scheme

A weighting scheme is used to differentiate between different features in a molecule, based on how important they are in determining the similarity of that molecule with another molecule. Certain molecular features can be emphasised by associating higher weights to them when calculating similarity.

Different types of statistical information can be extracted from computerised representations of molecules to form the basis for a feature weighting scheme. These are as follow:

1. Feature Frequency *(ff)*, the number of occurrence of a particular feature within a compound, with more frequently occurring features being given greater weights than those that occur less frequently.
2. Inverse Compound Frequency *(icf)*, the frequency of the feature in the whole compound collection, with less frequently occurring features being given a greater weight than those that occur more frequently throughout the molecule collection.
3. Compound size (compound length), the number of the unique features assigned to a compound, with features in a smaller compound being assigned a greater weight than the same features in a larger compound.

The assignment of weights has been used at the National Cancer Institute (NCI) (*16*). Willett and Winterman (*17*) found that giving more weight to features that occur more frequently in a molecule did seem to give good results and other weighting schemes had little significance.

Interpretation of Inference Network

The conditional probability and the Bayes rule play a central role in our inference model. The topology of the inference network model is intended to capture all of the significant probabilistic dependencies among the variables represented by nodes in the entire network. Once the Bayesian network has been created, it can be used to predict the values that certain variables can take. Given the prior probabilities associated with the compounds and the conditional probabilities associated with the interior nodes, we can compute the posterior probability or belief associated with each node in the network.

The main aim of this model is to obtain the probability of biological similarity of each compound in the collection to a given query. When the query network is first built and attached to the compound network, we compute the belief associated with each node in the query network. The initial value at the node representing the activity-need is the probability that the activity-need is met given that no specific compound in the collection has been observed to be more similar to the query compared to the other compounds. If a single compound c_j is instantiated and evidence is attached to the network asserting $c_j = true$ with all remaining compound nodes set to *false*, we can compute a new belief for every node in the network given $c_j = true$. In particular, we can compute the probability that the activity-need is met given that c_j has been observed in the collection. We can now remove this evidence and observe another compound c_i, where $i \neq j$. By repeating this process, we can compute the probability that the activity-need is met given each compound in the collection and then rank the compounds accordingly. Here, we consider only compounds in isolation for simplicity reasons. The compound network is built once for a given collection. Given one or more queries representing the activity-need, we then build a query

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

network that attempts to characterize the dependence of the activity-need on the collection.

Encoding the Probabilities Using Link Matrices

Once the structure of the network has been created, the information will be propagated toward the node represented by the activity-need. The process of propagation is known as inference. For this process, we need to estimate the strength of the relationships represented by the network. This process involves estimating and encoding a set of conditional probability distributions. For any of the non-root nodes A in the network, the dependency on its set of parent nodes $\{P_1, P_2, ..., P_n\}$, quantified by the conditional probability $P(A|P_1,P_2,..,P_n)$. The conditional probability can be estimated by many types of weighting schemes. This estimation can be encoded using the link matrices form. Unfortunately, evaluation of the link matrix for node $\boldsymbol{A}$ with n parents requires $O(2^n)$ floating-point operation and space for all combination of parent values. However, a family of link matrices exists that allow this evaluation to be done efficiently, so that the space and time complexity is reduced to $O(n)$ (*8*).

More specifically, we use the *weighted-sum* and *weighted-max* canonical link matrices to implement a variety of weighting schemes, including feature frequency, inverse compound frequency, compound size or any combination . We assign a weight to the child node $\boldsymbol{A}$, which is, in essence, the maximum belief that can be associated with that node. Moreover, weights are also assigned to its parents, reflecting their influence on the child node. Consequently, our belief in the node depends on the specific parents that are true. To illustrate how the link matrix L_A can implement various weighting schemes, let node $\boldsymbol{A}$ have only two parents P_1 and P_2, and let w_1 and w_2, be the parent weights, and let w_A be the child weight $\boldsymbol{A}$ and $P(P_1 = true) = p_1$ and $P(P_2 = true) = p_2$, , then the full 2×2^n link matrix L_A is as follows:

$$L_A = \begin{bmatrix} 1 & 1 - w_2 w_A & 1 - w_1 w_A & 1 - (w_1 + w_2) w_A \\ 0 & w_2 w_A & w_1 w_A & (w_1 + w_2) w_A \end{bmatrix} \quad (1)$$

In this representation the values of first row corresponds to the case that $\boldsymbol{A}$ = *false* and the second row corresponds to $\boldsymbol{A}$ = *true*. We use the binary representation of the column number to index the values of the parents, so that the highest order bit reflects the value of the first parent, the second highest order bit the value of the second parent and so on. The w_1, w_2 and w_A substitute by any one of weighting schemes. Evaluation of this link matrix form results in the following.

$$\begin{aligned} P(A = true) &= w_2 w_A p_1^{-} p_2 + w_1 w_A p_1 p_2^{-} + (w_1 + w_2) w_A p_1 p_2 \\ &= (w_1 p_1 + w_2 p_2) w_A \\ P(A = false) &= 1 - (w_1 p_1 + w_2 p_2) w_A \end{aligned} \quad (2)$$

In case of a node $\boldsymbol{A}$ has n parents, the link matrix at Equation 1 becomes **NP-**hard, therefore the derived link matrix can be evaluated using the following closed form expressions:

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

$$bel_{wsum}(A) = w_A \sum_{i=1}^{n} (w_i p_i) \tag{3}$$

$$bel_{\mathrm{wmax}}(A) = \max \{w_1 p_1, w_2 p_2, ..., w_n p_n\} \tag{4}$$

Probabilities Estimation

Given the link matrix form, we need to provide estimates that characterise the dependence of the random variables (non-root nodes) in our model. The roots in Figure 1 are compound nodes, with the prior probability associated with these nodes set to 1/(collection size). Estimates are required for three different types of nodes: features, queries and activity-need.

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

Feature Nodes

Compound and query feature nodes are viewed as identical under the assumption that the user knows the set of compound features and can formulate queries using the compound features directly by using similar molecular descriptors. For the features involved in compound and not in query, we assign "false" beliefs, to achieve the identical assumption. Each feature node contains a specification of the conditional probability associated with node given its set of parent nodes. While in principle, computation of this probability would required $O(2^n)$ floating-point operation and space for a node with n parents, since we only consider one compound at a time, a simple estimation formula can be used. This estimate is given by the following equation:

$$P(f_i | c_j = true) = \alpha + (1-\alpha) \times \frac{ff_{ij}}{ff_{ij} + 0.5 + 1.5 \times \frac{cl_j}{avg\_cl}} \times \frac{\log\left(\frac{m+0.5}{cf_i}\right)}{\log(m+1.0)} \tag{5}$$

where $\boldsymbol{\alpha}$ is a constant and experiments using the inference network show that the best value for $\boldsymbol{\alpha}$ is 0.4, ff_{ij} is the frequency of the i^{th} feature within j^{th} compound, cf_i is the number of compounds containing the i^{th} feature, cl_j is the number of the unique features assigned to j^{th} compound, *avg_cl* is the average compounds length (over collection), and m is the number of compounds in the collection.

Query Nodes

We need to encode the dependency of each query formulation upon the feature nodes. To encode this probability, we use a *weighted-sum* link matrix form, as described in Equation 3. By using a *weighted-sum* link matrix, we assign a weight to each of the n parents of the query node, reflecting their influence on the query node. The parents with larger weights have more influence on our belief *bel(q)*. The belief in the query node *k* is then determined by the parents that are true and evaluated as

$$bel(q_k | f_{i..n} = true) = \frac{c_{jk}}{cl_j} \times \sum_{i=1}^{n} \left(\frac{ff_{ik}}{\max ff_k} \times \frac{\log\left(\frac{m+0.5}{cf_i}\right)}{\log(m+1.0)} \times p_i \right) \qquad (6)$$

where c_{jk} is the set of feature in common between j^{th} compound and k^{th} query, cl_j is the number of the unique features assigned to j^{th} compound, ff_{ik} is the feature frequency of the i^{th} feature within k^{th} query, $\max ff_k$ is the maximum frequency of occurrence in k^{th} query and p_i is the estimated probability at the i^{th} feature node (p_i computed at Equation 5).

Activity-Need Node

To encode this probability, we use *weighted-sum* or *weighted-max* canonical link matrices form, as described in Equations 3 and 4. By using these link matrices forms, we assigned a weight to each of the *n* parents of the activity-need node, reflecting their influence on the activity-need node. The parents with larger weights have more influence on our belief *bel(A)*. The belief in the activity-need node is then determined by the parents that are involved and evaluated as

$$bel_{wsum}(A) = \sum_{k=1}^{r} \left(\frac{c_{jk}}{ql_k} \times p_{jk} \right) \qquad (7)$$

$$bel_{w\max}(A) = \max\left\{ \frac{c_{j1}}{ql_1} \times p_{j1}, \frac{c_{j2}}{ql_2} \times p_{j2}, \ldots, \frac{c_{jk}}{ql_k} \times p_{jk}, \ldots \frac{c_{jr}}{ql_r} \times p_{jr} \right\} \qquad (8)$$

where c_{jk} is the number of common features between j^{th} compound and k^{th} query, ql_k is the number of the unique features assigned to k^{th} query, p_{jk} is the estimated probability that the k^{th} query is met by the j^{th} compound and *r* is the number of queries. In case of a single query used, the belief in the activity-need node then coincides with the belief in the query node.

Experimental

Our experiments have used the most popular chemoinformatics database: the MDDR (*14*). The database was first filtered using a set of filters from the Pipeline Pilot software (*18*) to remove duplicate compounds and those that could not be processed. The remaining database comprised 58693 compounds, including 6804 compounds belonging to 12 different activity classes. Details of these classes are given in Table I, which also contains numeric estimates of the level

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

of structural diversity in each activity class, this being based on the pair-wise Tanimoto similarities calculated using the ECFP_6 fingerprints from SciTegic (*19*), where it can be seen that the renin inhibitors are the most homogenous and the cyclooxygenase are the most heterogeneous.

Table I. MDDR structure activity classes

Code	*Activity class*	*Actives*	*mean*	*SD*
5H3	5HT3 antagonists	213	0.8537	0.008
5HA	5HT1A agonists	116	0.8496	0.007
D2A	D2 antagonists	143	0.8526	0.005
Ren	Renin inhibitors	993	0.7188	0.002
Ang	Angiontensin II AT1 antagonists	1367	0.7762	0.002
Thr	Thrombin inhibitors	885	0.8283	0.002
SPA	Substance P antagonists	264	0.8284	0.006
HIV	HIV-1 protease inhibitors	715	0.8048	0.004
Cyc	Cyclooxygenase inhibitors	162	0.8717	0.006
Kin	Tyrosin protein kinase inhibitors	453	0.8699	0.006
PAF	PAF antagonists	716	0.8669	0.004
HMG	HMG-CoA reductase inhibitors	777	0.8230	0.002

In order to make the evaluation of an approach independent of the characteristics of the specific fingerprint design, we included six different fingerprints in our experiments: atom type extended-connectivity counts (ECFC), functional class extended-connectivity counts (FCFC), atom type atom environment counts (EEFC), functional class atom environment counts (FEFC), atom type hashed atom environment counts (EHFC), and functional class hashed atom environment counts (FHFC) from SciTegic (*19*). The experiments here used the ECFC_4, FCFC_4, EEFC_4, FEFC_4, EHFC_4, and FHFC_4, where the numeric code denotes the diameter in bonds up to which features are generated. To make the computational task manageable, we employed a diameter size of four for all fingerprint types in this study, and the fingerprint types are folded to a fixed length of 1024 bits. All the fingerprint types above were generated by Pipeline Pilot software (*18*) from SciTegic.

To provide a basis of comparison for the BIN searches, analogous experiments were carried out using a conventional, Tanimoto-based similarity searching system (TAN). This system is a well established method in ligand-based virtual screening and therefore used as reference. In addition, the BIN method is compared to a popular technique for similarity searching using multiple bioactive reference structures, the data fusion (DF) (*20*) approach in combination with Tanimoto-based similarity searching system (*21*). Our application of DF involves fusing the

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

similarity scores yield from similarity searches of a chemical database against each member of the reference set. In particular, the MAX and SUM fusion rules, for the maximum of the similarity scores and the sum of the similarity scores respectively, were used.

For each of the 12 activity classes, 10 different sets of 10 active compounds were randomly selected as the reference sets. Each searching method was repeated 10 times using 10 different reference sets for each type of fingerprint. For each combination of a fingerprint and activity class, the different methods were applied and the average percentage of the active structures at the top 1% and the top 5% of the ranking list were generated. Finally, the results presented in this study are the mean and standard deviations for these recall values, averaged over each set of the 10 searches. TAN and DF methods were applied in combination with non-binary Tanimoto coefficient to compute the similarity scores.

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

Results and Discussion

Our experiments were carried out in two different ways. First, the BIN and TAN methods conducts an individual similarity search for each active reference structure and then the results averaged over all of them. Second, the BIN and DF methods conduct an individual similarity search for each active reference structure, and then combine the resulting similarity scores using *weighted-sum* (WSUM), *weighted-max* (WMAX) link matrices for BIN and MAX, SUM fusion rules for DF method.

Table II presents the results of the BIN and TAN methods when searches were carried out with a single reference structure. We can readily see the recall rates are different for the 12 activity classes. Inspection on the results reported in Table II show that BIN obtained average recall rates of 10-79%, higher than the TAN method, with 22% performance improvement in overall average recall rates compared to the TAN method. In only a single instance, for activity class Cyc, BIN produced a slightly lower recall rates (23.72%) than the TAN approach (24.44%). It is noticeable that this inferior result is associated with the most diverse datasets. Results reported in Table II shows that, recall rates for classes Kin, Cyc, and PFA were consistently lower than other classes, which could at least in part be attributed to the high diversity of Kin, Cyc, and PAF classes, whereas recall rates for classes Ren and Ang were consistently higher than other classes, which could at least in part be attributed to the low diversity of Ren and Ang classes.

The results in Table III suggest that WMAX and MAX, on average, produced the highest recall rates followed by WSUM and SUM, respectively. Investigation on the results reported in Table III reveal that the BIN approach is superior to the DF approach. The BIN (WMAX and WSUM) obtained the highest recall rates for all activity classes than DF approach, with 36% and 52% performance improvement in overall average recall rate compared to the MAX and SUM, respectively. The superiority of combined scores resulting from BIN over the DF approach is ascribed to the fact that, the BIN approach uses weights expressing the importance of each score. In addition, an individual similarity search for each active reference structure, which generated the ranked lists, is enriched with active structures more than those generated by the TAN approach.

Results reported in Tables II-III reveals the benefit that can be achieved using multiple reference structures. The values under the mean columns in Tables II-III show that the expected recall rate using a single reference structure are much lower than the results reported in Table III for the BIN and DF approaches, with 79%, 78%, 61% and 43% performance improvement in the overall average recall rate when multiple reference structures used rather than just one reference structure.

The results presented here included only the top 5% experiments using EHFC_4 fingerprints that because the conclusions that can be drawn from these results are the same as those that can be drawn from the top 1% experiments and other fingerprint types. Similar comments apply to experiments in which we evaluated the various approaches in terms of the recall of active Murcko scaffolds (*22*), rather than of active molecules.

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

Table II. Comparison of the average of active compounds recalled over the top 5% using BIN and TAN approaches

	BIN		*TAN*	
Activity	*mean*	*SD*	*mean*	*SD*
5H3	32.73	3.26	29.43	4.15
5HA	35.28	2.95	27.04	3.2
D2A	26.05	1.62	23.35	2.33
Ren	85.89	5.85	78.07	9.85
Ang	69.91	3.78	59.44	4.74
Thr	34.60	4.61	19.38	3.67
SPA	40.46	5.08	35.67	5.09
HIV	43.70	5.25	39.38	3.95
Cyc	23.72	2.63	24.44	1.7
Kin	24.96	6.40	22.06	5.45
PAF	19.50	2.05	14.58	1.56
HMG	57.70	2.49	33.32	3.12
Average	41.21	3.83	33.85	4.07

In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

Table III. Comparison of the average percentage of actives compounds recalled over the top 5% using BIN and DF approaches

	WMAX		*WSUM*		*MAX*		*SUM*	
Activity	*mean*	*SD*	*mean*	*SD*	*mean*	*SD*	*mean*	*SD*
5H3	70.35	5.18	69.36	4.50	40.20	5.34	42.42	13.07
5HA	76.60	7.58	77.55	9.45	57.83	8.03	40.75	10.37
D2A	66.32	6.22	64.81	6.25	49.17	7.13	39.70	6.12
Ren	95.62	1.21	95.73	0.68	91.24	4.20	92.33	2.35
Ang	94.70	2.25	96.52	0.67	75.06	3.87	76.01	2.95
Thr	66.50	8.59	66.32	8.26	26.99	8.60	31.58	7.17
SPA	79.76	6.01	77.95	7.90	72.88	7.05	57.72	11.36
HIV	75.15	4.81	73.80	3.44	59.28	4.31	54.04	5.13
Cyc	63.75	8.19	64.21	8.48	57.24	9.57	39.01	8.66
Kin	50.95	8.22	51.58	8.77	35.37	8.42	33.32	10.11
PAF	55.84	8.84	52.42	7.84	33.65	7.84	19.46	4.75
HMG	91.74	1.56	91.15	2.75	55.54	6.88	54.75	7.42
Average	73.94	5.72	73.45	5.75	54.54	6.77	48.42	7.46

Conclusion

One of the disadvantages in simple similarity searching is that molecular features or descriptors that are not related to the biological activity carry the same weights as the important ones. To overcome this limitation, we introduced a novel approach based on Bayesian inference network where the features carry different statistical weights. Features that are statistically less relevant are de-prioritized. In addition, we investigated similarity searching based on a Bayesian inference network and conventional similarity searching approaches when multiple reference structures are available. The Bayesian inference network approach was found to outperform the conventional similarity searching approaches. Our results suggest that, the Bayesian inference network provides an interesting alternative to existing tools for ligand-based virtual screening.

Acknowledgments

This work was funded by a grant no 79304 from the Malaysian Ministry of Science, Technology and Innovation.

References

1. Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
2. Sheridan, R. P.; Kearsley, S. K. *Drug Discovery Today* **2002**, *7*, 903–911.
3. Nikolova, N.; Jaworska, J. *QSAR Comb. Sci.* **2003**, *22*, 1006–1026.
4. Bender, A.; Glen, R. C. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.
5. Maldonado, A.; Doucet, J.; Petitjean, M.; Fan, B.-T. *Mol. Diversity* **2006**, *10*, 39–79.
6. Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Kluwer Academic Publishers: London, 2003.
7. Howard, R. T.; Croft, W. B. *Comput. J.* **1992**, *35*, 279–290.
8. Howard, R. T. Ph.D. Thesis, University of Massachusetts, 1991.
9. Berthier, A. N. R.; Richard, M. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, Association for Computing Machinery, 1996.
10. Luis, M. d. C.; Juan, M. F.-L.; Juan, F. H. *Int. J. Approximate Reasoning* **2003**, *34*, 265–285.
11. Willett, P. *Inf. Res.*. **2000**, *5*. http://www.webcitation.org/5a2tL872j.
12. Abdo, A.; Salim, N. *ChemMedChem* **2009**, *4*, 210–218.
13. Abdo, A.; Salim, N. *QSAR Comb. Sci.* **2009**, *28*, 654–663.
14. The MDL Drug Data Report Database. MDL Information Systems, Inc. http://www.mdli.com.
15. Chen, B.; Mueller, C.; Willett, P. *J. Cheminf.* **2009**, *1*, 5.
16. Hodes, L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 66–71.
17. Willett, P.; Winterman, V. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18–25.
18. *Pipeline Pilot Basic Chemistry Component Collection*, v6.1; SciTegic, Inc.
19. SciTegic Accelrys, Inc. http://www.SciTegic.com.
20. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
21. Willett, P. *J. Med. Chem.* **2005**, *48*, 4183–4199.
22. Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1996**, *39*, 2887–2893.

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch004

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

Chapter 5

A Computational Fragment Approach by Mining the Protein Data Bank: Library Design and Bioisosterism

F. Moriaud,[*,1,2] S. A. Adcock,[1,2] A. Vorotyntsev,[2] O. Doppelt-Azeroual,[2] S. B. Richard,[2] and F. Delfaud[1,2]

**[1]Felix Concordia SARL, 400 av Roumanille Bât. 7,
BP 309 06906 Sophia-Antipolis, France
[2]MEDIT SA, 2 rue du Belvedere, 91120 Palaiseau, France
[*]E-mail: fmoriaud@medit.fr**

Through database mining of the Protein Data Bank (PDB), protein pocket similarities and 3D structural alignments of similar pockets can be performed. These 3D structural alignments can serve as guides in drug design. The commercial MED-SuMo software performs superimposition of PDB ligands based on the ligand-binding corresponding pockets' and subpockets' 3D similarities. Subpockets are occupied by fragment-like molecules or portion of ligands. The mining of such fragments' interaction with the macromolecule surface serves as both a target-based and fragment-based computational method for PDB mining. In this work, we describe two practical applications: (1) a ligand-based drug design technique for bioisosteric replacement and compound library design and (2) a computational fragment-based drug design protocol for target-based drug design scenarios : ligand design, ligand decoration and compound library design. The bioisosteric approach is based on a database of bioisosteric replacement rules which were dervived from the entire PDB and are applicable to any ligand with a known or predicted 3D bound conformation. We present two successful applications: the design of alkenyldiarylmethane ligands for HIV-RT, and the design of a small compound library for HSP90. A case study using the computational Fragment-Based Drug Design

© 2011 American Chemical Society
In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

approach, was applied to the design of compounds for three types of protein target: Protein kinase, GPCR and kinesin.

1. Introduction

There are more than 72,000 macromolecular structures in the Protein Data Bank (*1*). The PDB has been growing at a rate of 13% per year over the last 5 years. This is an invaluable source of information available for understanding macromolecule interactions and ligand binding. Comparison of a protein pocket to all PDB pockets, as defined by ligand occupancy, enables pocket mining, pocket detection, functional annotation, drug repurposing and off-target identification. Some methods simply detect pocket similarities and others go a step beyond by generating 3D alignments of similar pockets (*2–10*). Those 3D alignment methods enable additional applications such as pocket characterization and drug design. While the original SuMo heuristic (*3*) was designed to detect convergent and divergent biochemical functional evolution between protein families, the commercial MED-SuMo software allows superimposition of any PDB ligands based on their corresponding pockets' 3D similarities and in addition to their subpockets' 3D similarities. In this study, we describe two practical applications: (1) a ligand-based drug design technique for bioisosteric replacement and compound library design and (2) a computational fragment-based drug design protocol for target-based drug design scenarios: ligand design, ligand decoration and compound library design (*3*, *11–14*).

Bioisosteres are compounds that, despite being structurally different, share similar physical properties and chemical interactions and therefore exhibit similar biological activities. This concept is relevant during the lead optimization stage of a drug-design programme, as bioisosteres can offer improved physical, chemical or toxicological properties while maintaining the desired biological activity. They are also of increasing value as alternative structures to overcome synthetic or patent-related obstacles to drug commercialization. Bioisosteric replacement is the process through which bioisteres are created by the replacement of substructures within the reference compound.

The bioisosteric approach is based on an automatically generated database of bioisosteric replacement rules derived from the entire PDB and applicable to any ligand in a 3D conformation. In the first application, we used the ester/benzo[d]isoxazole bioisosteric replacement observed in the Heat Shock Protein family and applied it to alkenyldiarylmethane ligands of HIV-RT. In the second application, we generated small compound libraries for HSP90.

The computational Fragment-Based Drug Design protocol (*15*, *16*) was used to design compounds that would bind to different targets: Protein kinase, GPCR and kinesin. In contrast to the bioisosteric approach, this is a target-based drug design technique which requires the target structure as input. This input structure can be any macromolecular model containing protein and/or oligonucleotides. A structure with a bound ligand is not required.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

2. Target-Based Alignment of PDB Ligands

Methods based on 3D comparison of binding sites enable the 3D alignment of the protein environments (*2–10*) and as a consequence, the 3D alignment of co-crystallized ligands in those pockets. This is referred to as target-based alignment of PDB ligands. There are 13,503 co-crystallized ligands with amino acid chains in the PDB (from X-ray diffraction data) having a molecular weight between 300 and 550 Da (*1*) (April 19th 2011). Only one occurrence of each PDB code & Ligand identifier is considered. Roughly half of them, 6,634 ligands, are co-crystallized in structures with a resolution of 2.0 Å or better. This resolution might be considered as a safe cut-off to exploit the 3D atomic positions of ligand atoms.

The description of the protein has a strong effect on the results and on the predicted similarity. There are two main classes of methods: those which are atom based (*5*, *7*) and in some cases including alpha carbon only (*10*), and those which use a pharmacophoric-like description (*2–4*). MED-SuMo belongs to the later class (*11–14*): the amino acids and nucleotides are transcribed using a dictionary of Surface Chemical Features (SCFs) capable of describing any macromolecule, whether protein, DNA, RNA or any combination of these. The SCFs can be directionless objects, vectors or planes. This customizable dictionary of features allows conversion of the amino acids and nucleotides into a set of user-defined SCFs. The SCFs are capable of encoding the alpha carbon and/or the sidechains, including non-exhaustively, hydrophobic, aromatic, formal charges and H-bond SCFs.

Once the ligands are aligned in 3D space, they can be combined by hybridization (*15*, *17*) or used as a source for matching fragments to act as bioisosteric replacement pairs (*18*). Relevant superposition of PDB ligands are obtained when very similar binding sites are superposed correctly. When the whole PDB is mined, the incorrect superposition should be discarded (described hereafter).

3. Ligand-Based Drug Design: Exploring Bioisosteric Replacements Derived from PDB Data

3.1. Introduction

As previously reviewed (*19*, *20*), the replacement of a given fragment in a 3D molecule may be used for bioisosteric replacement and scaffold hopping. The first case assumes that the fragment is exchanged with another fragment with similar potential interactions. Such interactions might include, for example, hydrophobicity, stacking, H-bonds and formal charges. Rather strict isosteres are preferred in this application. The second case involves the replacement of the entire scaffold while retaining the substituents optimized for the interaction with the target. Scaffolds with the same attachment point can be decorated with the same chemical substituents and are considered initially to be non-specific to the target and can therefore be generated *de novo*. Changing the scaffold (i.e. increased rigidity) can modify the binding affinity, improve drug-like properties, and increase the binding affinity because it may lose some conformational entropy upon binding. Therefore changing any significant fragment of a compound can significantly change its affinity. This is the reason why substitution rules should be established with an effort to maintain the original information in the protein environment responsible for the ligand binding. A third application mentioned in this review is one of compound library design. Here efficient navigation within the virtual substituent and functional group space is required, and is achieved by exhaustive bioisostere enumeration.

These three applications can be carried out using the FC-Bioisostere software described below. The replacements are defined from 3D aligned PDB ligands (using their target-bound 3D structure alignment) and are applicable to any original ligand. In other words, we asume that a bioisosteric pair can be defined within a protein family and applied to another protein family. We refer to this as "Bioisosteric rules hopping". This is more efficient for exploring the chemical space around an original ligand than using only the rules that would have been obtained by superimposing ligands from the same protein family. For compound library design for a given target, all the original ligands bound to that specific target, or a subset of those, are used to generate bioisosteres. The resulting bioisosteres are collated into a single file and constitute a compound library now specific for this target. Examples are provided in the results section for Heat Shock Protein 90 (HSP90).

In order to build the bioisostere database, the PDB was broadly mined using a diverse set of selected pocket queries. For each query, on the order of tens of ligands are superimposed. This approach provides the user with a diverse set of fragments of various sizes that could be replaced. The simplest case of replacement is changing a substituent with a single attachment point. This can be achieved using a database of unique replacements with a single optimized superposition. In our case, a pair of substituents can be superimposed with diverse conformations, as observed in the PDB, leading to what appears to be duplicates. Despite being duplicate chemicals, they are considered unique due to their different bound geometries. In fact, this approach provides suggestions of replacement independently of the number of attachment points and can therefore solve more complicated cases than a simple substituent replacement. The assumption of fragment replacement is not based on attachment points matching but rather on the fact that these two fragments were found to be overlaid in two very similar superposed binding sites. A limitation of this method is that the binding affinity of the ligand is not known for the complexes or, even if known, is not exploited here, as the contribution of any specific fragment's contribution to the global ligand binding affinity is not known. Therefore, the relative affinity of the two fragments is not known experimentally and is at best evaluated as not preventing binding.

3.2. Method

The method described here corresponds to the recent R&D developments of Felix Concordia SARL and implemented in the FC-Bioisostere software, where most of the functionalities are implemented.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

Bioisostere Database (DB)

The DB Generation is a multi-step process, and the earlier steps are performed with the MED-SuMo technology. In essence, the early steps simply create a list of ligand molecules that are each overlaid with a series of other ligand molecules that exhibit the same binding modes. The MED-SuMo algorithm finds and aligns molecules with similar binding modes onto the input ligand(s) by pocket mining over the whole PDB. The PDB ligands are standardized according to definitions in the Ligand Expo dictionary (*21*) Using this data set, the DB generation tool locates potential bioisosteric replacements and populates the DB with this new data.

Our overall protocol closely resembles that of Kennewell *et al.* (*18*), but the individual steps are significantly different. The aligned set of PDB ligands are selected from very similar 3D binding pockets, identified using the MED-SuMo protocol, rather than only using amino acid sequence similarity as the selection criteria. Protein structures with the same sequence are very likely to have an identical pocket (conformation may differ) and therefore ligands are superimposed in the same environment. Each query ligand protein environment is compared, using MED-SuMo, against a subset of the PDB binding sites containing ligands with an HAC between 15 and 65 (filter available in the MED-SuMo server) that were considered of interest in this work. This subset contains only those ligands manually selected as likely being of synthetic chemistry origin (not endogenous). We did that selection to keep the database small and relevant for medicinal chemistry projects.

The MED-SuMo parameters used to describe the pockets are graphs of SCFs containing the 4.5Å environment of the ligand and a high density of triangles of chemical features (20-60) (*3*). This MED-SuMo database is optimized for this bioisosteric application and has a virtually zero rate of false positives. The detected structural alignment is retained only if the pockets have a strong similarity, *i.e.*, MED-SuMo score above 4.0. This set of 3D aligned ligands therefore exploits most of all the relevant experimentally validated 3D superimposable ligands. False positives are defined here as non-relevant superpositions, either from less highly similar sites or from highly similar sites that are incorrectly superimposed. In this case, a correct superposition of the pockets implies a relevant superimposition of the bound ligands. The aim here is not to find pockets which are only partially similar, but similar as a whole with possibly a few residues that differ in the neighborhood of the ligand. These differences are tolerated in our bioisosterism definition as it is an additional source of potential bioisosteric pairs and a reasonable assumption because the ligands were designed for very similar pockets. In fact, even in the situation of identical pockets, it is not certain that all fragments of the ligand were optimized for binding and/or in some cases they may not even favor binding and act only as linkers (spacers), so replacement defined from identical pocket are also an approximation in some cases. In virtually all cases of relevant superpositions with MED-SuMo, the local protein folds are very similar and therefore a relevant superposition is a local superimposition of the local scaffolds. As a consequence, fragments of superimposed ligands are likely to be exchangeable in this particular protein environment. In cases where the protein conformation differs significantly in a part of the binding site where a bioisosteric rule could be potentially derived, then it is likely that the ligands will not overlap there. Therefore no bioisosteric rule will be derived from this part of the ligand, as the SEAL score cut-off verifies the fragment overlap.

In this study, 1945 ligands with a molecular weight between 300 and 500 Da, co-crystallized with a X-Ray resolution of 2.0Å or better and preferably not endogenous (like ATP, NAD), were selected from the PDB. Each ligand was used to define a MED-SuMo query launched against the subset of PDB binding sites described above. A consequence is that this DB is meant to search only replacements in ligands mostly derived from synthetic chemistry. We do not expect the replacements to be transferable to other ligand sources. To investigate replacement for endogenous ligand, we've generated a larger database (10000 query ligands instead of 1945) containing also endogenous ligands to evaluate the effectiveness of this approach in finding bioisosteres of natural compounds (not described here).

5.5 million pairs are stored in this DB with query fragments ranging from a heavy atom count (HAC) of 2 to a maximum of 65. Replacement of fragments consisting of a single atom are therefore not included in the DB. However cases of replacement of a single atom are found in the DB through replacements of more than one atom, *e.g.*, replacement of an ether (C-O-C) by a thioether (C-S-C), where the resulting bioisostere will differ by only one atom. The pairs are evaluated with the SEAL function (*18*) to select only highly overlapping fragments, not necessarily having the same 3D coordinates for most atoms. A SEAL score greater than 0.75 was selected in this work and the fragments were defined using the sectioning algorithm (*18*) described below. For each "aligned pair", the two molecular fragments, the SEAL score, the transformation matrix, the MED-SuMo Score and the SCF count are stored in the DB, the last 3 values being calculated for the superimposed ligands, query and one ligand hit.

Bioisosteric Replacement Rule Elucidation

As stated, the overall algorithm used for locating likely bioisosteric replacements resembles the one described by Kennewell *et al.* (*18*). Nonetheless, not only may default parameters differ but we perform fragmentation differently. We also explicitly record attachment points that indicate where bonds were broken during fragmentation. The user provides a series of ligands ("query ligands"), each with one or more overlaid ligands ("hit ligands"). In the work presented, the fragmentation method was a substructure pattern matching using a file of desirable fragments in SMARTS format. We used a file containing PDB ligand fragments derived from the whole PDB in order to match substructures of PDB ligand efficiently. These fragments are diverse in size and chemistry and are potentially applicable to any original ligand: substituents, linkers, pairs of rings, pairs of ring with their substituents, also more complex fragments such as ring assemblies with and without their substituents, and finally the scaffold with or without their exocyclic double bonds and linker double bonds. The fragmentation is therefore optimized for the fragmentation of PDB ligands, and the query

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

ligands are fragmented into many overlapping fragments in this way. The hit ligands overlaid are then fragmented using the sectioning algorithm presented by Kennewell *et al.* (*18*). Wherever the resulting hit ligand section fragment matches the query ligand fragment with a SEAL score of greater than 0.75, those fragments are considered an equivalent pair and are stored in the bioisostere database. The two fragments forming a pair are known as the "query fragment" and the "replacement fragment". The SEAL score above 0.75 allows keeping replacements which range from isosteres to less similar fragments allowing replacements such as methyl ester to benzoxazole.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

Bioisosteres Generation

The bioisostere generation described herein occurs in the FC-Bioisostere GUI though an interactive process. In principle, it will also be available in a FC-Bioisostere CLI for use in batch processes.

The usual workflow consists of a few steps: the user loads a ligand ("original ligand") in its target bound conformation. This original ligand is fragmented into "original ligand substructures". These ligand substructures are located in the DB as fragments, and the corresponding replacement fragments will be used to generate bioisosteric molecules from the original ligand. The fragmentation scheme is the same as the one used to fragment the query ligand while building the DB. Therefore, the fragmentation is less optimal for original ligands which are not present in the PDB. However it is likely that there is a significant overlap between those fragments and any ligand from synthetic chemistry. We found that it is indeed the case frequently, and we demonstrate it here in the particular case of ADAM ligands (see results).

Query Fragment and Replacement Fragment Selection

The possible replacements to explore for the query substructures are selected from the bioisostere database according to a number of simple rules. For each query substructure, all query fragments in the DB that have the same topological structure (*i.e.*, the same type of atoms connected with the same pattern of bonds) are found. Any query fragments that have fewer attachment points than the query substructure are disregarded, but any surplus attachment points are accepted. These query fragments are optimally overlaid onto the query substructure using Kearsley's superimposition algorithm (*22*). Those fragments which have a RMSD exceeding a given RMSD are discarded. A RMSD value of 1.0 Å was used in this work.

The replacement fragments found in the bioisostere DB for the set of acceptable query substructures was used in the recombination process, that is the process by which the query ligand undergoes bioisosteric replacements. "DbCount" is the term for the number of acceptable query fragments found. A final filtering step is applied to the replacement fragments after alignment onto the corresponding query substructures. Those without an attachment atom within a given distance threshold to each attachment point of the query substructure are discarded. A value of 1.5 Å was used in this work. "ReplacementCount" is the term for the number of acceptable replacement fragments found after filtering.

Any additional attachment atoms in the replacement fragment are disregarded for the purpose of filtering but their locations and connectivities are stored. These surplus attachment points may be viewed as potential sites for optimizing the molecular structure in the final generated bioisosteres, if desired.

Bioisostere Enumeration

All possible bioisosteric molecules of the original ligand, given the set of potential replacements, may be enumerated using a systematic and exhaustive search algorithm. This algorithm proceeds via a depth-first, backtracking tree search. When invalid replacements are found, those search branches are terminated. If disconnected structures are created during the search, these branches continue to be followed as, in some cases, a subsequently attempted replacement may happen to resolve the disconnection.

Through this algorithm every possible valid combination of replacements on the original ligand will be found. In the results presented herein, only a single replacement in the original ligand is applied. The diversity of the bioisosteres is therefore rather limited to the close space around the original ligand.

Bioisostere Reconstruction

Bioisostere Reconstruction is the process by which a query substructure is removed from the original ligand (or intermediate bioisostere) and a replacement fragment is inserted in its place to reconstitute a putative bioisostere. This process is performed for every step of the bioisostere enumeration search. The specific algorithm used in this study, referred as "Two-Way Attachment Recombination", is one of a few alternatives. In this, all atoms comprising the query substructure are simply deleted from the original ligand, leaving attachment points where remaining atoms lose bonds. Attachment points representing the broken bonds from the original ligand from the original source of the replacement fragment are compared. Wherever possible a broken bond will be reconstituted at these points. A bond is considered good if one attachment is within a specified distance of an atom to which the other attachment point is bonded, and *vice versa*. Matching the tolerance described above, an identical value of 1.5Å is used in this study. To reduce the distortions inevitably seen in the recombined molecules, the coordinates of the newly bonded atoms are adjusted according to the weighted mean location of the real atoms with a multiplier of two and the attachment atoms with a multiplier of one. This algorithm is considered a pure bioisostere generation method, as only the originally broken bonds can be recreated. Alternative algorithms, not presented for this study may form *de novo* bonds, may delete clashing atoms, or even add additional atoms to reduce excessive strain where the replacement fragment is linked into the original molecule.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

3.3. Bioisosteric Replacement: Case of ADAM Ligand

The alkenyldiarylmethanes (ADAMs) are a class of potent and highly specific HIV non-nucleoside reverse transcriptase inhibitors (NNRTIs). Unfortunately, most of the ADAMs are too unstable toward hydrolysis in blood plasma to be considered as potential therapeutic candidates. A series of alkenyldiarylmethanes (ADAMs) with a benzo[d]isoxazole ring in place of the metabolically unstable methyl ester moiety and an adjacent methoxyl group were synthesized by Deng *et al.* (*23*). In that study, the authors' initial results demonstrated that the benzo[d]isoxazole ring is an effective bioisosteric replacement of the metabolically labile methyl ester moiety in ADAMs. The replacement of methyl esters with fused benzo[d]isoxazole could prove to be generally useful in situations that require alternatives to hydrolytically unstable methyl esters. See (*23*) and references therein.

We applied our protocol to explore bioisosteres of one ADAM molecule cotaining the previously described ester substituent. The 3D model of alkenyldiarylmethanes (ADAM) 28a, deposited by the same authors (*23*), was retrieved from BindingDB (named BindingDB_2786) (*24*). This ADAM ligand was used as the starting point for bioisostere enumeration. 272 bioisosteres, excluding duplicates, were generated using all query substructures. Thirty-nine of those bioisosteres were identified as replacing the methyl ester substructure. Only one of the methyl esters is described here as the results are virtually identical for the second ester. Among them, the benzo[d]isoxazole is found (shown in Figure 1). It is ranked 26th when ranked according to a SEAL score of 0.80 and ranked 29th according to MED-SuMo Score of 7.3.

In the next few paragraphs, we describe the origins of this methyl ester/benzo[d]isoxazole pair in the Bioisostere DB. This pair originates from a MED-SuMo run with the query pocket of RDA in the 2FXS PDB entry (*25*), a HSP82 structure. This MED-SuMo query, one of the 1,942 that were launched to generate the Bioisostere DB used in this study, generated the superimposition of 135 pockets (hits) similar to the query pocket and, as a consequence, the superimposition of 135 co-crystallized ligands with the query ligand RDA. Looking more closely, these ligands are all from proteins of the GHKL fold (*26*) sharing a high local protein fold and pocket similarity (HSP82, HSP83, HSP90, GRP94, MutL, pyruvate dehydrogenase kinase). By manual inspection, there are no false positives in these results. The MED-SuMo hit, from which the hit fragment was extracted (3BMY CXZ) (*27*) is ranked 35th, according to MED-SuMo score of 7.0. This illustrates that replacements are obtained not only within the same protein family (HSP82 in this case) but also within very similar pockets, independently of sequence identity.

In Figure 2, the MED-SuMo superimposition of RDA and the CXZ PDB ligands is shown. The 3D superposition of the ligand from the 2D structure would be very difficult without knowledge of the binding modes; in contrast, the target-based superimposition is very accurate and unambiguous. It is also worth noting that the methyl ester is a substituent but the benzo[d]isoxazole is part of the scaffold, thus exemplifying the potential diversity of the bioisosteric pairs in the DB. This information could not have been obtained, in this case, by considering only the substituents. Another key point is that the pair was generated from HSP proteins and applied to HIV-RT. The protocol makes exhaustive use of the entire wealth of knowledge available in the PDB, as opposed to using structural information from only the protein family of the original ligand (HIV-RT).

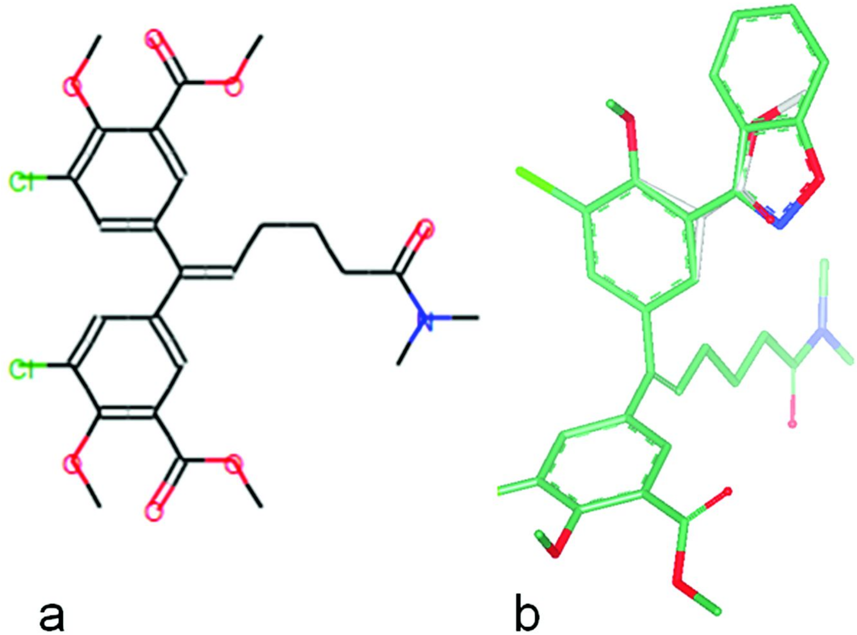


Figure 1. (a) 2D depiction of the original ligand alkenyldiarylmethanes (ADAM) 28a; (b) 3D view of the same ligand (carbon atom rendered in light grey) and its bioisoster with the benzo[d]isoxazole substitution (carbon atom rendered in green) 158 x 122mm (96 x 96 DPI). (see color insert)

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

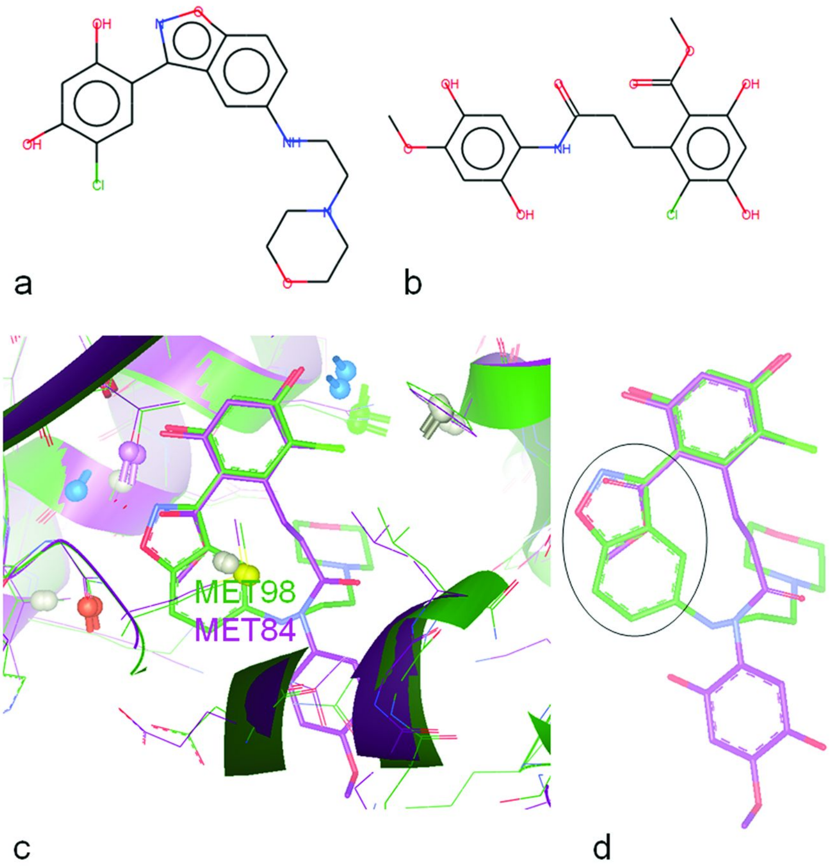


Figure 2. (a) 2D depiction of the PDB ligand CXZ; (b) 2D depiction of the PDB ligand RDA; (c) Screenshot from MED-SuMo GUI showing the superimposition of the pocket of CXZ in the PDB HSP90 file 3BMY (rendered in green) and the pocket of RDA in the PDB HSP82 file 2FXS (rendered in violet). One of the matching residues, a methionine, is labelled with the residue number. The Surface Chemical Features used to generate the 3D superposition of the proteins are shown (balls and ball&sticks); (d) Same superimposition as in (c) but only the superimposed ligand. RDA is rendered in violet and CXZ in green. The area used to generate the pair (methyl ester/benzo[d]isoxazole) is schematically shown within the ellipse. 224 x 232mm (96 x 96 DPI). (see color insert)

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

3.4. Small Compound Molecule Libraries: Explore Bioisosteric Replacements in HSP90 PDB Ligands

A list of HSP90 PDB ligands was obtained using the PDB website (*1*), using the Seq. Similarity search feature, from an HSP90 alpha structure and selecting 90% sequence identity as cut-off. 68 ligands of the ATP site were collected and downloaded with their 3D structure (bound conformation). In case of multiple occurrences within a single PDB file, the first ligand was selected.

Our aim was to generate a library of compounds that are likely to bind to HSP90 and then search for exact matches in BindingDB (*24*), in the PDB ligand expo (*21*) and in the PubChem Compound libraries (*28*) (these databases were all accessed on April 29th, 2011). For this purpose, we ran the MED-Search module as described in Moriaud *et al.* (*15*).

In Figure 3, four example bioisosteres are shown together with the bioisosteric pair from which they were proposed: (a) a sulfonamide/amide pair from RNA-directed RNA polymerase displaying very similar fragments, (b) a carboxylate/tetrazole pair from beta lactamase demonstrating a classical case of bioisosterism, (c) a case of scaffold hopping using a bioisosteric pairs from the Protein kinase family, (d) an amide/flurophenyl pair from serine proteases where the carbonyl and the fluoro atom are both interacting with an H-bond donor. In HSP90 the fluoro atom is facing the amine group of a lysine residue and, therefore, also with the H-bond donor chemical feature. These four examples show the diverse origin of the pairs in terms of protein families.

In total, 930,986 bioisosteres were generated. They are new compounds compared to the 68 original ligands. Of these, 16,657 are unique (*i.e.*, after duplicate removal). All matches in the PDB are HSP90 ligands. Matches in BindingDB are compounds which are known to bind to HSP90 (73), other HSP proteins (4), Estrogen Receptor (13), Endoplasmin (1), Arachidonate lipoxygenase (2) and adenosine receptor (4). This suggests that the bioisosteres generation explores a focused region of biochemical space around the PDB HSP90 ligands and are not in the chemical space of other targets. That implies that most of the 16,657 bioisosteres are likely to be specific binders of HSP90. There are 361 unique matches in Pubchem Compounds (from a total of 30.3 million compounds, only those with a HAC≥15 are considered) using a Tanimoto cut-off of 0.9 and 283 exact matches. This demonstrates that a small molecule library for HSP90 can be built using our bioisosteric approach using a single fragment replacement and exact matches. Looking at the examples shown in Figure 3, we can observe that the bioisosteres are designed in a rational way using replacements based on reliable experimental structural data. Using two replacements instead of one in this case study would increase greatly the diversity of the bioisosteres and would be an option to generate a library with thousands of compounds. Experimental testing would give a prospective validation of this approach. Interestingly the bioisosteres are posed in the reference frame of the original ligand (original ligands are all PDB ligands in this HSP90 case study), therefore their geometries could be optimized *in situ* and prioritized for onward testing and/or synthesis, according to their scoring using, for example, a PLP intermolecular term (*16*). This was not done in this study, but previously described in our recent paper (*16*). That elaboration of the protocol was not required here as the bioisosteres rarely have a high strain energy, because only one fragment replacement was made for each proposed bioisostere and only and because only original bonds were reconstituted.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

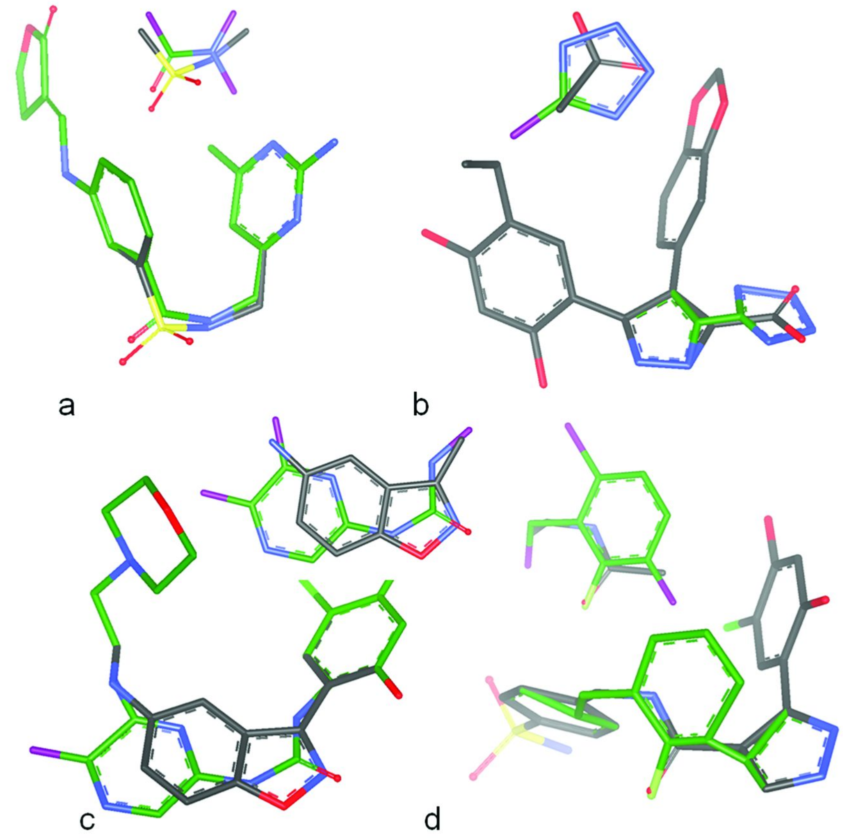


Figure 3. Superpositions of original PDB HSP90 ligand (carbon atoms rendered in grey) and bioisostere (carbon atoms rendered in green). Also shown the query fragment and hit fragment superposition of the bioisosteric pair used to generate the bioisostere from the original ligand. For each of the four panels PDB ID, PDB Ligand ID and Protein name are given for: the original ligand; the query fragment; the hit fragment (a) 2QG0 A94 HSP90; 2D3Z FIH RNA-directed RNA polymerase; 2GIR NN3 RNA-directed RNA polymerase (b) 1YC1 4BC HSP90; 3HLW CE3 Beta lactamase; 3G32 3G3 Beta lactamase (c) 3BMY CXZ HSP90; 3E93 19B P38 Protein kinase; 2E9V 85A CHK1 Protein kinase; (d) 2BYI 2DD HSP90; 1W13 SM1 Urokinase type plasminogen activator; 2R2M I50 Thrombin. 251 x 253mm (96 x 96 DPI). (see color insert)

4. Target-Based Drug Design: Fragment-Based Approach from PDB Data

4.1. Introduction

Obtaining experimental structural information on fragments or ligands complexed to a target protein is a key element, and also a major limitation, to the number and types of targets that are amenable to fragment-based drug discovery. Consequently, computational methods play a crucial role in deriving structural information for designing compounds that fit a particular site on a given protein. If the three-dimensional structure of the protein is known, this information can be directly exploited for the retrieval and design of new ligands.

Here, we review the key points of the work done at MEDIT on this Fragment-Based computational application (*15*, *16*). This is a target-based approach that requires the 3D structure of the target. However, no ligand bound complex structure is needed to design the ligand. This is in contrast with the above description of the bioisosteric approach where only the 3D structure of a ligand in its bound conformation (whether known or predicted) is needed. Another difference is that the entire PDB is considered: ligands bound to RNA and DNA are also considered. This is possible because the macromolecules, proteins and oligonucleotides, are described using the same Surface Chemical Features (SCFs) (H-bond donor, H-bond acceptor, hydrophobic, ring stacker, *etc.*), though there are features unique to proteins such as thiol. Also, formal charges are only relevant for proteins.

Our aim in using this approach is to detect protein local similarities which are smaller in volume than whole binding site similarities. In the bioisosteric approach only high similarities between whole sites are retained. Seeking pocket similarities is efficient both with protein families (intrafamily hits) and in between protein superfamiles (interfamily hits). Examples of similar pockets across protein superfamilies are rare. They occur when pockets bind similar ligands with similar binding modes. Seeking protein local similarities provides more hits across protein superfamilies than simple pocket mining does. Local similarities are exploited to repurpose fragments of any PDB ligands and in particular of drugs. Repurposing ligands as described above is limited to similar binding sites. Therefore, potential repurposing fragments of ligands is more likely because similar sub-pockets are more often found than similar binding sites.

In contrast to docking/scoring protocols, the pose of the fragments does not rely on the exploration of all possible binding modes nor on a scoring function to rank the poses. Mining within protein superfamilies like the protein kinases allows identification of the chemical moieties from a ligand that are the most likely to target-hope from one kinase to another. One such example is the chemical moiety bound to the hinge of kinases and can be in most cases transferred from the different conformations of DFG-out to DFG-in conformations. This is not the case for the fragments of the ligand in the allosteric pocket. Therefore this approach is well suited to mine all binding sites, including flexible sites with compounds bound. The interfamily hits are for example fragments binding to

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

hinge-like motifs found in other proteins as in the case of the protein kinase ATP site and the non-nucleoside HIV reverse transcriptase (*14*).

4.2. Method

This FBDD protocol is based on the assumption that similar protein surfaces are likely to bind the same fragment with the same pose. The large volume of protein-ligand structures now available in the PDB enables applications of the protocol for diverse fragments and for many protein families (*15*). The PDB is encoded as a database of MED-Portions, where a MED-Portion is a structural object encoding protein-fragment binding sites. MED-Portions are derived from mining all available protein-ligand structures with any library of small molecules by the MEDP-fragmentor software. They contain atoms, dummy atoms keeping track of where the bonds were cut to make the portion (that is the substructure or fragment) of the PDB ligand. Mined with the MED-SuMo software to superpose similar protein interaction surfaces, pools of matching MED-Portions can be determined for any binding surface query. A typical MED-Portions database contains one million of portions of PDB ligands. To generate hit-like molecules from fragments in each MED-Portion, MED-Portions are combined in 3D with the MED-Ligand toolkit.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

4.3. Hybrids: Scoring and Ranking

This fragment-based drug design protocol generates hybrids from a set of MED-Portion chemical moieties selected with several criteria (as described above). These hybrids are thus likely to have:

(1) chemically reasonable structures, since they are generated from chemically accessible molecules,
(2) to fit in the binding site, since the selected MED-Portions chemical moieties have been selected to have a maximum number of tolerated steric clashes, and
(3) potentially favorable interactions with the protein since they have been co-crystallized with a protein containing locally similar biochemical features.

In our previously published study on kinesin allosteric pockets (*16*), the generated hybrids were analyzed and scored using an *in situ* energy minimization step prior to the computation of standard scoring functions. We found that the intermolecular term of the PLP scoring function was simple, fast and efficient at retrieving known actives at the top of the list. This scoring function is validated on this target and on other targets (data not published) and is the recommended method to rank such hybrids (and bioisosteres when generated from an original ligand bound to its target). There is no need to do *in situ* minimization of the hybrids as long as only the PLP intermolecular term is used to score the hybrids. The PLP intramolecular term is usually very high and not reliable because the hybridization leads to bond length and bending angles which are not optimal, though an *in situ* minimization would reveal in most cases that the conformation is close to a minimum in terms of RMSD.

4.4. Library Design

This protocol was used to generate ligands for a VEGFR2 protein kinase. In the first step, fragments of PDB ligands derived from the protein kinase family (intrafamily 75%) and from other families (interfamily 25%) are aligned within the binding site of interest. These fragments are combined through hybridization. To avoid generating millions of compounds, we focused the design towards ligands having a phenylamide moiety close to the gatekeeper in the same way as the GIG ligand in the 2OH4 structure (*29*). *In situ* hybridization of these fragments leads to 220k hybrids which represent 10,000 scaffolds. 175 scaffolds match the scaffolds of PDB ligands, therefore: (1) PDB Protein Kinase scaffolds are retrieved and (2) most of the scaffolds are new compared to the ones of the PDB, and are then likely to be original compounds, at least for structural studies.

In the GPCR study, there are virtually no intrafamily hits as there are very few GPCR structures in the PDB. Using only interfamily hits, the shape of known ligands was retrieved and some of the hybrids matched known beta-adrenergic ligands. These hybrids contained compounds similar to the initial 3 ligands (PDB codes 2rh1, 3d4s, 2vt4). We also obtained 11 other ligands (CGP12177, ICI-118551, SR59230A, alprenolol, carvedilol, pindolol, NIP, bevantolol- S, nebivolol, timolol, bucindolol. This shows that the protocol can generate molecules similar to known active ligands. This is a significant retrospective validation as these GPCR ligands are not present in the PDB (*15*).

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

5. Summary

We presented a new protocol to predict bioisosteric structures based on the wealth of 3D protein structures now available both publicly and, in principle, within pharmaceutical research organizations. This method is complementary to the more usual literature-based and de novo approaches. Our database offers some convincing advantages: We can link the resulting hypothetical bioisosteres back to the original 3D structures corresponding to the suggested replacements. This leads to higher implicit confidence in the predictions, which is extremely desirable to the scientists involved in its use. Furthermore, our method can be built to hold not only generic sets of replacements, but also therapeutic target-specific replacements.

References

1. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–42.
2. Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2000**, *323*, 387–406.

3. Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137–145.
4. Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607–633.
5. Gold, N. D.; Jackson, R. M. SitesBase: A database for structure-based protein−ligand binding site comparisons. *Nucleic Acids Res.* **2006**, *34*, D231–234.
6. Debe, D. A.; Danzer, J. F.; Goddard, W. A.; Poleksic, A. STRUCTFAST: Protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring. *Proteins* **2006**, *64* (4), 960–7.
7. Ramensky, V.; Sobol, A.; Zaitseva, N.; Rubinov, A.; Zosimov, V. A novel approach to local similarity of protein binding sites substantially improves computational drug design results. *Proteins* **2007**, *69*, 349–57.
8. Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**Jun, *71* (4), 1755–78.
9. Brenke, R.; Kozakov, D.; Chuang, G. Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **2009**, *25* (5), 621–7.
10. Feldman, H. J.; Labute, P. Pocket similarity: are alpha carbons enough? *J. Chem. Inf. Model.* **2010**, *50* (8), 1466–75.
11. Doppelt, O.; Moriaud, F.; Bornot, A.; de Brevern, A. G. Functional annotation strategy for protein structures. *Bioinformation* **2007**, *1* (9), 357–9.
12. Doppelt-Azeroual, O.; Moriaud, F.; Adcock, S. A.; Delfaud, F. A review of MED-SuMo applications. *Infect. Disord.: Drug Targets* **2009**, *9* (3), 344–57, review.
13. Doppelt-Azeroual, O.; Delfaud, F.; Moriaud, F.; de Brevern, A. G. Fast and automated functional classification with MED-SuMo: an application on purine-binding proteins. *Protein Sci.* **2010**, *19* (4), 847–67.
14. Moriaud F.; Richard S. B.; Adcock S. A.; Chanas-Martin L.; Surgand J.-S.; Ben Jelloul M.; Delfaud F. Identify drug repurposing candidates by mining the Protein Data Bank. In *Briefings Bioinf.* **2011**, DOI: 10.1093/bib/bbr017, accessed online April 21, 2011.
15. Moriaud, F.; Doppelt-Azeroual, O.; Martin, L.; Oguievetskaia, K.; Koch, K.; Vorotyntsev, A.; Adcock, S. A.; Delfaud, F. Computational fragment-based approach at PDB scale by protein local similarity. *J. Chem. Inf. Model.* **2009**Feb, *49* (2), 280–94.
16. Oguievetskaia, K.; Martin-Chanas, L.; Vorotyntsev, A.; Doppelt-Azeroual, O.; Brotel, X.; Adcock, S. A.; de Brevern, A. G.; Delfaud, F.; Moriaud, F. Computational fragment-based drug design to explore the hydrophobic sub pocket of the mitotic kinesin Eg5 allosteric binding site. *J. Comput.-Aided Mol. Des.* **2009**, *23* (8), 571–82.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

17. Pierce, A. C.; Rao, G.; Bemis, G. W. BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, p38, and HIV protease. *J. Med. Chem.* **2004**, *47*, 2768–75.
18. Kennewell, E. A.; Willett, P.; Ducrot, P.; Luttmann, C. Identification of target-specific bioisosteric fragments from ligand–protein crystallographic data. *J. Comput.-Aided Mol. Des.* **2006**, *20* (6), 385–94.
19. Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric replacement and scaffold hopping in lead generation and optimization. *Mol. Inf.* **2010**, *29* (5).
20. Meanwell, N. A. Synopsis of some recent tactical application of bioisosteres in drug design. *J. Med. Chem.* **2011**, *54* (8), 2529–91.
21. Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: A data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004**, *20* (13), 2153–5.
22. Kearsley, S. K. On the orthogonal transformation used for structural comparisons. *Acta Crystallogr., Sect. A* **1989**, *45*, 208–210.
23. Deng, B. L.; Zhao, Y.; Hartman, T. L.; Watson, K.; Buckheit, R. W., Jr.; Pannecouque, C.; De Clercq, E.; Cushman, M. Synthesis of alkenyldiarylmethanes (ADAMs) containing benzo[d]isoxazole and oxazolidin-2-one rings, a new series of potent non-nucleoside HIV-1 reverse transcriptase inhibitors. *Eur. J. Med. Chem.* **2009**Mar, *44* (3), 1210–4.
24. Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R. N.; Gilson, M. K. BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.
25. Immormino, R. M.; Metzger, L. E.; Reardon, P. N.; Dollins, D. E.; Blagg, B. S.; Gewirth, D. T. Different poses for ligand and chaperone in inhibitor-bound Hsp90 and GRP94: Implications for paralog-specific drug design. *J. Mol. Biol.* **2009**, *388*, 1033–1042.
26. Dutta, R.; Inouye, M. GHKL, an emergent ATPase/kinase superfamily. *Trends Biochem. Sci.* **2000**, *25* (1), 24–28.
27. Gopalsamy, A.; Shi, M.; Golas, J.; Vogan, E.; Jacob, J.; Johnson, M.; Lee, F.; Nilakantan, R.; Petersen, R.; Svenson, K.; Chopra, R.; Tam, M. S.; Wen, Y.; Ellingboe, J.; Arndt, K.; Boschelli, F. Discovery of benzisoxazoles as potent inhibitors of chaperone heat shock protein 90. *J. Med. Chem.* **2008**, *51*, 373–375.
28. The PubChem Project. http://pubchem.ncbi.nlm.nih.gov.
29. Hasegawa, M.; Nishigaki, N.; Washio, Y.; Kano, K.; Harris, P. A.; Sato, H.; Mori, I.; West, R. I.; Shibahara, M.; Toyoda, H.; Wang, L.; Nolte, R. T.; Veal, J. M.; Cheung, M. Discovery of novel benzimidazoles as potent inhibitors of TIE-2 and VEGFR-2 tyrosine kinase receptors. *J. Med. Chem.* **2007**, *50*, 4453–4470.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch005

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Chapter 6

A Fragment-Based Docking Engine: eHiTS

Exhaustive Fragment Pose Enumeration and Matching Using Surface Contact Based Statistical Score

Zsolt Zsoldos, Ph.D.*

**Chief Scientific Officer, SimBioSys Inc., 135 Queen's Plate Dr.,
Toronto, ON, Canada M9W 6V1
*E-mail: zsolt@simbiosys.ca**

There are numerous methods for flexible ligand docking, including stochastic and systematic methods to sample the conformational and pose space of the ligand. The eHiTS docking engine described in this chapter uses a fragment-based approach closely resembling the experimental fragment-based design techniques that flood the active site cavity with small binding fragments independently from each other and then look for ways to link up the fragments. eHiTS is a deterministic, exhaustive flexible docking method that systematically covers the part of the conformational and positional search space that avoids severe steric clashes, producing highly accurate docking poses at a speed practical for virtual high throughput screening. A new scoring function has been developed as part of the eHiTS flexible ligand docking software. The method has a unique approach to combine the strengths of the statistical and empricial scoring functions. Statistical information was collected from a large number of crystal structures considering the full distribution of interaction geometries as described by the temperature factors associated with every atom in the crystal structrues. Empricial functions are derived from the statistical data to define the final scoring function terms.

© 2011 American Chemical Society
In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Introduction

The process of finding the binding conformation and pose of a chemical structure in the active site of a receptor macromolecule is called flexible ligand docking. It can be considered as an energy minimization problem, however most of the available molecular mechanics programs are too sensitive to local minima to find the appropriate docking poses (*1*).

Various stochastic search methods can be used to tackle this problem, including Simulated Annealing (e.g. AutoDock2 (*2*), Dockvision (*3*), MCDOCK (*4*)), Genetic Algorithms (GOLD (*5*), AutoDock3 (*6*)), Tabu Search (ProLeads (*7*)), etc. They have been reported to be successful in reproducing the experimental binding conformations of some ligand receptor complexes (*5*). The search algorithm in these methods is a random probing technique, driven solely by a scoring function. However, the search space is vast (see details in consequent subsections), thus these methods can not guarantee to find the optimal solution in finite time.

There are also systematic methods, including incremental construction (FlexX (*8*), Hammerhead (*9*), DOCK4) and multiple conformer rigid body docking (e.g. FLOG (*10*), DOCK3 (*11*) or FRED (*12*)). Due to the vast search space size, these systematic methods employ various heuristics and sampling limits to avoid combinatorial explosion. It is difficult to strike a balance in the sampling such that the conformational and pose space is searched exhaustively within reasonable CPU time. The incremental construction methods employ a coarse sampling of conformations using a small number of discrete rotomers. The multiple conformer rigid docking systems use a few hundred low energy conformers of the ligand.

Statistical analysis of experimental data from bound ligand conformations was performed and the findings indicate that sampling of low energy conformers is insufficient to reproduce protein-ligand binding geometries, a much more exhaustive search is required. eHiTS (electronic High Throughput Screening) offers a truly exhaustive systematic search algorithm that considers all poses with a fine resolution sampling to guarantee sufficient accuracy.

The accuracy of the eHiTS algorithm is demonstrated on reproducing known bound conformations and poses of ligands from co-crystallized proteins. The program's ability to enrich database selections with actives is also measured as well as the scoring function's ability to reproduce experimentally measured binding affinities.

The Fragment Docking Problem

There are some notable differences in the problem of docking small molecule fragments into actives sites compared to docking full ligands. The fragments are typically smaller and have few or no rotatable bonds. Smaller size means that the fragments can fit in many more poses and orientations, thus the pose space is significantly larger, on the other hand the conformational search problem is trivial within the fragments.

Pose and Conformational Sampling Requirements

A fundamental design goal of the eHiTS system is to provide an exhaustive systematic search of the part of the conformational and pose space that avoids severe steric clashes with sufficiently fine sampling to reproduce experimentally observed binding modes. In theory, a truly exhaustive search should explore the infinite continuum of rotational and translational space. In practice, discrete sampling is acceptable ***if*** it is fine enough to ***not*** miss a solution.

Statistics on hydrogen bond geometry in small molecular crystal structures (*13*) show a range of 1.6Å to 2.2Å distance between the hydrogen and the acceptor atoms, i.e. it can be described as 1.9Å ± 0.3Å.

Hydrophobic contacts are observed (*14*) in the range of 3.2Å to 4.2Å between the atom centers of two carbons, i.e. 3.7Å ± 0.5Å.

Aromatic π stacking interactions and metal ion interactions also have their own ranges of acceptable geometry with similar tolerances. It is clear that a half Angstrom difference in atom positions may mean losing a crucial hydrogen bond or cause a severe steric clash instead of a perfect van der Waals contact. Therefore, we define sufficient sampling to mean that atom positions must be sampled at least every half an Angstrom.

This definition of sufficient sampling for atom displacements implies a requirement for rigid fragment rotation and dihedral angle sampling. A simple trigonometric calculation shows that a tangential movement of 0.5Å is caused by rotation of about 5° at a radius of 7Å. Drug-like ligands can easily reach or exceed the size of 7Å, therefore rotations and dihedral angles must be sampled at least every 5 degrees.

Similarly, there are ligand structures, e.g. the ligand from PDB code 1CX2, that contain rotatable bonds that influence the position of atoms 7Å away from the axis of rotation. If the dihedral angle is changed by 5°, then the the atom far away from the axis would move by more than 0.5Å Angstroms. If a sampling algorithm missed the correct dihedral by 10 or 15 degrees then such atoms would end up severely clashing with the receptor site instead of creating a perfect hydrophobic surface contact. In other words, a 15-30+ degree sampling is far too crude to be useful for a docking program that aims to be exhaustive.

Bound conformations of 5000 ligands in high resolution (less than 2.5Å) crystal structures from the RCSB Protein Data Bank (PDB) have been analyzed to collect statistical data on the dihedral angles of rotatable bonds. Table 1 shows how many ligands have ***all*** their dihedral angles within the given ± range to either a staggered or a gauche value, i.e. how many bound ligand conformations would be found within a given error if only those dihedral angles were sampled. Another important data point is that about 10\% of the bound ligand conformers exhibit at least one eclipsed dihedral angle, i.e. 0±5° between sp3 centers each bearing one additional heavy-atom neighbor. 97% of X-ray conformations in this set deviate by more than 5° from conformations generated by sampling the dihedrals of each rotatable bond every 60° . It is clear from the data that it is necessary to include conformations which in an isolated molecule would be of high energy, in order to sample the conformations adequately for docking.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Table 1. Statistics concerning staggered and gauche conformers in X-ray structures of bound ligands from 5000 PDB entries with 2.5Å or better resolution. The Table shows how many ligands have *all* their dihedral angles within the given ± range to either a staggered or a gauche value

Error limit	*Number of ligands*	*Percentage*
5°	108	2.2%
10°	211	4.2%
15°	315	6.3%

Many researchers have performed similar analysis (*15–18*) and come to essentially the same conclusions regarding high-energy conformations of bound ligands.

Search Space Size

The size of the search space can easily be calculated from the sampling requirement defined above. For an average sized ligand with six rotatable bonds, the following formula is computed:

- Translations along 3 axes: every 0.5Å in 10Å box, i.e. 20^3
- Orientations about 3 axes: every 5° in 360°, i.e. 72^3
- Dihedral angle sampling: every 5° in 360°, i.e. 72^6
- Total number of poses: $20^3 \times 72^3 \times 72^6 \sim 10^{20}$

This number (ten to the power twenty) is so huge that brute force evaluation of all those poses with a relatively fast scoring function -- that can process 2 thousand poses per second -- would take 3 billion years on a single CPU. Using large supercomputers or distributed clusters (e.g. Grid computing) with hundreds of thousands of CPUs, it would still take more than tens of thousands of years to dock a single ligand.

Stochastic methods that employ fine enough sampling, do search this same vast space, but instead of systematic sampling, they employ random walks. Decisions are made based on a goal function evaluation and some stochastic decision process whether or not to keep a given trial pose. However, new trial poses are selected by some random alteration of an already tested pose. There is ***no*** driving force employed towards new areas of the search space that are yet unexplored. Therefore the poses examined by stochastic methods, if represented as points in N-dimensional space, are comparable to Brownian movement. Such random walks are known (*19*) to over-sample some regions while leaving some large areas completely unexplored. The flexible ligand docking pose space is 10-20 dimensional (depending on the number of rotatable bonds) and the sampling problems of random walks are much more severe than they are in low dimensional problem space.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Our goal was to develop an intelligent exhaustive method that can limit the fine sampling of the search space to areas of interest where good scoring solutions may reside, while eliminating large portions of the vast search space where it is guaranteed that no good scoring position can be found.

The eHiTS Pose Generation Algorithm

As demonstrated above, brute force evaluation of *all possible poses and conformations* with sufficiently fine sampling is not feasible within practical CPU time limits. Therefore, the search space must be reduced. One reduction applied by eHiTS is to limit the search to conformations and poses that avoid severe steric clashes between receptor and ligand, i.e. where geometric fit is possible.

In order to explore the vast search space exhaustively in an efficient manner, our approach involves sub-division of the task into smaller partial problems that are easier to solve. However, unlike DOCK or FLexX, eHiTS does ***not*** use an incremental construction method, but instead attempts to find the global optimum by enumerating combinations of independent partial structure dockings.

eHiTS has a novel flexible ligand docking method that is exhaustive on the conformations and poses that avoid severe steric clashes between receptor and ligand. The algorithm generates all major docking modes that are compatible with the steric and chemistry constraints.

First the binding pocket is determined by building a steric grid for the whole receptor, dividing regions into separate pockets and identifying the possible interaction sites. Then, a cavity description is built that consists of thousands of geometric shapes (polyhedra).

The ligand is divided into rigid fragments and connecting flexible chains. eHiTS docks ***all*** rigid fragments to ***all*** possible places in the cavity ***independently*** of each other. This is ***not*** an incremental construction, all rigid fragments are docked to every possible place regardless of the other fragments. Although, the poses are scored, no local (biased) decision is made to reject any stericaly feasible pose for any rigid fragment based on interaction score.

An exhaustive matching of compatible rigid fragment pose sets is performed by a rapid hyper-graph clique detection algorithm. This may yield a few hundred (small pocket, few rigid fragments) to several million (large pocket, many small rigid fragments) acceptable combinations of poses. However, at this point, the scores for each component have been evaluated, so it is possible to make a ***global*** decision as to which fragment pose combination is the best.

The flexible chains are then fitted to the specific rigid fragment poses that comprise a matching pose set. The reconstructed solutions define a rough binding pose and conformation of the ligand. These poses are refined by a local energy minimization in the active site of the receptor, driven by the scoring function. Figure 1. shows a greatly simplified schematic example of the process, how the ligand is fragmented, the rigid fragments are placed into the cavity in multiple poses and then a suitable set of ligand poses can form a solution.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

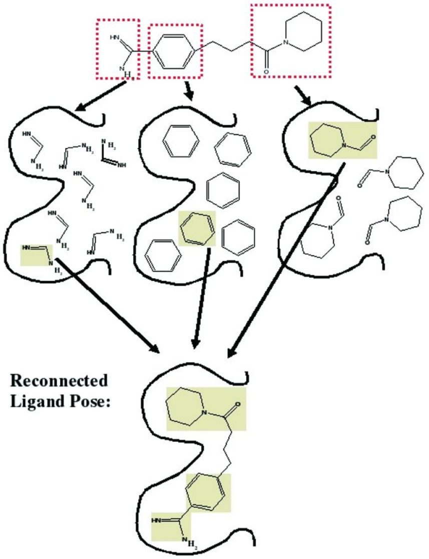


Figure 1. A simplified schematic example showing the processing stages of the eHiTS pose generation algorithm.

Geometric Shape and Chemical Feature Graph

The fragmentation of the ligand is focused on separating rigid fragments from the flexible linkers. All ring systems are considered rigid and their conformation is preserved as given in the input. Therefore it is desirable to use multiple ring conformers (e.g. chair, boat and twist boat for a cyclohexane) for complete conformational sampling.

Acyclic fragments with double or normalized (resonance) bonds and sp2 hybridized atoms are also considered rigid, e.g. including the amide functional group. Figure 2 shows an example of the fragmentation of a ligand. Whenever a bond is broken during this fragmentation, both atoms of the bond are duplicated, i.e. they appear both in the rigid fragment as well as in the flexible chain fragment. These are referred to as the ***join atoms***. Distances of the join atoms are used to determine the compatibility of rigid fragment poses in the pose-match phase of the algorithm. The join atom positions serve the end point constraints of the flexible chain fitting, furthermore they are used to define the overlay transformation in the reconstruction of the complete solution poses before optimization.

Both the cavity and the candidate ligands are described by a Geometric Shape and Chemical Feature graph, herein referred to as GSCF graph. The nodes of the GSCF graph represent a rigid shape by a simplified geometric hull. It is derived from regular polyhedra and then distorted to *shrink-wrap* the actual molecular fragment or cavity region (see detailed explanation of the shape generation below separately for the cavity and ligand fragment case).

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

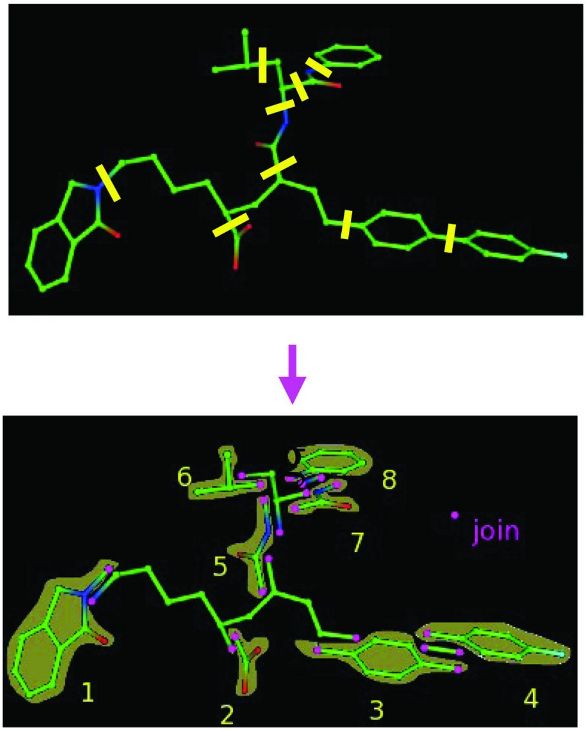


Figure 2. The ligand is broken into rigid fragments and flexible chains.

Chemical feature flags are associated with each vertex of the polyhedron. The edges of the GSCF graph define the connectivity between the nodes, including distance boundaries for the acceptable relative positions of the nodes.

Feature-Graph Representation of the Cavity

The cavity is described as a set (thousands) of geometric shapes, polyhedra. These polyhedra are generated by picking center points on a regular 0.5Å spacing 3D grid, placing a tiny polyhedron on the grid, then "stretching" its vertices out from the center until they reach the surface of the cavity. The center points are selected such that are suitable to place the center of mass of a rigid fragment there. Grid cells that either violate the receptor boundary or are too close to it, are not suitable as center points. The distance from the boundary must be at least an atom radius. The space is measured in various directions from those centers and the regular polyhedra is distorted so that the vector length from the center point matches the distance measured, thus building polyhedra that represent the shape of the available space around the center. Figure 3 demonstrates the generation of a cavity node using a 2D cartoon for sake of simplicity. The 3D polyhedra overlap with each other and fill the whole cavity space. Chemical feature flags are assigned to the vertices of the polyhedra.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

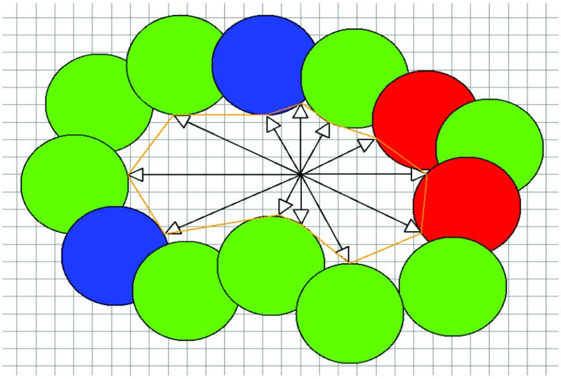


Figure 3. Simplified 2D cartoon demonstrating the generation of a cavity descriptor polygon using 12 vectors in 30 degree increments. The available space around grid points is measured in directions dictated by regular polyhedra shapes, then chemical property flags are assigned to the end points based on the chemical activity of the closest receptor atoms.

The distance measurement from the center to the receptor boundary is performed using a 3D steric grid, which is generated within a bounding box of the binding site. This bounding box also acts as an artificial closing of any binding pocket that is open to the solvent water. If no receptor boundary is hit by the scanning ray that is measuring the empty space in the direction of a vector, the the bounding box terminates the ray placing a practical limit on the polyhedron vector length.

The current version of eHiTS does not consider protein flexibility, while it handles ligand flexibility exhaustively. In order to handle protein flexibility, this feature graph representation would have to be extended either by allowing alternative vector size sets corresponding to each center, or allowing ranges of possible vector lengths and using some probability function within the range during the matching process.

Feature-Graph Representation of the Ligand

The rigid fragments are also wrapped into polyhedra described by directional vectors from their centers of mass. Again the vectors from the center to the vertices of the polyhedra are scaled to match the distance from the center of mass of the ligand fragment to the van der Waals surface in the direction of the vector. Figure 4 shows an example of how a ligand is divided into rigid fragments and shrink-wrapped into a polyhedron shape. The polyhedron is color coded to represent the chemical features assigned to the vertices.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

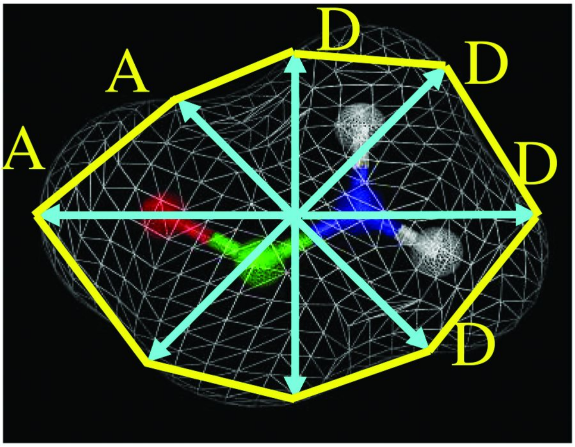


Figure 4. The ligand is broken into rigid fragments and each fragment is wrapped into a polyhedron shape with chemical properties assigned to the vertices of the polyhedron.

The polyhedrons are created by shrinking the vector lengths from the center to the vertices, but the directions are maintained, therefore the angles between them are not changed either. Consequently, if the self symmetric transformations of the regular polyhedra are applied to these polyhedra, then each directional vector from center-to-vertex will be overlayed on another such vector by the transformation. Each transformation can be described as a specific permutation of the vertices.

Rigid Fragment Docking

The rigid fragment docking proceeds by placing the rigid fragment polyhedra inside the cavity polyhedra. All combinations are explored (each rigid fragment polyhedron with each cavity polyhedron) and all orientations of the polyhedra. We use directional vectors based on the vertices of an icosahedron and a dodecahedron combined. These regular polyhedra have 60 self-symmetric transformations each, so we use those to orient the rigid fragment polyhedra inside the cavity polyhedra.

The polyhedron representation allows a very rapid enumeration of all fitting poses using the following method. The GSCF graph nodes contain the length of the directional vectors to each vertex, and they also contain the decreasing order of these lengths.

1. The lengths of the ligand node vectors are checked against the cavity vector lengths in decreasing order. If any ligand vector is larger than its corresponding cavity vector plus ε grid-tolerance, then it is impossible to fit the rigid fragment node into that cavity node in any orientation, therefore no detailed orientation check is necessary, so the whole loop of the following step can be skipped without any loss of solution.
2. All 60 self-symmetric transformations of the regular polyhedra (dodecahedron and icosahedron) are stored in the form of a permutation

table of their vertices. A loop is run to test each of the 60 orientations, using the permutation table, in each execution of the loop. The vertices of the ligand polyhedron are mapped to the vertices of the cavity polyhedron via the permutation table. The directional vector lengths are compared and the pose is rejected if the ligand vector is longer by more than ε grid-tolerance for any vertex.

3. For each vertex map that passes the vector length based steric check, the chemical feature flags of each vertex pair are scored and summed up to give a complete chemical fit score of the given ligand fragment pose.
4. The 3D coordinates are computed for the acceptable poses based on a transformation matrix that is pre-computed and stored for each row of the permutation table.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Note that in steps 1 and 2 a specific grid-tolerance value must be applied to the comparison of the vector lengths, i.e. allow the ligand vector to be longer than the corresponding cavity vector by a small amount and reject the pose ***only*** if the ligand vector is longer than cavity vector plus ε. This ε grid-tolerance depends on the resolution of the 3D grid that is used to generate the cavity center points (by default *a*=0.5Å resolution is used, but it is a user adjustable parameter, higher accuracy can be reached at the expense of more CPU time if this size is reduced). The reason is that cavity graph nodes are generated at discrete locations controlled by the grid, and it is possible that if the center is shifted by a fraction of a grid cell, then a larger fragment may fit. However, this sampling error is limited by the largest possible distance of the ideal position to the grid cell corner:

$$\epsilon = a\sqrt{3}/2$$

All of the chemical property flags that apply are assigned to each vertex of the polyhedron, both on the cavity and the ligand fragments. A scoring matrix is defined for the flags which contains a score for each flag-to-flag interaction pair (more details on the flag based scoring are given later in the scoring section). The score of a rigid docking pose is computed by summing all the scores of any flag pairs present on matched-up vertices between cavity and ligand.

For some larger rigid fragments, the 32 vectors of the combined polyhedra will produce a surface sampling where distance between surface points is larger than the desired 0.5Å. However, this does not limit the sampling precision of the docking, because multiple cavity polyhedra (partially overlapping each other) are used for the mapping, so there are target positions for each ligand vector with sufficient density. The cavity polyhedra are generated on a 0.5Å spacing grid with multiple orientations considered for the same center.

Typically, the program evaluates several million mappings of the rigid fragment polyhedra to cavity polyhedra. The ones that do not fit geometrically (steric violations) are rejected and the score is computed for those that do fit. Typically, there are tens of thousands of fitting poses (10-20 thousand for small pockets and large fragments, 60-100 thousand for small fragments in large cavities).

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

When the number of acceptable poses is too large to handle during the next (pose matching) phase, a clustering algorithm is applied to group the poses that are close to each other in RMSD metric space and a single representative is kept from each cluster. The diversity of the poses and their coverage of the cavity site is maintained during this clustering step.

This clustering step could potentially compromise the exhaustiveness of the search if the cluster representatives do not cover the pose space with sufficient resolution. The maximum number of cluster representatives is controlled by a user adjustable parameter and by default it is set to a value that achieves a fast (sub-second) PoseMatch run-time with an average separation between representatives of about 1-1.5Å RMSD. In terms of search space sampling, this means that a sampling pose is generated within √3/2 times the separation distance from any query pose (in the worst case), while the average error from the X-ray pose can be estimated to be about 0.43Å-0.65Å. This range goes slightly higher than our desired precision, but the parameter can be adjusted to achieve more precise sampling at the cost of CPU time. There is another tolerance applied during the PoseMatch phase that is computed from the actual average separation distance between the poses. That tolerance is applied to the compatibility check, i.e. comparison between join point distances and connecting chain lengths. The tolerance is dependent on the actual average pose separation, so that it counters the loss of precision, allowing the selected poses to represent their whole cluster (within the radii) for the purpose of matching instead of considering strictly the particular pose. Thus the algorithm maintains the exhaustive coverage via the use of this calculated tolerance and the ability to refine the search by adjusting the control parameters of the clustering.

It is very important to keep fragment poses that do not get good scores, because even for high affinity ligands it is possible that some fragments are acting simply as spacers and are not contributing much to the binding. In fact, analysis of the X-ray complexes in the test set shows that many contain fragments that either do not make any interaction with the protein, or even make clearly repulsive interactions. Of course, the energy loss due to the “bad” interactions must be compensated by some strong attractive interactions formed by other fragments of the ligand.

All acceptable poses of the rigid fragments are computed regardless of other fragments in the ligand. Therefore, the information about the acceptable poses of a given fragment can be reused when another ligand containing thesame fragment is docked to the same receptor. This situation occurs very frequently during a virtual screening study when many thousands (or even millions) of drug-like ligands are docked to a given target receptor, because such ligands often contain some typical functional groups. The DockTable extension of eHiTS makes use of the repeating fragments to speed up the screening process by using an SQL database to store all the results of the rigid fragment docking phase. An efficient hash key (canonical name) is used for indexing the database to retrieve the previous results. If no results are stored for the given rigid fragment yet, then the docking proceeds as described above in this section, then the results are deposited to the database.

It is sufficient to store the 3D transformation and the score for each pose, therefore a space efficient storage can be achieved that requires about 1MB disk space per rigid fragment for the DB (this size does not depend on the size of the fragment but it does depend on the size of the cavity). We have run experiments screening various ligand libraries against various receptor targets and observed the speed-up curve of the docking time per ligand as well as the number of fragments deposited to the database. Significant speed-up is observed during the first few hundred to few thousand ligand docking runs, but the speed tends to level out between 5 and ten thousand ligands (the speed is 2-4 fold faster at that point than docking speed without the SQL DB). The number of commonly re-used fragments is in the order of a few thousands, therefore a limit of ten thousand fragments has been implemented in the DockTable extension of eHiTS. This limits keeps the disk space requirement under 10GB per receptor regardless of the size of the ligand library docked.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Pose Matching

There are several thousand alternative poses generated and scored at the rigid docking step for each rigid fragment. The next task is to select pose-sets containing a single pose for each ligand rigid fragment such that the distances between them are compatible with sizes of the flexible chains that should connect them. In addition, they must not bump into each other.

One can think of this task as mapping the ligand graph (where each node represents a rigid fragment) on to the receptor cavity graph (where each node represents a possible placement position and orientation of a ligand rigid fragment). Such graph-mapping problems are often solved by graph algorithms operating on a hyper-graph rather than on the graphs to be mapped. The hyper-graph is a higher order graph, where nodes represent mappings between the original graphs.

This task is solved by clique detection on the following hyper-graph. Each node of the ligand graph is represented by a set of hyper-graph nodes, one corresponds to every accepted rigid fragment pose, i.e. the nodes of the hyper-graph represent individual mappings of ligand graph nodes to a cavity graph nodes. There are edges between those node pairs where all the following conditions hold true:

a) The nodes correspond to poses of different ligand fragments,
b) There is no steric clash between the two poses, and
c) The distance between the join points of the fragments in the given poses is compatible with the length of the chain that should connect them, i.e. it is within the interval that is possible to span by the given chain.

Maximal cliques of this hyper-graph should consists of as many nodes as the number of rigid fragments in the ligand (number of nodes of the ligand graph). Each maximal clique defines a unique docking solution. By enumerating the maximal cliques we can find all distinct docking modes of the ligand in the receptor cavity.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

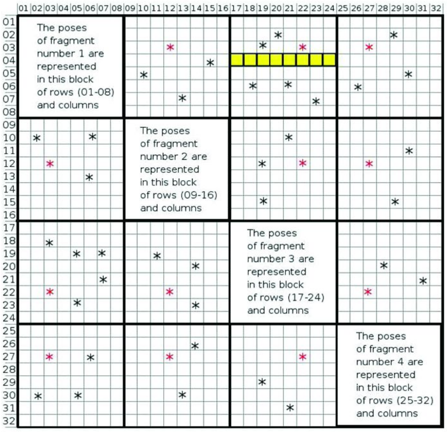

01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
The poses of fragment number 1 are represented in this block of rows (01-08) and columns
The poses of fragment number 2 are represented in this block of rows (09-16) and columns
The poses of fragment number 3 are represented in this block of rows (17-24) and columns
The poses of fragment number 4 are represented in this block of rows (25-32) and columns

Figure 5. Example adjacency bit matrix of the hyper-graph corresponding to a ligand that consists of 4 rigid fragment. For each rigid fragment, there are 8 poses represented in the matrix. Rows and columns 1-8 correspond to poses of rigid fragment 1, 9-16 belongs to rigid fragment 2, 17-24 fragment 3, 25-32 fragment 4. The stars represent fragment pose pairs that are compatible, i.e. not bumping into each other and placed at a distance that can be spanned by the connecting chain fragments.

Figure 5 shows a simple example of an adjacency bit matrix of such a hyper-graph. The matrix *M* can be divided into blocks representing pose combinations between the poses of two specific rigid fragments. The example matrix corresponds to a ligand that contains 4 rigid fragments, and for the sake of simplified example we assume only 8 poses for each fragment. Rows (and columns) 1 to 8 correspond to rigid fragment number 1, rows 9-16 correspond to rigid fragment number 2, etc. The stronger lines indicate the boundary between the blocks that correspond to different rigid fragments. The stars (*) mark the bits that represent edges, i.e. where the column and row index corresponds to compatible pose pairs. The diagonal blocks are empty, because they would correspond to alternative poses of the same node, so they are not compatible, i.e. only one pose can be selected for each node. The task is to find an *S* set of 4 indices such that:

For all i,j in S, $i \neq j$: $M_{i,j} = 1$

The highlighted stars mark the solution maximal clique $S=\{3,12,22,27\}$.

The clique detection algorithm described by Bron and Kerbrosh (*20*) was used as the basis for the pose matching implemented in eHiTS. The original algorithm was improved using the extra information available about the blocked nature of the adjacency bit matrix of our hyper-graph. Note, that if any row *i* contains an empty segment corresponding to any rigid fragment, i.e. If

$$\exists r \in \{0, ..., 3\}, \forall j \in [8r, 8(r+1)-1] : M_{i,j} = 0,$$

then the pose corresponding to row *i* cannot be part of any solution, because there is no suitable pose for rigid fragment *r* that would be compatible with pose

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

i. Rows 1,2,4,5,6,7 and 8 are all examples of such unusable rows (e.g. row 4 has no star in columns 17 through 24 that correspond to the third rigid fragment, this segment of row 4 is highlighted by yellow on the figure). Such rows can all be deleted to reduce the problem size before the recursive (back-track) algorithm is started. Furthermore, during the back-track algorithm, a bit row is maintained that contains the logical *and* operation of the matrix rows corresponding to the currently selected poses. If this bit-row contains an empty segment corresponding to any rigid fragment not yet represented in the clique, then it is not possible to find a completion to the current set, so the whole search tree branch can be cut and the algorithm steps back to choose a different candidate pose for an earlier rigid fragment.

With this problem specific optimization, the algorithm becomes very efficient. In fact the worst case complexity is no longer exponential as it was for the general case, but a polynomial bound can be defined, where the degree of the polynomial is equal to the number of rigid fragments.

Each maximal clique found in the hyper-graph defines a different docking solution by selecting a pose for all the rigid fragments of the ligand in such a way that they do not bump into each other and the distances between them are compatible with the lengths of the flexible chains. The 3D coordinates of all atoms within rigid fragments are defined for every solution and the sum of the scores of the rigid fragments give a very good indication of the total interaction score that can be achieved by each solution. Even though the number of solution cliques may be large (it is several million for some examples), global scoring information is available for them at very low cost (summing up a handful of pose scores), so it is feasible to evaluate them all and select the most promising candidates for further processing.

Note, that selecting a subset of solutions at this point in the process does not compromise the exhaustiveness of the algorithm since the selection is based on global scoring information. All solutions are enumerated exhaustively, the number of PoseMatch solutions is the total number of distinct docking modes possible. The search engine must be exhaustive in order to be able to present all potential solutions to the scoring function for evaluation, as achieved here.

As explained in the scoring section, the full detailed and sensitive scoring function is not employed at this phase, but a faster, crude (greedy) function is employed. The final scoring function has also been tested in the rigid docking phase, but it was found to be inferior to the crude function in selecting the correct poses. This result can be explained by the fact that the final scoring function is too sensitive to precise interaction geometries, therefore it can only differentiate and rank optimized poses correctly.

Flexible Chain Fitting

Following the rigid fragment pose set selection, it becomes necessary to deal with the rotatable bonds joining them, i.e. the challenge of flexible chain fitting. However, this task is much simpler than is the case in the general flexible docking problem, because two atom positions at each end of the chain are already fixed, as they are given by the join atoms of the selected rigid fragment poses.

The task is to find a dihedral angle sequence that will lead from the given starting points to the given end points while avoiding steric clashes with the receptor boundary and the rigid fragments along the way. For smaller chain lengths, even analytical calculation of the complete algebraic solution space would be feasible without considering the steric boundary conditions.

A more general approach has been chosen to find a suitable set of dihedral angles that bridge the distance between the atom pairs and avoids steric clash with the receptor while preferring angles near low energy rotomers. First, a lookup table is used to select initial candidate chain conformers that consist of low energy dihedrals that have ending atom pair distances similar to those required. Then a local minimization is performed to tweak the dihedrals to reach the exact required distances.

For the lookup table, a double diamond lattice is used, which contains all pathways consisting of staggered and gauche dihedrals up to the desired number of bonds. The lattice is positioned on the starting atom pair, then the ending atom pair positions are used to locate nearby atoms in the lattice. The lookup table associated with the diamond lattice contains information about the path lengths (number of bonds from the starting atom) for each atom of the lattice. Any path with the required number of bond that ends within 3Å of the desired 3D coordinates will be considered. A deterministic minimization, based on the partial least squares fit method, is applied to tweak the chain until the end points match precisely and no severe boundary violations occur. This tweaking method may produce *any* dihedral necessary to reach the end points and resolve clashes -- even the highest local energy eclipsed conformation is allowed, if necessary. However, the local optimization starts out with low energy rotomers and will only apply the minimum necessary distortion to resolve steric clashes and bring the end points closer to the goal, so the tweaking process stops with a chain conformation with the lowest energy dihedrals that are suitable for the requirements.

There is no discrete sampling applied in this dihedral refinement process, the precision is only limited by the floating point representation of the computer. Therefore the dihedral angle sampling of eHiTS is practically equivalent to continuous (infinitesimally small) sampling.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Reconstruction and Optimization

When all the flexible chains have been fitted to the rigid fragment poses, the complete ligand is reconstructed from the fragments.

Each hyper-graph clique defines a separate solution. Each solution is constructed by pair-wise joining of the rigid fragments in the selected pose with the flexible chains fitted to them. The mapped pose of each rigid fragment and the resulting conformation of the flexible chain fitting are overlayed using the two atoms that form the broken bond. These two atoms were replicated in both the rigid fragment and the flexible chain, so they can be used to drive the reconstruction.

The flexible chain fitting minimization process attempts to position the last two atoms of the chain to overlay with the target rigid fragment, however, it is not guaranteed that perfect (zero distance) match can be achieved. In other words, the join atoms on the rigid fragments and those on the flexible chain may have different coordinates. Small transformations are carried out on the fragments to achieve complete overlay of the join atoms prior to reassembly of the complete ligand. This step ensures that all bond lengths and angles are maintained from the input structure.

A continuous local energy minimization which only allows torsional changes and rigid body transformations (rotations and translations) is applied to the complete ligand to refine binding geometries and resolve any sampling roughness from the initial polyhedron based rigid fragment positioning. A steepest descent downhill optimization is applied on $6+n$ variables (where n is the number of rotatable bonds) to improve the scoring function value using the modified Powell's algorithm (*21*). The free variables of the optimization correspond to 3 degrees of translation, 3 degrees of rigid body rotation and n degrees of torsional conformation freedom.

The precision of atom positions obtained in this phase are not limited to any discrete sampling, they are again limited only by the precision of the floating point representation of the computer. The optimization is terminated when the scoring function value does not improve in any direction in the $6+n$ dimensional transformation space, i.e. local minimum is reached.

The objective function includes interaction scoring components between the receptor and the ligand, as well as internal intra-molecular interaction components within the ligand and conformational strain energy for the sub-optimal dihedral angles. As a result, eHiTS is capable of generating strained dihedral angles, where necessary, when compensated by the interaction energy - as observed in many experimental crystal structures. However, the program will prefer the low energy conformers when they are suitable for the docking pose.

There is no stochastic element in this applied optimization technique, because the goal is to find a *local* minimum of the objective function for every particular solution. The global coverage of the search space is guaranteed by the full cavity coverage of the rigid fragment docking step and the exhaustive algorithm of the pose matching step.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Protonation Handling

The issue of protonation state is very important to the docking problem. Ligands and receptors with different protonation states can have dramatically different binding poses. However, it is common practice for many docking programs to ignore this issue and require that the user define a particular protonation state prior to running a docking experiment.

Protonation states of ligands and receptors are determined by the interaction between the two. Thus for any particular receptor-ligand pair there will generally be one correct protonation state. However for a different ligand, the protonation state of the receptor may be altered, to reflect the characteristics of the ligand. If a docking program were to pre-set the protonation state of the receptor then possible interactions with a ligand could be lost. Similarly, presetting the protonation states of ligands in a library would produce incorrect results with respect to certain

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

receptors. A better solution, with a more appropriate score, can be found only if the program is run with various protonation states (not necessarily the neutral or the normally lowest energy form of the receptor or ligand on its own or in solvent, but the form required to reach the lowest energy for the complex).

The molecule in Figure 6 has 150 possible protonation states if all combinations are considered for the 4 functional groups that may change protonation states. Figure 7 shows the 5 possible protonation states for functionla groups A and D, 2 for B and 3 for C, combined this leads to 5*5*2*3=150 different possible protonation states. Although, two pairs of states for A and D can be considered equivalent via rotations about the bond to R (swapping the roles of the 2 oxygen atoms), so a flexible docking program could work using only 3 protonation states for those fragments giving a total of 3*3*2*3=54 instead of 150. Most docking programs would need to dock all 150 (or at least 54) combinations separately to evaluate the different possibilities, not even considering different protonation states of the receptor.

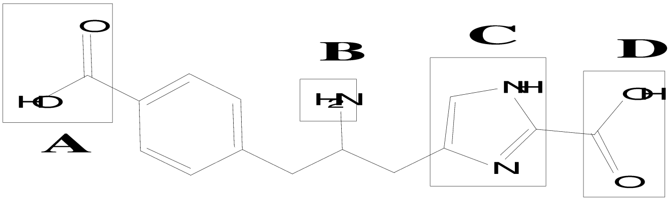


Figure 6. Example ligand with multiple functional groups that may change protonation state upon binding to a receptor site.

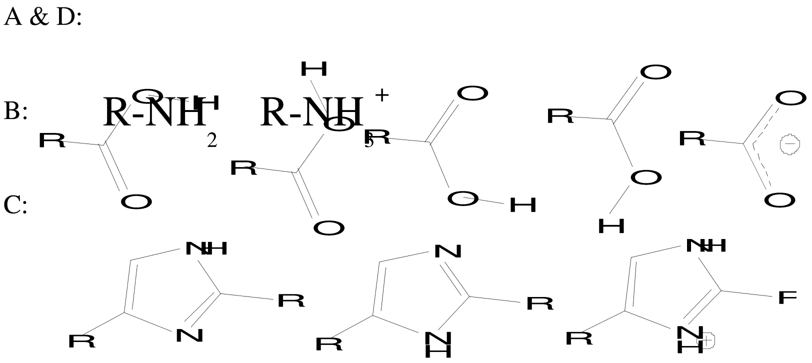


Figure 7. Possible protonation states of the functional groups A,B,C and D.

eHiTS takes a unique approach to the protonation problem by systematically evaluating all possible protonation states for both the receptor and ligand efficiently in a single run. Ambiguous properties flags are assigned for positions that could be either protonated or deprotonated (i.e. have a lone pair). Then during the docking algorithm both states of such surface points are evaluated and scored, selecting the best protonation state for each individual interaction independently, thus avoiding the combinatorial effect of multiple functional groups with variable protonation states. The results of a single eHiTS run using the ambiguous properties flags contain the cummulative results that would be achieved by running many individual docking runs with fixed protonation states considering all ligand protonation states (150 in the above example) against all receptor protonation states (usually an even higher number).

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Scoring Function

Scoring functions in docking programs make assumptions and simplifications in the effort to reach a balance between computational time and accuracy of the results. Essentially there are three classes of scoring functions used in docking programs (*22*, *23*), force-field based, empirical, and knowledge based.

Overview of Different Scoring Approaches

Knowledge based (aka. statistical) scoring functions use statistics collected from experimentally determined protein-ligand complexes to extract rules on preferred and non-preferred atomic interactions. They are designed to reproduce binding poses rather than binding energies. Rules are interpreted as pair-potentials that are subsequently used to score ligand binding poses. Common examples of knowledge based scoring functions include PMF (*24–27*), DrugScore (*28*), SoftScore (*29*), PAS-Dock (*30*) and SmoG (*31*).

Empirical scoring functions consist of the sum of a set of parameterized functions with weights and parameters set to reproduce experimental data, such as binding energies or conformations. The idea is that binding energies can be approximated by a sum of individual uncorrelated terms. The weights of these terms are assigned by regression methods that are used to fit the experimentally determined values found in a training set of protein-ligand complexes. The interaction terms typically have some physical meaning, such as Van der Waals, electorstatics interactions and hydrogen bonds. ChemScore (*32*, *33*), LUDI (*34*), SCORE (*35*), X-Score (*36*), GlideScore (*37*), FlexX (*7*) F-Score (*7*), PLP (*38–40*), SlideScore (*41*), LigScore (*42*) and Fresno (*43*) are all examples of empirical scoring functions.

Force-field based scoring functions are similar to empirical scoring functions, in that they attempt to predict binding energies of ligands by adding individual contributions from different types of interactions. However, force-field based scoring functions use interaction terms derived from physical chemical phenomena as opposed to experimental affinities. Some examples of force-field based scoring functions include D-Score, G-score (*44*), GoldScore (*4*), AutoDock (*45*), Glide-

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Emodel (*46*) and DOCK Energy (*11*, *47*, *48*). These methods rely heavily on correct assignments of partial charge values, which itself is a rather non-trivial task (*49–51*) therefore Shoichet's group has employed a finite-difference Poisson-Boltzmann approach to model electrostatic potential (*52–54*).

Many reviews have been published on the use of various scoring functions in docking programs (*55–63*). The general consensus is that there is currently no universal scoring function that works well across a wide range of protein families and structurally diverse set of ligands.

A new scoring function has been developed with a unique approach to combine the strengths of the statistical and empirical scoring functions. First, an overview of the new scoring method is presented, then the next sub-section describes how statistical information is collected from a large number of crystal structures considering the full distribution of interaction geometries as described by the temperature factors associated with every atom in the crystal structures. Section "Fitting empirical functions to the statistical data" describes how empirical functions are derived from the statistical data to define the final scoring function terms. Validation results are presented in the subsequent section to evaluate the pose ranking and binding energy prediction capabilities of the new scoring function.

The eHiTS Scoring Function

The knowledge based (statistical) scoring functions associate energy with the probability of various interaction patterns based on the Boltzman principle (*64*). Some methods consider only heavy atom distances when collecting the statistics. However positions of hydrogen atoms and directionality is well known to be crucial factors determining the strength of hydrogen bonding interactions. Therefore, it is desireable to collect more detailed geometric information including angles (directionality) and try to determine the probability of various geometric arrangements. Empirical functions typically include angular dependency in their hydrogen bonding terms.

Directionality may also change the nature of the preferred interaction by the same heavy atom. Let's consider a nitrogen atom in an aromatic ring system. The atom presents a strongly polar, hydrogen bonding activity at the edge of the ring, along the direction of its lone electron pair or attached hydrogen atom (one or the other depending on protonation state). Perpendicular to that direction, i.e. above and below the plane of the aromatic ring, the same atom presents a hydrophobic and aromatic π stacking interaction activity.

In order to capture such differences in the scoring function, it was decided that specific interacting surface points (ISP) along with their normal vectors will be used to express the interactions instead of the heavy atom positions. Instead of atom types, the various interaction patterns are categorized based on the set of surface point types listed in Table 2. The surface points are placed on the van der Waals surface of the corresponding heavy atom along the direction of the interaction, e.g. in the direction of the hydrogen atom attached or the central direction of the space occupied by a lone electron pair. In case of a π electron (e.g. ISP-type AromP) the direction is choosen to be perpendicular to the plane of the

sp^2 hybridized electron pairs, e.g. above and below the plane of an aromatic ring. There is an associated direction for each ISP, which is the normal vector of the van der Waals sphere at the given point, i.e. the direction from the heavy atom center towards the ISP.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Table 2. Interaction Surface point type (ISP-type) definitions used in the eHiTS scoring function

ISP-type	***Definition***
Metal	positively charged metal ion interaction point
DonH$^+$	positively charged hydrogen bond donor, e.g. Arginine
Amine	primary amine hydrogen/lone-pair, e.g. -NH$_2$
Don-H	strong (primary) hydrogen bond donor H (polar-atom-H)
WSdon	weak (secondary) hydrogen bond donor H (polarized C-H)
PO$_3^-$	lone pair of negatively charged group
AcidL	lone pair of an acidic functional group, e.g. carboxylate
AccLp	strong (primary) hydrogen bond acceptor lone pair
WS-Lp	weak (secondary) hydrogen bond acceptor lone pair
Ambiv	donor H or acceptor Lp depending on protonation state
Rot-H	rotatable-hydroxy donor H
RotLp	rotatable-hydroxy acceptor Lp
Lipo	H on sp^3 hydrophobic carbon
AromH	H on hydrophobic carbon in aromatic ring (non-polarized)
WSlip	H on weak secondary hydrophobic atom (e.g. carbon next to polar)
Neutr	H/Lp on neutral atom (no recognized activity)
AromP	π electron of an aromatic ring
Res+-	π electron on polar atom (N/O) in resonance chain, e.g. amide
Res-C	π electron on carbon atom in resonance chain, e.g. amide
Sp2+-	π electron on sp^2 polar atom (N/O) (non-resonating, non-aromatic)
Sp2-C	π electron on sp^2 carbon atom (non-resonating, non-aromatic)
Halog	lone electron pair of a halogen atom (F,Cl,I,Br)
Sulfu	lone electron pair of a sulfur atom

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

The interaction geometry between a ligand and a receptor ISP is fully described by four parameters (*65*), see Figure 8 below:

- the distance between the two surface points: d
- the angle between the normal vector of the ligand ISP and the axis connecting the two surface points: α
- the angle between the normal vector of the receptor ISP and the axis connecting the two surface points: β
- the torsion angle (dihedral) between the two normal vectors along the heavy atom axis: δ

The distance and one angle parameter is typically included in the hydrogen bonding term of most empirical scoring functions as well as force fields. Using a single hydrogen bonding angle in the term expresses the importance of the hydrogen directionality while ignoring lone pair directionality, however a recent study (*66*) demonstrated that the later is also important.

For sake of generality, all four geometric parameters are used between any two types of ISP in the scoring function, even though, it is likely that some interaction types (e.g. hydrophobic) do not have significant dependence on the directionality. The actual shape of the geometric dependence function is determined by the statistical data, therefore directionality is not artificially enforced, but simply dictated by the observed data if it is applicable to a given pair of ISP-types.

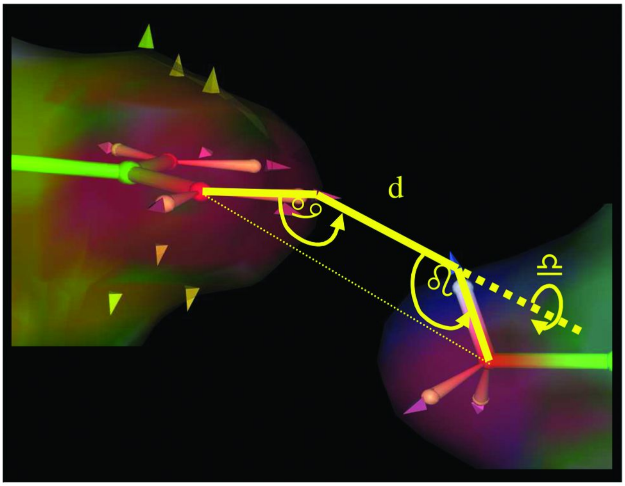


Figure 8. The four values that describe the geometry of a hydrogen bond.

Statistical Data Collection

Interaction geometry statistics has been collected from a set of nearly 2500 high resolution (less than 2.5Å) crystal structures of protein-ligand complexes from the RCSB Protein Data Bank (PDB). The set of complexes (PDB codes) contain the list published in the PDB-bind (*67*) database (1091 codes) for which experimental binding energy is also available. The set also contains the PDB codes from the published Astex validation set (*68*). The ligands in these sets are available in separate files in mol2 format, thus it can be verified that the correct ligand and corresponding binding site is identified by the split utility of eHiTS from the PDB complex. The proteins have been clustered into families based on residue motifs that appear in the binding sites. At least 5 residues match within each family including the distances between the alpha carbons of the residues within 3.0Å tolerance. This clustering method has identified 97 protein families in the set and several hundred singletons that did not match up with any other protein. Then each identified family has been extended from the rest of the PDB using only complexes where the ligand matched all the Lipinski rules for drug likeness (*69*) and the crystal resolution was within the 2.5Å limit.

In standard PDB files, there is a temperature factor (B) associated with the position of each atom. That factor correlates to the probability of that atom having the stated coordinates. The atomic coordinates stated in a PDB file are an average of all "observed" poses and conformers of the protein in the crystal. Each atom may have a slightly different position in each copy of the protein within the crystal. The temperature factor is an indication of how much the atom varies from the mean. Some have a very precisely confined position, while others are more loosely defined. If statistics were collected ignoring the temperature factor information and all atom coordinates were treated as equally significant, then the data would have been misinterpreted failing to take advantage of all the information provided. For statistics collection on, for example, H-bond geometries, it is very important to recognize which arrangements are well defined with low temperature factors and which geometries are mere averages of wide variations. Some geometries occur with high frequency, but if they always occur with high temperature factors, then it does not mean that the specific geometry is really preferred. Imagine a situation where two geometric arrangements are equally likely (e.g. due to two different protonation states) one with a separation distance of 2.8Å and another with a separation distance of 3.6Å. The PDB complex file may contain the *mean* position of the atom at a separation distance of 3.2Å which *never* actually happens in any of the real structures (because it is impossible to have a corresponding "middle" protonation state that would allow that geometry) and so the position considered would have an error of 0.4Å from any feasible, realistic structure. Therefore, it is very important to interpret the crystal structure data as the mean position of a Gaussian probability distribution curve along with the temperature factor that indicates the shape of the curve (*70*). In the given example, the shape of the curve would be flat indicating that the real 3D positions are just as likely as the average position in between - which is still not the true picture, but at least it is much closer to it and the correct position is also considered in the statistics in addition to the artificial average. Furthermore, the probability of the artificial (fake) average

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

position is significantly lower than the probability of another interaction where no such ambiguity occurs. The complete 3D density distribution could be better described by Anisotropic Temperature Factors (*71*), however such data is typically not available in standard PDB files.

Based on the assumed Gaussian distribution of atom positions, the probability of displacement *u* (in any direction) at temperature factor *B* from the given mean position is (*72*):

$$p(u) = \left(\frac{B}{4\pi}\right)^{-3/2} \exp\left(-\frac{1}{2}\frac{8\pi^2 u^2}{B}\right)$$

From this, we can express the probability of an atom being on a sphere surface at given radius *d* around the mean position using an integral on the sphere surface *A*, which can be computed using spherical coordinate system (α planar angle and β azimuth for a spheric point (x_s, y_s, z_s)):

$$\begin{aligned}
p(d) &= \int_A p(r_A)\, d^2A \\
x_s &= d\sin(\alpha)\cos(\beta) \\
y_s &= d\sin(\alpha)\sin(\beta) \\
z_s &= d\cos(\alpha) \\
r_s^2 &= x_s^2 + y_s^2 + z_s^2 \\
p(d) &= \int_0^{\pi}\int_0^{2\pi} p(r_s) \left|\frac{\partial r_s}{\partial \alpha} \times \frac{\partial r_s}{\partial \beta}\right| d\alpha\, d\beta
\end{aligned}$$

The determinant of the partial derivates can be easily computed and the probability function for the displacement substituted, so the probability is expressed as:

$$p(d) = \left(\frac{B}{4\pi}\right)^{-3/2} \int_0^{\pi}\int_0^{2\pi} \exp\left(\frac{-4\pi^2 r_{\alpha\beta}^2}{B}\right) d^2 \sin(\alpha)\, d\alpha\, d\beta$$

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Let us introduce the following auxilary notations to simplify the forthcoming formulae:

$$\begin{aligned} F_0 &= \left(\frac{B_0}{4\pi}\right)^{-3/2} \\ F_1 &= \left(\frac{B_1}{4\pi}\right)^{-3/2} \\ C_0 &= \frac{-4\pi^2}{B_0} \\ C_1 &= \frac{-4\pi^2}{B_1} \\ r_0^2 &= (x-x_0)^2+(y-y_0)^2+(z-z_0)^2 \\ r_1^2 &= (x+x_s-x_1)^2+(y+y_s-y_1)^2+(z+z_s-z_1)^2 \end{aligned}$$

Using the above equations and notations, the probability of a given pair of interacting atoms (with mean positions (x_0,y_0,z_0) and (x_1,y_1,z_1), temperature factors B_0 and B_1 respectively) being at a distance d can be expressed via the following volumetric integral:

$$P(d) = \int_V \int_0^{\pi} \int_0^{2\pi} F_0 \exp(C_0 r_0^2) F_1 \exp(C_1 r_1^2)\, d^2 \sin(\alpha)\, dV\, d\alpha\, d\beta$$

The volumetric integral can be expressed as three independent coordinate integrals from negative to positive infinity on x,y,z. The constants F_0 and F_1 can be brought outside the integrals, and the integration order can be changed, so that the following integration fact can be used via variable substitution:

$$\int_{-\infty}^{\infty} \exp\left(\frac{-1}{2}at^2\right)\, dt = \sqrt{\frac{2\pi}{a}}$$

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Let us reorder the integration so that the integral on variable x is the innermost and factor out everything that is not dependent on x from the integral

(Let F= $F_0\, F_1\, d^2$):

$$
\begin{aligned}
P(d) &= F \int\!\!\int \int_0^{\pi} \int_0^{2\pi} \sin\alpha \exp(-C_0((y-y_0)^2 + (z-z_0)^2) \\
&\quad -C_1((y+y_s-y_1)^2 + (z+z_s-z_1)^2)) f_x(d)\, dy\, dz\, d\alpha\, d\beta \\
f_x(d) &= \int \exp(-C_0(x-x_0)^2 - C_1(x+x_s-x_1)^2)\, dx \\
a &:= C_0 + C_1 \\
b &:= C_1(x_s - x_1) - C_0 x_0 \\
t &:= x + \frac{b}{a} \\
f_x(d) &= \exp(\frac{b^2}{a} - C_0 x_0^2 + C_1(x_s - x_1)^2) \int \exp(-at^2)\, dt \\
f_x(d) &= \exp(\frac{b^2}{a} - C_0 x_0^2 + C_1(x_s - x_1)^2) \sqrt{\frac{\pi}{a}}
\end{aligned}
$$

Using similar rearrangament and substitution steps for variable y and z, the closed form of their integrals can also be computed, so that we can express $P(d)$ with only 2 integrals on variables α and β with the constant expressions simplified to the form:

$$
P(d) = \left(\frac{4\pi}{B_0 + B1}\right)^{\frac{3}{2}} d^2 \int_0^{\pi} \int_0^{2\pi} \exp\left(\frac{-4\pi^2}{B_0 + B_1} \|P_0 - P_1 + P_s\|^2\right) \sin\alpha\, d\alpha\, d\beta
$$

Where $P_0=(x_0,y_0,z_0)$ is the mean position of the first atom with temperature factor B_0, $P_1=(x_1,y_1,z_1)$ is the mean position of the second atom with temperature factor B_1 and $P_s=(x_s,y_s,z_s)$ is a point of a sphere with radius d identified by spheric coordinates α and β as given in the earlier equations.

Similarly, the probability of interactions to occur at given angle and dihedral parameter values can also be expressed with volumetric integrals considering the Gaussian probability distribution of the atom positions according to the temperature factors given.

The statistics collection was performed on all pairs of interacting atoms from the nearly 2500 protein ligand complexes. There are several hundred interacting atom pairs in each complex giving rise to more than a million interactions in total. For each ISP pair, a four dimensional probability array was accumulated by evaluating the volumetric integrals for every possible parameter quadraple value set (distance, 2 angles and a dihedral) within the full range of angle values and distance values ranging from zero to 5.6Å using a fine resolution sampling for the numeric integration.

Fitting Empirical Functions to the Statistical Data

The four dimensional probability array accumulated from the observed experimental data defines the shape of the geometric dependency functions for any given ISP-type pair. However, using the accumulated data directly for scoring is impractical for both memory and CPU resource requirement reasons, i.e. the data tables are too large to fit into physical memory and the lookup process would require expensive interpolations on the 4D array cells.

Many graphical plots of the collected data were examined and various statistical analysis techniques were applied, then it was determined that the data can be approximated with relatively simple analytical functions. The following parameterized formula was chosen to represent the geometric dependency terms of the interactions in the eHiTS scoring function:

$$g(d,\alpha,\beta,\delta) = e_0\; s(d)\; l(\alpha)\; r(\beta)\; t(\delta) \tag{1}$$

$$s(d) = p_{10}d + p_{11}d^2 + p_{12}d^3 + p_{13}a(d) + p_{14}c(d) + p_{15} \tag{2}$$

$$a(d) = p_8(d - p_9)^2 \tag{3}$$

$$c(d) = \begin{cases} \cos(a(d)) & \text{if } a(d) > -\pi \wedge a(d) < \pi \\ -1 & \text{otherwise} \end{cases} \tag{4}$$

$$l(\alpha) = p_0 \cos\alpha + p_1 \cos^2\alpha + p_2\sqrt{\cos\alpha} + p_3 \tag{5}$$

$$r(\beta) = p_4 \cos\beta + p_5 \cos^2\beta + p_6\sqrt{\cos\beta} + p_7 \tag{6}$$

$$t(\delta) = p_{16} \cos\delta + p_{17} \cos^2\delta + p_{18}\sqrt{\cos\delta} + p_{19} \tag{7}$$

In the above scoring function formula, e_0 is the energy coefficient for the given interaction, which depends on the pair of ISP-types that participate in the interaction, while parameters $p_0,\ldots,p_{19}$ are fitted to reproduce the accumulated statistical probability data as closely as possible. The fitting process was carried out using a simulated annealing method combined with a simplex minimizer (*21*) and followed by a modified Powell local minimizer.

The values for the energy coefficient e_0 in the ISP-type matrix has been determined based on the Boltzmann principle, i.e. the ratio of observed interactions versus the probability of such interaction occurring by random placement of the atoms is calculated, then the Maxwell-Boltzmann exponential distribution function was used to convert the probability into energy value.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

The pairwise interactions between interacting surface points (ISP) are scored using the empirical function $g(d,\alpha,\beta,\delta)$ which is dependent on the four variables that describe the interaction geometry as shown in Figure 8. The score is accumulated by summation over all interacting point pairs including all types of interactions listed in Table 2:

$$E_{inter} = \sum_{i \in ISP_L} \sum_{j \in ISP_R} g(d(i,j), \alpha(i,j), \beta(i,j), \delta(i,j))$$

Additional Terms of the Scoring Function

The above described interaction scoring terms account for all kinds of interactions between the receptor and the ligand, including hydrogen bonding, metal ion interactions, hydrophobic interactions, aromatic π stacking, etc. However, the eHiTS scoring function also contains additional terms to account for de-solvation effects, steric clashes, topological positioning of the ligand in the binding pocket, entropy effects, ligand conformation strain energy and intramolecular interactions within the ligand. The calculation methods of these additional terms are described in this section.

De-Solvation Term

The de-solvation term of the eHiTS scoring function is based on a continuous solvation model. It is assumed that each ISP was interacting with a solvent water molecule prior to complex formation, thus a solvent interaction energy value can be associated with each surface point. To form the receptor-ligand complex, de-solvation of both molecules had to occur thus the energy components representing the solvent interactions should be removed, i.e. subtracted from the total score:

$$E_{desolv} = - \sum_{i \in ISP_L} e_{solv}(i_{type}) - \sum_{j \in ISP_R} e_{solv}(j_{type})$$

The function $e_{solv}(t)$ associates a different constant energy value with each ISP type, i.e. it is recognized that different solvation energy corresponds to different surface point types. For example, polar and hydrogen bonding ISP types would have a favourable interaction with solvent water while hydrophobic ISP types would have an energy constant with opposite sign representing unfavourable interaction.

The solvation energy constants per ISP type were also determined using the Bolztmann formula based on statistics collection of the ISP occurances that are exposed to solvent in the protein ligand complexes. The total number of surface points of each type was counted, as well as the number of solvent exposed surface points of each type. The ratio of exposed counts versus the total counts reflects the probability of a certain type of surface point being exposed which is then converted into energy value based on the Boltzmann formula.

Van der Waals Term

The usual Lennard-Jones 6-12 potential is used to estimate the van der Waals energy term between receptor and ligand atoms (where r_a and r_b refers to the van der Waals radii of atoms a and b respectively):

$$\begin{aligned} E_{vdw} &= \sum_{a \in L, b \in R} v(a,b) \\ v(a,b) &= \left[\left(\frac{r_a + r_b}{d(a,b)} \right)^{12} - \left(\frac{r_a + r_b}{d(a,b)} \right)^{6} \right] \end{aligned}$$

Due to approximations in docking pose generation and imprecisions in crystallographic data, the repulsive term of the 6-12 potential can be prohibitively overwhelming, limiting the practical use of the scoring function to clean, well optimised structure complexes. This limitation can be overcome by the application of cut-off value on the E_{vdw} component and the use of a less prohibitive steric clash penalty function. A zero cut-off has been chosen for the van der Waals term thus limiting its meaning to attractive contribution. The repulsive effects of too close contacts are handled by a separate term described below.

Steric Clash Penalty Term

One problem with the repulsive term of the Lennard-Jones 6-12 potential is explained above, i.e. it is too steep and easily overwhelms all other score components even for relatively small steric clashes that may occur in docking poses due to approximate placements as well as in X-ray crystallographic data due to poor resolution of the experimental data. This problem could be solved by using a function potential with a lower power, e.g. Cubic or quadratic.

A second problem is the large number of local minima generated by the repulsive term of the Lennard-Jones 6-12 potential due to the spatial arrangement of receptor atoms. It is a simple mathematical fact, that every 4 receptor atoms that are not co-planar would generate a local minimum at the center of the tetrahedron formed by the 4 atoms. An average protein structure in the PDB consists of over 4000 heavy atoms, picking a random 4 of them is most likely not coplanar and there are

$$\binom{4000}{4} \approx 10^{13}$$

ways to pick 4 out of 4000. Consequently, such potential generates several trillions of local minima in the scoring function landscape which makes local energy minimization (to find the optimal pose with the best scoring function value) practically hopeless, i.e. equivalent to random trial and error search in a vast space.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

To eliminate both problems at once, steric clash term of the eHiTS scoring function uses the distance-square from the Connolly surface (C) of the receptor:

$$\begin{aligned} E_{clash} &= \sum_{a \in L} s(a, C) \\ s(a, C) &= \begin{cases} d^2(a, C) & \text{if } a \text{ is violating surface } C \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Mathematically, this function has infinite number of local minima, since all points without steric violation have the minimum value of zero (which is also the global minimum). However, all the minimum points form a single continuous region (the binding pocket) for most practical cases, thus it can be considered as a single minimum (not a point but a 3D region). The main advantage of this type of clash function is that for any ligand atom position that has steric violation, the gradient of the penalty function points towards the closest point of the Connolly surface where the violation can be resolved with the smallest amount of atom movement, thus it helps to direct the local minimisation process to resolve all violations.

In contrast, the repulsion term of the Lennard-Jones 6-12 potential can “trap” atoms if the steric violation is so severe that the ligand atom centre moves beyond the plane of 3 receptor atoms in close vicinity.

Pocket Depth Term

Typically many docking poses are generated by eHiTS, some of them deep inside the binding pocket while others are binding on the outer surface or shallow indentations of the receptor. It has been observed that other scoring terms are often not able to distinguish the correct deep binding pose from some of the well scoring surface binding poses. Therefore, another scoring term was introduced to reflect the topology of the ligand binding pose with respect to the pocket depth.

$$\begin{aligned} E_{depth} &= \sum_{a \in L} h(a, C) \\ h(a, H) &= \begin{cases} -d(a, H) & \text{if } a \text{ is inside convex hull } H \\ +d(a, H) & \text{otherwise} \end{cases} \end{aligned}$$

The surface \$H\$ in the formula above represents the convex hull of the receptor (protein) structure.

Family Coverage Term

Statistical information is collected per protein family during protein family based training (see next subsection for details). The information includes data about surface point type coverage pattern, i.e. what percentage of various ligand

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

surface point types are interacting with each receptor surface point type. This statistical information is used in scoring to compute the family coverage term:

$$\begin{aligned} E_{family} &= \sum_{j \in ISP_R} f(j, i_j) \\ i_j &= \{i \in ISP_L | i\text{interacts with} j\} \\ f(j,i) &= \{\text{frequency of } (j,i) \text{ interactions in the family}\} \end{aligned}$$

This scoring component is ignored (has a value of zero) if the protein receptor is not recognised as belonging to any of the trained protein families.

Ligand Strain Energy Term

There is substantial statistical evidence (*15–17*) that ligands often bind in conformations that differ significantly from the lowest energy conformation of the unbound ligand (either in vacum or in aquaous solution). The docking engine of eHiTS is capable of generating practically any conformation necessary to satisfy the goal interactions picked by the rigid fragment docking phase. However, the scoring function must account for the energy penalty (strain energy) associated with any given conformation during the final local minimisation phase (which alters the conformation of the ligand) as well as the strain energy of the final docking pose:

$$\begin{aligned} E_{strain} &= \sum_{b \in L_{rot}} dih(b) + \sum_{a,b \in L} nbv(a,b) - E_{base} \\ L_{rot} &= \{\text{rotatable single bonds of the Ligand}\} \\ dih(b) &= \{\text{strain energy of dihedral angle of } b\} \\ L_B &= \{\text{bonds of the Ligand}\} \\ nbv(a,b) &= \begin{cases} 0 & \text{if}(a,b) \in L_B \vee \exists c \in L : (a,c) \in L_B \wedge b(c) \in L_B \\ v(a,b) & \text{otherwise} \end{cases} \end{aligned}$$

The function *dih(b)*, which determines the strain energy of the dihedral angle of the bond *b* is also based on statistical data collected from the Protein DataBank (PDB) (73,74). All bound ligand conformations of the PDB has been analyzed to collect statistics on the dihedral angles of single rotatable bonds. Data has been clustered considering the hybridisation states of the atoms at the end of the bond as well as the number and type of heavy atom neighbours of the atoms. Within each cluster, the dihedral angle with the highest occurance frequency is considered to have the lowest energy (ground state). Strain energy for other dihedrals is assigned based on the frequency ratio between the given dihedral and the most frequent one again using the Boltzmann formula to convert probability into energy value.

The formula for E_{strain} in the previous equation contains the base energy E_{base} for the ligand, because the two sums would not yield a zero value for the lowest energy conformation of the ligand. The score component needs to be adjusted to establish a zero baseline for the term, otherwise weighting of the term relative to the other components in the full scoring function becomes dependent on the actual ligand structure. To establish the value of E_{base} for each ligand, several hundred conformations are generated by systematically sampling the conformational space using multiple low energy dihedral values (e.g. considering gauche dihedral angles in addition to staggered ones), then a local minimisation is performed from each generated conformation, with a goal function similar to the strain equation, except omitting the E_{base} term. The lowest energy found by this search is chosen for the value of E_{base} for the given ligand. This procedure is executed once for each ligand during the preprocessing phase, then the computed E_{base} value is used for scoring of various docking poses during the eHiTS pose generation and final optimisation.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Ligand Intra-Molecular interactions

Intra-molecular interaction may occur in certain conformations of bound ligands. Such interactions can stabilise the conformation and contribute to the total energy of the protein-ligand complex, therefore it is important to consider them in the scoring function. The eHiTS scoring function contains a term E_{int} for this purpose, and it is computed the same way as the receptor-ligand interaction term, except that both ISPs are taken from the ligand in the sum.

Ligand Entropy Term

The binding free energy of the protein ligand complex is by definition the difference between the total energy of the complex and the sum of the energy of the receptor and ligand in solution. However, all of these energy terms consist of two components: enthalpy and entropy. So far, all the scoring function terms described deal with the enthalpy component, but we also need to consider the change of entropy upon binding. The entropy change of the ligand can be attributed mostly to the loss of entropy upon binding due to freezing of freely rotatable single bonds in the ligand:

$$E_{entropy} = \sum_{b \in L_{rot}} e_{rot}$$

Currently, the eHiTS scoring function uses a simple constant value for e_{rot}, although the real entropy loss may vary for various rotatable bonds due to conformational constrains. Furthermore, not every rotatable bond is necessarily fully frozen upon binding, partial or even full freedom of movement may remain for some of the rotatable bonds.

The entropy change of the protein receptor is much harder to estimate and would require costly entropy analysis of both the bound and unbound conformations. There is no scoring term in the eHiTS scoring function to estimate the entropy change of the receptor.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Protein Family Recognition and Clustering

There is a component (E_{family}) in the eHiTS scoring function that requires the recognition of the protein family that the receptor structure belongs to. The protein family classification is automated in eHiTS and is based on the geometric pattern of residues at the active site. The amino acid type of the protein residues forming the active site is collected for every protein and the 3D coordinates of the center of mass of each residue is used to compute a distance matrix between the participating residues. Clustering of proteins into families is performed based on the similarity of the distance matrices. The minimum criteria for two proteins to fall into the same family is to have at least 5 residues with all their pairwise distances compatible, the tolerance for the distance difference is 3.0Å. There are two control parameters for the clustering:

- number of residues required to match within a family (default 5)
- tolerance for the distance difference between residue pairs (3.0Å)

Choosing different values for these two parameters yields different number of families as a result of clustering and the population of the families also varies. A large range of values have been tested for both parameters and the clustering results were analyzed with respect to the categorization stated in the comment and header sections of the PDB entries. The mentioned default values (5 and 3.0Å respectively) were chosen, because these values result in clusters with close agreement with the stated classification.

The family knowledge base has been prepared based on about 2500 PDB codes, yielding 97 distinct families with at least 5 members in each family. There were also 349 singletons that did not fit any of the identified families and did not form large enough clusters to designate them as additional families. When the eHiTS software is run for a given protein receptor, the family of the receptor is determined by the same procedure as the clustering was performed, i.e. The residues at the active site surface are identified and their distance matrix is computed. The distance matrix is checked against the matrices stored in the family knowledge base. If a match is found according to the parameters, then the family is recognized and the corresponding statistical data and weight set will be used for the scoring. Otherwise the global weight set is used which has zero weight associated with the E_{family} term.

Tuning the Weight Parameters

The full scoring function is composed as a weighted sum of the statistically derived empirical interaction term and the various other terms detailed in the previous section:

$$E = w_0 E_{inter} + w_1 E_{desolv} + w_2 E_{vdw} + w_3 E_{clash} + w_4 E_{depth} + \\ w_5 E_{family} + w_6 E_{strain} + w_7 e_{intra} + w_8 E_{entropy}$$

The weights $w_0,...,w_8$ are tuned so that the resulting value of E corresponds to an estimated binding affinity $log(K_i)$ value. The tuning could be performed using simple linear regression technique to fit the score of X-ray ligand poses to experimentally measured binding affinity data. However, the purpose of the eHiTS scoring function is to evaluate large number of docking poses of a ligand and select the best pose in addition to providing a binding energy estimate that can be used for screening ligand databases. Furthermore, the scoring function is also driving the local optimisation of the solution poses and conformations in the final phase of the eHiTS docking system. Therefore, the weight set should be optimised considering all of the following goals:

1. Local minimisation on the scoring function of a set of poses generated within a small RMSD radius around the X-ray pose should converge to single location as close as possible to the X-ray pose, i.e. the scoring function should have a funnel shape with a local minimum near the X-ray pose.
2. When all the optimized solution poses are ranked by the scoring function, the best pose should have as small RMSD as possible from the X-ray pose of the ligand, i.e. the scoring function should be able to identify the correct pose.
3. The score value (of the X-ray pose) should have a good correlation with the experimental binding energy.
4. The score values generated for known active ligands should be superior to the score values of docked inactive ligands (decoys).

If the weight values are tuned specifically for any of the listed four goals, they will usually not yield good results from the perspectives of the other goals. Therefore, the weight tuning procedure has to consider all objectives at once, i.e. the goal function of the weight tuning must include terms to reflect each of the above listed goals. The correlation requirement is the only one easily expressed with linear function, but the others (e.g. with the RMSD requirements) are inherently nonlinear. Therefore, instead of the linear regression technique, the modified Powell optimisation engine was selected to perform the weight tuning. A stochastic method based on simulated annealing was also tested but it did not produce better results than the Powell engine and the run time was considerably longer.

The tuning was performed once for all the 2500 PDB codes together to generate the default (global) weight set, then separate weight tuning runs were performed for all identified 97 protein families individually. Not all training data provide equally valuable information, because of variations in crystallographic resolution as well as experimental binding energy. Furthermore, some of the ligands are drug-like according to the Lipinski rules of 5 (about 60% of the data), while others are not likely to be of interest to the pharmaceutical industry. An importance weight was associated with each PDB code in the training data to reflect these differences.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Results

To test the accuracy of the pose generation engine and the ability of the scoring function to recognise the correct pose, the Astex diverse validation set (*68*) was chosen, which contains a single representative PDB complex for each protein family covered by the set. This set is also filtered for crystallographic accuracy, therefore the data can be considered more reliable than a random subset of the PDB.

It is important to note that no manual preprocessing was performed on any of the selected PDB complexes. The protonation states, cofactors, counter-ions, solvent molecules, partial charge assignment, etc. were all handled by eHiTS without user intervention. This automation makes eHiTS very user-friendly and capable of automated processing.

The ligands were docked into the original protein binding site (as provided in the X-ray structure) and the accuracy was measured by calculating the root-mean-squared deviation (RMSD) between the coordinates of the heavy atoms of the ligand in the eHiTS docked pose and those in the crystal structure. The results are summarized in Figure 9, which shows a receiver operating characteristic (ROC) type of curve corresponding to the success rate achieved: the X axis corresponds to the RMSD accuracy values, and the plot shows the success rate in percentage on the Y axis. The higher curve shows the success rate of the pose closest to the X-ray structure out of the 32 generated output poses, while the lower curve shows the success rate of the top-ranked pose. A typical cut-off value for successful pose generation is considered to be 2Å. The top-rank pose is within 2Å for 85% of the cases, while the closest is for 95%.

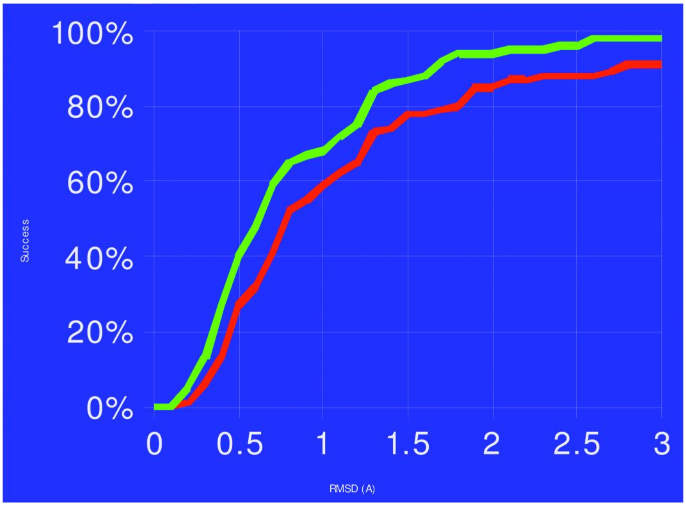


Figure 9. RMSD-success rate curve on the Astex diverse 85 set.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Enrichment Results

The most practical use of the docking methods is for virtual screening of large libraries of ligands. The screening performance of the docking/scoring protocols is often expressed as the enrichment of actives in a selected portion of the database, or the percentage of known actives found in the top few percent of the ranked list of all ligands. These numbers do not provide comparable absolute measures, but highly dependent on the actual data set. We have measured eHiTS performance on 2 published datasets. Figure 10 shows the screening results of eHiTS for the data set published by Hongming Chen *et.al.* (*73*) – the publication reports results of 7 other screening techniques with average enrichment factors ranging from 1.44 (Gold) to 7.43 (ICM). The average enrichment factor produced by eHiTS on the same dataset is 7.56, i.e. Better than all the reported methods.

The results of eHiTS on the published Surflex data set are shown in Figure 11. This dataset contains 869 decoys plus actives specific for each family (ranging from 5 to 20 molecules). The results show a remarkable enrichment across a wide range of receptor families with an average recovery rate of ~80% of all actives in the top 10% of the ranked database.

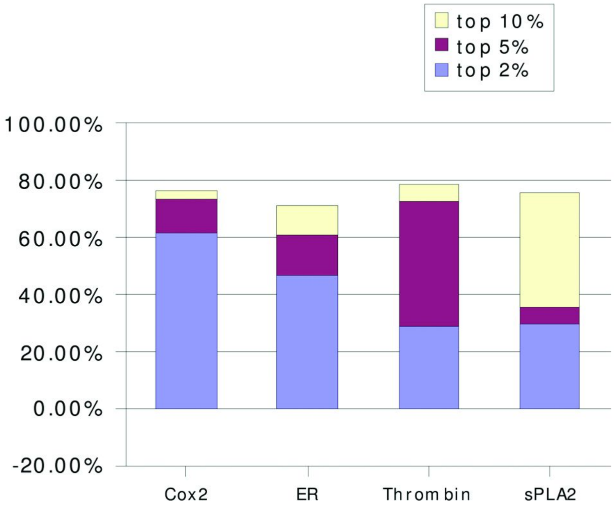


Figure 10. Enrichment results on Hongming Chen et.al. (73) data.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

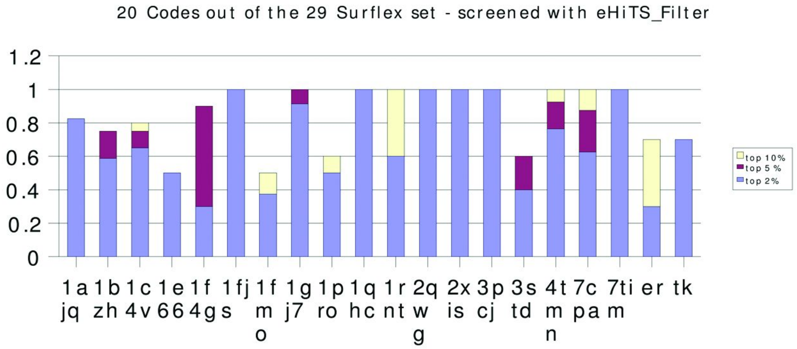


Figure 11. Enrichment results in the Surflex data set.

Correlation with Experimental Binding Energy

One of the most important measure of success for a scoring is the correlation between the calculated score value and the experimental binding energy. It is also the most difficult aspect of scoring. Table 3 contains the results for various test data sets: untrained data for comparison base-line, results of all data after global training, results of splitting the data into disjunct training set and test set. There is an additional line (Xray optimised) that corresponds to the result obtained when the training is done purely for this measure, i.e. the goal function is limited to optimizing the correlation and ignore the other 3 aims. In this case, significantly better correlation can be achieved, but the weight set obtained this way is less suitable for use in the docking software. The last column of the table reports the root mean square error of the estimate in pK_i units.

Figure 12 shows a scatter plot of experimental binding affinity versus scores of docked poses and X-ray ligand pose as a result of global training with all four goal components considered (funnel shape, ranking ability, correlation and enrichment). The X-ray ligand pose data of the plot corresponds to the second data line of Table 3.

Table 3. Correlation of score values and the experimental binding energy

Test data set	***Correlation (R)***	***Error (rms)***
All untrained X-ray	0.122	6.157
All trained X-ray	0.539	2.233
Training set (half of cases) X-ray	0.564	2.297
Cross validation set X-ray	0.511	2.431
Xray optimised	0.751	1.613

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

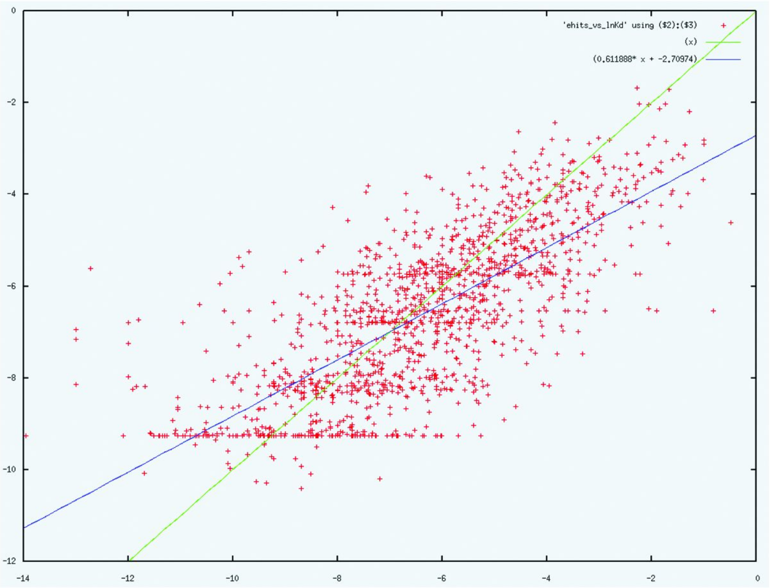


Figure 12. Correlation of eHiTS-score with experimental binding data.

Conclusions

The eHiTS flexible ligand docking engine has been described along with a new statistical based scoring function. The search engine is based on exhaustive positioning and then re-linking of rigid fragments that often correspond to chemical functional groups. This approach provides the means to achieve computationally feasible *complete* search of the conformational and pose space with sufficient resolution, providing high accuracy.

The presented scoring function is capable of selecting a good representative pose from the generated candidates, although not always the best. It is also capable of identifying active compounds, leading to good enrichment. However, accurate estimation of the binding free energies remains a challenging problem, needing further research and improvements.

References

1. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Discovery* **2000**, *3*, 935–949.
2. Goodsell, D. S.; Olson, A. J. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 195–202.
3. Hart, T. N.; Ness, S. R.; Read, R. J. *Proteins* **1997**, *Suppl. 1*, 205–209.
4. Liu, M.; Wang, S. *J. Comput.-Aided Mol. Des.* **1999**, *1*, 435–451.
5. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727–748.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

6. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
7. Westhead, D. R.; Clark, D. E.; Murray, C. W. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 209–228.
8. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A. *J. Mol. Biol.* **1996**, *261*, 470–89.
9. Welch, W.; Ruppert, J.; Jain, A. N. *Chem. Biol.* **1996**, *261*, 449–462.
10. Kearsly, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 565–582.
11. DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L. *J. Med. Chem.* **1988**, *31*, 722–729.
12. McGann, M.; Almond, H.; Nicholls, A.; Grant, J. A.; Brown, F. *Biopolymers* **2003**, *68*, 76–90.
13. Jeffrey, G. A.; Saenger, W. *Hydrogen Bonding in Biological Structures*, Springer Verlag: Heidelberg, 1991.
14. Cambridge Structural Database. http://www.ccdc.cam.ac.uk/products/csd/.
15. Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. *Bioorg. Med. Chem.* **1995**, *3*, 411–428.
16. Bostrom, J.; Norrby; Per-Ola; Liljefors, T. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 383–396.
17. Bostrom, J. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137–1152.
18. Perola, E.; Walters, W. P.; Charifson, P. S. *Proteins: Struct., Funct., Genet.* **2004**, *56*, 235–249.
19. Doob, J. L. *Ann. Math.* **1942**, *43*, 352–369.
20. Bron, C.; Kerbosch, J. *Commun. ACM* **1973**, *16*, 575–577.
21. Press, W. H.; Teukolsky, S. A.; Flannery, W. T.; Vetterling And, B. P. *Commun. ACM* **2002**, *16*, 575–577.
22. Cornell, W. D. *Commun. ACM* **2006**, *2*, 297–323.
23. Ajay, J. N. *J. Comput.-Aided Mol. Des.* **2006**, *7*, 407–420.
24. Muegge, I.; Martin, Y. C. *J. Med. Chem.* **1999**, *42*, 791–804.
25. Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik, S. W. *J. Med. Chem.* **1999**, *42*, 2498–2503.
26. Muegge, I. *Perspect. Drug Discovery Des.* **2000**, *20*, 99.
27. Muegge, I. *J. Comput. Chem.* **2001**, *22*, 418–425.
28. Gohlke, H.; Hendlich, M.; Klebe, G. *J. Mol. Biol.* **2000**, *295*, 337–356.
29. Yang, C. Y.; Wang, R.; Wang, S. 2005.
30. Tondel, K.; Anderssen, E.; Drablos, F. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 131–144.
31. DeWitte, R. S.; Shakhnovich, E. I *J. Am. Chem. Soc.* **1996**, *118*, 11733–11744.
32. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins* **2003**, *52*, 609–623.
33. Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 593–606.
34. Wang, R.; Liu, L.; Lai, L.; Tang, Y. *J. Mol. Model.* **1998**, *4*, 379–394.
35. Wang, R.; Lai, L.; Wang, S. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11.
36. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Blanks, J. L. *J. Med. Chem.* **2004**, *47*, 1739–1749.

In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

37. Gelhaar, D. K.; Rejto, G. M.; Verkhivker, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. *Chem. Biol.* **1995**, *2*, 317.
38. Gelhaar, D. K.; Bouzida, D.; Rejto, P. A.; Parril, L.; Reddy, M. R. *Chem. Biol.* **1999**, *2*, 292–311.
39. Verkhivker, G. M.; Bouzida, D; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731–751.
40. Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 883–902.
41. Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. *J. Mol. Graphics Modell.* **2003**, *21*, 289–307.
42. Rognan, D.; Lauemoller, S. L.; Holm, A.; Buus, S.; Tschinke, V. *J. Med. Chem.* **1999**, *42*, 4650–4658.
43. Kramer, B.; Rarey, M.; Lengauer, T. *Proteins* **1999**, *37*, 228–241.
44. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelly, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739–1749.
45. Meng, E. C.; Gschwend, D. A.; Blaney, J. M.; Kuntz, I. D. *Proteins* **1993**, *17*, 266–278.
46. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411–428.
47. Rappe, A. K.; Goddard, W. A., III *J. Phys. Chem.* **1991**, *95*, 3358–3363.
48. Ogawa, T.; Kitao, O.; Kurita, N.; Sekino, H.; Tanaka, S. *Chem-Bio Inf. J.* **2003**, *3*, 78–85.
49. Kitao, O.; Ogawa, T. *Mol. Phys.* **2003**, *101*, 3–17.
50. Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. *J. Mol. Biol.* **2002**, *322*, 339–355.
51. Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. *J. Mol. Biol.* **2004**, *337*, 1161–1182.
52. Nicholls, A.; Honig, B. *Science* **1995**, *268*, 1144.
53. Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. *J. Med. Chem.* **2005**, *48*, 962–976.
54. Krovat, E. M.; Steindl, T.; Langer, T. *Curr. Computer-Aided Drug Des.* **2005**, *1*, 93–102.
55. Mohan, V.; Gibbs, A. C.; Cummings, M. D.; Jaeger, E. P.; DesJarlais, R. L. *Curr. Pharm. Des.* **2005**, *11*, 323–333.
56. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorth *J. Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.
57. Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. *J. Med. Chem.* **2004**, *47*, 558–565.
58. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. *J. Med. Chem.* **2004**, *47*, 3032–3047.
59. Schulz-Gasch, T.; Stahl, M. *J. Mol. Model* **2003**, *9*, 47–57.
60. Perola, E.; Walters, W. P.; Charifson, P. S. *Proteins: Struct., Funct., Genet.* **2004**, *56*, 235–249.
61. Bohm, H. J.; Stahl, M.; Lipkowitz, K. B.; Boyd, D. B. *Proteins: Struct., Funct., Genet.* **2002**, *18*, 41–87.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

62. Sippl, M. J. *J. Comput.-Aided. Mol. Des.* **1993**, 473–501.
63. Chen, Y.; Kortemme, T.; Robertson, T.; Baker, D.; Varani, G. *Nucleic Acids Res.* **2004**, *32*, 5147–5162.
64. Kortemme, T.; Morozov, A. V.; Baker, D. *J. Mol. Biol.* **2003**, *326*, 1239–1259.
65. Wang, R.; Fang, Y. Lu; Wang, X. S. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114–2125.
66. Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 457–471.
67. Lombardo, C. A.; Lipinski, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delevery Rev.* **2001**, *46*, 3–26.
68. Dunitz, J. D.; Schomaker, V.; Trueblood, K. N. *J. Phys. Chem.* **1988**, *92*, 856–867.
69. Schneider, T. R. *J. Phys. Chem.* **1996**, *92*, 133–144.
70. Willis, B. T. M.; Pryor, A. W. *J. Phys. Chem.* **1975**, *92*, 133–144.
71. Abola, E. E.; Bernstein, F. C.; Koetzle, T. F. *J. Phys. Chem.* **1985**, *92*, 139–144.
72. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.
73. Chen, H.; Lynn, P. D.; Giordanetto, F.; Lovell, T.; Li, J. *J. Chem. Inf. Model.* **2006**, *46*, 401–415.

Downloaded by PENNSYLVANIA STATE UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch006

Chapter 7

Fragment-Based High-Throughput Docking and Library Tailoring

Peter Kolb[*]

Department of Pharmaceutical Chemistry, University of California,
San Francisco, San Francisco, CA 94158
[*]E-mail: kolb@docking.org

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

Fragments, i.e. small and simple molecules, have garnered a lot of attention in recent years. They are not only promising starting compounds in biochemical and biophysical assays, but also lend themselves to the development of novel concepts in computational chemistry. The key advantage of fragments in this area is that they can be treated at relatively low computational cost. Among the novel concepts, fragment-based docking is one of the most successful ones. It uses the poses of fragments that have been obtained by decomposing molecules to guide the placement of the entire molecule. This strategy leads to high complementarity between all the subunits of a molecule and the receptor. Along the same lines, anchor-based library tailoring has been developed. This method reduces the size of a molecular library by keeping only molecules containing a fragment that interacts favorably with the receptor upon docking. This chapter will describe all techniques in more detail and highlight the most important applications.

Introduction

Contrary to the most common use of the term these days, this chapter will for most part use an alternative definition of “fragment”. Usually, this term refers to small molecular entities with a molecular weight below 250 g/mol and less than three hydrogen bond donors and acceptors, respectively (*1*). This subset of chemical compounds has been investigated more and more in recent years, due to attractive features such as an increased likelihood to bind to a receptor (*2*) and

© 2011 American Chemical Society
In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

more complete coverage of chemical space on the level of fragments (*3*, *4*). On the other hand, fragments can be defined as parts – or subgraphs – of molecules. It makes sense to think about drug-sized molecules in terms of their constituting fragments, as the latter are much easier to treat computationally because they are less flexible and less complex. For years, organic chemists have generated this kind of fragments by hand when they tried to determine the building blocks of molecules. Computers have more and more taken over this task and fragment molecules by applying certain rules to determine which bonds to cut (*5*). For the remainder of this text, "fragments" will refer to these (computer-generated) molecule parts. This chapter is organized as follows: first, I will describe DAIM, a software that has been developed to decompose molecules, and reflect on some of the potential applications; then the various software that has been written in order to process and dock drug-sized molecules using their fragments as anchors will be described; third, I will detail the fragment-based library tailoring procedure that we have developed; and finally, some applications of the method will be higlighted.

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

Computational Fragment Generation

DAIM

DAIM (**D**ecomposition **a**nd **I**dentification of **M**olecules) is a computer program to automatically and specifically fragment chemical compounds into their constitutive small components (*6*). Its original purpose was to obtain small, rigid fragments to be docked with the program SEED (*7*, *8*) (**S**olvation **E**nergy for **E**xhaustive **D**ocking; described in the next chapter). This is also reflected in the original basic set of rules that was created to obtain fragments with few, if any, internal degrees of freedom (*6*).

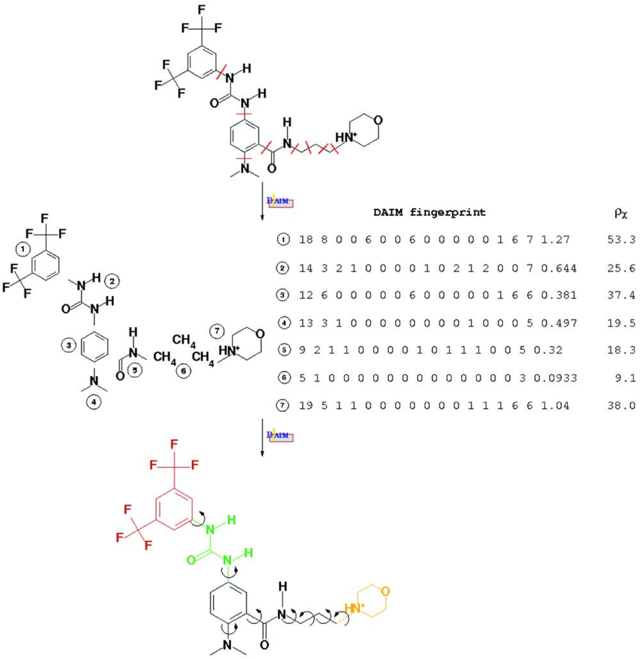
Decomposition Rules

The decomposition of a molecule proceeds in four phases (Figure 1): ring identification, initial fragment definition, functional group merging, and completion of valences.

(i) *Ring identification.* Rings are identified by a modified breadth-first search. All neighbors, i.e. directly covalently bound atoms are enumerated and a neighbor with an already assigned number indicates a ring closure, with the corresponding ring size being the sum of the order numbers of the two atoms. (ii) *Initial fragment definition.* A fragment is defined as a set of atoms connected by unbreakable bonds. The basic definition of unbreakable bonds includes terminal, double, triple, and aromatic bonds and bonds in rings. In the original study (*6*), an extended definition of unbreakable bonds was used since with the basic definition single bonds of groups that form chemical entities would be cut (e.g., in a sulfonamide group, the bond between sulfur and nitrogen is formally a single bond and would thus be cut). This extended list includes amide, phosphate group, and sulfonamide bonds, as well as the single bonds in conjugated systems, and the single bond connecting an amidine group. (iii) *Functional group merging.*

To form chemically relevant fragments and avoid the generation of many small groups, simple functional groups (e.g., -OH, -CH_3, -CX_3 [where X can be any halogen], -SO_3, -CHO, -NO_2, -NH_2, and -SH) are merged with the fragment they are connected to. Unbreakable bonds and functional groups (points ii and iii, respectively) can be defined by the user. (iv) *Completion of valences.* In the final step, missing atom neighbors are added. An atom will lack a neighbor atom where the bond connecting them has been cut. These missing neighbors are replaced by hydrogen atoms to reconstitute the correct valence for every atom. A methyl group is used to fill valences where a hydrogen atom would result in an unwanted additional hydrogen bond direction (e.g., a hydrogen replacing a carbon atom bound to an sp^3 nitrogen).

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007



	DAIM fingerprint	ρ_χ
①	18 8 0 0 6 0 0 6 0 0 0 0 0 1 6 7 1.27	53.3
②	14 3 2 1 0 0 0 0 1 0 2 1 2 0 0 7 0.644	25.6
③	12 6 0 0 0 0 0 6 0 0 0 0 0 1 6 6 0.381	37.4
④	13 3 1 0 0 0 0 0 0 0 0 1 0 0 0 5 0.497	19.5
⑤	9 2 1 1 0 0 0 0 1 0 1 1 1 0 0 5 0.32	18.3
⑥	5 1 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0.0933	9.1
⑦	19 5 1 1 0 0 0 0 0 0 0 1 1 1 6 6 1.04	38.0

Figure 1. Compound 1 of reference (9) is used as an example of a DAIM decomposition and triplet selection. (Top) Compound 1 is shown with the covalent bonds that are cut by DAIM marked with cross lines. (Middle) The fragments identified by DAIM are shown together with their DAIM fingerprints and chemical richness ρ_χ. Note that ρ_χ is evaluated by summing over all values in the fingerprint but neglecting hydrogen atoms or CH_3 groups added by DAIM (e.g., the CH_3 group on the nitrogen of the morpholine in fragment 7). (Bottom) The fragment triplet suggested for docking by DAIM is shown in color. The trisubstituted benzene is considered "central" and is not suggested as anchor. Curly arrows denote rotatable bonds. Reproduced with permission from reference (6). Copyright 2006 American Chemical Society.

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

Comparison to Manual Decomposition

It is interesting to compare the compositions of two libraries on the fragment level. In that way, it becomes apparent whether the two libraries originate from the same set of building blocks and represent only different ways of connecting them or whether the libraries are genuinely different in terms of fragment chemotypes. We did this analysis for a set of 1.85 million unique molecules from the ZINC 2005 (*10*) database of compounds and compared the resulting fragments and their frequencies with the well-known study on drugs by Bemis and Murcko (*11*, *12*). Of course, the two libraries are based on somewhat different assumptions: ZINC is an unbiased collection of small molecules for the purpose of virtual screening; many of the compounds therein do not have pharmaceutically favorable properties. On the other hand, the MDDR (MDL Drug Data Report) set used in (*11*, *12*) consists of all the approved drugs at that time. Yet, the comparison is still meaningful as the promise of ZINC is that it contains lead molecules that might lateron be developed into drugs. The second difference is that Bemis and Murcko used somewhat different decomposition rules by ignoring element types and treating the molecules as graphs. As an example, DAIM will always separate rings connected by a linker, whereas they are treated as one scaffold (“framework”) in reference (*11*) (e.g., benzylbenzene, the third most frequent framework in known drugs [frequency of 68/5120] is decomposed into two benzene rings by DAIM). It is then not surprising that benzene, which is the most frequent fragment in both databases, has a much larger frequency in ZINC (42.2%) than in known drugs (8.5%) (*6*). When one looks at less common fragments like naphthalene and pyridine, it turns out that they have comparable frequencies (1.88% and 3.66%, respectively, in ZINC and 0.59% and 0.82% in the known drugs (*6*)). The most important difference between ZINC and the known drugs is the occurrence of aromatic heterocyclic five-rings: there is only one such scaffold among the 41 most frequent frameworks in known drugs (imidazole, frequency of 19/5120) whereas ZINC contains three such rings with a frequency of ≥ 1%. Conversely, the subset of ZINC that we used lacks steroid-derived scaffolds, despite the fact that there are five among the 41 most frequent frameworks in known drugs. The situation is different for the acyclic fragments (“side chains” in (*12*)) and the overlap between types and frequencies between the DAIM-generated fragments and the ones obtained by Bemis and Murcko is much larger (*6*). It can be speculated that this similarity originates from the facile synthetic accessibility of certain functional groups, and thus is characteristic of synthesized compounds. In summary, despite the differences between the DAIM decomposition and the approach used in the previous analysis of known drugs, frameworks and side chains in commercially available molecular libraries reflect the chemical features present in drug molecules. Such bias towards known chemical space is advantageous (despite all concerns about novelty of molecules): it has been shown that it is much more likely to find binders in libraries that are biased towards known ligands (*13*) and, ultimately, natural products (*14*).

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

Fragment-Based Docking

Fragment-based docking is based on the assumption that each characteristic subunit of a molecule should be involved in at least some favorable interactions upon binding. While this does not necessarily mean that the pose of an individual fragment will coincide with the location of the subunit of the parent molecule it corresponds to, it is reasonable to assume that each fragment will occupy a pose in which it has a favorable interaction energy. The first point has been demonstrated by Babaoglu and Shoichet in a study in which they investigated the binding modes of an inhibitor and its individual subunits by x-ray crystallography (*15*). They show that the poses of the subunits F1, F2 and F3 do not overlap with the respective portions of the original inhibitor L1 (Figure 2).

This does not invalidate the basic assumption of fragment-based docking, however, since most small fragments have more than one binding mode in a protein binding site, a point which has first been brought up in a theoretical paper by Hann *et al.* (*2*). So even if a subunit of a molecule is not at the location in which the individual fragment has the most favorable interaction energy, and which is thus the crystallographically dominant one, it will probably be at a position where it interacts favorably. Consequently, determining a number of favorable poses for each fragment should give enough possibilities so that every subunit can be placed in an appropriate position.

Besides the DAIM-SEED-FFLD package, which will be described in more detail in the following, another program that follows a similar approach is eHiTS (*16*, *17*). The main difference is that eHiTS follows a divide-and-conquer strategy by docking fragments and then reconnecting them to form the entire ligand. FFLD, on the other hand, always treats the entire molecule and just uses the fragment positions as anchors to place the ligand.

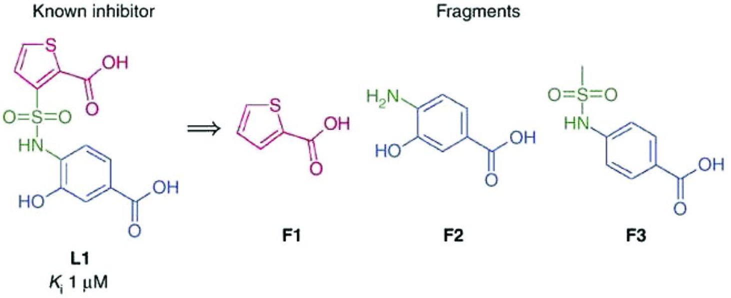


Figure 2. The known inhibitor L1 of (15) was divided into three commercially available fragments, F1, F2 and F3, each containing an aryl carboxylate. None of them bound in the same location as for the original molecule. Reproduced with permission from reference (15). Copyright 2006 Nature Publishing Group.

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

DAIM

The mechanics of DAIM have already been detailed above. Its main purpose in the context of this suite of programs is to generate fragments that are as rigid as possible so that docking them without sampling their internal degrees of freedom is justifiable. At the same time, DAIM determines the basic properties of a fragment and thus whether the docking in SEED will use the polar or apolar vectors to dock it (see below). Importantly, the poses calculated for a certain fragment can be recycled, *viz.* the favorable positions of a benzene fragment will be the same, regardless of which parent molecule it originated from. DAIM thus has to keep track of the fragments it has generated, such that every chemotype is docked only once. For that purpose it uses the internal fingerprints, comparing them with the Tanimoto coefficient.

DAIM Internal Fingerprints

The fingerprints used in DAIM are simple and human-readable structural keys that are generated for each molecule and its fragments. Their main aim is to provide a numerical identifier for a chemical structure that allows fast comparisons. The DAIM fingerprint is made up of 17 fields that are counts of atomic and chemical features (Figure 3): field 1: number of atoms; fields 2, 3, 4, 5, 6, 7: number of carbon, nitrogen, oxygen, halogen, phosphorus and sulfur atoms, respectively; fields 8, 9, 10, 11: number of aromatic, double, triple and amide bonds, respectively; fields 12, 13: number of hydrogen bond acceptors and donor directions, respectively; field 14: number of rings; field 15: number of heavy atoms in rings; field 16: length of the longest chain of atoms in the molecule; field 17: Wiener Index 4 (*18*), modified to take into account the covalent radii of the atoms instead of their maximum principal quantum numbers (*6*), and divided by 1000.

Instead of the substructures of a molecule they use only chemical elements, which are easily and quickly counted. Furthermore, the entries of such a fingerprint consisting of chemical element counts can be combined to estimate molecular descriptors, such as the log P (octanol/water partition coefficient), which can be calculated by atom-additive methods (*19*, *20*). In DAIM, the fingerprints are used to decompose a library into a set of unique fragments and to choose the three anchor fragments necessary for docking with FFLD (*21*, *22*) (**F**ast **F**lexible **L**igand **D**ocking).

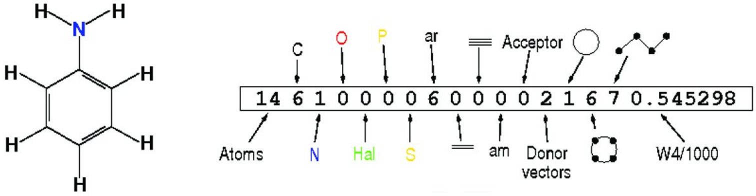


Figure 3. Aniline with its DAIM fingerprint. Reproduced with permission from reference (6). Copyright 2006 American Chemical Society.

Selection of Fragments as Anchors

As will be explained below, FFLD translates each molecule conformation into a triangle which it uses to match the conformation onto the appropriate SEED points. The corners of such a triangle are defined by the geometrical centers of the fragments (Figure 5, top right). Most lead-like or drug-like molecules will be decomposed into more than three fragments, however, which leaves us with the question how to choose the most appropriate ones. The most suitable anchor fragments for fragment-based docking are those that form highly favorable interactions with the protein upon binding. In other words, these fragments likely will have the greatest influence on the final pose of the ligand. DAIM selects the fragment triplets in a three-step selection process (*6*). In the first step, the "chemical richness" ρ_χ of a fragment is evaluated by summing over all values in the fingerprint (Figure 3) but neglecting hydrogen atoms or -CH_3 groups which have been added by DAIM to fill valences. The assumption behind this simple sum is that both size features and functional groups are encoded in the internal fingerprint. A fragment's size will determine in which pockets of a binding site it can fit, whereas functional groups are likely to form directional interactions and thus determine the orientation of the fragment. Since the DAIM fingerprint consists of feature counts, the fragments with high values of ρ_χ are more likely to contain many such functional groups.

All fragments with a value of ρ_χ lower than ten are discarded to increase computational efficiency. This value was chosen to exclude small apolar fragments such as methane ($\rho_\chi = 9.09$), because they would be too frequent otherwise. Methanol ($\rho_\chi = 14.18$) is still a viable selection with this threshold, however, and is arguably more important in terms of potential interactions with the receptor. In the second step, highly substituted fragments are eliminated. These "central" fragments can not form significant interactions with the protein for steric reasons. For a cyclic fragment, the number of substituents (n_{subst}) and the number of rings (n_{rings}) are counted, and the cyclic fragment is rejected because of being "central" if $n_{subst} \geq k_r \times (n_{heavy\ atoms\ in\ ring} - n_{rings})$. Using a value of 1/1.75 for the constant k_r, a disubstituted benzene is retained, whereas a trisubstituted one is considered "central" and is therefore not used as anchor (as is the case for the central benzene of the compound depicted in Figure 1). An acyclic fragment is deselected if $n_{subst} \geq k_l \times n_{heavy\ atoms}$. The default value of 0.5 for k_l allows for the selection of terminal amide groups (i.e., connected to one other fragment) but rejects amide groups originating from within the chain (i.e., connected to two fragments). After the preceding two steps have been passed, the three fragments with the highest ρ_χ values are chosen as anchors. Figure 1 shows a β-secretase inhibitor as example, together with its DAIM-derived fragments, and their chemical richness.

SEED

SEED (**S**olvation **E**nergy for **E**xhaustive **D**ocking) places small, rigid fragments in a protein binding site with exhaustive sampling of a fragment's poses and evaluates the interaction energy taking into account the contribution

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

of bulk solvent (*7*, *8*). It uses polar and hydrophobic vectors as anchors to orient the fragments. The polar vectors are distributed around hydrogen bond donors and acceptors pointing in directions such that hydrogen bonds originating from them are within the extended range of 180±50°. Apolar vectors are used to mark hydrophobic regions; those are obtained by placing a low dielectric sphere (methane) at equal intervals on the solvent accessible surface of the protein. Points that have a favorable interaction energy are retained and the vectors are defined by joining each point with the corresponding atom center. During docking, every vector is matched to the complementary vectors on the fragments and the fragments are rotated exhaustively around these vector-defined axes (Figure 4 shows an example with pyrrole as the fragment). For each fragment position around each vector, a binding energy which includes electrostatic solvation is evaluated.

Thus, if the fragments are rigid, as is the case for small molecules and aromatic systems, conformational strain can be neglected and the most favorable poses of a certain fragment can be determined with high accuracy. The information calculated by SEED is reduced through energy-weighted geometrical clustering from 10^5-10^6 poses to around 100 poses per fragment. Of these 100, the geometrical centers of the top 20 cluster representatives (according to energy) are passed on as possible corner points of the placement triangle used in the last step (see below). For each fragment, the 20 points define a “map” which contains the important information from SEED but is still diverse enough to offer useful anchor points. Diversity, i.e. clustering, is especially important because using only the top-ranked poses of the fragments does not always lead to the solution. This is due to the fact that the binding mode of the entire ligand is a compromise that tries to satisfy most of the fragments.

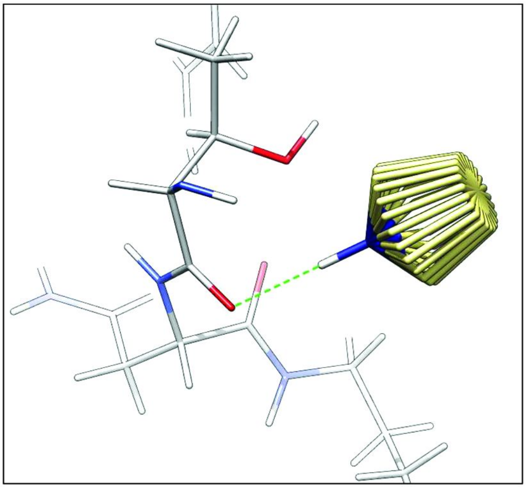


Figure 4. Pyrrole (gold carbons) is rotated around its hydrogen bond (green dashed line) with the backbone carbonyl. It is also sampled in all other possible hydrogen bond directions and around all hydrogen bond acceptors of the binding site.

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

FFLD

The last step is the docking of the complete putative ligand. This is done with the program FFLD (**F**ast **F**lexible **L**igand **D**ocking), which uses a scoring function consisting of ligand dihedral and van der Waals energy, and protein-ligand polar and van der Waals contributions (*21*, *22*). Ligand conformations are generated and optimized by a genetic algorithm with local optimization (*22*), which encodes the torsional angle values of the rotatable bonds. A ligand conformation is placed in the binding site by matching the geometrical centers of the subunits to the corresponding geometrical centers of the fragment maps calculated by SEED. In any given ligand conformation, the three fragments define a triangle and based on the side lengths of this ligand triangle, FFLD finds those SEED points that form triangles of approximately the same shape (Figure 5).

The ligand triangle is matched to each of the possible SEED triangles using a least-squares-fitting method (the Kabsch algorithm (*23*)). At this point, the top 10% individuals of a population (i.e. the 10% conformations with the most favorable interaction score) are locally optimized. FFLD uses the Solis and Wets algorithm (*24*) for this task and the resulting improved conformations are re-encoded as chromosomes in the genetic algorithm if they are not too similar to conformations already present in the population. This local optimization together with the exclusion of similar conformations dramatically improves the performance of the genetic algorithm (*22*, *25*). The output of FFLD consists of the final poses for all conformations, usually 100-200 in total. This is a strength of the program as a variety of different poses with similar interaction energies can be used as different starting points for modifications.

In its current implementation, FFLD takes between 30 s and one minute per ligand. The calculation times for DAIM and SEED are negligible in the context of the docking of a large library and usually on the order of minutes to hours.

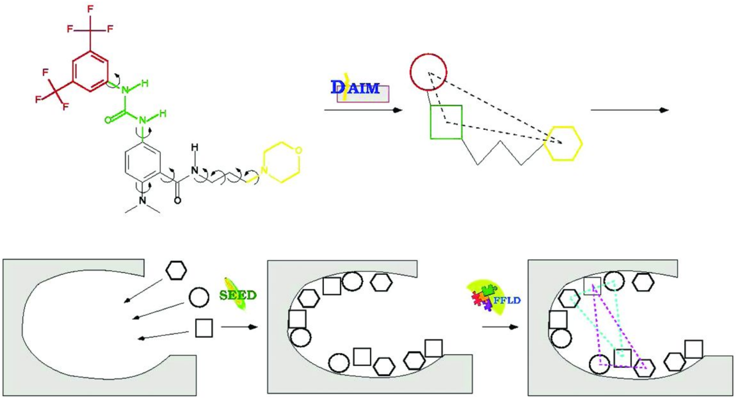


Figure 5. Schematic representation of the DAIM-SEED-FFLD process of fragment-based docking.

Anchor-Based Library Tailoring

ALTA (**A**nchor-based **L**ibrary **T**ailoring) was developed to exploit the fragment generation capabilities of DAIM and the fine-grained sampling of SEED to customize large libraries and to discard molecules with a low chance of binding early on in a docking run (*26*). On the protein side, ALTA can make use of prominent pharmacophoric features that the user deems as important for ligand binding. During tailoring, first, all molecules in a large library are decomposed into their fragments (Figure 6, Step 1). These are then filtered to extract those fragments that are compatible with the pharmacophore (Figure 6, Step 2): e.g. aromatic fragments to complement hydrophobic patches on the protein; fragments with a positively charged group that can form an ionic interaction with a negatively charged sidechain; etc. This will already substantially reduce the number of feasible fragments. Those that are accepted are then docked with SEED (Figure 6, Step 3). In the post filtering step, only fragment poses that fulfill the pharmacophore are kept and ranked according to their SEED score (Figure 6, Step 4). Depending on the amount of molecules that shall finally be docked, only the top N of these fragments are propagated to the next stage. There, all the molecules that contain at least one of the top N fragments are retrieved (Figure 6, Step 5) and then fed into the standard fragment-based docking pipeline (Figure 6, Step 6).

In that way, only molecules that in principle can fulfill the pharmacophoric constraints of the binding site will be docked. This results in shorter docking times, but, more importantly, in less noisy docking runs as there are fewer molecules that will rank highly for random reasons. It is also important to note that ALTA can be run without applying a pharmacophore constraint, in which case one would just skip all the pertaining filter steps.

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

Applications

In this chapter, I would like to present several successful applications of the DAIM-SEED-FFLD suite of software. The hit rates range between 3 and 40%, with an average of 16% over all docking campaigns (cf. Table 1 of (*27*)).

EphB4

The kinase EphB4 is a promising antiangiogenic target in prostate and other cancers. We applied the ALTA method to it, using the well-described double hydrogen bond that most kinase inhibitors form with the hinge region as the pharmacophoric constraint (*26*). Table 1 shows the statistics for this docking campaign: from 728202 molecules at the outset, the size of the libraries was reduced to 21418 molecules that had to be docked. Intermediately, 13533 fragments were docked with SEED, but docking a rigid fragment is in general far less CPU intensive than docking a flexible ligand.

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

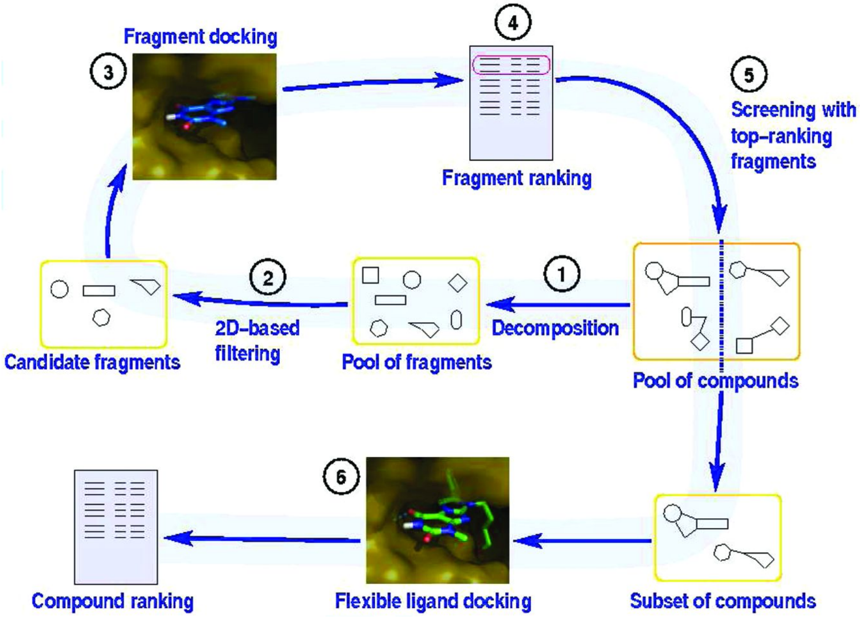
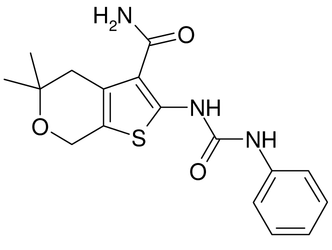
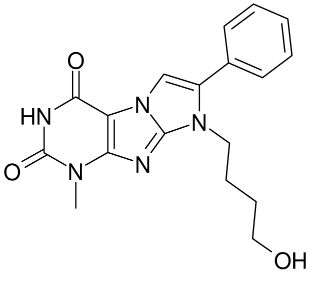


Figure 6. Graphical representation of the workflow of the ALTA procedure (top) and its application to EphB4 (bottom). The first step is the automatic decomposition of a library of compounds (right middle rectangle) to obtain the pool of fragments. Afterwards, fragments selected based on the binding site features are docked and ranked according to their binding energy. Poses for molecules that contain at least one of the top-ranking fragments are then generated by flexible-ligand docking. Reproduced with permission from reference (26). Copyright 2006 Wiley-Liss, Inc.



1



2

*Figure 7. Compound **1** and **2** of ref. (26). Compound **2** has successfully been progressed to a single-digit nanomolar binder (28).*

Forty compounds from the high-throughput docking of the focused library of 21418 molecules were selected after visual inspection (Step 6 in Figure 6) and tested in a Förster-resonance energy transfer (FRET)-based enzymatic assay. Ten of these interfered with the fluorescence read-out and could thus not be measured.

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

Table 1. Application of the anchor-based library tailoring approach to EphB4

Step	*Outcome*	N_{mol}[a]
	Original libraries	728202
1	Fragments obtained by decomposition	35513
2	Fragments remaining after 2D-based filtering	13533
3	Fragments forming two hydrogen bonds with hinge	5235
4	Anchor fragments selected upon energy ranking	1205
5/6	Molecules docked using the "mis"[b]	21418
5/6	Molecules docked using the "mds"[c]	8849
6	Molecules forming one or two hydrogen bonds with hinge	9960

[a] Number of fragments/compounds processed in the individual steps. Docking (Steps 3 and 6) was carried out in parallel on two structures of EphB4 differing only in the orientation of the hydroxyl group of Thr693 in the ATP binding site. The value of N_{mol} is the number of unique fragments (in Steps 3–5) or unique molecules (in Step 6) originating from the docking into the two structures. [b] Most interesting set (mis): flexible-ligand docking using the three fragments with the highest chemical richness (*6*) as anchors. [c] Maximum diversity set (mds): flexible-ligand docking using the three fragments which are most dissimilar to each other as anchors. The compounds docked using the mds are a subset of the compounds docked using the mis. Reproduced from ref. (*26*).

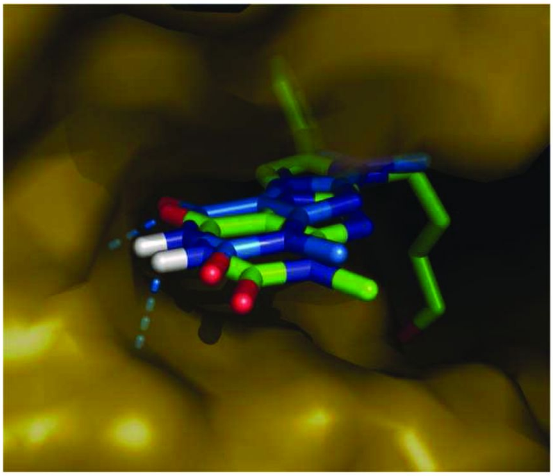


Figure 8. Predicted binding mode of compound **2** *(carbon atoms in green) and its anchor fragment (carbon atoms in light blue) in a homology model of EphB4 (gold surface). The hydrogen bonds with the hinge region are shown in light blue dashes. Note the significant overlap in the binding mode of compound and fragment. Figure prepared with PyMOL (DeLano Scientific, USA). Reproduced with permission from reference (26). Copyright 2006 Wiley-Liss, Inc.*

A compound with a phenylurea anchor (**1** in Figure 7) showed an IC_{50} of 76 µM in the FRET-based enzymatic assay. The most potent compound (**2** in Figure 7) consisted of a three-ring system, which had also been its anchor, and showed a K_i of about 1.6 µM, with a molecular weight of only 353 Da. Five more compounds with different anchors inhibited the activity of EphB4 by 15–40% at a concentration of 125 µM (data not shown). The ligand efficiency ($LE = -\Delta G^{\exp}_{binding} \big/ HAC$, where HAC is the number of heavy atoms (*29*)) of compound **2** is excellent with a value of 0.3 kcal/mol per heavy atom suggesting that it is an interesting compound for further development. To evaluate its cell permeability and cellular activity, compound **2** was tested in CHO cells for inhibition of EphB4 autophosphorylation in a mammalian cell-based environment. It showed only mild inhibitory effects in CHO cells at a concentration of 20 µM. Several derivatives of compound **2** were better able to permeate into cells, however, and recently a single-digit nanomolar potency was reached (*28*).

Thus, with a low number of docked molecules and a very low number (thirty) of tested ones, we were able to identify two promising lead molecules, which corresponds to a hit rate of 6%. Moreover, the more potent scaffold could successfully be progressed (*28*) without changing its anchor fragment. Lastly, we asked the question whether our initial assumption that the pose of the anchor fragment was predictive of the pose of the whole molecule and it was thus justifiable to use ALTA as a selection procedure, was true. The overlay of the docked poses of the anchor fragment and compound **2** (Figure 8) indeed show good correspondence of the heavy atom positions, indicating that at least on the level of docking, choosing a molecule because it contains a good anchor fragment makes sense.

Comparing the CPU time required for the preparation and docking of the focused library with docking of all compounds in the three libraries illustrates that the ALTA approach required about 6500 hrs (on a Linux cluster with CPUs with clock speeds of 1.7 GHz): 2 hrs for decomposition into fragments, 2200 hrs for fragment docking, 1000 hrs for the substructure search, and 3300 hrs for flexible-ligand docking and CHARMM (*30*) minimization. The focused library contains only 1/34th of the initial collection of compounds and only about 1/3rd of the fragments. This corresponds to a total speedup by a factor of about 20. Whereas the actual computation times per compound will naturally be different for other docking programs, the speedup achieved by library preprocessing with the ALTA procedure will remain significant.

Select Other Targets

Besides EphB4, multiple other proteins were targeted using the fragment-based docking approach and in all cases, several ligands were discovered (*27*). An outstanding example is *β-secretase*, a key target in Alzheimer's disease, where the Caflisch lab identified three series of novel inhibitors: phenylurea derivatives (*9*) (the compound presented in Figure 1 originates from this screen); triazine derivatives (*31*); and a set of five cell permeable low-micromolar inhibitors with a different scaffold (D. Huang and A. Caflisch, unpublished results). These screens

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

yielded a total of 27 active compounds, with the most potent having an IC_{50} of 3.0 μM. Furthermore, more than half of the compounds were also active in at least one of two different mammalian cell-based assays with EC_{50}s below 20 μM.

The fragment-based docking approach was also successful with the West Nile virus *NS3 protease*. This virus and its close relative, the Dengue virus, cause encephalitis and other fatal diseases, and are a major problem in tropical regions. The NS3 protease was targeted in two screening campaigns, one against the recently solved (*32*) x-ray structure (*33*), the other against a snapshot from a 1 ns-long explicit solvent molecular dynamics simulation that was selected based on its ability to accommodate three molecular fragments representative of known drugs and key substrates (*34*). In addition to high hit rates of 5 and 40%, the most potent lead compound binds with a IC_{50} of 2.8 μM. Importantly, this compound is a good candidate for further development as it occupies only two of the three subpockets forming the NS3 binding site.

The last example that shall be presented here is a screen against *cathepsin B*, a protease that is involved in cancer and rheumatic disorders (*35*). It is a unique enzyme, as it can display endopeptidase, peptidyldipeptidase as well as exopeptidase activity. The mode of activity is governed by the “occluding loop” which acts as a lid that can block part of the binding site. If the loop is open and the binding site is thus accessible, cathepsin B will function as an endopeptidase, while it acts as exopeptidase when the loop is in its closed state. Indeed, we were able to find a reversible inhibitor binding to the active site in a library of 48000 compounds. Through kinetic studies, it was demonstrated that this inhibitor interacts with the occluding loop, stabilizing it in the closed conformation, leading to reduced endoproteolytic activity.

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

Conclusions

Although the “sum is more than the parts”, docking based on the fragments of a ligand is a successful strategy, especially for proteases where the subpockets of a binding site are usually filled with small chemical entities. In spirit, the fragment-based docking approach described here is very similar to the method used in DOCK (*36*, *37*): the geometrical centers generated by SEED and used by FFLD are the equivalent to the matching spheres employed by DOCK. A key difference is that the SEED/FFLD geometrical centers are specific for a chemotype and used for the positioning of an entire fragment, whereas the matching spheres are general (with the exception of “colored spheres”) and used as anchor points for atoms.

As a second strategy, the fragments obtained by decomposing a molecular library can be used to quickly and efficiently scout a protein binding site and dock only those molecules that contain one of the fragments that interact with the protein in a favorable way. Thus, the number of molecules that actually have to be docked is substantially reduced. While speed is less and less of an argument considering the ever-increasing power of computers, library tailoring offers the advantage of fewer docked molecules, and consequently less noise, as well as more ways to filter out molecules with undesired properties.

In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Given their attractive features such as small size, high rigidity, and higher coverage of chemical space, fragments will likely continue to play an important role in computational chemistry. They will be the tool compounds that will allow us to ask questions about true hit rates, the usefulness of our attempts to strive for maximal coverage of chemical space, and the feasibility of achieving highly potent ligands by connecting fragments.

Acknowledgments

I thank all colleagues who have contributed to the studies that have been described in this book chapter. My special thanks go to Prof. Amedeo Caflisch and Dr. Danzhi Huang for many interesting discussions. Moreover, I thank them and Dr. Denise Teotico for helpful comments on this manuscript.

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

References

1. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A ’Rule of Three’ for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876–877.
2. Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–864.
3. Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem., Int. Ed. Engl.* **2005**, *44*, 1504–1508.
4. Fink, T.; Reymond, J. L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342–353.
5. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP - Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
6. Kolb, P.; Caflisch, A. Automatic and efficient decomposition of two-dimensional structures of small molecules for fragment-based high-throughput docking. *J. Med. Chem.* **2006**, *49*, 7384–7392.
7. Majeux, N; et al. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins: Struct., Funct., Bioinf.* **1999**, *37*, 88–105.
8. Majeux, N.; Scarsi, M.; Caflisch, A. Efficient electrostatic solvation model for protein-fragment docking. *Proteins* **2001**, *42*, 256–268.
9. Huang, D. Z.; et al. Discovery of cell-permeable non-peptide inhibitors of β-secretase by high-throughput docking and continuum electrostatics calculations. *J. Med. Chem.* **2005**, *48*, 5108–5111.
10. Irwin, J. J.; Shoichet, B. K. ZINC: A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

11. Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
12. Bemis, G. W.; Murcko, M. A. Properties of known drugs. 2. Side chains. *J. Med. Chem.* **1999**, *42*, 5095–5099.
13. Kolb, P.; et al. Structure-based discovery of β_2-adrenergic receptor ligands. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 6843–6848.
14. Hert, J.; et al. Quantifying biogenic bias in screening libraries. *Nat. Chem. Biol.* **2009**, *5*, 479–483.
15. Babaoglu, K.; Shoichet, B. K. Deconstructing fragment-based inhibitor discovery. *Nat. Chem. Biol.* **2006**, *2*, 720–723.
16. Zsoldos, Z.; et al. eHITS: An innovative approach to the docking and scoring function problems. *Curr. Protein Pept. Sci.* **2006**, *7*, 421–435.
17. Zsoldos, Z.; et al. eHiTS: A new fast, exhaustive flexible ligand docking system. *J. Mol. Graph.* **2007**, *26*, 198–212.
18. Yang, F.; Wang, Z. D.; Huang, Y. P. Modification of the Wiener index 4. *J. Comput. Chem.* **2004**, *25*, 881–887.
19. Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for 3-dimensional structure-directed quantitative structure-activity relationships. 1. Partition-coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–577.
20. Broto, P.; Moreau, G.; Vandycke, C. Molecular structures: Perception, auto-correlation descriptor and SAR studies. Auto-correlation descriptor. *Eur. J. Med. Chem.* **1984**, *19*, 66–70.
21. Budin, N.; Majeux, N.; Caflisch, A. Fragment-based flexible ligand docking by evolutionary optimization. *Biol. Chem.* **2001**, *382*, 1365–1372.
22. Cecchini, M.; Kolb, P.; Majeux, N.; Caflisch, A. Automated docking of highly flexible ligands by genetic algorithms: A critical assessment. *J. Comput. Chem.* **2004**, *25*, 412–422.
23. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A* **1976**, *32*, 922–923.
24. Solis, F. J.; Wets, R. J. B. Minimization by random search techniques. *Math. Oper. Res.* **1981**, *6*, 19–30.
25. Morris, G. M.; et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
26. Kolb, P.; Berset Kipouros, C.; Huang, D.; Caflisch, A. Structure-based tailoring of compound libraries for high-throughput screening: Discovery of novel EphB4 inhibitors. *Proteins* **2008**, *73*, 11–18.
27. Huang, D.; Caflisch, A. Library screening by fragment-based docking. *J. Mol. Recognit.* **2009**, *23*, 183−193.
28. Lafleur, K.; et al. Structure-based optimization of potent and selective inhibitors of the tyrosine kinase EphB4. *J. Med. Chem.*, *52*, 1737–1746.
29. Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: A useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430–431.
30. Brooks, B. R.; et al. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

31. Huang, D. Z.; et al. In silico discovery of β-secretase inhibitors. *J. Am. Chem. Soc.* **2006**, *128*, 5436–5443.
32. Erbel, P.; et al. Structural basis for the activation of flaviviral NS3 proteases from dengue and West Nile virus. *Nat. Struct. Mol. Biol.* **2006**, *13*, 372–373.
33. Ekonomiuk, D.; et al. Discovery of a non-peptidic inhibitor of West Nile virus NS3 protease by high-throughput docking. *PLoS Neglected Trop. Dis.* **2009**, *3*, e356.
34. Ekonomiuk, D.; et al. Flaviviral protease inhibitors identified by fragment-based library docking into a structure generated by molecular dynamics. *J. Med. Chem.* **2009**, *52*, 4860–4868.
35. Schenker, P.; et al. A double-headed cathepsin B inhibitor devoid of warhead. *Protein Sci.* **2008**, *17*, 2145–2155.
36. Kuntz, I. D.; et al. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
37. Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking. *Protein Eng.* **1993**, *6*, 723–732.

Downloaded by CORNELL UNIV on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch007

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

Chapter 8

In Silico Fragment-Based Generation of Drug-Like Compounds

Peter S. Kutchukian, David Lou, and Eugene I. Shakhnovich*

Department of Chemistry and Chemical Biology, Harvard University,
12 Oxford Street, Cambridge, Maine 02138
***E-mail: shakhnovich@chemistry.harvard.edu**

During virtual library construction, the ability to focus the potential combinatorial explosion of generated molecules on a desired region of chemical space is paramount. As such, *de novo* molecule generating programs must strike a balance between the freedom to explore new chemical space and the limitations that must be imposed on growth in order to achieve desired features in the generated compounds, such as stability in water, synthetic accessibility, or drug-likeness. With this in mind, the Fragment Optimized Growth (FOG) algorithm was developed to statistically bias the growth of molecules with desired features. At the heart of the algorithm is a Markov Chain which adds fragments to the nascent molecule in a biased manner, depending on the frequency of specific fragment -fragment connections in the database of chemicals on which it was trained. We demonstrate that FOG generates synthetically feasible compounds, and that it can be trained to grow new molecules that resemble desired classes of molecules such as drugs, natural products, and diversity-oriented synthetic products. In addition to generating virtual libraries of compounds, FOG is well suited to expand experimental fragment hits during lead optimization.

Introduction

There has recently been a surge in the application of computational *de novo* drug design tools in the discovery of experimentally validated ligands, as

© 2011 American Chemical Society
In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

measured by the number of publications reporting the successful use of these programs (*1*). Their application will no doubt escalate, with the ever increasing number of macromolecular crystal structures and amount of computational resources. Furthermore, the field is set for these programs to take an even greater place of prominence in ligand discovery as these tools are especially suited to assist in the elaboration of fragments discovered in fragment-based drug discovery campaigns (*2*). This overlap of computational and experimental techniques underscores the need for algorithms that are especially relevant to practical experimental work. Our goal here was to develop a *de novo* algorithm that would produce synthetically accessible molecules occupying a desired chemical space, such as drug-like or natural product-like.

De novo methods have been the subject of a number of reviews (*1*, *3–7*), so only features especially relevant to the current work will be highlighted. In all *de novo* growth applications, great care must be taken to focus the generation of new molecules that occupy useful chemical space, since the potential combinatorial space when generating new molecules is vast (*8–13*). When the goal is to develop new therapeutically relevant molecules, it is essential to focus that space on compounds that will bind their target adequately, are synthetically feasible, and possess drug-like properties. Shape and energetic complementarity to binding pockets was elegantly examined by most early *de novo* methods (*14–23*), while synthetic tractability was only coarsely addressed, for example by penalizing connections between heteroatoms (*24*), only allowing new bonds to form between carbons when linking functional groups together (*25*), only allowing functional groups to be connected to sp^3 carbons (*26*), disallowing certain connections between atoms (*27–30*), or by disallowing certain connections as well as sequences of connections between fragments to avoid generating unstable moieties such as acetals (*31*). When specific classes of molecules were grown, such as peptides (*32*, *33*), it was unnecessary to develop rules to connect organic fragments since, in this case, amino acids were incorporated as building blocks. Later methods began to address the synthetic feasibility more carefully when generating molecules (*6*, *34–39*) or have added synthetic accessibility scores to prioritize generated candidates (*40*). It is also possible to prioritize compounds post-generation by employing stand-alone programs that score synthetic accessibility (*41*, *42*). Drug-likeness, on the other hand, remains only crudely addressed by *de novo* methods, for example by only using scaffolds and appendages commonly found in drugs (*43*), by using drug-like fragments (*37*, *38*), or by applying penalties when the cutoffs implied by the Lipinski "Rule of Five" (*44*) are violated by grown molecules (*45*). It has become commonplace to use Lipinski's "Rule of Five" as a filter to exclude nondrug-like compounds from chemical libraries, although a number of studies imply that if used to classify drugs versus nondrugs, extremely poor - nearly random – accuracies are obtained (*46*), and there are a number of more sophisticated machine learning algorithms that might be applied post-generation of compounds (*1*).

Here we describe an algorithm, FOG (Fragment Optimized Growth) (*47*), which grows molecules by sequentially adding fragments to a nascent molecule in a statistically biased manner. It should be pointed out here that in the field of cheminformatics and molecular modeling, fragments refer to substructures of

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

compounds, that might be generated, for example, manually or by cleaving a library of authentic compounds. They can range in size from a single atom to a few rings with linkers and side chains, depending on the application. In the field of fragment-based drug discovery, on the other hand, they refer to low molecular weight compounds (MW < 300 Da), whose affinity for a target might be enhanced by further chemical elaboration or by fusion with another fragment that has affinity for the same target. FOG uses fragments, as described by the former definition, to generate synthetically tractable molecules, as deemed by synthetic chemists and synthetic accessibility prediction software (*42*). In addition, the chemical and topological features of compounds grown by FOG are similar to a desired class of chemicals, such as natural products (NP), diversity-oriented synthesis (DOS) products, or drugs, used to train the algorithm. For example, if trained on a NP database, our algorithm would be able to generate new natural product-like compounds with features such as polyphenol moieties that are typical of many of the chemicals in the authentic database, while being devoid of moieties like triazole rings which might be found in DOS compounds. We developed an algorithm capable of classifying compounds, for example as drugs or non-drugs, in order to validate that our algorithm produced compounds that occupied a desired chemical space. Our classification algorithm, TopClass (Topology Classifier) (*47*), exploits the statistical bias of fragments and fragment connections (2D metrics), as well as coupled 1D metrics (such as number of atoms and rotatable bonds). The accuracy of TopClass compares favorably with methods reported in the literature (*48*). In addition, since TopClass is transparent in the features that it classifies compounds by, it was used to identify salient features of drug-like compounds.

Method

1.1. Calculating Transition Probabilities

At the heart of our algorithm is a Markov Chain. Each fragment is considered a “state,” and during growth, transition probabilities are used when selecting subsequent fragments. As such, it is necessary to calculate transition probabilities for each fragment-fragment connection. We first calculated the probability that two fragments were connected in an authentic database of compounds (such as drugs or natural products), and then converted these frequencies into transition probabilities, as detailed in our original publication (*47*). In all database searches that were performed, SMARTS (*49*) strings were used in order to query the desired fragment or substructure, as implemented in ChemAxon’s jcsearch (*50*). One could imagine collecting similar statistics by exhaustively enumerating fragments from a particular database using specific cleavage rules, rather than performing substructure searches. Keeping with the sequential growth of fragments joined by single bonds employed by a number of *de novo* algorithms (*23*, *29–32*, *45*) fragments are attached to each other by removing a hydrogen from each fragment, and subsequently connecting the atoms that were attached to those hydrogens to each other. Fragments are now added, however, in a statistically biased way, depending on the growth fragment. In the current version, rings are only generated by adding ring fragments to the growing molecule, and two non-ring

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

fragments already connected to a growing molecule cannot connect to form a ring. The fragments included in our first version of the program are depicted in Figure 1. To evaluate how well our algorithm reproduced the probabilities that specific fragments would be bonded to each other in grown databases, we compared these probabilities (for example, how likely a benzene ring is bonded to an amine fragment) with the probabilities obtained from the training database, as discussed in the results. We defined our connectivity propensities as the number of times fragment *i* is connected to fragment *j* (N_{ij}^{D}) divided by the number of times fragment *i* is connected to all other fragments ($\sum_{s}\left(N_{is}^{D}\right)$):

$$P_{ij} = \left(\frac{N_{ij}^{D}}{\sum_{s}\left(N_{is}^{D}\right)} \right) \tag{1}$$

1.2. Growth

To initiate growth a fragment is chosen either randomly or based on its frequency in the training database. All subsequent fragments are added in the following manner. Throughout growth fragments present in the growing compound are assigned to one of three lists based on what type of growth is available from that fragment: linear (only one existing connection to another fragment), branch (at least two existing connections to other fragments), or none (all growth sites have been filled). The population of these three lists determines the possible growth modes that are available (linear, branch, or none). A user defined branching probability is used to select one of the modes (linear or branching), if both are available. If only one of the modes is detected, that mode is automatically selected. If all growth sites are saturated, then growth is terminated and the molecule is discarded. Unless stated otherwise, the branching probability $P(B)$ was set to 0.5 for our experiments. This was to access moderately branched structures, while avoiding highly branched structures that might be synthetically inaccessible (*51*). Once a growth mode has been selected, a growth fragment is then chosen from the appropriate list, and a growth site on that fragment is randomly selected.

The selection of the fragment that will be connected to the current growth fragment is then made. This can either be done by using the transition probability of the growth fragment to select the subsequent fragment, or by first deciding to select a ring or non-ring fragment prior to using transition probabilities to select the next fragment. When the latter method is used, a ring non-ring decision is made based on how often the growth point of the fragment is connected to a ring or non-ring in the training database. Alternatively, the user can provide a ring/non-ring transition probability that will be used for all fragments. Once a decision to grow to a ring or non-ring is selected, the correct type of fragment is then selected based on the growth fragment's transition probabilities. It should be noted that the transition probability matrix in this case is split into two matrices, one for transitions to rings, and one for transitions to non-rings, and normalized accordingly.

This process then repeats itself until all growth sites are saturated, a user defined maximum number of fragments have been added, or a maximum molar mass has been obtained. The molecule is written to file as a SMILES string (*52*). This process is illustrated in Figure 2.

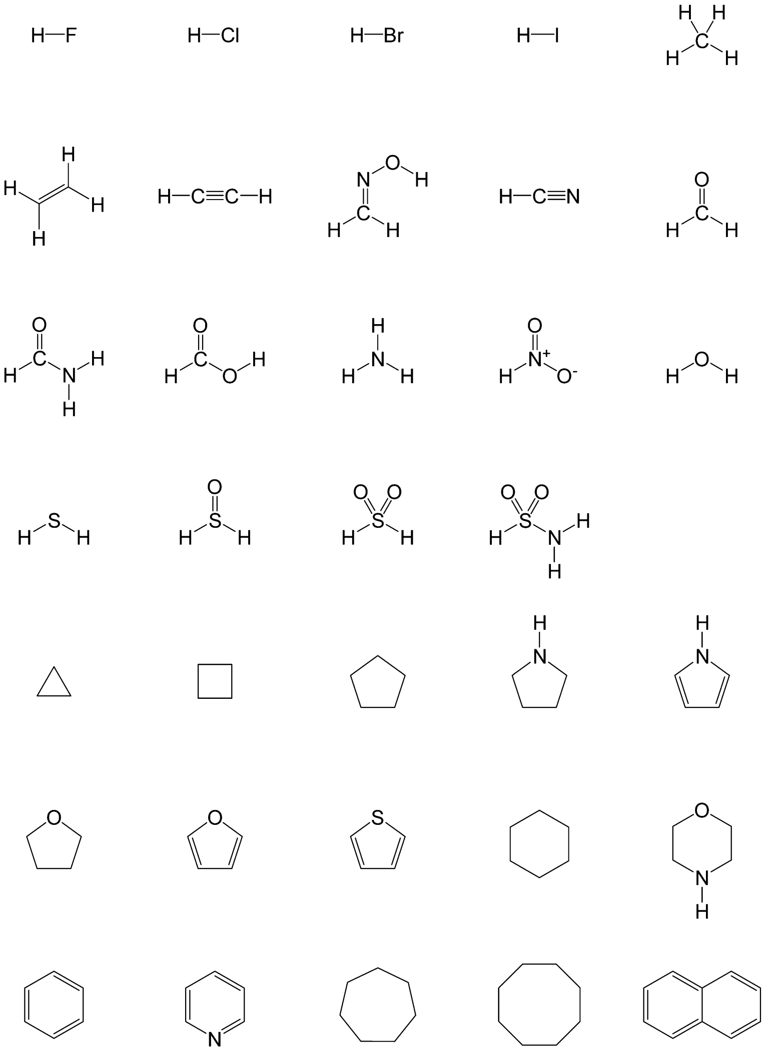


Figure 1. Fragments used by FOG algorithm during growth.

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

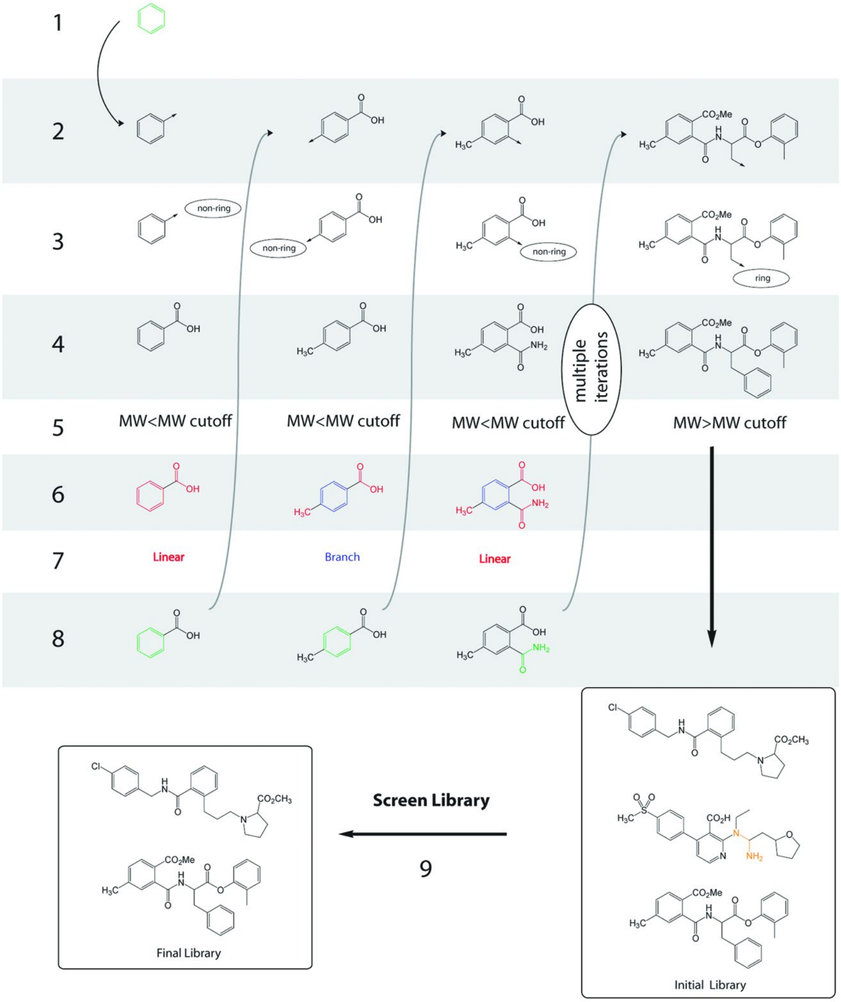


Figure 2. Fragment Optimized Growth (FOG) illustration. ***1****: Select initial growth fragment based on frequency in training database p$_i$.* ***2****: Select growth atom on fragment.* ***3****: Decide to transition to ring or non-ring fragment based on growth atoms ring/non-ring probability.* ***4****: Select next fragment to be added with transition probability $P_{i\rightarrow j}$, and connect two fragments.* ***5****: If current MW > MW_{cutoff} or if the number of fragments > fragment cutoff, write molecule to file and start over.* ***6****: Designate fragments in nascent molecule as linear (red), branch (green), or no growth sites. 7: Select mode of growth (linear or branch) based on branching probability P(B).* ***8****: Attempt to select growth fragment (green) that fits current growth mode. If mode is not available, try other mode of growth. If no growth mode available, discard growing molecule and start over. Repeat from step* ***2****.* ***9****: Remove all molecules containing disallowed 3mers (orange). (see color insert)*

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

1.3. 3mer Screen

Our growth algorithm is capable of stringing together a sequence of 3 fragments that might be synthetically unfeasible or chemically unstable, since it only employs information about the current growth fragment when adding a new fragment, and is "unaware" of any other fragments that might already be connected to it. A geminal diol (Figure 3, top) which in most cases would convert to a ketone in aqueous conditions is an example of such a substructure. One might employ a second order Markov Chain where transition probabilities are based on the current growth fragment and all fragments already connected to it, in order to avoid such substructures. We decided to use a simpler approach that entails removing all compounds that contain disallowed 3mers post-generation. Two sources for disallowed 3mers are implemented in FOG. First, any 3mer sequence that is not observed in the training database is considered disallowed. We do this by searching our training database for all 3mer sequences that can be composed of our fragments. Rings are treated very generally in order to avoid being too stringent. For example, a specific SMARTS string representing a ring carbon might match all sp^3 carbons that are in a ring, but it would not be sensitive to the type of ring that the sp^3 carbon belongs to. The user can manually supply a second set of disallowed 3mers. These are added to avoid chemically unstable moieties such as acetals, ketals, aminals, and iminals (Figure 3), or other substructures that the user might want to avoid. The user defined disallowed 3mers might be similar to the "disallowed angles" in the chemical rules employed by GroupBuild (*31*). One could imagine using more stringent or higher order screens (4mer, 5mer, etc.), but we chose not to do this as we suspected that it would hinder the algorithm's ability to generate novel compounds while not significantly increasing the likelihood of generating synthetically feasible compounds.

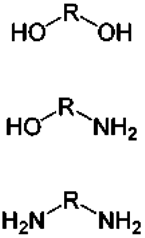


Figure 3. User defined disallowed 3mers. R is any ring or non-ring sp^3 carbon. Hydrogens can also be occupied by any non-hydrogen atom.

2. Classification Algorithm: TopClass

We developed the classification algorithm TopClass in order to evaluate the output of our growth algorithm. A number of individual components that assess different features of a molecule are combined to generate the final TopClass score. Each measure is based on the difference in probabilities or log odds score of observing some feature in a given database *A* versus *B*. They return a positive or negative value depending on whether the scored molecule is deemed more representative of one class of molecules or the other. The total score is a linear summation of the individual scores, and a molecule is classified based on the sign of the final score. Two of the components measured 2D descriptors: the fragment frequency score ($L_1^{'}$) and the fragment-fragment connection score ($L_2^{'}$).

$$L = \alpha_1 L_1^{'} + \alpha_2 L_2^{'} \qquad (2)$$

We refer to this score as "2D" in Tables 1-2. These were combined with three 1D descriptors that assessed the joint probability of hydrogen bond donors and acceptors ($D_{P(don,acc)}$),rotatable bonds and atoms ($D_{P(atoms,rbonds)}$), and rings and atoms ($D_{P(atoms,rings)}$):

$$L = \alpha_1 L_1^{'} + \alpha_2 L_2^{'} + \alpha_3 D_{P(don,acc)} + \alpha_4 D_{P(atoms,rbonds)} + \alpha_5 D_{P(atoms,rings)} \qquad (3)$$

We refer to this score as "2D + c1D" in Tables 1-2. The coefficients (α_1 - α_5) were chosen in order to yield the best separation, without over-fitting to the training set. Details of how these descriptors and coefficients were obtained are provided in the original text (*47*).

3. Separation Algorithm: *D*(min) or *D*(ave)

The minimum Tanimoto dissimilarity D(min) as computed by Chemaxon's Compr (*50*)of a test set compound was calculated in respect to the two training sets that it was being compared to. The test chemical was then classified according to whatever training database it had the lowest *D*(min) for. The average dissimilarity *D*(ave) between a test compound and training database compounds was also used in a similar manner to assign test molecules to database A or B. For the drug/nondrug separation, the *D*(min) score was combined with the coupled 1D topology metrics (the last three terms in equation 4), as described previously (*47*), which we called the *D*(min) + c1D score Tables 1-2.

4. Lipinski and Veber Screens

For the Veber oral bioavailability screen, jcsearch (*50*) was used to identify all molecules that had a rotatable bond count of 10 or less, and had a polar surface area (PSA) (*53*) of 140 Å^2 or less. For the Lipinski screen with no violations, jcsearch (*50*) was used to identify all molecules with MW ≤ 500, logP ≤ 5, H-bond donors ≤ 5, and H-bond acceptors ≤ 10. For the Lipinski screen with 1 or two violations, the MW, logP, H-bond donors, and H-bond acceptors were calculated for each entry using cxcalc (*50*), and an in-house perl script was used to determine how many members in a library passed with 1 violation, and how many members passed with two violations.

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

5. Authentic Compound Libraries

The drug test (218) and training sets (2,495) as well as the non-drug test (110) and training sets (1,263) have previously been described by Hutter (*48*). The DOS and Natural Product (NP) libraries are from the Forma Collection compiled at the Broad Institute (*54*). About 10% of the DOS and NP compounds were randomly selected for use as the test sets (673 for DOS, 230 for NP), and the remaining compounds were used as the training sets (5,950 for DOS, 2,247 for NP).

Results

A Markov Chain approach with branching, treating each growth fragment as the current state, and selecting subsequent fragments based on transition probabilities, was employed in an effort to develop an algorithm that generates novel small molecules that resemble but are not identical to known compounds. In a preliminary study, the ChemBank Bioactives (4,669 compounds) (*54*) were used to train these transition probabilities for a diverse set of fragments (Figure 1). The ChemBank Bioactives database is relatively small and contains chemically reasonable molecules capable of perturbing biological systems, making it an attractive choice for our initial studies. We compared grown compounds (10,000) with the ChemBank Bioactives by comparing their connectivity statistics (the probability that a given fragment i is connected to fragment j in the database, Eqn 1). The similarity of the two databases was evidenced by the excellent agreement we observed (R^2=0.90 for all points, R^2=0.76 when values <0.1 were removed, Figure 4). Compounds resulting from unbiased growth, on the other hand, did not resemble the ChemBank Bioactives (R^2=0.01). Similar results were obtained when larger training databases were employed (NCI Open Database Aug00 (*55*), 250,251 compounds, R^2=0.92 for all points, R^2=0.81 when values <0.1 were removed).

Visual inspection of the transition probability matrix (Figure 5) obtained after training on the ChemBank Bioactives revealed that it was in good agreement with chemical intuition. The ring→ring transition probabilities are in general lower in magnitude than the ring→non-ring transition probabilities, and the ring→ring region of the matrix is also sparse compared to the ring→non-ring region. This might be because ring-ring connections are often difficult to synthesize. The matrix as a whole is also relatively sparse, revealing that many fragments are never connected in the training database. This undoubtedly helps focus combinatorial growth. It is also apparent that transitions to specific fragments, are especially high – most notably the methyl and benzene fragments (denoted by the black and red asterisks, respectively, Fig 5). The prominence of high transition probabilities to the methyl group is not surprising since sp^3 carbons often serve as part of the framework of organic compounds. Benzene chemistry is very well established, and facile substitutions and transformations of appendages allows for diverse groups being connected to benzene (*56*).

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

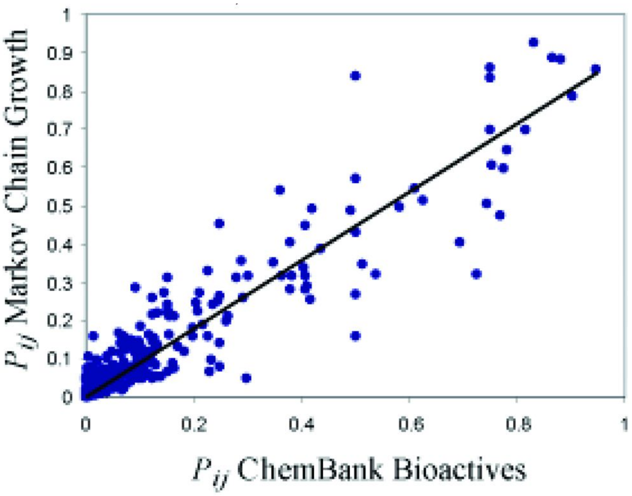


Figure 4. The probability that fragment i is connected to fragment j (P_{ij}, Eqn. 4) in a database of molecules grown with a Markov Chain versus P_{ij} of ChemBank Bioactives used in training the Markov Chain. The probability of branching P(B)=0.5. (see color insert)

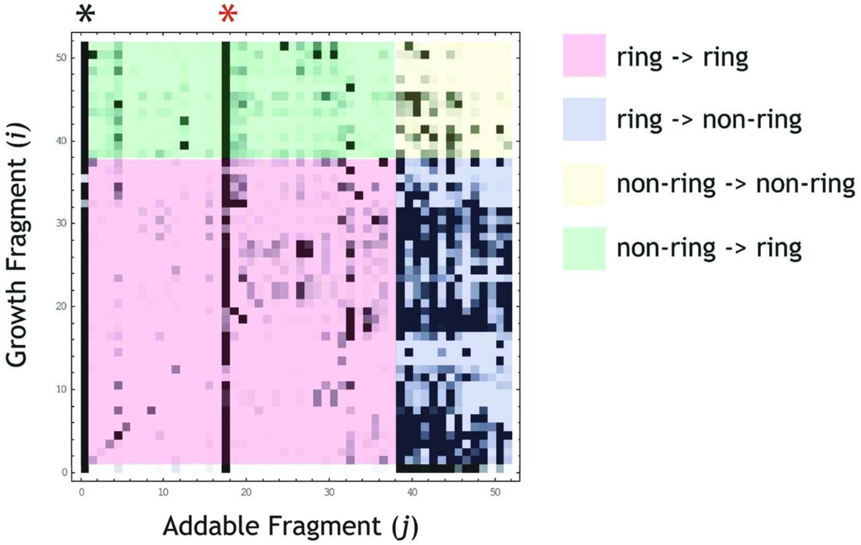


Figure 5. Transition probability matrix used in Markov Chain growth. High probability transitions are depicted as black, while low probability transitions are clear. Certain transitions are highlighted with color. Transitions from methyl to other fragments (bottom row) and from other fragments to methyl (first column) are not highlighted. Asterisks are used to denote the columns representing transitions to methyl (black) and benzene (red). (see color insert)

We next assessed how the number of fragments added and the branching probability impact how grown molecules compare to the training database. The corresponding SMARTS (*49*) of molecules of various sizes (1-11 fragments) and grown with different branching probabilities ($P(B)$ = 0.0-1.0) were used as substructure search strings on the original training database (Figure 6). The branching probability did not have a significant effect on the percentage of substructure hits (*47*). The number of hits fell quite rapidly, however, as the number of fragments increased (Figure 6, FOG (MC)). Even so, the probability to grow substructures present in the training database was much higher for FOG versus unbiased growth (Figure 6, No Bias). This implies that when a few fragments are added with FOG, it is likely that they yield a substructure of a molecule in the original database. An entirely new molecule is accessed as more fragments are added, but it is likely that it is composed of one or more substructures that can be found in the database.

We then assessed whether the FOG compounds were synthetically accessible. We asked organic chemists to judge the synthetic accessibility of compounds that were grown with and without a statistical bias (*47*). Apparently, statistically biasing the addition of fragments with a Markov Chain approach was not sufficient to produce synthetically feasible compounds, since molecules generated with a statistical bias were just as likely to be scored as unsynthesizable or unstable as compounds grown with no bias (FOG (MC) versus No Bias, Fig 7). The following improvements were implemented in FOG after visually inspecting what molecules were deemed unstable.

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

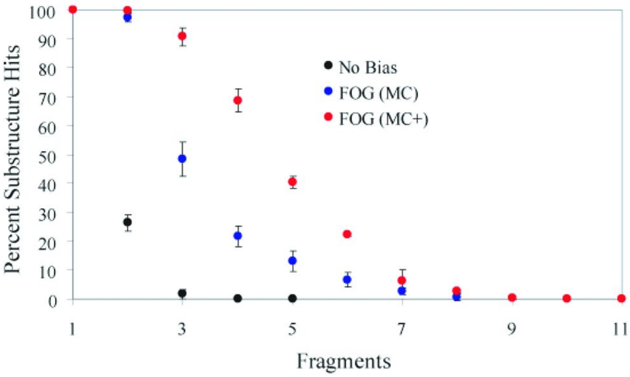


Figure 6. The probability of a grown molecule to be a substructure hit of a compound in the training database. Molecules are either grown with no statistical bias (No Bias), or with FOG using a Markov Chain (MC) or a Markov Chain employing a ring/non-ring transition probability as well as a disallowed 3mer screen (MC+). Errors bars reflect the standard deviation of 3 sets of 100 grown molecules. (see color insert)

An initial observation was that the probability that a fragment is connected to a ring in the grown molecules (11%) was less than in the training set (24%). This might be due to our fragment pool under-representing ring fragments in the training database. We reasoned that this would lead to fragments transitioning to non-ring fragments more often than they would if more ring fragments were included in our growth fragments. This being the case, we incorporated a ring/non-ring transition probability. The algorithm first decides whether the next fragment should be a ring or a non-ring, whenever a fragment is about to be added, based on how often the growth fragment is connected to rings in the training database. A specific fragment is then selected from the pool of rings or non-rings based on renormalized transition probabilities. The second modification we made was to add a post-generation disallowed 3mer screen. The FOG algorithm is capable of forming 3mers that are chemically unstable or synthetically demanding, since it adds fragments based on the current growth fragment, and not on fragments that might be connected to the current growth fragment. To remedy this, compounds that contain 3mers that are undesired by the user (such as acetals), or any 3mer substructures that were not detected in the training database are removed post-generation (step 9, Figure 2).

We observed that ring propensities in the grown molecules (21%) were similar to those observed in the training set (24%) when our modified algorithm was employed. Furthermore, the new algorithm was more likely to grow substructures present in the training database (Figure 6, FOG (MC+)). Surveys of organic chemists demonstrated that FOG did not grow a single molecule that was deemed unsynthesizable or unstable (Figure 7). The difficulty of synthesis, on the other hand, remained similar to molecules grown without any bias (Figure 8). Evaluation of the synthetic accessibility with SYLVIA (*42*) suggested that the grown compounds' synthetic accessibility was similar to that of the chemicals FOG was trained on, and that it was slightly more accessible than compounds grown with no bias (*47*).

We then assessed FOG's ability to grow classes of molecules. We first sought to develop a classification algorithm capable of accurately categorizing molecules, in order to evaluate the output of FOG. We initially employed an algorithm that classified compounds based on statistical biases in the fragments that they were composed of, and how they were connected (Eqn 2). Using such an algorithm, authentic DOS products could be accurately separated from natural products (2D, Table 1). After training FOG on either DOS or NP compounds, we generated libraries of putatively DOS and NP-like molecules, respectively. Our classification algorithm scored 100% of grown DOS compounds and 88% of grown natural products as belonging to their intended molecule class. We were also able to separate authentic DOS compounds from NP compounds with high accuracy using an alternative separation algorithm based on the minimum Tanimoto dissimilarity when a test compound is compared to training set compounds (*D*(min), Table 1). In contrast, using the average dissimilarity in chemical fingerprints as a metric to classify compounds gave poor separation of classes (*D*(ave), Table 1). Although our molecules were more often scored as belonging to the database FOG was trained on using *D*(min) as a classifier, the enrichment was more moderate than our earlier assessment would suggest (63.0% DOS, 78.0% NP).

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

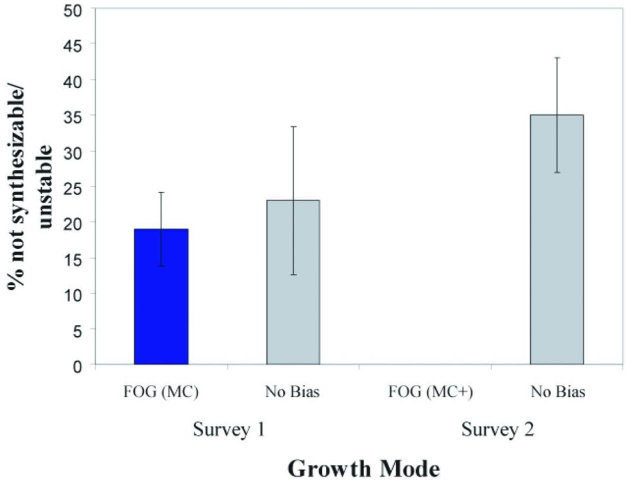


Figure 7. The percent of de novo grown compounds deemed not synthesizable or unstable by organic chemists. Molecules are either grown with no statistical bias (No Bias), or with FOG using a Markov Chain (MC) or a Markov Chain employing a ring/non-ring transition probability as well as a disallowed 3mer screen (MC+). Errors bars reflect the average deviation of responses for Survey 1 (N=5) and Survey 2 (N=8). (see color insert)

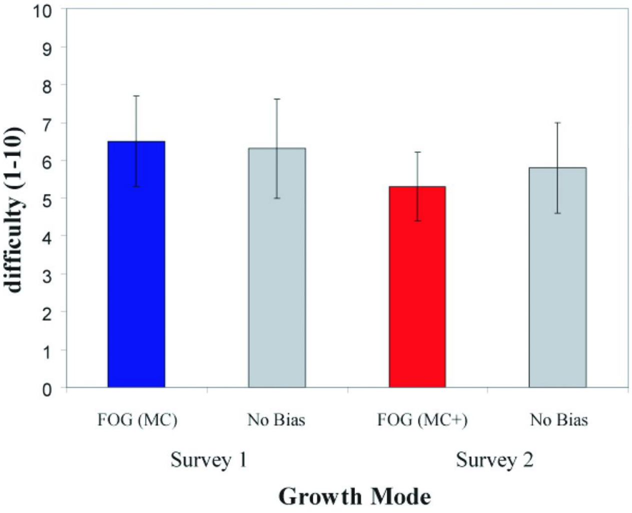


Figure 8. The average synthetic difficulty of de novo grown compounds ranging from 1 (easy) to 10 (difficult) as judged by organic chemists. Molecules are either grown with no statistical bias (No Bias), or with FOG using a Markov Chain (MC) or a Markov Chain employing a ring/non-ring transition probability as well as a disallowed 3mer screen (MC+). Errors bars reflect the average deviation of responses for Survey 1 (N=5) and Survey 2 (N=8). (see color insert)

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

Table 1. Evaluation of separation algorithms. We used fragment and fragment connection biases (2D) as well as coupled 1D metrics such as H-bond donor/H-bond acceptors in addition to the 2D descriptors(2D + c1D). In addition D(min) and D(ave) of chemical fingerprints compared to training sets were used for classification. We also used a combination of D(min) and the coupled 1D metrics (D(min)+c1D). Test sets were evaluated (DOS vs NP, drug vs nondrug), as well as molecules grown with the FOG algorithm (DOS grown, NP grown).

		Classification Method (% correct)				
Compound Set	*Comp.*	*2D*	*2D + c1D*	*D(min)*	*D(ave)*	*D(min) + c1D*
DOS test	673	79.3		99.7	96.3	
NP test	230	90.0		97.8	56.1	
DOS grown	100	100.0		63.0	100.0	
NP grown	100	88.0		78.0	17.0	
drugs test	218	80.3	80.7	94.5	98.6	92.7
nondrugs test	110	58.2	62.7	68.2	9.1	73.6

We then aimed to separate authentic drugs from nondrugs. Our initial results for identifying drugs (80.3%) and nondrugs (58.2%) compared favorably to accuracies reported in literature for the same databases (71.1% for drugs, 40.9% for nondrugs) (*48*). Fragments characteristic of drugs and nondrugs were identified by our method (Figure 9). The top three fragments overrepresented in drugs, for example, are methyl, amide, and non-ring tri-substituted sp^3 carbon, while the alkene, non-ring sp^2 oxygens and fused benzene rings (as in naphthalene) are overrepresented in nondrugs. It should be noted that these relative fragment propensities are sensitive to the composition of the nondrug database (*57*).

We then added three 1D coupled topology metrics to our classification algorithm in an attempt to improve the separation accuracy. These metrics score the differences between two databases in joint probabilities for two variables. They are depicted as heat maps in Figure 10, and inspection of the plots yields valuable information concerning drug-like features. For example, the drug-like region (red) of the donor-acceptor plot resides above the nondrug-like (blue) region. Also of note, we see that molecules with ~3-7 more acceptors than donors are scored drug-like. The bulk of both the drug and nondrug-like regions lie within the Lipinski cutoffs (<10 acceptors, <5 donors), while some of the drug-like region lies outside of these cutoffs. The atoms-rings plot reveals that highly fused structures (high ring:atom ratio) as well as large molecules without any rings (low ring:atom ratio) lie in the nondrug-like region. Similarly, drug and nondrug-like regions are separated for rotatable bonds versus atoms. Larger molecules (>40 atoms) are predominantly scored as drugs, while smaller molecules (<25 atoms) are predominantly scored as nondrugs. Intermediate sized molecules (~30-40 atoms) that are either highly flexible or extremely rigid tend to be non-drugs, while those with intermediate flexibility tend to be scored as drugs. Using the modified algorithm, TopClass (Topology Classifier), there was an improvement in the classification of nondrugs (62.7%) while the accuracy in classifying drugs was maintained (80.7%). We also applied an alternate separation strategy based on the minimum Tanimoto dissimilarities *D*(min) of chemical fingerprints of a test compound compared to the training set compounds. High accuracies separating the drugs (94.5%) and nondrugs (68.2%) test sets were obtained. By combining the *D*(min) score with our three coupled 1D metrics the overall accuracy was slightly improved (92.7% drugs, 73.6% nondrugs).

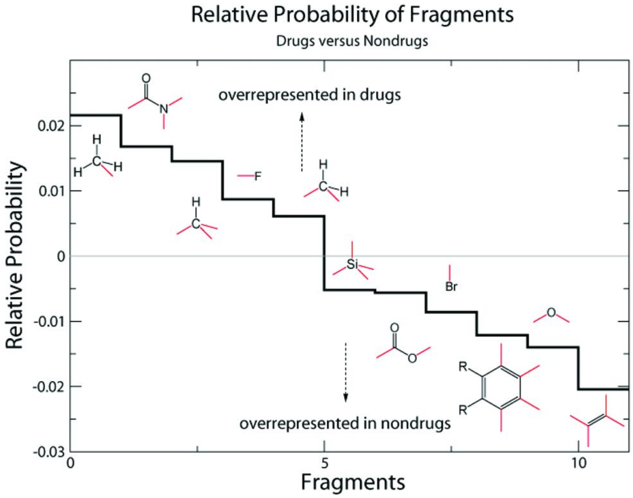


Figure 9. Relative probability of fragments in drugs versus nondrugs. Red bonds indicate connections to any atom including hydrogen, except for the sp^3 carbons where the number of attached hydrogens in explicitly defined. The benzene ring with two R substituents searches for fused benzene rings as in napthalene. Only fragments with large positive or negative values are depicted for clarity. (see color insert)

We devised the following two step screen in order to ascertain how enriched in drug-likeness our grown molecules were compared to those grown with no bias (Figure 11). First, molecules were classified as “no bias” or drugs. The training set for no bias compounds (10,000) was generated by growing compounds with no bias in the transition probabilities. Compounds classified as drugs by the first step were then passed to the second step, and classified as drugs or non-drugs. The entire screen was performed using three different classification algorithms: TopClass, *D*(min), or *D*(min) as well as the coupled 1D metrics from TopClass (*D*(min)+c1D). When 200 compounds grown with no bias were subjected to the first screen using TopClass, not a single molecule was classified as drug-like (Table 2). When molecules grown with FOG (previously trained on the drug database), were subjected to the same screen, however, 83.0% remained after the first step, and 81.5% of the initial 200 remained after both steps. Slightly different results were obtained using the other two classification algorithms (Table 2).

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

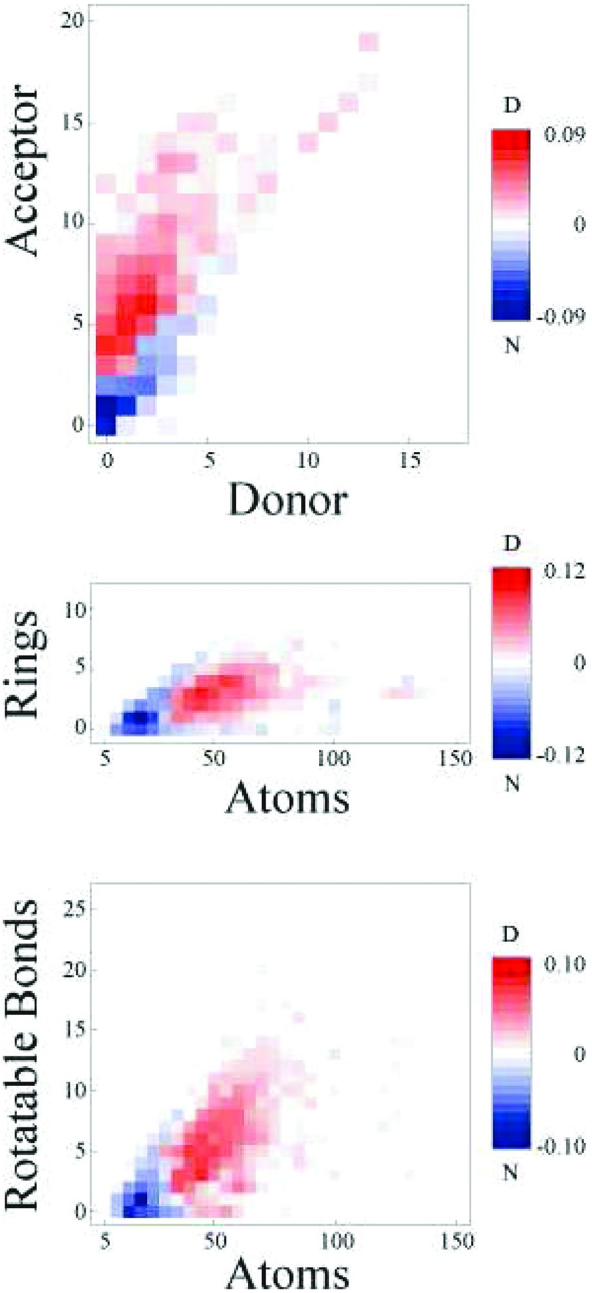


Figure 10. The differences in joint probabilities of 1D topology descriptors between drugs and nondrugs reveal drug-like regions (red) and nondrug-like regions (blue). Atom counts are binned with increments of 5. (see color insert)

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

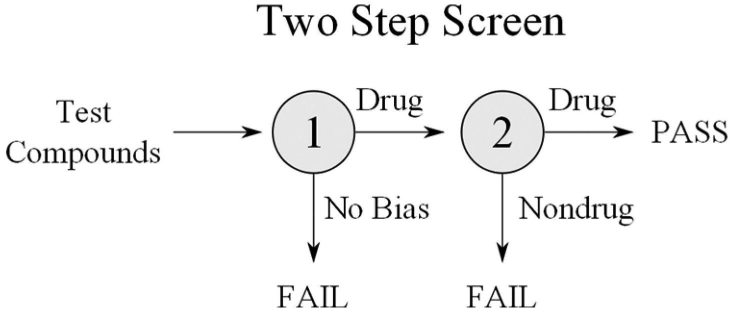


Figure 11. Two step screen to identify drug-like compounds. In step 1, test compounds are classified as drugs or no bias compounds. Compounds that pass step 1 are then classified as drugs or nondrugs in step 2. Three classification algorithms were employed independently using this framework: TopClass, D(min), and D(min)+c1D.

The drug-likeness of our grown molecules was also assessed using two popular oral bioavailability screens: the Lipinski (*44*) and Veber (*58*) screens. To first assess the behavior of these screens, the authentic drugs and nondrugs were subjected to them (Table 2). Although the majority of drugs pass these screens, the majority of nondrugs pass as well. This further supports previous findings that the Lipinski rule of 5 is a poor discriminator of drugs versus nondrugs (*46*, *59*). Even so, we applied them to our grown molecules, to demonstrate the weaknesses inherent in relying on these screens during *de novo* design. A significant amount of compounds grown with FOG pass the Lipinski (*44*) (55%) and Veber (*58*) (80%) screen. Remarkably the majority of the compounds grown with no bias pass the Lipinski screen (79.5%) or the Veber screen (80.0%) even though only 0-2% passed our two step screens.

We then assessed the ability of FOG to generate “privileged” or biologically active scaffolds (*60*). Several scaffolds were investigated, that were composed of 2-4 fragments (Figure 12). Illustrative examples of drugs, natural products, or bioactive compounds containing these scaffolds are depicted in Figure 12 (*61–69*). A library of ~500,000 compounds was generated with FOG, after first training FOG on the drugs training set (2,495 compounds). The frequency of each scaffold’s presence in our FOG library, as well as in the training database of drugs was assessed. FOG was able to generate each of the scaffolds that were evaluated. Notably, FOG generated scaffolds that were not present in the training database of drugs (**3c**-**d**). Furthermore, FOG was capable of generating large scaffolds, composed of 4 fragments (**4a**-**b**).

We sought to balance the ability to grow “drug-like” molecules with the ability to access synthetically accessible *new* molecules. The synthetic accessibility of our grown drugs was similar to that of authentic drugs of similar molecular weight as judged by SYLVIA (*42*), and it was slightly more accessible than compounds grown with no bias (Figure 13A). We calculated the minimum Tanimoto dissimilarity when comparing each of our grown drugs’ chemical fingerprints with the entire training database of authentic drugs (Figure 13B).

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

Table 2. Percentage of compounds scored as drugs using the two step screens as well as popular oral bioavailability screens such as Lipinski (L) and Veber (V)

		Drug Screen (%)[a]						
Compound Set	*Compounds*	*2 step TopClass*	*2 step D(min)*	*2 step D(min) + c1D*	*L(2)*	*L(1)*	*L(0)*	*V*
drugs test	218				100.0	93.1	85.3	84.8
Nondrugs test	110				99.1	99.1	88.2	93.6
drugs grown	200	81.5	39.5	46.5	100.0	99.5	54.5	79.5
no bias grown	200	0	2.0	2.0	100.0	99.0	79.5	80.0

[a] Two step screens were based on TopClass, *D*(min), or *D*(min) and coupled 1D descriptors (*D*(min)+c1D). Oral bioavailability screens such as Lipinski (L) with 2, 1, or 0 violations allowed and Veber (V) are also reported.

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

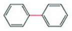
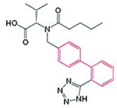
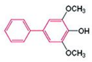
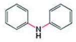
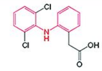
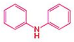
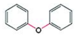
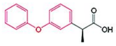
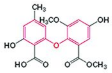
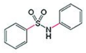
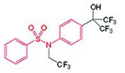
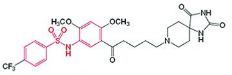
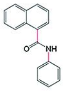
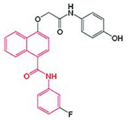
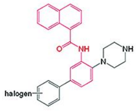
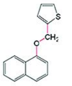
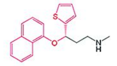
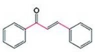
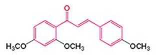
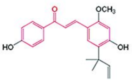
Scaffold	Examples		Drugs (~2.5 K)	FOG (~500 K)
 **2a** Biphenyl	 Valsartan (Diovan) Angiotensin receptor blocker (ARB)	 Aucuparin *Rosaceae* Phytoalexin	23 (0.9%)	108,223 (20%)
 **3a** Diphenylamine	 Diclofenac (Voltaren) NSAID	 Diphenylamine *Onion, Tea* Antihyperglycemic	12 (0.5%)	24,068 (4.5%)
 **3b** Diphenyl ether	 Fenoprofen NSAID	 Asterric Acid	6 (0.2%)	18,492 (3.5%)
 **3c** *N*-Phenylbenzene-sulfonamide	 T0901317 Liver X Receptor agonist	 RS-102221 5-HT$_{2C}$ Receptor antagonist	0	13,908 (2.6%)
 **3d**	 antagonist of HOX/PBX	 BACE-1 inhibitor	0	422 (0.08%)
 **4a**	 (S)-Duloxetine SNRI		1 (0.04%)	6 (0.001%)
 **4b** Chalcone	 Vesidryl choleretic, diuretic	 Licochalcone A *Chinese licorice*	1 (0.04%)	156 (0.03%)

*Figure 12. The occurance of bioactive scaffolds in Drugs (~2.5 K compounds) and in a FOG generated library (~500K compounds). Scaffolds are composed of two (**2a**), three (**3a-d**), or four (**4a-b**) fragments. Connections between fragments in the scaffolds are colored red. Examples of drugs, biologically active compounds, and natural products that contain the scaffolds are depicted with the scaffold highlighted in red. The occurance of the scaffold in the training set of drugs (~2.5 K compounds) or in FOG generated compounds (~500 K compounds) is reported. (see color insert)*

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

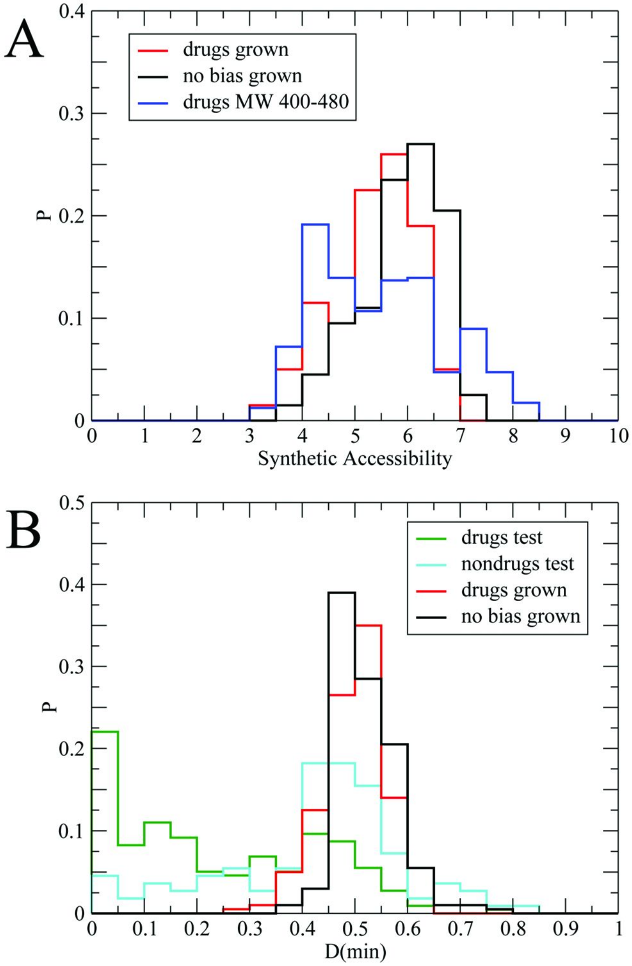


Figure 13. (A) Histogram of synthetic accessibility (1=easy, 10=difficult) of drugs grown (N=200), no bias grown (N=200), and authentic drugs with MW of 400-480 (N=410) as assessed by SYLVIA. (B) Histogram of minimum dissimilarity D(min) of chemical fingerprints of the drugs test (N=218), nondrugs test (N=110), drugs grown (N=200), and no bias grown (N=200) libraries compared to the authentic drugs training set. Chemical fingerprints were generated with GenerateMD (Chemaxon), and D(min) was calculated with Compr (Chemaxon). (see color insert)

We followed this procedure to ensure that we were accessing new molecules. For comparison, we also calculated the minimum dissimilarity of the drugs test, nondrugs test, and no bias grown libraries. For most of our grown compounds, a minimum dissimilarity of ~0.4-0.6 was obtained, ensuring that we were indeed generating novel compounds.

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

Discussion

FOG generates new compounds in a chemical space that is similar to that of the compounds that it was trained on, whether they are drugs, natural products, or DOS compounds. This is achieved by constraining the sequential growth of small molecules by the transition probabilities of the growth fragment. Our method is in contrast to programs that sequentially grow small molecules by selecting new fragments randomly, or by selecting them based on a user defined frequency (*30*, *70*) or their frequency in a database (*71*), rather than based on the frequency of their *connections* to the growth fragment in a database. Of note, libraries grown with our transition probabilities reproduce the frequencies in connections between fragments (Figure 4). Our algorithm can be used as a stand-alone program to generate a virtual library of compounds of specific classes, or it can be easily be incorporated into existing *de novo* design programs that employ sequential growth of fragments. Furthermore, it could be used to explore chemical space around experimentally determined fragment-based screening hits.

One of the features of FOG is that it was designed to be flexible: it can be trained on a particular class of chemicals, and produce compounds that occupy similar chemical space. We demonstrated this by generating compounds that resemble drugs, natural products, and DOS compounds. This is in contrast to programs that are tailored to produce specific classes of compounds, such as peptides (*32*, *33*) or natural products (*72*, *73*).

Another aspect of FOG that is worth noting is that its building blocks are fragments. In the current version, we used functional groups, rings, and tetrahedral carbon (Fig 1). Specific fragments have been shown to preferentially interact with particular protein residues (*74*)(*75*). In addition, fragment-based drug design efforts often identify fragments with high ligand efficiency for a particular binding pocket. This information can be used to seed FOG with a particular fragment, in order to access focused libraries that are extensions of these fragments.

We have also developed TopClass, a linear scoring compound classification algorithm. The transparency of our algorithm has allowed us to investigate interesting features that distinguish drugs from nondrugs. For example, the H-bond donor and acceptor plot (Figure 10) indicated that drugs tend to have ~3-7 more acceptors than donors. This observation leads one to question whether the opposite trend (donors>acceptors) would be observed in the binding sites of proteins, or whether it is because of some other physical or biological reason. It does not seem to be a synthetic bias since nondrugs tended to have the same number of donors and acceptors. We also used a complementary classification method based on Tanimoto dissimilarities *D*(min). This approach proved extremely accurate in classifying test sets of drugs and nondrugs, and helped inform how drug-like our grown molecules were.

Application of the TopClass separation algorithm in a two step screen (Figure 11) demonstrated that FOG molecules did indeed occupy the chemical space that was intended (81.5 % deemed drug-like). Alternate classification methods (*D*(min) or *D*(min)+c1D) corroborated this finding. Furthermore, privileged scaffolds were grown by FOG (Figure 12), even if they were not present in the training database. Generating drug-like virtual libraries has been challenging in the past. For example, when fragments were randomly combined to generate compounds, <0.1% were selected when they were screened for similarity to known drugs and predicted biological activity (using trend vector analysis) (*71*). When scaffolds and appendages commonly found in drugs were randomly combined to generate 30 compounds, only 33% were scored as drug-like (*76*). Likewise, using a similar method to generate 10^6 compounds, only 7% were considered CNS-active with a high degree of confidence (*77*). When all chemically stable combinations of C,N,O, or F containing 11 atoms or less were virtually generated (26.4 x 10^6 million compounds), only ~0.16% were deemed as having GPCR, kinase, or ion channel blocking activity when screened with a Bayesian ANN (*12*). We found that when molecules were built by the sequential addition of fragments without any bias it was nearly impossible to access drug-like compounds (0% with TopClass, 2% with *D*(min) or *D*(min)+c1D), although the majority of these compounds passed popular oral bioavailability filters such as Lipinski (80%) or Veber (80%). When FOG was employed, a much higher fraction of generated compounds were scored as drugs with various separation algorithms in our two step method (81.5% Topclass, 39.5% *D*(min), or 46.5% *D*(min)+c1D). This signifies a huge enrichment in drug-like character in the resulting virtual library. It also exposes the potential liability of generating molecules that lie outside of drug-like space when compounds are generated without any connectivity bias followed by an oral bioavailability filter (such as LigBuilder (*45*)), although a user of these methods may have a false sense of focusing the combinatorial space with the oral bioavailability filters. It is conceivable that unbiased growth when constrained by the geometric and electrostatic environment of a protein binding pocket may result in drug-like compounds, but this has yet to be demonstrated. Our findings strongly suggest that implementing our growth algorithm in *de novo* methods could greatly improve the chance of identifying interesting lead compounds by focusing the potential combinatorial explosion on compounds that occupy relevant chemical space.

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

Acknowledgments

We thank Michael C. Hutter for providing us with the drug and nondrug training and test sets that he assembled, Paul Clemons of the Broad Institute for providing us with the Natural Product and Diversity Oriented Synthesis libraries, Winston Wang for technical programming assistance, and each of the expert chemists that completed our surveys (Supporting Information of original text (*47*)).

References

1. Kutchukian, P. S.; Shakhnovich, E. I. De novo design: Balancing novelty and confined chemical space. *Expert Opin. Drug Discovery* **2010**, *5* (8), 789–812.

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

2. Loving, K.; Alberts, I.; Sherman, W. Computational approaches for fragment-based and de novo design. *Curr. Top. Med. Chem.* **2010**, *10* (1), 14–32.
3. Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4* (8), 649–63.
4. Todorov, N. P.; Alberts, I. L.; Dean, P. M., De Novo Design. In *Comprehensive Medicinal Chemistry, II*; Triggle, D. J., Taylor, J. B., Eds.; Elsevier Science: Amsterdam, The Netherlands, 2006; Vol. 4, pp 283−305.
5. Mauser, H.; Guba, W. Recent developments in de novo design and scaffold hopping. *Curr. Opin. Drug Discovery Dev.* **2008**, *11* (3), 365–74.
6. Hartenfeller, M.; Schneider, G. De novo drug design. *Methods Mol. Biol.* **2011**, *672*, 299–323.
7. Hartenfeller, M.; Schneider, G. Enabling future drug discovery by *de novo* design. *Wiley Interdiscip. Rev.: Comp. Mol. Sci.* **2011**, *1* (3).
8. Cayley On the mathematical theory of isomers. *Philos. Mag.* **1874**, *47*, 444–446.
9. Trinajstic, N.; Nikolic, S.; Knop, J. V.; Muller, W. R.; Szymansky, K., *Computational Graph Theory: Characterization, Enumeration and Generation of Chemical Structures by Computer Methods*; Ellis Horwood: New York, 1991.
10. Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16* (1), 3–50.
11. Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem., Int. Ed. Engl.* **2005**, *44* (10), 1504–8.
12. Fink, T.; Reymond, J. L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47* (2), 342–53.
13. Blum, L. C.; Reymond, J. L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131* (25), 8732–3.
14. Miranker, A.; Karplus, M. Functionality maps of binding-sites: A multiple copy simultaneous search method. *Proteins: Struct., Funct., Genet.* **1991**, *11* (1), 29–34.
15. Gillet, V. J.; Johnson, A. P.; Mata, P.; Sike, S. Automated structure design in 3D. *Tetrahedron Comput. Methodol.* **1990**, *3*, 681–696.
16. Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. SPROUT: A program for structure generation. *J. Comput.-Aided Mol. Des.* **1993**, *7* (2), 127–53.
17. Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. SPROUT: Recent developments in the de novo design of molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (1), 207–17.
18. Bohm, H. J. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6* (1), 61–78.

19. Verlinde, C. L.; Rudenko, G.; Hol, W. G. In search of new lead compounds for trypanosomiasis drug design: A protein structure-based linked-fragment approach. *J. Comput.-Aided Mol. Des.* **1992**, *6* (2), 131–47.
20. Rotstein, S. H.; Murcko, M. A. GenStar: a method for de novo drug design. *J. Comput.-Aided Mol. Des.* **1993**, *7* (1), 23–43.
21. Pearlman, D. A.; Murcko, M. A. Concepts: New dynamic algorithm for de-novo drug suggestion. *J. Comput. Chem.* **1993**, *14* (10), 1184–1193.
22. Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B.; Westhead, D. R. PRO-LIGAND: An approach to de novo molecular design. 1. Application to the design of organic molecules. *J. Comput.-Aided Mol. Des.* **1995**, *9* (1), 13–32.
23. DeWitte, R. S.; Shakhnovich, E. SMoG: de novo design method based on simple, fast and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733–11744.
24. Gehlhaar, D. K.; Moerder, K. E.; Zichi, D.; Sherman, C. J.; Ogden, R. C.; Freer, S. T. De novo design of enzyme inhibitors by Monte Carlo ligand generation. *J. Med. Chem.* **1995**, *38* (3), 466–72.
25. Miranker, A.; Karplus, M. An automated method for dynamic ligand design. *Proteins* **1995**, *23* (4), 472–90.
26. Roe, D. C.; Kuntz, I. D. BUILDER v.2: Improving the chemistry of a de novo design strategy. *J. Comput.-Aided Mol. Des.* **1995**, *9* (3), 269–82.
27. Nishibata, Y.; Itai, A. Automatic creation of drug candidate structures based on receptor structure: Starting point for artificial lead generation. *Tetrahedron* **1991**, *47* (43), 8985–8990.
28. Nishibata, Y.; Itai, A. Confirmation of usefulness of a structure construction program based on 3-dimensional receptor structure for rational lead generation. *J. Med. Chem.* **1993**, *36* (20), 2921–2928.
29. Bohacek, R. S.; Mcmartin, C. Multiple highly diverse structures complementary to enzyme binding-sites: Results of extensive application of a de-novo design method incorporating combinatorial growth. *J. Am. Chem. Soc.* **1994**, *116* (13), 5560–5571.
30. Luo, Z. W.; Wang, R. X.; Lai, L. H. RASSE: A new method for structure-based drug design. *J. Chem. Inf. Comp. Sci.* **1996**, *36* (6), 1187–1194.
31. Rotstein, S. H.; Murcko, M. A. GroupBuild: A fragment-based method for de novo drug design. *J. Med. Chem.* **1993**, *36* (12), 1700–10.
32. Moon, J. B.; Howe, W. J. Computer design of bioactive molecules: A method for receptor-based de novo ligand design. *Proteins* **1991**, *11* (4), 314–28.
33. Bohm, H. J. Towards the automatic design of synthetically accessible protein ligands: Peptides, amides and peptidomimetics. *J. Comput.-Aided Mol. Des.* **1996**, *10* (4), 265–72.
34. Makino, S.; Ewing, T. J.; Kuntz, I. D. DREAM++: Flexible docking program for virtual combinatorial libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13* (5), 513–32.
35. Schneider, G.; Lee, M. L.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput.-Aided Mol. Des.* **2000**, *14* (5), 487–94.

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

36. Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F.; Heeres, J.; Koymans, L. M.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem.* **2003**, *46* (13), 2765–73.
37. Fechner, U.; Schneider, G. Flux (1): A virtual synthesis scheme for fragment-based de novo design. *J. Chem. Inf. Model.* **2006**, *46* (2), 699–707.
38. Fechner, U.; Schneider, G. Flux (2): Comparison of molecular mutation and crossover operators for ligand-based de novo design. *J. Chem. Inf. Model.* **2007**, *47* (2), 656–67.
39. Schneider, G.; Geppert, T.; Hartenfeller, M.; Reisen, F.; Klenner, A.; Reutlinger, M.; Hahnke, V.; Hiss, J. A.; Zettl, H.; Keppner, S.; Spankuch, B.; Schneider, P. Reaction-driven de novo design, synthesis and testing of potential type II kinase inhibitors. *Future Med. Chem.* **2011**, *3* (4), 415–424.
40. Boda, K.; Johnson, A. P. Molecular complexity analysis of de novo designed ligands. *J. Med. Chem.* **2006**, *49* (20), 5869–79.
41. Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discovery Des.* **1995**, *3*, 34–50.
42. Boda, K.; Seidel, T.; Gasteiger, J. Structure and reaction based evaluation of synthetic accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21* (6), 311–25.
43. Jorgensen, W. L.; Ruiz-Caro, J.; Tirado-Rives, J.; Basavapathruni, A.; Anderson, K. S.; Hamilton, A. D. Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorg. Med. Chem. Lett.* **2006**, *16* (3), 663–7.
44. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Del. Rev.* **1997**, *23*, 3–25.
45. Wang, R.; Gao, Y.; Lai, L. LigBuilder: A multi-purpose program for structure-based drug design. *J. Mol. Model.* **2000**, *6*, 498–516.
46. Frimurer, T. M.; Bywater, R.; Naerum, L.; Lauritsen, L. N.; Brunak, S. Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (6), 1315–24.
47. Kutchukian, P. S.; Lou, D.; Shakhnovich, E. I. FOG: Fragment Optimized Growth algorithm for the de novo generation of molecules occupying druglike chemical space. *J. Chem. Inf. Model.* **2009**, *49* (7), 1630–42.
48. Hutter, M. C. Separating drugs from nondrugs: A statistical approach using atom pair distributions. *J. Chem. Inf. Model.* **2007**, *47* (1), 186–94.
49. Daylight Theory Manual. Daylight Chemical Information Systems, Inc., 2008. http://www.daylight.com/dayhtml/doc/theory/index.html.
50. ChemAxon. http://www.chemaxon.com (accessed July 6, 2011).
51. de Silva, K. M.; Goodman, J. M. What is the smallest saturated acyclic alkane that cannot be made? *J. Chem. Inf. Model.* **2005**, *45* (1), 81–7.
52. Weininger, D. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comp. Sci.* **1988**, *28* (1), 31–36.

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

53. Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43* (20), 3714–7.
54. ChemBank. http://chembank.broad.harvard.edu/welcome.htm (accessed July 6, 2011).
55. NCI Open Database. http://cactus.nci.nih.gov/ncidb2/download.html (accessed July 6, 2011).
56. Badger, G. M. *The Structures and Reactions of the Aromatic Compounds*; Cambridge University Press: Cambridge, 1954.
57. Viswanadhan, V. N.; Rajesh, H.; Balaji, V. N. Atom type preferences, structural diversity, and property profiles of known drugs, leads, and nondrugs: A comparative assessment. *ACS Comb. Sci.* **2011**.
58. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–23.
59. Schneider, N.; Jackels, C.; Andres, C.; Hutter, M. C. Gradual in silico filtering for druglike substances. *J. Chem. Inf. Model.* **2008**, *48* (3), 613–28.
60. Welsch, M. E.; Snyder, S. A.; Stockwell, B. R. Privileged scaffolds for library design and drug discovery. *Curr. Opin. Chem. Biol.* **2010**, *14* (3), 347–361.
61. Curtis, R. F.; Hassall, C. H.; Jones, D. W.; Williams, T. W. The biosynthesis of phenols. 2. Asterric acid, a metabolic product of *Aspergillus terreus* Thom. *J. Chem. Soc.* **1960**Dec, 4838–4842.
62. John, V.; Beck, J. P.; Bienkowski, M. J.; Sinha, S.; Heinrikson, R. L. Human beta-secretase (BACE) and BACE inhibitors. *J. Med. Chem.* **2003**, *46* (22), 4625–30.
63. Karawya, M. S.; Abdel Wahab, S. M.; El-Olemy, M. M.; Farrag, N. M. Diphenylamine, an antihyperglycemic agent from onion and tea. *J. Nat. Prod.* **1984**, *47* (5), 775–80.
64. Erdtman, H.; Forsen, S.; Eriksson, G.; Norin, T. Aucuparin and methoxyaucuparin. 2. Phenolic biphenyl derivatives from heartwood of *Sorbus aucuparia* (L.). *Acta Chem. Scand.* **1963**, *17* (4), 1151−1156.
65. Ji, T.; Lee, M.; Pruitt, S. C.; Hangauer, D. G. Privileged scaffolds for blocking protein-protein interactions: 1,4-disubstituted naphthalene antagonists of transcription factor complex HOX-PBX/DNA. *Bioorg. Med. Chem. Lett.* **2004**, *14* (15), 3875–3879.
66. Bonhaus, D. W.; Weinhardt, K. K.; Taylor, M.; DeSouza, A.; McNeeley, P. M.; Szczepanski, K.; Fontana, D. J.; Trinh, J.; Rocha, C. L.; Dawson, M. W.; Flippin, L. A.; Eglen, R. M. RS-102221: A novel high affinity and selective, 5-HT2C receptor antagonist. *Neuropharmacology* **1997**, *36* (4-5), 621–9.
67. Dimmock, J. R.; Elias, D. W.; Beazely, M. A.; Kandepu, N. M. Bioactivities of chalcones. *Curr. Med. Chem.* **1999**, *6* (12), 1125–1149.
68. Natori, S.; Nishikawa, H. Structures of osoic acids and related compounds, metabolites of Oospora sulphurea-ochracea v. Beyma. *Chem. Pharm. Bull.* **1962**, *10* (2), 117−124.
69. Repa, J. J.; Turley, S. D.; Lobaccaro, J. M. A.; Medina, J.; Li, L.; Lustig, K.; Shan, B.; Heyman, R. A.; Dietschy, J. M.; Mangelsdorf, D. J. Regulation of

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

absorption and ABC1-mediated efflux of cholesterol by RXR heterodimers. *Science* **2000**, *289* (5484), 1524–1529.

70. Ho, C. M.; Marshall, G. R. DBMAKER: A set of programs to generate three-dimensional databases based upon user-specified criteria. *J. Comput.-Aided Mol. Des.* **1995**, *9* (1), 65–86.
71. Nilakantan, R.; Bauman, N.; Venkataraghavan, R. A method for automatic-generation of novel chemical structures and its potential applications to drug discovery. *J. Chem. Inf. Comp. Sci.* **1991**, *31* (4), 527–530.
72. Yu, M. J. Natural product-like virtual libraries: Recursive atom-based enumeration. *J. Chem. Inf. Model.* **2011**, *51* (3), 541–557.
73. Zotchev, S. B.; Stepanchikova, A. V.; Sergeyko, A. P.; Sobolev, B. N.; Filimonov, D. A.; Poroikov, V. V. Rational design of macrolides by virtual screening of combinatorial libraries generated through in silico manipulation of polyketide synthases. *J. Med. Chem.* **2006**, *49* (6), 2077–2087.
74. Chan, A. W. E.; Laskowski, R. A.; Selwood, D. L. Chemical fragments that hydrogen bond to Asp, Glu, Arg, and his side chains in protein binding sites. *J. Med. Chem.* **2010**, *53* (8), 3086–3094.
75. Wang, L.; Xie, Z.; Wipf, P.; Xie, X. Q. Residue preference mapping of ligand fragments in the protein data bank. *J. Chem. Inf. Model.* **2011**, *51* (4), 807–15.
76. Ajay, A.; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41* (18), 3314–24.
77. Ajay; Bemis, G. W.; Murcko, M. A. Designing libraries with CNS activity. *J. Med. Chem.* **1999**, *42* (24), 4942–51.

Downloaded by CORNELL UNIV on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch008

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

Chapter 9

Fragment-Based Drug Discovery for Diseases of the Central Nervous System

Vicki L Nienaber*

Zenobia Therapeutics, 505 Coast Blvd. South, Suite 111, La Jolla, CA 92037
***E-mail: Vicki@zenobiatherapeutics.com**

Although diseases of the central nervous system are among the most devastating to patients and their families, disease modifying treatments have lagged behind other therapeutic areas. Current treatments were primarily discovered by serendipity and address disease symptoms. In the genomic era, understanding of CNS biology and disease associated mutations is growing thereby identifying a new series of putative targets. As CNS biology matures, there is growing need for a discovery paradigm that addresses the unique needs of CNS therapeutics, namely the ability of compounds to cross the blood-brain-barrier. The physiochemical properties of CNS therapeutics have been identified based upon historic data and may be used to guide discovery efforts. One notable variable is that the compounds should be low molecular weight. In this chapter, we discuss the merits of fragment-based lead discovery and how it may be used to address the challenges of CNS drug discovery. We also summarize practical strategies for library design and screening. Finally, we summarize examples of how fragments may be optimized into lead compounds.

Central nervous system (CNS) disorders comprise the second largest area of need in the drug discovery industry behind cardiovascular disease (*1*). These diseases are among the most devastating for patients. In fact, dementia and psychosis are ranked in the top five most disabling conditions in the world (*2*). CNS disorders are also among the most expensive for patient care. For example, Alzheimer's disease which affects over 37 million people worldwide (*3*) has an

© 2011 American Chemical Society
In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

annual estimated cost to society of >$100 billion in the US alone (*4*). Despite the undisputed need for treatments and disease modifying therapies for CNS disease, this area has lagged behind other therapeutic areas. In fact, many current treatments were discovered over 50 years ago by serendipity (*5*). Why is this? It has been attributed to a combination of factors. One confounding factor is the complicated biology and likely multipronged basis for these diseases. Another is the requirement that compounds access the CNS by crossing the blood-brain barrier (BBB).

The genomic era has facilitated a new level of understanding of the regulatory processes in cell signaling and biological disorders. Proteins upregulated, downregulated or mutated in disease have been identified. Specifically for CNS disease, efforts such as the Allen Brain Atlas (*6*) (www.alleninstitute.org) are mapping gene expression in mouse and human brain. In fact, commercial kits such as those available from 23andme (www.23andme.com) are available to the public allowing routine genetic testing for markers of CNS disorders such as Parkinson's disease (PD), Alzheimer's disease (AD), schizophrenia and bipolar disorder. One of these, LRRK2 kinase has a series of activating point mutations associated with increased risk of PD (*7–9*) and is a popular drug discovery target. Of course, understanding genetic markers for a disease is only the first step, for example, the genetic mutation basis for Huntington's disease (HD) has been understood since 1993 (*10*) but it is only recently that targets have been brought forward for development of a therapeutic agent. These targets are being curated and made publically available by the Cure Huntington's Disease Initiative (CHDI, http://www.hdresearchcrossroads.org/). The Michael J. Fox Foundation also has a significant effort in identification of targets for PD (http://www.michaeljfox.org/) as does the Alzheimer's Research Forum (http://www.alzgene.org/). These efforts combined with biological validation studies in cells and animals are bringing forward a new generation of targets for CNS drug discovery.

The Blood Brain Barrier: A Unique Consideration for CNS Drug Discovery

It is estimated that 98% of potential drug molecules are excluded from the brain (*11*) which presents a considerable challenge for discovery of CNS therapeutics. For compounds to access the brain, they must cross the blood-brain barrier (BBB) which is formed by endothelial cells of cerebral blood vessels characterized by tight junctions that are present in the brain and at the interface between the blood and cerebro-spinal-fluid (CSF) (*12*). This barrier maintains cerebral homeostasis and has evolved to protect the brain. Transport systems exist to allow nutrients and amino acids into the brain and efflux transport systems of the ATP binding cassette (ABC) family transport lipophilic molecules, such as xenobiotics, out of the brain. While some CNS penetrable compounds effectively utilize active transport systems to access the brain [e.g. L-dopa for PD], the vast majority of compounds enter the brain by passive diffusion. The efflux transporter most relevant to CNS drug discovery is the P-glycoprotein (P-gp) pump (*12*). The

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

BBB is considered one of the unique challenges in discovery of a successful CNS therapeutic versus other disease areas.

Although predictive computational models for compounds that are subject to Pgp efflux are in the early stages of development, general guidelines have been assembled and may be incorporated into the drug discovery process. One such guideline is the "rule-of-four" (*13*) which states that compounds subject to Pgp transport (Pgp +) will have total hydrogen bond acceptors (N+O) ≥ 8, molecular weight (MW) > 400 and an acidic pKa > 4. Compounds resistant to Pgp transport (Pgp -) will have total hydrogen bond acceptors (N+O) ≤ 4, MW <400 and basic pKa < 8. Hence a correlation has been drawn between number of hydrogen bond donors and propensity for Pgp efflux. More specifically defined criteria may include a structural motif which has two H-bond acceptors 4.6 Å apart or three H-bond acceptors 2.5 Å apart (*14*). These guidelines are in agreement with the observed chemical properties of marketed CNS therapeutics and properties for passive diffusion into the brain (see Table 1).

As stated above, most compounds cross the BBB by passive diffusion through the lipid membrane. Unlike for active transport, the physiochemical properties for passive diffusion have been characterized and guidelines for successful CNS clinical candidates have been defined. As expected, these rules are more stringent than for peripheral indications. Lipinski modified his rule of 5 for CNS indications (see Table 1) by effectively lowering the number of hydrogen bonds (HBD < 3, HBA < 7) and the molecular weight (< 400) (*15*). A similar set of rules was reported by Pajouhesh (*16*) which includes a polar surface area cut-off of less than 60Å^2, less than eight rotatable bonds and a non-acidic pKa range. Many CNS drugs are basic and exist in equilibrium between their charged and neutral states at physiological conditions or are amphiphilic if they also possess an acidic group. A positive charge at pH 7-8, in particular a tertiary nitrogen, shows a relatively high degree of brain penetration while strong bases and acids, in particular carboxylates, do not (*17*). These observations are in agreement with the parameters described for minimization of Pgp efflux. Furthermore, because the plasma membrane of the brain is largely composed of negatively charged head groups, weakly positively charged compounds are thought to interact favorably, thereby increasing the local concentration at the membrane surface and promoting passive diffusion. Approximately 75% of the most prescribed CNS drugs are basic, 19% are neutral and 6% are acidic (*18*) Polar surface area is also a well accepted parameter for predicting brain penetration and is generally lower for CNS drugs than for other indications (*11*). Leesen (*19*) analyzed this by comparing the % PSA for CNS drugs versus all drugs launched post-1983. Leesen found that % PSA (polar surface area/total surface area) is significantly lower for CNS drugs than for all marketed drugs (16 versus 21%). Clark and Lobell also report an additional parameter accounting for ClogP and number of hydrogen bond acceptors: ClogP-(N+O) > 0.

An increased understanding of the physiochemical requirements for brain penetrable compounds provides a unique opportunity in the field of drug discovery both in designing screening libraries and in the lead optimization phase. A defining characteristic for CNS penetrable and Pgp pump resistant compounds is low molecular weight (< 400) indicating that the molecular weight of screening

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

compounds should be even less. One may estimate this starting point because it has been shown that lead optimization adds approximately 100 Daltons to the initial hit (*23*) yielding an upper cut-off of 300 for a CNS screening library. If one uses the average molecular weight of marketed CNS drugs (310), then the desired starting point is even lower at ~200. Screening low-molecular weight scaffolds (<200-300 MW) is a relatively new area of drug discovery termed fragment-based lead discovery (FBLD). Below we discuss the basic principles of FBLD and how it the method may be optimized for CNS drug discovery.

Table 1. Summary of chemical properties of CNS drugs

	Properties of Marketed CNS Drugs (19)	*Pgp Efflux (13)*		*Passive Diffusion*		
		Pgp -	*Pgp +*	*Lipinski (20) (CNS)*	*Pajouhesh (16)*	*Clark/ Lobell (21, 22)*
Molecular Weight	310	< 400	> 400	< 400	< 450	< 450
clogP	2.5	NR	NR	< 5	< 5	1-3
H-bond donors	1.5	NR	NR	< 3	< 3	NR
H-bond acceptors (N+O)	2.1	≤ 4	≥ 8	< 7	< 7	< 6
Rotatable Bonds	4.7	NR	NR	NR	< 8	NR
Polar Surface Area (PSA, Å^{3})	NR	NR	NR	NR	< 60-70	< 60-70
% Polar Surface Area	16	NR	NR	NR	NR	NR
pKa	NR	Basic < 8	Acidic > 4	NR	NR	NR

Fragment-Based Lead Discovery

FBLD was reported nearly 15 years ago (*24–26*) and has evolved significantly resulting in multiple clinical candidates over the past 10 years (*27*, *28*). It is now viewed as a solid alternative to high-throughput screening (HTS) and has become increasingly popular in recent years (*27*). One of the primary advantages of FBLD for CNS drug discovery is that unlike HTS, early leads and hits tend to be lower molecular weight. Furthermore, as these small fragments are optimized, one may closely track and monitor the physical properties of future compounds to keep them within the acceptable parameters for a CNS therapeutic agent.

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

Other practical advantages are that because of the lower molecular weight of fragments, fragment chemical space is smaller and therefore easier to sample than all of drug-like space (10^7 versus 10^{60} molecules (*29*) (*20*)). This in principle gives one access to a more diverse set of starting points than from an HTS screen. Additionally, Hann (*30*) showed that simpler molecules are more likely to bind productively to a target than more complex molecules. As a result, libraries composed of low molecular weight fragments need not be large to yield hits. A related advantage is that fragments bind more efficiently on a per atom basis than do HTS hits. The concept was described by Hopkins (*31*) who provided a formula for ligand efficiency (LE) which is the free energy of binding divided by the number of heavy atoms. Based upon Lipinski's rules, this number is ideally above 0.3 kcal/heavy atom. When viewed in total, one can devise a strategy applying FBLD to CNS drug discovery.

Fragment-Based Screening: Practical Application to CNS Disease

To conduct a fragment screen in its simplest form, one must have a target, a screening library and a method to screen it. The biological basis for target choice is beyond the scope of this review, although from a technical point of view, we tend to choose "druggable" protein classes which are tractable by x-ray crystallography. Our goal is to modify the FBLD method to meet the unique needs of CNS drug discovery. Our approach is described below with the general FBLD process summarized in Figure 1.

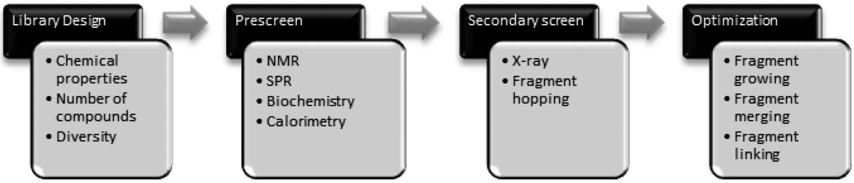


Figure 1. Summary of fragment-based lead discovery method.

Library Design

General library design considerations include the chemical properties of the compounds, the number of compounds in the library and the library's chemical diversity. These three parameters influence each other. For example, more stringent chemical properties filters will lower the number of compounds needed to cover chemical space. On the other hand, libraries with more liberal or fewer property filters yield larger libraries and may require higher-throughput screening methods. Most libraries will yield hits; the goal is to identify hits that may be rapidly optimized to early leads and eventually clinical candidates. We have designed our library with a bias towards CNS indications.

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

Most commercial and internally developed fragment libraries obey the rule of three (RO3) (*32*) which states: MW ≤ 300; cLogP ≤ 3, hydrogen bond donors ≤ 3 and hydrogen bond acceptors ≤ 3. Because our current efforts are focused on treating CNS disease, we have modified these rules to allow for the lower average molecular weight observed for CNS drugs (Table 1). Recall that the average molecular weight for marketed CNS products (*33*) is 310 which is nearly 100 Da lower than for marketed drugs in general. Hence, we believe that the standard RO3 molecular weight cut-off of 300 is too high for a CNS fragment library, especially when allowing for a ~100 Da increase in molecular weight during lead optimization. To compensate for this, our fragment library has an average molecular weight of 150 with a practical upper limit of about 225. In addition to the RO3, we impose a polar surface area restriction of < 60 $Å^2$ which limits the overall hydrophilicity of our compounds. We also impose general filters to remove reactive and known problem compounds from our library.

How many fragments adequately cover chemical space? In an attempt to estimate fragment chemical space, Fink and Reymond (*34*) enumerated that there are ~26 million possible fragments (limit to 11 heavy atoms: C, N, O and S). These compounds included 1028 ring systems with about half existing (or previously existing) as chemically synthesized or a natural product. Of the 26 million theoretical compounds, about half (13 million) meet the rule of three (*27*) but only ~26 thousand are commercially available. Of these 26 thousand, many are aliphatic and would not be included in a fragment library. About 70% of the theoretical compounds are chiral and may be excluded from a screening library. For example, we routinely exclude compounds with more than 2 chiral centers unless the compound represents a unique core amenable to follow-on chemistry. Another interesting observation from this study is that fragment space is biased by compounds found in nature representing a “chemical evolution.” One might imagine that it would be difficult to synthesize compounds that are very different from those found in nature and to carry that thought further, one might imagine that compounds that are evolved well beyond those found in nature are less likely to bind to our naturally occurring drug targets. That will be a debate for computational chemists of future generations.

Another approach for estimating commercially available fragment library size is to analyze available compounds and create a library based upon chemical property filters and diversity. Zenobia has assembled a database of 2.6 million commercially compounds. Of these, ~65,000 meet the rule of three but only 16,000 meet Zenobia’s more stringent cut-off for CNS screening. From here reactive, aliphatic and known problem compounds are removed limiting the library to ~5000 compounds.

For our libraries, we aim for maximum core diversity rather than overall diversity of the library. This minimizes representatives from each core class, even if that class is highly represented in the 5000 compound starting set. To accomplish this, compounds are clustered and library members hand-picked to represent core classes. Cores are biased slightly towards those found in marketed drugs. The cores in our library are also biased based on the number of commercial analogues available to facilitate rapid follow-up of hits through SARbyCatalogue. We generally follow a fragment screen by a fragment-hopping exercise where

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

additional scaffolds that maintain key interactions with the protein are purchased and tested. This increases the diversity of our starting points for synthesis (see below).

The general properties of Zenobia's fragment library are summarized below:

- MW: Avg ~155; upper limit 225
- cLogP: 1-3
- Hydrogen bond donors: ≤ 3
- Hydrogen bond acceptors: ≤ 3
- Polar surface area: <60 Å^2
- Primarily single ring aromatic
    - Includes fused ring
    - Some saturated or linked aromatic

- Compounds with more than two chiral centers are removed
- Reactive groups/problem compounds are removed
- Chemically accessible functionality
- Solubility at ~200mM in DMSO experimentally verified
- ~1000 compounds, > 60 cores

Screening Methods

For low molecular weight fragments, the primary screening method has been limited in those that to detect hits with low binding affinity (< 1-5mM for a fragment of 150 MW with LE > 0.3). Today, common methods include SPR (*35*), nuclear magnetic resonance (NMR) (*24*) and x-ray crystallography (*26*). Calorimetry is also a rapidly emerging method as higher throughput instruments are being developed with lower protein requirements including isothermal titration calorimetry and enthalpy arrays (*36*). Fluorometric detection of thermal melting shifts is also emerging as a method based upon the premise that ligand binding stabilizes the target to denaturation (*37*). With so many potential methods, how does one choose the optimal approach? Typically, we utilize orthogonal techniques where a binding method such as SPR, calorimetry or NMR is used for the primary screen and coupled with x-ray crystallography as a secondary screen. This approach provides both binding energy and detailed structural information for our fragment hits. In addition to binding methods, activity based methods such as biochemical screens may also be used for protein classes where fragments bind more potently than usual (e.g. kinases). At Zenobia, the primary screen varies by protein class and project. However, in all cases, either at the fragment screening or fragment optimization stage, biochemical assays are employed to confirm the functional significance of the hits or analogues.

In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

Fragment Hopping

Because we generate co-crystal structures for our fragment hits, common binding patterns such as hydrogen bonding motifs, electrostatic interactions or important van der Waals contacts begin to emerge as the hits are analyzed in total. These motifs facilitate definition of pharmacaphore models that can be used to rapidly mine commercial (or virtual) chemical space for additional fragment hits (fragment hopping). A practical consequence of this process is that by screening a small library of diverse chemotypes to coarsely sample chemical space, we can define the chemical space of our targets binding site that can be used in identifying additional fragments and in fragment optimization. Hence, since each hit represents a binding motif or chemotype and not necessarily a single compound, we are able to employ simple computational methods to expand the SAR efficiently. Fragment hopping may be used to find additional fragment hits with improved potency, chemical properties or to facilitate rapid synthetic optimization as discussed below.

Fragment Optimization: Options and Approaches

Fragment Optimization

During the fragment screening phase, the goal is to identify a collection of diverse fragment hits so that the best 3-5 can be taken forward into fragment optimization. There are a number of common methods for optimizing fragments into lead compounds (*27*). These include fragment growing, fragment merging and fragment linking as depicted in Figure 2. Which is the best method? We have found that there are multiple paths to a lead compound even through fragment screening and that the path taken is in part driven by the data and resources at hand. The aim is to arrive at the final compound as efficiently as possible and in the fewest number of steps.

Fragment merging can be useful early and late in a drug discovery program. It typically occurs when the structural information from multiple fragment hits are merged or a fragment hit is merged with an existing chemical series. The latter is a particularly powerful application of fragment-based optimization because it can rapidly advance an existing program into new chemical or intellectual property space. An example of fragment merging was first published by Nienaber (*38*) for the target, urokinase. Here, the first crystallographic fragment screen was conducted to identify a core to replace a napthamidine of the existing lead series. This napthamidine series had no oral absorption most likely due to its high basicity. The fragment screen identified a hit with a lower pKa that bound about 10-fold weaker than the napthamidine core (Figure 2A). Crystal structures were completed for the quinoline and the substituted napthamidine, and the two series were merged by attaching the amino-pyrimidine of the naphthamidine to the quinoline core (Figure 2A). Fragment merging resulted in a 100-fold increase in potency as was observed with the napthamidine series for a potency of 370 nM and a ligand efficiency of 0.41. The new fragment-merged compound had an oral bioavailability of 38%. This provided a new low molecular weight starting point for lead optimization.

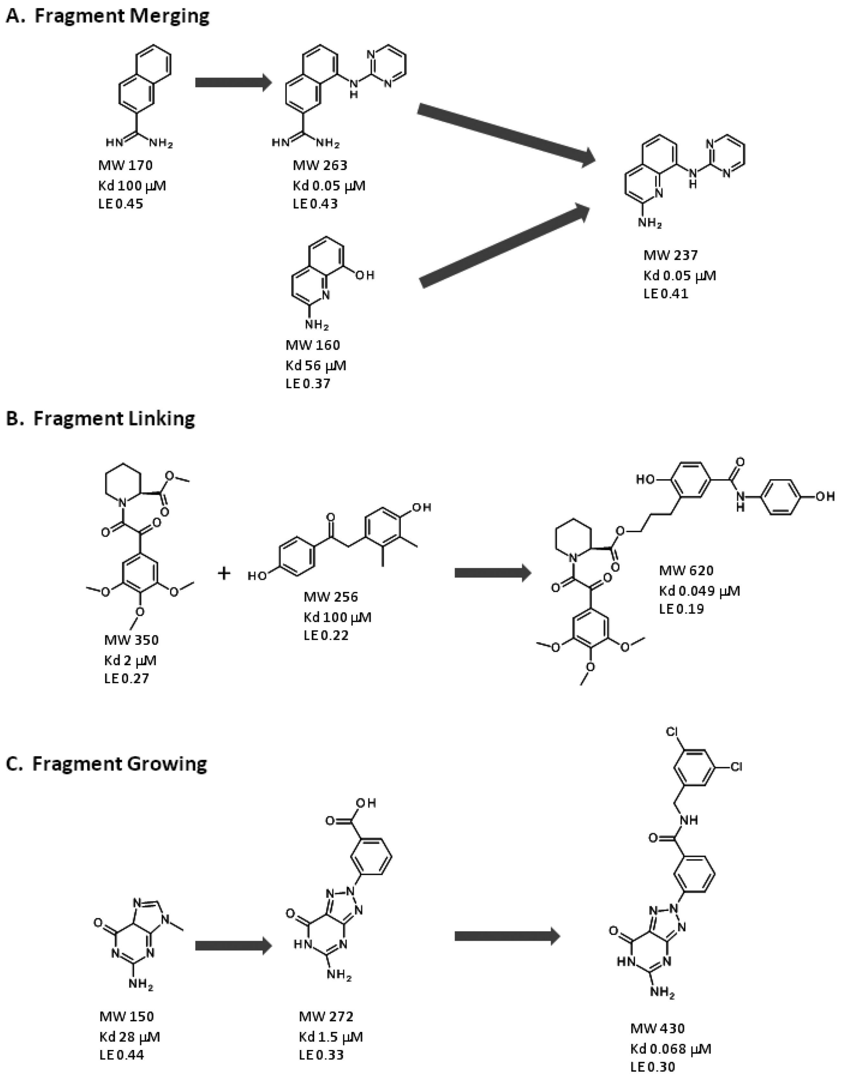


Figure 2. Examples of common fragment optimization approaches, including A. fragment merging (26); B. Fragment linking (24), and C. Fragment growing (38).

Fragment-linking also utilizes information from multiple fragment hits or existing SAR. This was the first application of FBLD published by Shucker et al., (*24*) using NMR as a screening method. Although in this early paper, the fragments are larger than those screened in our embodiment of the method, the success of fragment linking is clearly demonstrated. Here, fragment hits binding to FKBP with potencies of 2 μM and 100 μM were identified by NMR

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

screening and their NMR structures determined. The fragments were linked and lead compounds ranging from 19 nM to 228 nM identified. While fragment linking was one of the first fragment optimization methods implemented, it has not been used widely in the community. This is in part because of the difficulty in identifying fragments that occupy two sites simultaneously and the geometric challenges in linking them through medicinal chemistry. For CNS disease, if fragments are not bound very close in the binding site, linking may not be feasible in meeting the physicochemical requirements of brain penetration.

Fragment growing is one of the most popular methods for fragment optimization. In this method, one only requires a fragment hit and ideally a crystal structure of the hit bound to the target. By examining the binding mode of the fragment to the target, a site for chemical modification and growth into adjacent areas of the active site can be identified. An early example of fragment growing was published by Sanders et al., (*38*) for the target dihydroneopterin aldolase (DHNA). As show in Figure 2C, an initial fragment hit was identified with a Kd of 28μM and modified through a fragment hopping and growing approach to a new starting point with an IC50 of 1.5μM. This new starting point was chosen because it maintained potency and provided a handle for chemistry that was directed towards a binding groove on the protein. A small structure focused library was prepared to probe this site and a new hit identified with a potency of 68nM. When using crystal structures as a guide, fragment growing can be a very efficient method of fragment optimization to gain potency while keeping the chemical properties and molecular weight within an acceptable range.

The normal consequence of lead optimization is that molecular weight and lipophilicity increases while LE decreases (*15*, *39*). However, because of the stringent requirements for brain penetration, we closely monitor chemical properties throughout the fragment to lead and lead optimization process to keep them within the guidelines for CNS drugs. The goal of each design cycle is to increase potency with minimal increase in molecular weight. Ideally, each atom contributes to potency and is explored and optimized. From a practical point of view, this is not always possible due to the shape and characteristics of the target of interest but remains a goal during the process. General goals during fragment and lead optimization are summarized below:

- Keep MW low and ligand efficiency high
- Keep hydrogen bonds low, in particular h-bond donors
    - Add intramolecular H-bonds
    - Remove carboxylic acids
    - Aim for weakly basic pKa

- Reduce Pgp efflux and maximize passive diffusion by monitoring chemical properties throughout the screening and optimization cycle
- Increase lipophilicity while maintaining adequate solubility

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

Compounds that meet these criteria are not guaranteed to be a drug (e.g. stable, efficacious, non-toxic or even cross the BBB). However, as discussed above, compounds that don't meet these criteria do seem to have a higher probability of failing in the clinic making the criteria a guidepost in our drug discovery programs. Furthermore, it can focus efforts as non-drug-like properties are engineered out of the molecule.

Conclusions

As the 78 million baby-boomers (born 1946-1964) in the US age, incidence of neurodegenerative CNS disease is expected to double by 2050 (*40*). Hence, the need for drugs that treat these diseases is only expected to grow. Currently, there is no effective treatment for the most prevalent of these diseases, AD and no effective long-term treatment for PD. In fact, drugs currently on the market for CNS disease treatment including various psychiatric diseases alleviate symptoms, in some cases quite effectively, but do not halt or stop progression of the disease. There is a critical need for these disease modifying therapies.

Target-based therapies have eluded CNS disease in part because of the complicated biology and potential that each disease may have sub-variants requiring specialized therapies for different patient populations. Phenotypic screening is a proven method for identifying marketed CNS therapeutics providing symptomatic relief. However, identification of disease modifying therapies by this method has also proven challenging. This is again most likely due to the complicated biology and in many cases a lack of understanding of the mechanism of action for these compounds. Hence, to provide neuroprotective or disease altering therapies, targeted approaches perhaps in conjunction with the historically more common in vivo phenotypic screening may provide the best options for success.

For both phenotypic and target-based therapies, the chemical properties of compounds must be closely monitored, more closely than for other targeted therapies because these compounds must cross the blood brain barrier. One important property is the molecular weight of the compound which ideally should be below 400. Other chemical properties are summarized in Table 1. Because the properties of the final therapeutic candidate are important for success in the clinic, the properties of the starting compounds and screening libraries should be considered.

Here, we summarize one method for discovery of targeted CNS therapeutics, fragment-based lead discovery. This method starts with very low molecular weight fragments of drugs and then optimizes them, ideally with the use of structural information either through experimental methods (NMR, x-ray crystallography) or computational modeling. A number of biophysical methods have evolved for primary screening and the first clinical candidate for CNS disease derived from fragment-based lead discovery has been reported for the AD target BACE (*41*). As CNS biology is becoming better understood and drug targets identified, other focused targeted efforts using fragment-based lead discovery may provide a basis for a new discovery paradigm in CNS disease.

References

1. IMS Health Drug Monitor, 2003.
2. Ustun, T. B.; et al. Multiple-informant ranking of the disabling effects of different health conditions in 14 countries. WHO/NIH Joint Project CAR Study Group. *Lancet* **1999**, *354* (9173), 111–5.
3. Rafii, M. S.; Aisen, P. S. Recent developments in Alzheimer's disease therapeutics. *BMC Med.* **2009**, *7*, 7.
4. Meek, P. D.; McKeithan, K.; Schumock, G. T. Economic considerations in Alzheimer's disease. *Pharmacotherapy* **1998**, *18*, (2 Pt 2), pp 68−73; discussion 79−82.
5. Enna, S. J.; Williams, M. Challenges in the search for drugs to treat central nervous system disorders. *J. Pharmacol. Exp. Ther.* **2009**, *329* (2), 404–11.
6. Lein, E. S.; et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **2007**, *445* (7124), 168–76.
7. Hardy, J.; et al. The genetics of Parkinson's syndromes: A critical review. *Curr. Opin. Genet. Dev.* **2009**, *19* (3), 254–65.
8. Gandhi, P. N.; Chen, S. G.; Wilson-Delfosse, A. L. Leucine-rich repeat kinase 2 (LRRK2): A key player in the pathogenesis of Parkinson's disease. *J. Neurosci. Res.* **2009**, *87* (6), 1283–95.
9. Schapira, A. H. The importance of LRRK2 mutations in Parkinson disease. *Arch. Neurol.* **2006**, *63* (9), 1225–8.
10. Gusella, J. F.; et al. Molecular genetics of Huntington's disease. *Arch. Neurol.* **1993**, *50* (11), 1157–63.
11. Pardridge, W. M. The blood-brain barrier: Bottleneck in brain drug development. *NeuroRx* **2005**, *2* (1), 3–14.
12. Wolburg, H.; Lippoldt, A. Tight junctions of the blood-brain barrier: Development, composition and regulation. *Vascul. Pharmacol.* **2002**, *38* (6), 323–37.
13. Didziapetris, R.; et al. Classification analysis of P-glycoprotein substrate specificity. *J. Drug Target* **2003**, *11* (7), 391–406.
14. Seelig, A.; Landwojtowicz, E. Structure-activity relationship of P-glycoprotein substrates and modifiers. *Eur. J. Pharm. Sci.* **2000**, *12* (1), 31–40.
15. Lipinski, C. A.; et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46* (1-3), 3–26.
16. Pajouhesh, H.; Lenz, G. R. Medicinal chemical properties of successful central nervous system drugs. *NeuroRx* **2005**, *2* (4), 541–53.
17. Austin, R. P.; Davis, A. M.; Manners, C. N. Partitioning of ionizing molecules between aqueous buffers and phospholipid vesicles. *J. Pharm. Sci.* **1995**, *84* (10), 1180–3.
18. Liu, Y. Factors Affecting Total and Free Drug Concentration in the Brain. AAPS Conference: Critical Issues in Discovering Quality Clinical Candidates, Philadelphia, PA, 2006.

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

19. Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6* (11), 881–90.
20. Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432* (7019), 855–61.
21. Lobell, M.; Molnar, L.; Keseru, G. M. Recent advances in the prediction of blood-brain partitioning from molecular structure. *J. Pharm. Sci.* **2003**, *92* (2), 360–70.
22. Clark, D. E.; Grootenhuis, P. D. Predicting passive transport in silico: History, hype, hope. *Curr. Top. Med. Chem.* **2003**, *3* (11), 1193–203.
23. Carr, R. A.; et al. Fragment-based lead discovery: Leads by design. *Drug Discovery Today* **2005**, *10* (14), 987–92.
24. Shuker, S. B.; et al. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **1996**, *274* (5292), 1531–4.
25. Verlinde, C. L.; Rudenko, G.; Hol, W. G. In search of new lead compounds for trypanosomiasis drug design: A protein structure-based linked-fragment approach. *J. Comput.-Aided Mol. Des.* **1992**, *6* (2), 131–47.
26. Nienaber, V. L.; et al. Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nat. Biotechnol.* **2000**, *18* (10), 1105–8.
27. Congreve, M.; et al. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008**, *51* (13), 3661–80.
28. Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: Strategic advances and lessons learned. *Nat .Rev. Drug Discovery* **2007**, *6* (3), 211–9.
29. Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem., Int. Ed. Engl.* **2005**, *44* (10), 1504–8.
30. Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 856–64.
31. Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: A useful metric for lead selection. *Drug Discovery Today* **2004**, *9* (10), 430–1.
32. Congreve, M.; et al. A ’rule of three’ for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8* (19), 876–7.
33. Leeson, P. D.; Davis, A. M. Time-related differences in the physical property profiles of oral drugs. *J. Med. Chem.* **2004**, *47* (25), 6338–48.
34. Fink, T.; Reymond, J. L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47* (2), 342–53.
35. Myszka, D. G. Analysis of small-molecule interactions using Biacore S51 technology. *Anal. Biochem.* **2004**, *329* (2), 316–23.
36. Torres, F. E.; et al. Higher throughput calorimetry: Opportunities, approaches and challenges. *Curr. Opin. Struct. Biol.*, *20* (5), 598–605.
37. Ericsson, U. B.; et al. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal. Biochem.* **2006**, *357* (2), 289–98.

In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

38. Sanders, W. J.; et al. Discovery of potent inhibitors of dihydroneopterin aldolase using CrystaLEAD high-throughput X-ray crystallographic screening and structure-directed lead optimization. *J. Med. Chem.* **2004**, *47* (7), 1709–18.
39. Rees, D. C.; et al. Fragment-based lead discovery. *Nat. Rev. Drug. Discovery* **2004**, *3* (8), 660–72.
40. Hebert, L. E.; et al. Annual incidence of Alzheimer disease in the United States projected to the years 2000 through 2050. *Alzheimer Dis. Assoc. Disord.* **2001**, *15* (4), 169–73.
41. Wyss, D. F.; et al. Combining NMR and X-ray crystallography in fragment-based drug discovery: Discovery of highly potent and selective BACE-1 inhibitors. *Top. Curr. Chem.*.

Downloaded by STANFORD UNIV GREEN LIBR on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ch009

Downloaded by UNIV OF GUELPH LIBRARY on June 2, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ot001

Editor's Biography

Rachelle Bienstock

Rachelle Bienstock, a native New Yorker, received her undergraduate degree in Chemical Engineering from The Cooper Union in New York City and her PhD in Chemistry from The University of Michigan in Ann Arbor, Michigan. Following postdoctoral studies at the University of Texas Southwestern Medical Center (Dallas) involving NMR and molecular modeling of constrained peptide analogs and peptidomimetics, she joined The National Institute of Environmental Health Sciences, (NIEHS), Research Triangle Park, North Carolina, as a molecular modeler and computational chemist. Her main research interests are protein structure and protein complex prediction methodologies, computational and structure-based ligand design methods and protein−protein and protein−ligand docking studies.

© 2011 American Chemical Society
In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Subject Index

A

ABC. *See* ATP binding cassette (ABC)
Accelrys Discovery Studio software, 8
Activity-need node, 64
ADAM. *See* Alkenyldiarylmethanes (ADAM)
ADMET properties, 29
Alkenyldiarylmethanes (ADAM), 79
  2D depiction, 80*f*
AlleGrow, 11
ALTA. *See* Anchor-Based Library Tailoring (ALTA)
Alzheimer's disease, 3, 54
Alzheimer's Research Forum, 180
Anchor-Based Library Tailoring (ALTA), 7, 140
  EphB4 application, 142*t*
  graphical representation, 141*f*
Aromatic π stacking interactions, 93
ArQule kinase inhibitor, 4
Astex group, 48
  validation set, 112
Atom-pair 2 (AP2), 32
Atom-pair 3 (AP3), 32
ATP binding cassette (ABC), 180
Aurora kinase inhibitor, 3
Authentic compound libraries, 159

B

BACE-1, 3
Bait fragment, 12
Bayesian inference network (BIN), 7, 57
  DF approaches, comparison, 68*t*
  TAN approaches, comparison, 67*t*
Bayes rule, 61
BBB. *See* Blood-brain barrier (BBB)
Benzene chemistry, 159
BIN. *See* Bayesian inference network (BIN)
Bioisostere database (DB), 72, 75
  enumeration, 78
  generation, 77
  reconstruction, 78
Bioisosteric replacement
  case of ADAM ligand, 79
  HSP90 PDB ligands, 82
  rule elucidation, 76
Bioisosteric rules hopping, 74
BioSolveIT, 4
Blood-brain barrier (BBB), 46
  CNS drug discovery, 180
Boltzmann principle, 116, 117
Bound conformations of 5000 ligands, 93
  statistics, 94*t*
Branching probability, 154
Breaking into Restrosynthetically interesting Chemical Substructures (BRICS), 5
BREED, 11
BRICS. *See* Breaking into Restrosynthetically interesting Chemical Substructures (BRICS)
BROOD, 5
Brownian movement, 94
β-secretase, 143

C

Caflisch group, 9
CASD. *See* Computer-Assisted Synthesis Design (CASD)
CAS registry file, 29
Cathepsin B, 144
Central nervous system (CNS)
  blood-brain barrier, 179
  chemical properties of drug, 182*t*
  disorders, 179
  drug discovery, 180
  mapping gene expression, 180
  penetrable and Pgp pump resistant compounds, 181
Cerebro-spinal-fluid (CSF), 180
CHDI. *See* Cure Huntington's Disease Initiative (CHDI)
ChemAxon's jcsearch, 153
ChemBank bioactives, 159
Chemical Abstracts Service, 6
Chemoinformatics, 58
Chemotype Evolution, 12
CHOMP, 5
Click chemistry, 10
CNS. *See* Central nervous system (CNS)
CNS diseases, 46
Co-crystallized ligands, 73
CoLibri, 4
Completion of valences, 132
Compound network, 59
Compound tabulation, 50

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ix002

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ix002

Computer-Assisted Synthesis Design (CASD), 30
CONFRIM, 11
Connolly surface, 119
CrystaLEAD method, x-ray, 2
Crystallographic fragment screening
  library, 46, 48
  experimental data, 49
  exploiting information, 48
  rule-of-three (Ro3), 50
  without crystals, 49
CSF. *See* Cerebro-spinal-fluid (CSF)
Cumulative reaction types, 37*f*
Cure Huntington's Disease Initiative (CHDI), 180

D

DAG. *See* Direct acyclic graph (DAG)
DAIM. *See* Decomposition and Identification of Molecules (DAIM)
DAIM-SEED-FFLD package, 135
  schematic representation, 139*f*
Data fusion (DF), 65
DB. *See* Bioisostere database (DB)
1D coupled topology, 164
  differences in joint probabilities, 166*f*
De novo bonds, 78
De novo design
  discovering novel compounds, 29
  knowledge-base, 31
  knowledge-base of reactions, 38, 39*f*, 40*f*
  method, 6
  recent focus, 30
  transform approach, 31
  virtual screening, 29
deCODE Biostructures, 48
Decomposition and Identification of Molecules (DAIM), 132
  decomposition and triplet selection, 132, 133*f*
  fragments as anchors, 137
  internal fingerprints, 136, 136*f*
  mechanics, 136
  method, 4
Decomposition rules, molecule, 132
Dengue virus, 144
De-solvation term, 117
DF. *See* Data fusion (DF)
2D fingerprint methods, 2
DHNA. *See* Dihydroneopterin aldolase (DHNA)
Dihydroneopterin aldolase (DHNA), 188
Direct acyclic graph (DAG), 59
Disallowed angles, 157
Diversity-oriented synthesis (DOS) products, 152
DOS. *See* Diversity-oriented synthesis (DOS) products
Drug fragment set, 48
Drug-like compounds
  bioactive scaffolds, 169*f*
  histogram, 170*f*
  percentage of compounds scored, 168*t*
  two step screen, 167*f*
Drug-like ligands, 93
Drug-likeness, 2
2D Tanimoto similarity searches, 6

E

ECFC. *See* Extended-connectivity counts (ECFC)
ECFP_6 fingerprints, 64
eHiTS. *See* electronic High Throughput Screening (eHiTS)
eHiTS docking engine, 91
eHiTS flexible ligand docking software, 91
electronic High Throughput Screening (eHiTS), 92
  chemical feature graph, 96
  docktable extension, 101
  geometric shape, 96
  pose and conformational sampling requirements, 93
  pose generation algorithm, 95, 96*f*
  results, 124
  scoring function, 109
Empirical scoring, 108
Enrichment results, 125, 125*f*
  surflex data set, 126*f*
EphB4, 140
  application of the anchor-based library tailoring approach, 142*t*
  homology model, 142*f*
Experimental binding energy, correlation, 126, 126*t*
  eHiTS-score with experimental binding data, 127*f*
Extended-connectivity counts (ECFC), 65

F

Family coverage term, 119
Fast Flexible Ligand Docking (FFLD), 139
FBLD. *See* Fragment-based lead discovery (FBLD)

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ix002

FC-Bioisostere software, 74
FCFC. *See* Functional class extended-connectivity counts (FCFC)
Feature nodes, 63
Feature-graph representation
  cavity, 97
  ligand, 98, 99*f*
FFLD. *See* Fast Flexible Ligand Docking (FFLD)
Flexible chain fitting, 104
  minimization process, 105
Flexible ligand docking, 92
FlexNovo, 10
FOG. *See* Fragment Optimized Growth (FOG)
Forma collection, 159
Förster-resonance energy transfer (FRET), 141
Fragment
  anchors, 137
  approaches, 3
  commercial vendors, 50
  complexes, 2
  computational screening, 2
  "cut and fit" approach, 7
  defined, 2, 131
  designing and searching libraries, 4
  hopping, 186
  library parameters, 46
  linking, 187
  Markov chain approach, 12
  optimization, 186, 187*f*
  relative probability, 165*f*
  silico screening methods, 3
Fragment-based docking, 135
Fragment-based drug design
  computational screening: docking, 8
  difficult targets, 3
  discovery, 3
  examples, 12
  growing and linking, 10
  libraries, 4
  overview, 1
  privileged structure, 3
  protocol, 72
Fragment-based high-throughput docking
  overview, 131
Fragment-based lead discovery (FBLD), 45, 182
  CNS disease, 183
  development, 45
  fragment hopping, 186
  library design, 183
  methods, 185
  project, 47
  summary, 183*f*
Fragment docking problem, 92
Fragment hits, structure-based exploitation, 53
Fragment Optimized Growth (FOG), 12, 151
  algorithm, 151
  algorithm during growth, 155*f*
  discussion, 171
  illustration, 156*f*
Fragments of life, 48
FRET. *See* Förster-resonance energy transfer (FRET)
FT algorithm, 8
FTMAP method, 8
FTrees-FS software, 4
Functional class extended-connectivity counts (FCFC), 65
Functional group merging, 132
Fused benzene rings, 164

G

GANDI, 9
Gaussian probability distribution curve, 112
Geminal diol, 157
GPCR crystal structures, 13
Grid cells, 97
GroupBuild, 10
GrowMol, 11
GSCF graph, 97
GSK3β protein-ligand structures, 54

H

HAC. *See* Heavy atom count (HAC)
Hash code, 38
H-bond donor, 82
HD. *See* Huntington’s disease (HD)
Heat Shock Protein 90 (HSP90), 74
Heavy atom count (HAC), 76
High ligand efficiency, 2
High throughput screening (HTS), 1, 46, 182
  experimental methods, 2
  hit rates, 1
  screen library, 1
HSP90. *See* Heat Shock Protein 90 (HSP90)
HSP90 inhibitor, 3, 49
HSP90 ligands, 82
HSP proteins, 79
HTS. *See* High throughput screening (HTS)

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ix002

Huntington’s disease (HD), 180
Hybridization, 73
Hybrids: scoring and ranking, 85
Hydrogen bond
  geometry, 111
  pyrrole rotation, 138*f*
Hyper-graph, 103*f*, 104
  clique, 105
Hypothetical fragment hits, 49

I

Inference
  defined, 62
  network, 61
In silico fragment-based generation
  de novo methods, 152
  discussion, 171
  growth, 154
  methods, 153
  overview, 151
Initial fragment definition, 132
Interacting surface points (ISP), 109, 117
  eHiTS scoring function, 110*t*
ISP. *See* Interacting surface points (ISP)

J

Join atoms, 96

K

Kearsley’s superimposition algorithm, 77
Kinase CNS targets, 54
Knowledge base, reproducing reactions, 34

L

LE. *See* Ligand efficiency (LE)
LeadModel3D, 6
Lennard-Jones 6-12 potential, 118
Library optimizer using Feature Trees (LoFT), 5
Ligand-based drug design, 73
Ligand-based virtual screening
  experiment, 64
  overview, 57
  results and discussion, 66
  searching chemical databases, 58
  similarity inference network model, 59
Ligand efficiency (LE), 4, 45
Ligand entropy term, 121
Ligand Expo dictionary, 75
Ligand intra-molecular interactions, 121
Ligand strain energy term, 120
Lipinski
  cutoffs, 164
  rule-of-five (Ro5), 50
  screen, 158
LoFT. *See* Library optimizer using Feature Trees (LoFT)

M

Manual decomposition, 134
Markov chain, 151, 153
  bioactives used, 160*f*
  de novo grown compounds, 163*f*
  FOG using, 161*f*
  ring/non-ring transition probability, 163*f*
  transition probability matrix, 160*f*
MAX fusion rules, 66
Maxwell-Boltzmann exponential distribution, 116
MCSS method, 8
MDDR. *See* MDL drug data report (MDDR)
MDDR database, 57
MDL drug data report (MDDR), 58
  structure activity classes, 65*t*
MEDP-fragmentor software, 85
MED-Portions, 6, 13, 85
MED-SuMo, 6
  algorithm, 75
  parameters, 75
  software, 72
MERGE, 5
3mer screen, 157
  user defined disallowed, 157
Michael J. Fox Foundation, 180
Minimum Tanimoto dissimilarity D(min) separation algorithm, 158
Molecular inference network model, 60*f*
MSCS. *See* Multiple solvent crystal structures (MSCS)
Multiple solvent crystal structures (MSCS), 8

N

National Cancer Institute (NCI), 61
Natural Product (NP), 152, 159

In Library Design, Search Methods, and Applications of Fragment-Based Drug Design; Bienstock, R.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ix002

NCI. *See* National Cancer Institute (NCI)
N-dimensional space, 94
NNRTI. *See* Non-nucleoside reverse transcriptase inhibitors (NNRTI)
Non-nucleoside reverse transcriptase inhibitors (NNRTI), 79
NP. *See* Natural Product (NP)
NS3 protease, 144
Nucleoside 2-Deoxyribosyltransferase, 48

O

Organic reactions, success rates in reproducing different types, 36*t*
Overlaid ligands, 76

P

Parkinson disease target, 13
PASS method, 7
PDB. *See* Protein Data Bank (PDB)
P-glycoprotein (P-gp), 180
P-gp. *See* P-glycoprotein (P-gp)
Pharmacophore models, 2
Phenylurea anchor, 143
Phosphodiesterase 4, 3
PLP intermolecular term, 85
Pocket depth term, 119
Polyhedra
  2D cartoon demonstrating generation of cavity, 97*f*
  vector lengths, 99
Polyhedron, 97
Pose matching, 102
Powell local minimizer, 116
Probabilities estimation, 63
  activity-need node, 64
  feature nodes, 63
  query nodes, 63
Probabilities using link matrices, 62
Project-purposed screening libraries, 51
Protein crystallography
  concept, 46
  overview, 45
Protein Data Bank (PDB), 71, 93
  bioisosteric approach, 72
  2D depiction, 81*f*
  superposition of HSP90 ligand, 83*f*
  target-based alignment, 73
Protein family recognition and clustering, 122
  weight parameters, 122
Protein kinase, 72
  scaffolds, 86
Protonation handling, 106
  functional groups A,B,C and D, 107*f*
  ligand with multiple functional groups, 107*f*
Pseudo docking, 6
Pubchem compounds, 82

Q

QSAR models, 31
QSearch, 7
Query fragment, 76, 77
Query ligands, 76
Query network, 59, 60
Query nodes, 63

R

Reaction generation algorithm
  step 1, 34*f*
  step 2, 34*f*
Reaction vectors, 31
  Beckmann rearrangement reaction, 33*f*
  combination of atom-pairs, 32
  overview, 31
RECAP method, 4
ReCore, 11
Replacement count, 77
Replacement fragment, 77
Replacement fragment selection, 77
Resorcinol, 50
Rigid fragment docking, 99
Rigid fragments and flexible chains, 97*f*
RMSD metric space, 101
RMSD-success rate curve, 124*f*
RNA polymerase, 82
Route Designer method, 30
Rule-of-three (Ro3), 50

S

Scaffold compounds, 54
Scaffold hoping, 11
SCF. *See* Surface Chemical Features (SCF)
SciFinder, 6
SciTegic, 64
Scoring function
  additional terms, 117
  docking programs, 108
Screening compounds, 52

SDsearch software, 51
SEAL score, 76
Search space size, 94
SEED. *See* Solvation Energy for Exhaustive Docking (SEED)
SeeDs. *See* Structural Exploitation of Experimental Drug Startpoints (SeeDs)
Separation algorithms, evaluation, 164*t*
Similarity inference network model, 59
SMARTS patterns, 10
SMILES string, 51, 155
Soak protein crystals, 47
Solvation Energy for Exhaustive Docking (SEED), 132, 137
  ligand triangle, 139
SPLICE method, 11
SPR. *See* Surface plasmon resonance (SPR)
SPROUT, 10
SQL database, 101
STAR technology, 12
Statistical data
  collection, 112
  fitting empirical functions, 116
Steric clash penalty term, 118
Structural Exploitation of Experimental Drug Startpoints (SeeDs), 5
Structure generation algorithm, 32
Sufficient sampling, defined, 93
SUM fusion rules, 66
Surface Chemical Features (SCF), 73, 84
Surface plasmon resonance (SPR), 2
Surflex data set, 125
  enrichment results, 126*f*

T

TAN. *See* Tanimoto-based similarity searching system (TAN)
Tandem reaction, 41*f*
Tanimoto-based similarity searching system (TAN), 65
Target-based drug design
  hybrids: scoring and ranking, 85
  library design, 86
  method, 85
  overview, 84
Thrombin, 3
TopClass, 152
  classification algorithm, 157
Transition probabilities, calculations, 153
Trypanosoma brucei, 48
Two-way attachment recombination, 78

U

Urokinase, 3

V

Validation procedure, 37*f*
Van der Waals
  surface, 109
  term, 118
Various stochastic search methods, 92
Veber oral bioavailability screen, 158
VEGFR2 protein kinase, 86

W

Weighted-sum link matrix, 63
Weighting scheme, 62
West Nile virus, 144
WOMBAT. *See* World Of Molecular Bioactivity (WOMBAT) databases
World Of Molecular Bioactivity (WOMBAT) databases, 58

X

X-ray crystallography, 8
  data, 118
X-ray resolution, 76

Z

Zenobia Therapeutics, 6, 48
  fragment library, properties, 185
ZINC molecule, 134

Downloaded by 89.163.35.42 on June 1, 2012 | http://pubs.acs.org
Publication Date (Web): September 30, 2011 | doi: 10.1021/bk-2011-1076.ix002