Large-Scale Analysis of Protein-Ligand Binding Sites Using the Binding MOAD Database

by

Nickolay Khazanov

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2012

Doctoral Committee:

Professor Heather A. Carlson, Chair
Professor Brian D. Athey
Professor Daniel M. Burns Jr.
Associate Professor Yang Zhang
Assistant Professor Mathew Young

ACKNOWLEDGEMENTS

I would like to acknowledge the long term confidence and support of Dr. Heather Carlson. Her passion, dedication, and attention to detail have set the standard for my current and future scientific endeavors. I am proud to have been part of a team that was dedicated to exciting and quality research, and I will always be honored to have been a HACer. I humbly thank my dissertation committee and the Bioinformatics program directors for their time, insightful advice, and saintly patience and support all along my winding road through the program.

I would like to thank the current and past members of the Carlson lab for making my work possible. Binding MOAD would not be the great scientific tool it is today if it wasn't for tremendous efforts of Richard Smith and Mark Benson, my brothers-in-arms in the ongoing fight for more binding constants. The wonderfully-elegant HwRMSD project would not have been possible without Kelly Ganamet and timely help from Daniel Quang. I acknowledge Jim Dunbar for his sage advice, and many hours of helpful and engaging scientific discussion. Thank you to Jerome Quintero, Katrina Lexa, Anna Bowman, Michael Lerner, and Man-Un (Peter) Ung for their tremendous friendship and wiling participation in the many therapeutic lunch breaks and coffee-runs.

Many projects in this thesis would not have been possible without the tremendous work done by the open-source scientific community. I wish to acknowledge the MySQL consortium for their constant improvement of the MySQL database, the Python community for many great libraries like NumPy and Biopython, the Django Project for their wonderful web framework, Ethan Merritt and Jay Painter for their mmLib library, and Warren DeLano for the outstanding PyMol application. Thanks go out to John Beaver, Brandon Dimcheff and Peter Dresslar at Torrey Path LLC, whose BUDA application has saved me countless hours of laborious effort that can accompany the annual Binding MOAD update process.

I extend enormous thanks to my family for their patience and long-term support in my pursuit of this degree. Their thoughts of encouragement were with me even though they were many miles away. They never lost faith in my potential, even when my own confidence faltered. Thank you for all your kind words of support, the many phone calls and your patience during the long periods between my visits home.

Last, but definitely not least, I would like to acknowledge my friends, both in Ann Arbor, and in Edison. You have all provided me with limitless inspiration, motivation, and general good vibes. I will not forget the great memories made during my time at Michigan. The biggest thank you goes out to Yuri Ikeda, who has directly (as the Bioinformatics student services coordinator) and indirectly (as a personal friend and companion) sacrificed much of her time and energy to make this work possible. Her bright smiles of support in times of success, and her firm hand-to-the-forehead smacks in times of un-warranted despair, have been major motivating factors in the completion of this work.

The work in this thesis was supported by a Training Grant from the National Institutes of Health (GM070449-01A1) and NSF Career Award for HAC (MCB 0546073).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Large-Scale Analysis of Protein-Ligand Binding Sites Using the Binding MOAD Database.

by

Nickolay Khazanov

Chair: Heather A. Carlson

Current structure-based drug design (SBDD) methods require understanding of general tends of protein-ligand interactions. Informative descriptors of ligand-binding sites provide powerful heuristics to improve SBDD methods designed to infer function from protein structure. These descriptors must have a solid statistical foundation for assessing general trends in large sets of protein-ligand complexes. This dissertation focuses on mining the Binding MOAD database of highly curated protein-ligand complexes to determine frequently observed patterns of binding-site composition. An extension to Binding MOAD's framework is developed to store structural details of binding sites and facilitate large-scale analysis. This thesis uses the framework to address three topics. It first describes a strategy for determining over-representation of amino acids within ligand-binding sites, comparing the trends of residue propensity for binding sites of biologically relevant ligands to those of spurious molecules with no known function. To determine the significance of these trends and to provide guidelines for residue-propensity studies, the effect of the data set size on the variation in propensity values is evaluated. Next, binding-site residue propensities are applied to improve the performance of a geometry-based, binding-site prediction algorithm. Propensity-based scores are found to perform comparably to the native score in successfully ranking correct predictions. For large proteins, propensity-based and consensus scores improve the scoring success. Finally, current protein-ligand scoring functions are evaluated using a new criterion: the ability to discern biologically relevant ligands from “opportunistic binders,” molecules present in crystal structures due to their high concentrations in the crystallization medium. Four different scoring functions are evaluated against a diverse benchmark set. All are found to perform well for ranking biologically relevant sites over spurious ones, and all performed best when penalties for torsional strain of ligands were included. The final chapter describes a structural alignment method, termed HwRMSD, which can align proteins of very low sequence homology based on their structural similarity using a weighted structure superposition. The overall aims of the dissertation are to collect high-quality binding-site composition data within the largest available set of protein-ligand complexes and to evaluate the appropriate applications of this data to emerging methods for computational proteomics.

CHAPTER I

Introduction and Background

1.1 Overview

Proteins are ubiquitous in the cellular environment and are essential for biochemical functions that sustain life. To accomplish their task, proteins interact with a variety of entities in the cell, some macro-molecular (DNA, RNA, membranes) and others smaller (catalytic substrates, nucleotides, peptides, and man-made chemicals). The diverse interactions between proteins and their small molecule ligands are the focus of intense study, not only for elucidation of cellular mechanisms, but also for facilitating the design of drugs to modulate protein function in disease states. A significant portion of this design process is based upon the three dimensional structures of the protein and the ligand, in complex or separate.

The challenge of structure-based drug design (SBDD) is to correctly predict which small molecule would bind to a specific protein and what impact it might have on its function. SBDD involves an intimate understanding of how a ligand interacts with its binding site on the protein surface and using that information to predict what other ligands can bind and how strong the binding will be. Screening large databases of potential lead compounds against a structure of a protein can speed drug development by focusing the more resource-intensive experiments on a narrow set of compounds most likely to have activity.

In general, two major requirements for SBDD are the availability of a 3D structure of the desired protein target and the annotation of the ligand binding site(s) on that protein. In the recent years, SBDD received a major boost as the number of known protein structures has grown exponentially, thanks in part to numerous structural genomics projects aimed at obtaining X-ray crystal structures of proteins with unknown or poorly understood function [1]. While some contain co-crystallized ligands, most are un-liganded. This large number of “incomplete” protein structures has raised new challenges in predicting and characterizing protein-ligand interactions. One of the more fundamental challenges is the identification of pockets on the protein surface capable of binding a small molecule. This can be a relatively straightforward process if the protein in question is a member of a well-studied family because function and binding pockets can be identified through sequence homology to an existing protein structure. However, the aim of many of the structural genomics projects is to find new uncharacterized targets, so many of the proteins in the new structures may have low sequence homology to any other structure in the Protein Data Bank (PDB) [2]. In these cases, structure-based methods need to be employed. Methods for determining the location of binding sites include computational and experimental fragment-based screening, identifying structural similarity to a known binding site, or using a structure-based prediction tool.

Once a binding site is located, the subsequent challenge is to describe the functional significance and selectivity/specificity of the site for various small molecules. While classical SBDD methods such as *in silico* screening can be used to determine possible binding partners, the existence of extensive databases of protein-ligand binding pockets motivates the development of comparative methods to leverage this structural data. One way to use this data is to derive a “druggability” or “bindability” score that is trained on existing protein-ligand structures. A druggability score can be used to rank identified pockets by how likely each is to bind a drug-like ligand. The ‘druggability’ index approach is relatively popular, but it has difficulty separating pockets known to bind drugs from pockets not suited to drug design, known to bind only small organic molecules. This is likely due to the high similarity of such sites and the limited dataset used to train the score [3]. Comparing one binding pocket to another to find significant similarities is a knowledge-based way to infer the ligands capable of binding to a pocket. The challenge faced by this approach is usually determining the significance of the similarity by estimating a cutoff value on a sufficiently diverse set of example complexes. These knowledge-based approaches also have the potential to address significant challenges in the drug-design process, – such as predicting off-target effect of new drugs [4], finding secondary targets for existing drugs [5], and engineering novel proteins to bind specific ligands [6]. The development of such tools requires an intimate understanding of the variation within binding pockets across the known "pocketome" and their relative importance in partitioning the vast many-to-many interaction network of proteins and small molecules.

All of the above-mentioned challenges require methods that incorporate efficient yet informative descriptors of structural, physicochemical, and dynamic features of the protein surface. Also, they must have a solid statistical foundation for assessing similarities or differences between such regions. This dissertation focuses on mining the Binding MOAD database, a vast, curated set of protein-ligand complexes, for frequently observed patterns of binding-site composition with respect to ligand type. We then apply these patterns to the improvement of emerging binding-site prediction and comparison methods. After a short description of Binding MOAD and the extensions made to the database framework to facilitate binding-site analysis, the thesis addresses three topics. It first describes a strategy for determining over-representation of certain amino acids in ligand-binding sites. The significance of this over-representation is assessed by comparing binding pockets of biologically relevant ligands to those that bind spurious molecules with no known function. Significant trends in the propensities are described, and the effect of the data set size on the variation in propensity values is determined. Second, the propensities of residue occurrence are applied to improve the performance of a *de novo* binding-site prediction method. Propensity-based scores are used to rank predictions from the geometry-based SiteFinder algorithm. They are found to perform well on their own and in combination with the default SiteFinder score. The effect of predicted site size and protein size on prediction success is examined to identify cases where propensity-based scores are especially helpful.

Finally, this dissertation describes an evaluation of current protein-ligand scoring functions with a new criterion – the ability to discern biologically relevant ligands from "opportunistic binders," molecules present in many crystal structures due to their high concentrations in the crystallization medium. A diverse benchmark of both types of complexes is compiled from the PDB and evaluated with four different scoring functions representing different scoring approaches. Particularly challenging examples are examined, such as those where biologically relevant binding is weak or invalid ligands that mimic biologically relevant contacts in a known binding site. The final chapter presents a new method for better structural alignment of two homologous proteins based on the weighted superposition (wRMSD) technique [7]. This new HwRMSD method performs comparably to established structural alignment methods and is effectively used on proteins pairs with large structural changes. While HwRMSD does not utilize binding-site information, it has potential applications to structural comparison of ligand-binding regions. The overall aims of the dissertation are to collect high-quality data to describe the composition of binding sites within the largest available set of protein-ligand complexes and to evaluate the appropriate applications of this data to emerging methods for computational proteomics.

1.2 Understanding General Trends of Protein-Ligand binding

The understanding of protein-ligand binding has come a long way since the formulation of the lock-and-key hypothesis in 1984 by Hermann Fischer, who suggested the binding of a substrate to an enzyme is analogous to a key being inserted into a lock. This model of shape complementarity between the ligand and the receptor has since expanded to incorporate the flexibility of the receptor. More dynamic models of binding include the zipper model, the induced-fit model, and the conformational-selection model (reviewed in [8]). Considering that a majority of structural knowledge of proteins still comes from static structures obtained by x-ray crystallography [1] or homology modeling [9], it is still a challenge to identify binding regions that might only be present in certain protein conformations, yet are important for the function of the protein. Geometric complementarity is required but not sufficient for ligand binding [10]. Molecular recognition also relies on physicochemical complementarity – namely the various electrostatic, hydrogen-bonding, hydrophobic, and solvent-mediated interactions between the protein and the ligand. All of these features must combine to create an energetically favorable interaction for the ligand to enter and remain in the binding site long enough to affect protein function. Despite major efforts to develop computational methods that can describe these physicochemical interactions from "first principles" using sophisticated force-fields, a large portion of existing methods rely on knowledge-based or empirical approaches.

The knowledge-based methods are fast and efficient, but they suffer from over-reliance on their training sets, which can limit their generalizability [11]. Empirical methods parameterize their energy-estimation functions with existing protein-ligand data, and thus, offer an intermediate approach. However, none of these methods have been very effective in predicting experimentally-determined ligand affinities [12], illustrating the challenges inherent in understanding the full range of factors governing protein-ligand binding.

1.2.1 Binding-Site Shape

Looking at the variation of shapes, sizes, and composition of protein-ligand binding sites and the ligands they bind, it is easy to see why finding a general method for predicting their location and binding partners is such a challenge. Recent studies of thousands of human protein-ligand complexes found a complicated relationship between the similarity of protein sequences and the similarity of their pockets and bound ligands [13-15]. By clustering the proteins, ligands, and pockets separately, one study found many examples of highly related proteins binding varied ligands. Conversely, many ligands similar in structure, as measured by Tanimoto similarity, have unrelated protein partners, as measured by sequence similarity [13]. Moreover, it has been observed that binding sites of two proteins can be similar despite having different global folds [16], which is likely a result of convergent evolution.

The shapes of binding pockets range from small spherical invaginations to deep curved or linear clefts in the protein [17]. Catalytic sites are usually situated at the bottom of the deeper clefts, where hydrophobic residues can shield the catalytic residues from solvent while the latter perform the enzymatic reaction. However, other functional sites can vary greatly in size and depth. To complicate matters, the size of the binding site is not necessarily related to the size of the ligand it can accommodate [10, 18], and several binding regions can exist in close proximity, forming a large swath of ligand-binding surface with complex geometry. The average volume of a drug-binding pocket is between 600 - 900 Å^3, depending on the method used to delineate pocket boundaries [19-21], while that of a drug-like ligand is around 400Å^3. Since drug molecules are often designed to be as small as possible to improve their bioavailability, this range of binding site and ligand volumes only increases when the full variety of biologically-relevant protein-ligand complexes is considered. Despite the variation, some general trends have emerged from the recent studies. Ligands, and especially drugs, have been observed to bind into the largest and/or deepest concavity on the protein surface [18]. On average, that cavity might be three times as large as the ligand, indicating the presence of a large ”buffer” zone between the ligand and protein [10] and demonstrating the difficulty in defining the boundaries of what really constitutes the binding site.

1.2.2 Binding-Site Composition

Chemical complementarity may have general trends as well. In a study of ligand efficiency in enzymes and non-enzymes, high-affinity enzyme ligands were observed to be larger than those with low-affinity, indicating increasing ligand size can improve affinity. However, non-enzymes were observed to have high ligand efficiencies irrespective of ligand size, and the composition of their binding sites had greater influence upon modulating affinity [22]. Such relative trends between protein classes are helpful in guiding SBDD with a particular protein target in mind, but provide only broad-brushstroke insight into binding-site behavior. Many methods that delve into the details of the protein-ligand interactions have been developed to score potential matches between a specific ligand pose and its receptor [23-26]. These knowledge-based scoring functions look for general trends of atom-atom contacts between ligand and protein in large structure databases, and they reward frequently-seen interactions in the potential protein-ligand pair to be scored. Depending on the training set, the interaction trends captured by the function may not be generally transferable to a wide variety of proteins [27]. Moreover, the trends utilized by the scoring functions require the presence of both protein and ligand atoms in a structure. They are usually applied in cases where the relative location of the binding site has been narrowed down, and only the best binding mode is sought [28]. This limits their usefulness in understanding general trends of binding-site composition in the absence of a potential ligand.

Yet another class of studies has analyzed the composition of protein surfaces in general. This led to general insight that a mix of both hydrophobic and hydrophilic solvent-exposed residue exist on protein surfaces [29], and that there is limited correlation between residue hydrophobicity and its surface exposure patters [30] (i.e., its presence on the surface). Such insights further the understanding of the composition of protein surfaces, but they do not compare and contrast their findings with the regions of protein surface involved in ligand binding.

With the advent of thousands of protein-ligand complexes in the PDB, analyses of composition and conservation of residues in ligand binding sites are becoming more common. Such studies take a more asymmetric view of the protein-ligand interactions, focusing on the trends in protein composition independently of the details of ligand interaction. The trends are often linked back to significant or frequently observed interactions, but they can also be used to contrast ligand-binding regions of the protein to the rest of the protein surface or assign some manner of scores to protein residues in a structure without co-crystallized ligand.

One of the most detailed studies of ligand-interacting residues was carried out by Bartlett et al. on ~200 enzyme active sites [31]. A residue’s role in enzyme catalysis was confirmed by manual curation. His, Cys, Asp, and Arg were found to have the highest catalytic propensities (over-representation in catalytic site as compared to the rest of the protein). A variety of other detailed trends for the solvent exposure and biochemical function of the residues were also determined. The study provided a heuristic basis for predicting catalytic residues in enzymes, and it was the source for the Catalytic Site Atlas database. However, the focus on catalytic residues side-stepped the analysis of residues involved in non-catalytic interactions with the ligand. These residues likely provide energetically-favorable interactions that maintain the ligand in the correct binding mode while the catalytic residues perform their function; therefore, they are important to include in residue composition analysis.

A recent study by Davis and Sali examined the general residue composition of ligand-binding sites as compared to protein-protein binding sites and the protein surface with no known interactions [32]. They found residue composition in protein-protein binding sites resembles that of the general protein surface, while residues in protein-ligand binding sites had much different propensities. Residues involved in both protein and ligand binding (bi-functional) showed intermediate propensities. Largest propensities for ligand-binding sites were seen for Cys, Phe, His, Met, Trp, Val, and Ile residues. Cys, Ile, and Val were unique with respect to protein-protein or bi-functional positions. That study analyzed ~35,000 binding sites, but it was based on only ~1000 domain families, which is a very redundant set of structures.

A more focused study by Nayal and Honig extensively characterized the binding sites of 99 non-redundant, protein-ligand complexes as part of an effort to develop a binding-site detection algorithm [18]. Unfortunately, the authors focused on classification of binding sites as being drug-binding or not, and they did not report general trends of the binding pockets analyzed in the study. However, they found that Asn, Gln, and Glu were important residues in recognizing drug-binding sites among a set of sites binding various ligands. Aside from these representative efforts, relatively few surveys of large sets of protein complexes have been completed [33-37], and the ever-growing number of raw structural data and binding-site characterization methods promise greater insights into the general theories governing protein-ligand complexes. These insights will undoubtedly lead to improvements in SBDD strategies that make use of these theories [36, 38].

1.3 Methods to Identify Binding Sites *de novo*

Methods that aim to predict binding sites face several major challenges. Two of the largest are the change in the binding-site structure upon ligand binding (an implication of the induced-fit model of ligand binding) and the sheer variety of existing binding-site shapes and sizes needed to accommodate the various biologically-relevant ligands. Another challenge is detecting sites located at protein subunit interfaces, which are often omitted in the test sets used to develop and benchmark prediction methods. Even with a large and properly chosen evaluation set, there remains the challenge of precision and accuracy, i.e., identifying the precise region expected to bind a ligand without over-predicting. Overpredicting is problematic because classifying the entire surface of a protein as a binding site would certainly result in a match, so evaluating the success of a prediction must be done carefully. The field as a whole still lacks gold-standard sets or consistent metrics for calling “true” predictions. These challenges will need to be overcome for successful application of prediction methods in SBDD, which require a precise and accurate definition of the binding site to focus the search effort on relevant areas of the protein and reduce false-positive results.

Three major classes of tools have emerged to address binding-site prediction: 1) those that use the geometry of the protein surface to identify large concavities resembling a binding site, 2) those that use probe-based methods to identify regions of the surface capable of making energetically favorable interactions with a potential ligand, and 3) those that use knowledge-based methods to search a protein structure against a database of known binding sites. Representative tools in these categories are discussed below and summarized in Table I-1. Additional tools exist that use more complex methods, such as molecular dynamics, to identify the binding regions, but they are outside the scope of the current discussion.

Table I-1: Representative algorithms for binding site prediction.

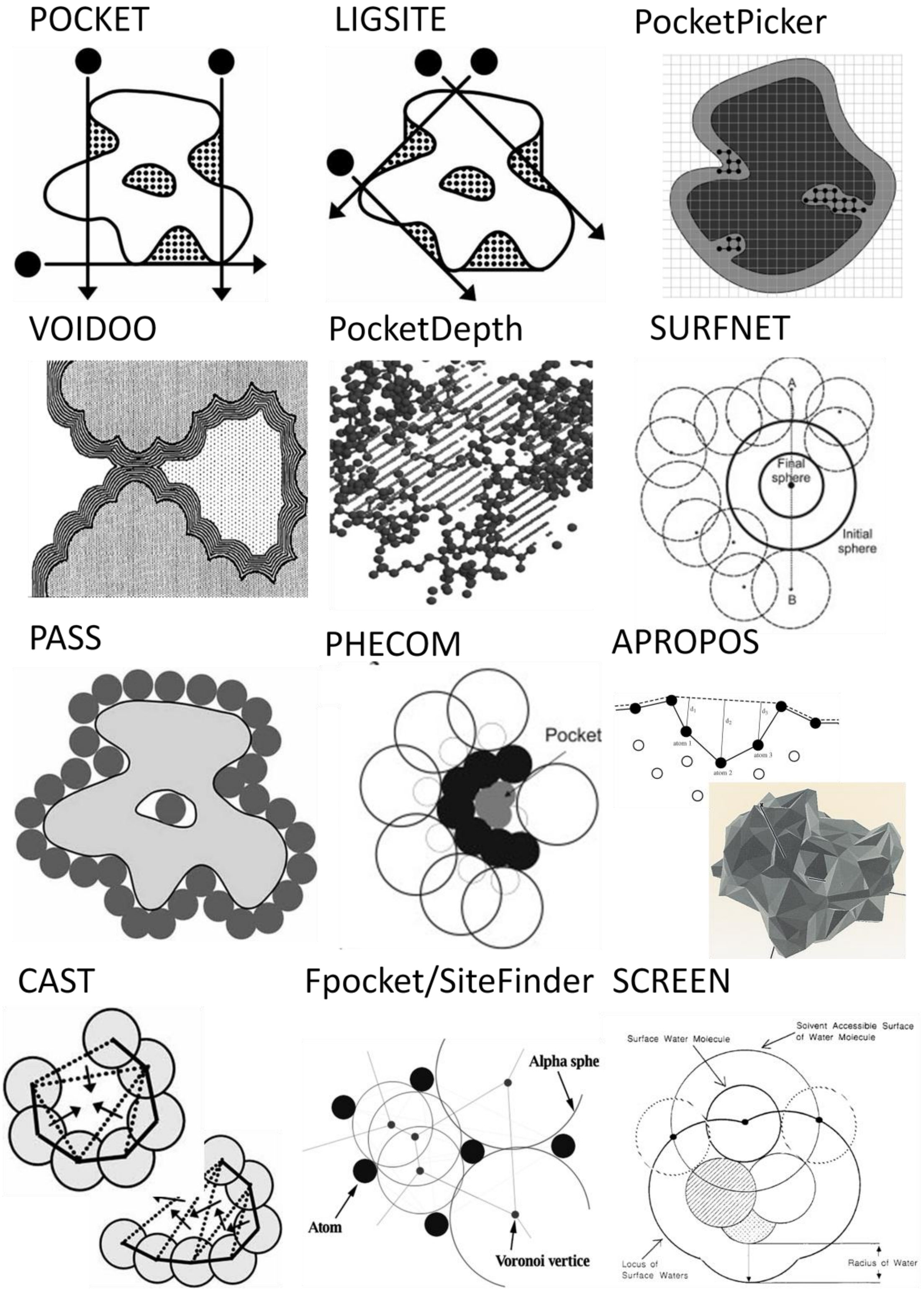
	Pub Date (Latest)	Type	Server	Download	Open Source
GRID	1985	Energy-based			
POCKET	1992	Geometric			
VOIDOO	1994	Geometric		✓	
APROPOS	1996	Geometric			
PASS	2000	Geometric		✓	
eF-Site	2002	Knowledge-based	✓		
Pocket-Finder	2005	Geometric	✓		✓
PocketFinder	2005	Energy-based	✓	$	
SiteFinder	2005	Geometric		$	
Q-SiteFinder	2005	Energy-based	✓		
ProFunc	2005	Knowledge-based	✓		
ConSurf	2005	Knowledge-based	✓		
SiteEngine	2005	Knowledge-based	✓		
LIGSITEcsc	2006	Geometric/Genomic	✓	✓	✓
SURFNETConSurf	2006	Geometric/Genomic	✓	✓	✓
CAST(p)	2006	Geometric	✓		

SCREEN	2006	Geometric		✓	
Pocket-Picker	2007	Geometric		✓	✓
PHECOM	2007	Geometric			
SiteMap	2007	Combination		$	
CavBase	2007	Knowledge-based	✓		
PocketDepth	2008	Geometric	✓		
FINDSITE	2008	Knowledge-based	✓		
Fpocket	2009	Geometric	✓	✓	✓
McVol	2010	Geometric		✓	✓
PROSITE	2010	Knowledge-based	✓		
ProBiS	2010	Knowledge-based	✓		

1.3.1 Methods Primarily Using Geometry of Static 3D Protein Structures

The earliest methods developed for prediction of surface pockets employed a grid-based approach, scanning the protein along grid lines using a geometric probe and detecting the regions where grid points lay outside of protein atoms (Figure I-1). In POCKET [39], the specific pockets were defined by protein-solvent-protein events, which are characterized as a series of grid points along the scan axes that alternate between being “inside” the protein to “outside” of the protein. Since this approach is sensitive to the relative orientation of the protein to the grid axes, it was extended in LIGSITE [40] to include scans along grid diagonals. LIGSITE was tested on a set of only 10 proteins in its initial publication, but at the time, it was a state-of-the-art method in terms of accuracy and efficient implementation. Owing to its success, it was further extended to incorporate the Connolly protein surface [41] in order to count surface-protein-surface events instead of protein-solvent-protein events. Recently, non-structural information was incorporated by considering the conservation of residues in the identified pockets. [42]. The extended version is named LIGSITE^{csc}. It was validated on a large set of 210 structures, which included a subset of matched *apo* and *holo* protein structures. It showed an improvement in top-ranked successful predictions from 67% to 75% as compared to LIGSITE. LIGSITE was also implemented as Pocket-Finder for comparison to energy-based methods [43].

Figure I-1: Illustration of several methods utilizing protein-structure geometry for binding-site identification. Each illustration is taken from the publication describing the respective algorithm.



LIGSTE served as a basis for the PocketPicker method, which used a finer grid representation and calculated an additional "buriedness" measure for grid points in the identified pockets [44]. The PocketPicker method performed worse (59% for top-hit success) than LIGSITE or LIGSITE^{csc} on the same test set of 210 complexes, a fact the authors attributed to the optimization of PocketPicker to identify smaller, more buried pockets useful for shape comparison. However, it did outperform alternative geometry-based methods PASS and SURFNET (see below). Conceptually, grid-based methods have limits due to their sensitivity to grid size, protein orientation, and the inability to define a cavity "ceiling" to delineate the outer limit of a pocket. VOIDOO was an early method that sought to rectify the limitations of grid-based approaches by detecting cavities through the expansion of atomic van der Waals (vdW) boundaries [45]. The premise was that deep pockets can be delineated by progressively expanding the vdW radii of all protein atoms until invaginations on the surface of the protein get "pinched off" by the vdW surfaces colliding at the narrowest point, i.e. the "mouth" of the pocket (Figure I-1). Although the method provided a way to outline the cavity and measure its volume independent of a grid, it could not detect shallow or broad pockets that cannot be closed off by increasing the vdW radii.

A slightly different use of grids was recently proposed in the PocketDepth method [46]. It uses the fine grid of points placed on a protein to calculate pair-wise measures of depth between pairs of points flagged as being on the surface of the protein. The points are identified as being on the surface by considering the density of protein atoms in the vicinity of a point. Depth measurements passing through a protein atom are not considered. Subspaces of the grid are evaluated based on the density and magnitude of the depths measurement among the points in the subspace. Complex clustering and filtering steps allow the algorithm to identify grid subspaces that have large numbers of 'deep' points in close spatial proximity, and thus delineate the predicted pocket shape based on these subspaces. The algorithm was extensively evaluated on a set of 1091 proteins from the PDBBind database [47], where it achieved a success rate of 55% among its top-ranked predictions. This performance was similar to or better than LIGSITE^{CSC} and Q-SiteFinder methods, respectively, based on the benchmark datasets for those methods (see below).

An early, alternative geometry-based method was implemented in SURFNET [48], which used an algorithm that fits spheres of varying size between all pairs of relevant protein atoms to find the pocket-like cavities. The radius of a sphere placed between several atoms is iteratively reduced until no overlap with protein atoms is achieved (Figure I-1). All indentations on a protein are packed with spheres of radii varying from 1 to 4 Å, defining the outlines and volumes of prospective binding sites. Ranking the pockets by size, SURFNET was found to correctly predict the ligand-binding pocket in 83% of the top-ranked results for a set of 67 enzyme complexes [49]. In a more recent test on 210 bound complexes from Huang & Schroeder, SURFNET underperformed in relation to LIGSITE and PASS methods, achieving a success rate of only 42% for its top-ranked predictions [42]. SURFNET was later expanded to SURFNET-ConSurf, which incorporates evolutionary sequence information to trim the predicted pocket size based on the conservation rate of its component residues.

The PASS method [50], whose top-ranked predictions achieved a performance of 54% in the Huang et al. evaluation on 210 complexes, is another popular tool that uses a probe-packing algorithm to detect protein pockets. In this algorithm, probes are packed on the protein surface so that each probe touches a triplet of adjacent protein atoms. Probes that clash with protein atoms are then removed, and the remaining probes are filtered by their degree of burial, estimated from the count of protein within 8 Å of the probe. Subsequent rounds of packing build up additional layers of probes until no more buried probes can be placed. The probes are clustered into "active site points" that define the predicted pockets. PASS was originally evaluated on 32 complexes, where its top-predicted pockets successfully identified 60% of the known sites, and on a set of 21 *apo* structures, where the success rate was 57%. A more recent method, PHECOM, used two sets of large and small spheres to pack the protein surface, and determined pockets by identifying small spheres that packed between the surface and the large spheres [20]. Probe packing methods might have difficulty detecting wide cavities, which would require very large spheres to define, and would limit the accuracy of pocket-volume estimation based on sphere volumes.

A series of algorithms utilize Voronoi diagrams [51] and Delaunay triangulation [52] for geometric surface representation. These techniques effectively "shrink wrap" the protein surface in a mesh of triangles and allow for complex geometric algorithms to calculate various shape properties of this tiling. APROPOS [53], one of the earliest methods in this category, creates a Delaunay representation of the protein and then generates an α-shape, which is conceptually similar to rolling a probe with a radius α to erase edges of the Delaunay triangles. The vertices of the triangles are located at atomic centers and unaffected by the rolling-probe step. By varying the α parameter, a series of surfaces can be generated, with the most extreme value (α = ∞) generating a convex hull of the protein. Pockets of various sizes can be detected by comparing the shape of the convex hull to the alpha shapes generated by alpha values corresponding to the radii of an oxygen atom or a methyl group. The algorithm initially achieved a 95% success rate in its initial test on 200 monomer complexes, but it has not been extensively compared to other methods. CAST is a similar method that combines the Delaunay representation with discrete flow theory to identify concave pockets [35]. It first identifies the alpha-shape of the protein and defines Delaunay triangles with one or more omitted edges as "empty". The neighboring empty triangles are then combined in the "discrete-flow" method to define continuous voids on the protein surface (Figure I-1). An obtuse empty triangle flows into its neighbor, while an acute empty triangle acts as a sink to collect the flow of its neighbors. If the flow is directed out of the protein the pocket is ignored, otherwise it is considered a putative binding site. The algorithm was tested on the 51 of the 67 proteins used by SURFNET, and it achieved a success rate of 74% (lower than SURFNET). However, the authors determined that the differences in nature of the pockets predicted by the two methods prevent fair comparison. The α-shape and secrete flow approaches might be limited in their applicability to very open pockets, since they are optimized to detect pockets whose "mouth" is smaller than the cross sections through the rest of the pocket [54]. Alpha-sphere methods, discussed below, are less sensitive to such pocket geometry.

More recent methods utilizing a surface-based approach include Fpocket [55] and SiteFinder [56]. Both use the concept of α-spheres [54] andVoronoi tessellation of the protein. An alpha sphere is a sphere contacting four protein atoms but no internal atoms (Figure I-1). Centers of alpha spheres correspond to the vertices of the Voronoi tessellation of the protein, and both Fpocket and SiteFinder algorithms determine the set of alpha spheres based on this tessellation. Since alpha-sphere radii scale with the curvature of the plane of the four atoms they contact, small alpha spheres are located in tightly packed areas of the protein while those with larger radii are located in cavities towards the surface. Thus, clusters of alpha spheres with a desired range of radii can be used to identify and define surface cavities. Fpocket clusters the spheres by proximity, density, and size. It then prunes uninteresting sphere clusters based on cluster size and some basic definitions of polarity with respect to the atoms the spheres touch. The sites are then ranked according to a scoring scheme that estimates the "bindability" of the site based on a set of geometric and physicochemical pocket descriptors. SiteFinder performs similar clustering and pruning based on the solvent exposure of the spheres and the hydrophobicity of the neighboring atoms. Clusters with at least one "hydrophobic" sphere are retained while the rest are discarded. Fpocket was tested on the smaller bound/unbound test set of structures used by PocketPicker (see above). It outperformed PocketPicker, LIGSITE^{csc}, CAST, PASS, and SURFNET on the bound set of 48 structures, with an 83% success rate among its top-ranked predictions [55]. On the unbound set of the same proteins, it performed as well as PocketPicker, achieving a 69% success rate. It slightly underperformed compared to LIGSITE^{scs} (71% success). In a recent large-scale comparison on a dataset of several thousand *apo* and *holo* structures, Fpocket performed similar into SiteFinder on the bound structure set, achieving top rank success of around 78%, compared to SiteFinder's 77%. However, it significantly underperformed on the unbound set, where it achieved a success rate of only 42% versus 62% achieved by SiteFinder [57].

Another way of defining a protein surface is by a rolling-sphere approach. The SCREEN method uses differences in two molecular surface representations of the protein to define protein cavities [18]. First a "tight" molecular surface is constructed by rolling a small sphere (1.4Å in radius) over the whole protein. Then, a low-resolution envelope surface is constructed by rolling a sphere closer in size to a ligand molecule (5.0Å in radius). The depth of the molecular surface is then computed with respect to the envelope surface, and cavity surfaces are defined as contiguous regions of the molecular surface that are below a certain depth. A sophisticated clustering method then merges continuous cavities to produce well-delineated, compact, predicted pockets.

This is the extent of the geometric component of SCREEN, which was further used to train a classifier for predicting druggability of the sites based on an array of physicochemical proteins of the predicted pockets.

Geometric methods are relatively simple and efficient, but they have no underlying physical meaning. The following section describes representative methods that apply a more physicochemical approach, or a combination of geometric and physical features, to scan the protein surface for potential ligand-binding sites.

1.3.2 Methods Primarily Using Energetic Mapping of 3D Protein Structure

Methods that evaluate the interaction energy between the protein and a chemical probe, fragment, or molecule, to determine favorable regions date back as early as 1985, when the GRID method was first published [58]. GRID calculates interaction energies between the protein atoms and probes placed on a grid superimposed on the protein. The interaction energy is composed of Lennard-Jones, Coulombic, and hydrogen-bonding terms, and the probe identity can be varied to represent different chemical fragments. The method was designed to identify regions of interest on the protein but not necessarily predict binding sites; as such, it generates a map of the protein surface as opposed to a set of pocket predictions. Although used heavily and successfully in SBDD of individual protein targets [59], it has not been extensively evaluated as a binding-site prediction algorithm. Probe interaction energies have been successfully combined with some measure of protein geometry in the recently-developed Q-SiteFinder [43]. The Q-SiteFinder algorithm relies on the GRID force field to calculate vdW interaction energies between the protein atoms and methyl probes placed on a fine grid. An interaction energy threshold is used to retain the most favorable probes. A clustering algorithm groups the probes according to their spatial proximity along the protein surface and uses the total interaction energy of the probes in a cluster to rank them relative to one another. The most favorably interacting clusters are presented as the predicted binding pockets. Q-SiteFinder was evaluated on a set of 134 proteins from the GOLD docking test set [60] and achieved a success rate of 71% in the top-ranked predicted sites [43]. The method also outperformed LIGSITE (as implemented in Pocket-Finder), which achieved a success rate of 48%. When evaluated on a set of 35 proteins with corresponding *apo* and *holo* structures, the Q-SiteFinder success rate in the top-ranked sites decreased from 80% in the bound structures to 51% in the unbound structures [43].

Shortly after the publication of Q-SiteFinder, an alternative energy-based method was proposed in PocketFinder [13] (not to be confused with Pocket-Finder, an implementation of LIGSITE). Like previous grid-based methods, PocketFinder calculates a grid potential map of vdW interaction energies using an implementation of the Lennard-Jones potential. The method deviates from its predecessors by smoothing this potential map to emphasize continuous regions of highly favorable vdW potential and then contouring the map at a level sufficient to identify putative ligand envelopes. Small envelopes are pruned, and the rest are ranked by their volumes. PocketFinder was tested on the largest set of any other method (5,616 bound and even more corresponding un-bound structures). It was found to perform better in predicting sites from ligand-bound structures than unbound structures. Since the authors chose to stratify their predictions by a site-coverage threshold, a direct measure of top-rank results is not available.

A more sophisticated approach similar to GRID and Q-SiteFinder is implemented in SiteMap, part of the Schrödinger software suite [61]. SiteMap first identifies potential sites by placing a grid of probe points over the protein and retaining those that are located “outside” of the protein, have a certain degree of enclosure, and obtain favorable vdW interaction energy with the protein. An agglomerative clustering step then groups the points by proximity, and point groups are further merged if an un-interrupted point-to-point traversal can be completed from one group to the other without encountering the protein. The predicted sites are then scored and ranked according to variety of aggregate physicochemical descriptors including size, enclosure, solvent exposure, tightness of non-bonded interactions, hydrophobicity, and hydrogen-bonding potential.

Energy-based methods tend to be much faster than geometric ones, but they are more sensitive to missing atoms and proper set-up of the protonation states and atom types of each protein [57]. Alternative energy-based methods include the fairly explicit and expensive computational solvent-mapping approaches [62, 63] and whole-protein "blind docking" methods such as MEDock [64] or an AutoDock-based method by Hetenyi et al. [65]. The former involve simulation of the protein surrounded by a bath of organic solvent to identify solvent-binding hot-spots. The later are based on existing docking and molecular dynamics software. Thus, both are beyond the scope of this introduction.

1.3.3 Methods Using Knowledge-Based Approaches to Identify Binding Sites

Knowledge-based methods used to predict ligand binding sites are comparative by definition. They make use of existing binding-site data obtained from X-ray and NMR experiments, biochemical data from site-directed mutagenesis studies, and sequence information associated with these data sources. Comparison of global or local structure of a protein of unknown function against existing databases of binding sites is the primary method for inferring the presence and/or function of the binding site. Sequence conservation or estimation of favorable energetic potentials is sometimes used to supplements the structural match. Unlike geometry- or energy-based methods, which report rank-ordered predictions, knowledge-based methods search against a large database of a known size, and then estimate the absolute statistical significance of their predictions. However, the obvious limitation of these methods is the difficulty of working with proteins that show little structural or sequence similarity to existing protein families. The power of knowledge-based methods will continue to improve as current sources of protein-structure data continue to grow.

There is an important distinction between predicting a binding site region on the surface of an uncharacterized protein and classifying the function of that region. Geometry-based and energy-based methods described in the previous sections focus on identifying the ligand-binding region, not on its classification. Although geometry-based methods such as SCREEN [18] follow up the binding-site identification step with a classification step, this classification step often relies on broadly-applicable binding-site properties to label the predicted regions as "druggable" or "bindable." In contrast, knowledge-based prediction methods compare a potential binding region to sites of known function, and consequently, try to assign a more specific functional class to their queries. Some knowledge-based methods that are considered binding-site *prediction* methods are actually binding site *classification* or *comparison* methods that rely on a geometry- or energy-based algorithm for predicting potential cavities to be used in querying known sites. For example, the popular Cavbase [66, 67] method actually relies on the geometry-based LIGSITE [68] cavity-prediction method to identify the potential pockets. The ProFunc [69] database uses the SURFNET [48] method to identify potential cavities. These predictions are then used for comparison to known ligand pockets using geometric and physicochemical criteria. The overview of the knowledge-based methods below focuses on the predictive aspect of these approaches. Binding-site similarity assessment is a broader extension of these methods that is outside of the scope of this chapter.

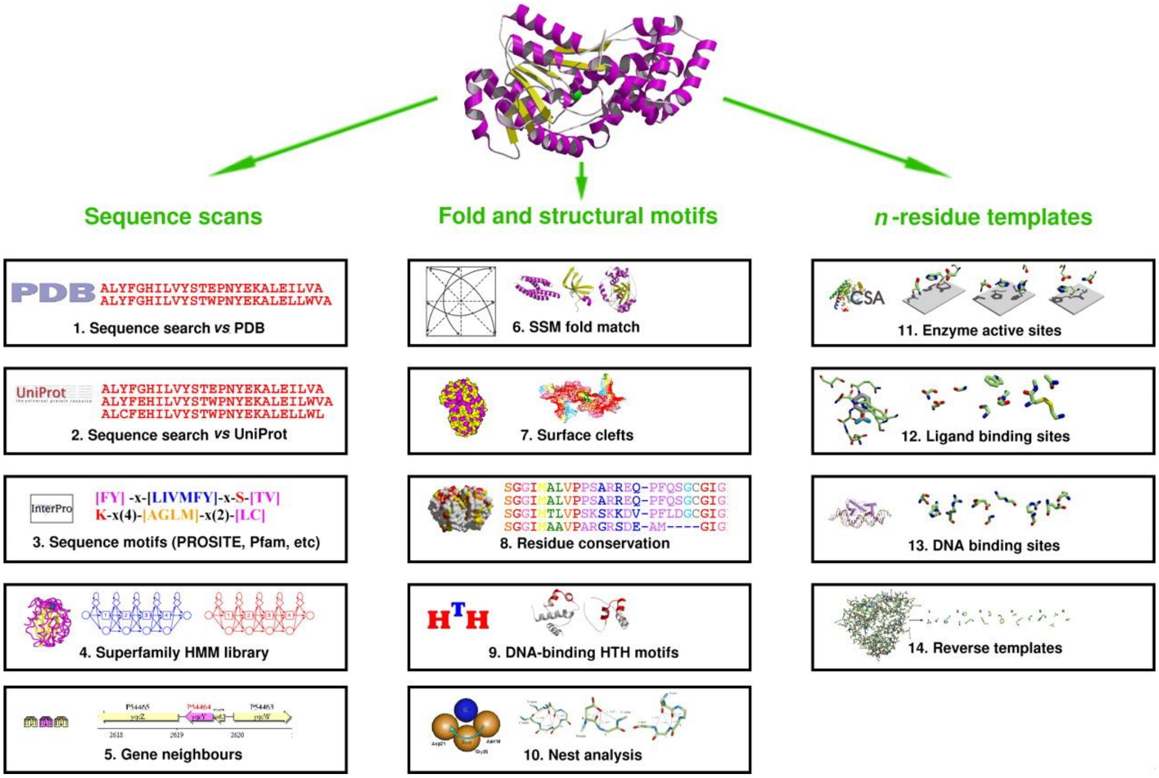
Evolutionary analysis can help identify patches of conserved residues on the protein surface that are essential to protein function. Conservation of residues in a family of proteins is obtained by a multiple-sequence alignment (MSAL), where family members are chosen by functional similarity, presence of a common ligand, or structural similarity. The Rate4Site method developed by Pupko et al. uses MSALs to estimate the rate of evolution among homologous proteins; it then maps the conservation data onto the 3D surface of the query protein [70]. Patches of conserved residues are assumed to have functional significance, but the interpretation of the specific binding-site location is left to the user. Bayesian inference used for the conservation calculations provides robust conservation scores that can differentiate highly conserved positions arising due to short divergence times between proteins, from less-conserved positions that deserve attention due to their relatively small changes across long phylogenetic lineages [71]. The ConSurf database uses the Rate4Site method to identify functionally important amino acids on a query protein surface by searching a large database of protein sequences to generate MSALs appropriate for estimating residue conservation. The self-professed bottleneck of this class of methods [72, 73] is the availability of sufficient sequence data. Too little variation in the MSAL due to insufficient diversity or too few sequences (usually < 10) can severely undermine the meaningful interpretation of the evolutionary relationships [74].

The PROSITE database also relies on sequence conservation to detect functionally important regions in proteins of unknown function [75]. It attempts to counteract the problems of insufficient sequence homology or small family size by considering short (10-20 residues), conserved sequence motifs in addition to full-length sequence profiles. A query sequence can be analyzed for the presence of a motif even it has no significant pair-wise sequence homology to known proteins. The short motifs can be constructed from full length MSALs in functionally-related proteins, such as enzymes or proteins containing prosthetic groups. Due to their size, individual patterns are not sufficient to infer function, but requiring the presence of several linked patterns in a query protein can boost their predictive power. [76]. If a significant match is found, the functionally-relevant residues can be identified. The exact functional relevance is dependent on the patterns used against the query. For example, patterns derived from nucleotide-binding sites will only be helpful in locating nucleotide-binding sites. Like any knowledge-based method, PROSITE performance is dependent on the presence of sufficient sequence information in its parent database – SwissProt. The diversity of motifs that can be generated from SwissProt bounds the type of functional residues that can be detected by this method.

When there is sufficient sequence homology between the query and a known protein structure, methods that combine the two sources of information might provide more effective functional site prediction than simple sequence comparison. FINDSITE is a method that can localize a ligand-binding site in a crystal structure or protein model by using a threading method commonly employed to build homology models of related proteins [77]. It can be used with query proteins that have < 35% sequence identity to the closest known reference structure, and it can tolerate differences of up to 10 Å RMSD between the query model and a known crystal structure.

The ProFunc database provides multiple sources of evidence for inferring the function of an un-characterized protein; many of which can be used to localize the functional site. It combines the largest-available variety of knowledge-based, functional prediction methods to take full advantage of an equally diverse set of data sources containing protein-function information [69] (Figure I-2). For example, ProFunc uses sequence-based searches against PROSITE and other structure/function databases to find evolutionary relationships to known functional classes. It also performs geometric comparison of local atom environments directly to 3D atom templates generated from known protein-ligand sites from the PDB[78]. The server uses the SURFNET program to generate query pockets to search against known structures and templates, and as such, does not constitute a stand-alone prediction method. However, the comprehensive combination of a multitude of search and comparison methods, make it a powerful knowledge-based tool for predicting functional sites in proteins.

Figure I-2: Schematic of different methods used by the ProFunc server to predict protein functional sites and infer protein function in general. The right-most column lists 3D template methods used to match potential cavities to existing PDB entries. Figure taken from Laskowski et al. [69]


Sequence scans
Fold and structural motifs
n-residue templates
PDB
ALYFGHILVYSTEPNYEKALEILVA
ALYFGHILVYSTWPNYEKALELLWVA
1. Sequence search *vs* PDB
UniProt
ALYFGHILVYSTEPNYEKALEILVA
ALYFEHILVYSTWPNYEKALEILWVA
ALCFEHILVYSTWPNYEKALELLWL
2. Sequence search *vs* UniProt
InterPro
[FY] -x-[LIVMFY]-x-S-[TV]
K-x(4)-[AGLM]-x(2)-[LC]
3. Sequence motifs (PROSITE, Pfam, etc)
4. Superfamily HMM library
P54465
P54464
P54463
yqcZ
yqcY
yqcW
2618
2619
2620
5. Gene neighbours
6. SSM fold match
7. Surface clefts
SGGIMALVPPSARREQ-PFQSGCGIG
SGGIMALVPPSARREQ-PFQSGCGIG
SGGIMTLVPSKSKKDV-PFLDGCGIG
SGGIMAAVPARGRSDE-AM----GIG
8. Residue conservation
HTH
9. DNA-binding HTH motifs
Asp21
Nest
Gly05
Asn18
10. Nest analysis
CSA
11. Enzyme active sites
12. Ligand binding sites
13. DNA binding sites
14. Reverse templates

Several knowledge-based methods are available that assume no sequence or fold similarity between the query and the known proteins. They solely rely on local structure similarity to locate potential ligand-binding sites. The algorithms at the heart of these methods can compare two sets of three-dimensional coordinates of atoms or residues to provide a measure of structural similarity between them. Since the location of a binding site on an un-characterized protein is not known, the methods need to employ clever optimization to search all possible regions of the query protein surface against the structures of known sites. As part of the optimization, matches of the sub-structure searches are often accompanied by a statistical measure of confidence in order to return only the most meaningful results. The global search across the entire protein makes these methods distinct from knowledge-based approaches that use geometric or energy-based methods to locate the potential pockets for use as structural queries. Of course, many of the methods can also be used as binding-site similarity tools to compare structural representations of known binding site, or predicted pockets.

The eF-site database was one of the first methods developed to search the molecular surface (MS) of a protein surface for potential functional sites [79]. In addition to using a geometric representation of MS, eF-site also calculates the electrostatic potential of the MS using the Poisson-Boltzmann equation. A clique-detection algorithm based on graph theory is used to compare the geometry and electrostatics of the entire MS of a query protein to a database of known ligand-binding proteins collected from the Protein Data Bank (PDB) [80]. This algorithm recognizes the region of the MS that matches known ligand-binding surfaces. It can be used to locate the binding site and assign potential function. The eF-site method uses a Connoly surface representation of the protein for its algorithm. The SiteEngine server developed by Shulman-Peleg et al. uses an alternative representation to achieve a similar high-performance search of a protein’s surface against a database of known ligand pockets [81]. Physico-chemical pseudocenters of surface residues are used to abstract the protein surface and reduce the dimensionality of the surface-comparison problem. Then, a geometric hashing algorithm is used to compare triangles of pseudocenters between the query and the database proteins. Enforcing the matching of physicochemical labels improves the algorithm efficiency by reducing the search space. Clustering and scoring steps help identify the regions of the query with the most relevant matches to a database ligand site. Like the eF-site method, SiteEngine is available via a server [82] capable of quickly searching a custom query structure against a database of pre-hashed sites from the PDB.

The recently-developed ProBiS algorithm [83], and its accompanying web server [84], employ yet another prediction method that detects structurally-conserved cliques of residues on the surface of a query protein. ProBiS represents the MS as a connected graph of functional groups derived from the surface residues. In searching a database of proteins, a maximum clique detection algorithm is used to detect subsets of functional groups that share geometric and physico-chemical complementarity. This algorithm effectively performs many local structural alignments between cliques of functional groups across the entire database. The results of these local alignments are aggregated to identify regions of strong structural conservation within the query structure and to select the most appropriate matching structure for further structural alignment and functional inference. The algorithm is fast enough to search a query structure against a pre-computed database of protein structures on–demand.

PINTS is another approach that uses a sophisticated algorithm to determine similarity between two sets of atoms regardless of the position of their residues in the sequence [85]. No information about the relationship between the atoms is needed aside from their 3D coordinates, and the search is not limited to protein surface atoms. A statistical model determines the significance of an RMSD between two sets of atoms and provides a confidence value that is dependent on the number and type of atoms, the obtained RMSD, and a background probability of a similar match against a large database of proteins [86]. This efficient algorithm can query a pattern of atoms against a database of protein structures in real-time, which allows searching an entire protein structure against the database without restricting the query to a small subset of atoms (although such restriction is also allowed).

The SiteEngine, ProBiS, and PINTS methods described above rely on their ability to quickly perform many small (local) pair-wise structure comparisons and then “grow” the individual matches to include larger protein–surface regions. These larger regions can be further compared to the database or simply reported to the user as the potential matches. When presented with the challenge of searching a whole protein surface against a database of such proteins, breaking down the problem into smaller, more efficient pieces is necessary. However, all methods utilize heuristics that help them prioritize and streamline which local surface regions are compared, and more heuristics to decide how they are eventually combined to delineate a matching site. Since no assumption is made about the location of a binding site in the query protein *a priori*, these methods would benefit from any information that will help them with such prioritization [85]. Trends of general binding-site composition can provide such heuristics, which can potentially improve the performance and accuracy of many of these methods.

1.3.4 Evaluation of Binding-Site Prediction Methods: Test Sets and Performance Metrics

The variety of proposed binding-site prediction methods is accompanied by a similar variety in metrics and datasets used for their evaluation. Most current methods have been developed in the past 10 years, and no gold standards have yet been established. Understandably, each publication proposing a new method aims to selects a bigger, better, and broader set of structures to highlight the strengths of their approach. Unfortunately, not many of these publications attempt to benchmark their tools on previously published datasets or to compare their novel method to existing ones. Notable exceptions to this are listed in Table I-2 in an attempt to highlight structure sets on which several methods have been evaluated. Many of these sets correspond to those discussed in the previous sections.

Table I-2: Notable test sets for binding-site prediction algorithms

Compiled By	**Used By**	**Evaluated algorithms**	**Basis**	**Size**	**Number of corresponding unbound structures**
Laskowski et al. (1995) [48]	Liang et al. (1998) [35]	SURFNET CAST	PDB [2]	67 proteins	--
Nissink et al. (2002)[60]	Huang & Schroeder (2006) [42] Laurie & Jackson (2005) [43] Le Guillaux et al. (2009) [55]	None in original paper	original GOLD validation set [87]	305 proteins	35 proteins
Laurie & Jackson (2005) [43]	Kalidas & Chandra (2008) [46]	Q-SiteFinder LIGSITE (Pocket-Finder) PocketDepth	subset of Nissink et al.	134 proteins	35 proteins
An et al. (2005) [13]	Schmidtke et al. (2010) [57]	PocketFinder Fpocket SiteMap SiteFinder	PDB	4,711 proteins 5,616 pockets	11,510 apo pockets (> 1 per holo pocket)
Huang & Schroeder (2006) [42]	Weisel et al. (2007) [44] Kalidas & Chandra	LIGSITE$^{(CSC)}$ PASS CAST SURFNET	subset of PLD [88] + subset of Nissink et al.	210 proteins	48 proteins

(2008) [46]	PocketDepth
Le Guilloux et	PocketPicker
al. (2009) [55]	Fpocket

Because binding-site prediction tools locate multiple candidate pockets on the protein surface, they usually return a ranked list of candidates for evaluation by the user. The number of predictions can be anywhere from 10 to 30+ potential sites, but every method strives to place the most confident prediction among the top ranks. Examining the top ranks and visually comparing the location of the predicted pocket to the location of a co-crystallized ligand is sensible for a manual evaluation of several structures. A more systematic approach for determining a true pocket match is required for a large-scale evaluation of an algorithm's performance. Since every method is first benchmarked on known protein-ligand complexes, the proximity of a predicted pocket to the ligand is a common criterion. For example: the authors of PASS used a distance metric measured from a predicted active-site point (APS) to the nearest ligand atom (D_{near}), and to the center of mass of the ligand (D_{COM}), to evaluate their results. An ASP was considered a "hit" if D_{near} was < 4.0 Å. D_{COM} was used to assess the accuracy of the predictions with respect to varied ligand sizes [50]. Laurie and Jackson used a precision threshold to evaluate Q-SiteFinder, where "precision" was defined as the percentage of the probe sites in a predicted pocket within 1.6 Å of a ligand atom [43]. A prediction was considered a hit if the precision was > 25%. In comparing the Q-SiteFinder tool to LIGSITE (as implemented in Pocket-Finder), the metric gave the authors a common measure of prediction accuracy. However, methods that do not use probes or the concept of an ASP for their pocket definition cannot be effectively evaluated with this metric. The authors of PocketDepth used a volume overlap with respect to the ligand as a measure to define their successful predictions. If the predicted site volume overlaps the ligand volume by more than 10%, the prediction was considered a hit. PocketFinder also defines a cavity as a bounded volume, but the authors of this algorithm chose to evaluate its predictions by the overlap of protein atoms in the vicinity of the predicted pocket with those in the vicinity of the ligand [13]. They define the relative overlap (RO) as $RO = (A_L \cap A_E) / A_L$ where A_L is the solvent-accessible surface area of the protein atoms within 3.5 Å of the ligand, and A_E is the solvent-accessible area of the protein atoms within 3.5 Å of the predicted envelope. A perfect prediction was considered RO = 1 and a failed prediction RO = 0, with predictions of RO > 0.5 being considered successful. This overlap criterion helped dissect the performance of PockerFinder with respect to the quality of its predictions. The disadvantage of this metric is that very large predicted pockets will achieve an RO = 1 while greatly over-predicting the binding-site region (defining the whole protein surface as a predicted binding site, for example, would achieve an RO of 1). Thus, Schmidtke et al. introduced a mutual overlap (MO) metric to quantify fraction of the predicted pocket that overlaps with the predicted site [57]. It is defined as $MO = (A_L \cap A_E) / A_E$. Similarly to RO, an MO close or equal to 1 signifies a better match to the known ligand binding pocket. Since a small ligand can bind in a large pocket, an MO of 1 is not to be expected for every binding site, but very low values of MO can flag predicted sites that extend too far beyond the known pocket. Taking into account both RO and MO makes it possible to better assess the accuracy of a prediction, and the authors use both metrics to compare four different algorithms against each other and identify the methods prone to over-predicting the size of the binding site.

Once a definition of a hit is determined, the algorithm's success on a large test set can be quantified by counting the number of successful predictions with respect to the rank of those predictions. A "Top-Rank" success rate could be the percentage of the predicted pockets with rank 1 that were successful hits. Conversely, the percent of the successful predictions that are ranked 1 can be counted as the Top-Rank success rate. The two measures are similar but give slightly different information since more than one site on a protein can be a hit due to the presence of multiple ligands. Care should be taken when comparing the performance of different algorithms, as authors often select the metrics that work best with their particular algorithm and test set.

1.4 Surface Area Calculations: NACCESS

Methods designed to detect protein pockets are usually optimized to report geometric pocket-delineating features, such as volumes, or simply the member residues comprising the protein edge of the pocket. Many do not report individual residue properties, such as degree of solvent exposure or even molecular surface area (MSA).

Several methods have been established as gold standard approaches for such calculations during the early emergence of computational SBDD field. Lee and Richard formally defined the concept of the accessible surface area (ASA) in 1971 [89], and it was promptly implemented in an efficient "probe rolling" algorithm by Shrake and Rupley in 1973 [90]. The algorithm lays down a grid of points equidistant from each atom of the protein and then calculates the points that can be considered solvent accessible. Points are checked against the vdW radii of each neighboring atom to determine if they are "outside" the protein. A similar check against the vdW radius plus a probe distance (usually 1.4Å, the radius of a water molecule) determines whether the point is solvent accessible. The surface points can then be used to calculate the ASA of each atom, and hence each protein residue. This method effectively represents a probe of a certain radius being rolled across the vdW surface of the atoms, with the probe center tracing out the ASA and the probe's surface tracing the MSA. The algorithm depends on the choice of correct vdW and probe radii and the density of grid points, but it tends to produce an accurate ASA with a relatively efficient numerical method. This thesis uses NACCESS, a popular implementation of the Lee and Richard's method by Hubbard and Thornton [91] to determine the ASA of each residue and define it as "surface" or "non-surface". Alternative methods also exist for calculating ASA. The Connolly surface method, described earlier, is effectively the inverse of the Lee and Richard method, defining the surface from the point of view of the solvent rather than the protein atoms. Numerous algorithms have since attempted to improve the speed of these approaches and provide additional surface metrics, but the conceptual definition of ASA has remained relatively unchanged [92, 93].

1.5 Scoring Protein-Ligand Binding

A fundamental result of understanding protein-ligand binding is the ability to predicting a binding mode and affinity of a ligand to a specific protein. The task of molecular docking is to do just that – predict the structure of the protein-ligand complex starting with the knowledge of the protein and ligand structures on their own. Docking has two primarily challenges. The first is to sample the conformational space of the protein-ligand complex to identify the possible binding modes of the ligand. The second is to score the quality of these conformations relative to each other. The goal of a scoring function is to rank these different modes, either by estimating the true binding affinity of the ligand or by computing a relative score to differentiate the modes containing favorable protein-ligand interactions.

The approaches to calculate an appropriate score come in three general classes. Force-field scoring functions like GOLD [94], DOCK [95], and AutoDock [96] use classical force fields like AMBER [97] or CHARMM [98] to obtain absolute binding energy values of the potential ligand poses. Empirical scoring functions decompose the energy functions found in the force-field methods into various classes of interactions, such as hydrogen bonding, ionic interactions, hydrophobic contacts, and penalties for entropic contributions. Although functional forms in empirical functions tend follow those in force-field based functions, they are usually simplified, and the relative weights of the interaction terms can be adjusted to fine-tune the scoring function performance using known binding data. The choice of which interaction terms to include and the relative weight of the terms can be an advantage by making the scoring functions more flexible, but it can also be a major hurdle in the development of a widely-applicable scoring method. Empirical scoring functions like F-Score [99] or X-Score [100] rely on large and diverse training sets with available binding data to appropriately parameterize their interaction terms.

Finally, knowledge-based scoring functions use statistical analysis of large sets of protein-ligand complexes to derive potentials of mean force between a ligand and protein. Protein-ligand atom pairs from the PDB are used to obtain pairing preferences and estimate their likelihoods. This approach is based on the theory that a Boltzmann-distribution rule for a single closed system held at fixed temperature is applicable to a database of structures [24]. The knowledge-based functions do not incorporate binding affinity data, which might make them more generally applicable due to the larger size of their training sets [24]. However, since these methods do not incorporate any first-principle energy calculations, they must be used with caution when attempting to estimate absolute binding energies. Examples of knowledge-based scoring functions include DrugScore [23] and ITScore [24]. Some approaches use a “consensus” method to combine scores from different methods in an attempt to capture complementary information inherent in the different classes of scoring functions. Scores in consensus methods may be combined using various schemes, such as voting, rank-voting, weighted-sum ranks, and other multivariate combination methods [101].

Scoring functions can be evaluated on various criteria. A recent comparison of 11 different scoring functions tested the ability of the functions to identify an experimentally observed ligand conformation from a series of permuted invalid conformations. Scoring functions from various classes performed equally well at identifying the conformation with F-Score, DrugScore, X-Score, and several others yielding success rates higher than the force-field based AutoDock [102]. Performance could be improved by combining several of the methods into a consensus scoring scheme. When the same functions were tasked with reproducing experimentally-determined binding affinities for a series of protein-ligand complexes, only a few, including X-Score and DrugScore, were able to give correlation coefficients that are better than random [102]. A recent large-scale exercise aimed at the protein-ligand scoring community tested the ability of 19 different methods to reproduce absolute binding affinities in a set of 343 high-quality protein-ligand complexes [103]. The exercise showed that the highest performance achieved a correlation (R^2) with experimental data of only 0.58. No family of methods was found to perform better than another, but subsets of complexes scored well/poorly by many methods were identified [12]. The difference between the sets of “easy” complexes and “hard” complexes were used to identify pitfalls common across all scoring approaches. These studies stress that despite significant progress, scoring functions used in molecular docking software remain an area with significant room for improvement [104, 105].

CHAPTER II

Updating and Extending the Binding MOAD Database

2.1 Introduction

Datasets of protein-ligand complexes with binding-affinity data were first used in the field of computational chemistry to develop scoring functions for small-molecule docking with the goal of improving the structure-based drug design (SBDD) process. In recent years, many sets of protein-ligand complexes have been accumulated in online databases to make them searchable, downloadable, and more useful to the scientific community. Some of the databases, like Binding MOAD, contain binding-affinity data for the complexes they collect; others focus on describing binding site properties or implement algorithms to assess binding site similarity. The largest and most relevant databases are outlined below, and their relative strengths and weaknesses with respect to Binding MOAD are highlighted. Our aim is to maintain Binding MOAD as the largest-possible collection of high-quality, protein-ligand complexes available from the Protein Data Bank (PDB) [2], augmented with the inclusion of binding-affinity data. Recently, we have undertaken an expansion of Binding MOAD to include information regarding which residues compose the individual binding sites. Also query tools have been developed for obtaining summary statistics of this data on large sets of protein-ligand pairs. These extensions will make Binding MOAD an even more powerful tool for characterization of protein-ligand binding.

The original release of Binding MOAD was created over 2001-2005, as described previously [106]. It has grown from 5331 protein-ligand complexes in its first release to its current collection of 14,720 complexes, with binding data available for 4782 (32%) of these complexes [107, 108]. The current release contains 7,064 unique ligands annotated with their biological relevance, and it is grouped into 4,618 protein families at

90% sequence identity. This chapter details the current state of the Binding MOAD annual update process and describes an extension of the database architecture to store, analyze, and display residue composition in individual sites and sets of complexes. Other similar resources are described to provide a scope of the field and the impact of Binding MOAD.

2.2 Protein-Ligand Databases

Large, well-curated databases are essential for analyzing protein-ligand binding. They can be used for a myriad of applications, from developing and testing scoring functions for docking and screening to mining structural and physicochemical properties for classification, prediction, and comparison of protein-ligand interactions. Consequently, each database tries to provide a unique perspective on the way the data is assembled, curated, and presented. Below are details of several leading databases that contain binding affinity data and/or provide easily accessible data on the interacting residues of protein-ligand binding sites.

BindingDB

BindingDB was developed by Michael Gilson in 2001, and it is currently hosted at the Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California, San Diego [109]. The database contains over 781,000 high-quality binding data for 6,448 protein targets (some are isoforms or mutants), and over 342,000 molecules. A significant portion of this data has been extracted from ChEMBL and PubChem databases. The data comes from isothermal calorimetry (ITC), K_i , EC50, and other experimental assays, all with annotated experimental conditions. BindingDB allows for deposition of novel binding data and has been growing at a steady pace. However, most complexes do not have associated crystal or NMR structures, as the focus is first and foremost on the binding data collection. Currently ~1500 BindingDB protein-ligand complexes can be un-ambiguously referenced to a PDB crystal structure (at 100% sequence identity). The database has excellent browse and search features and extensive cross-references. [url: http://www.bindingdb.org]

PDBbind

PDBbind was originally developed hee at University of Michigan under Shaomeng Wang. It is currently maintained and developed by Renxiao Wang at the Shanghai Institute of Organic Chemistry [110]. The latest PDBbind release contains structures and binding data for 7,986 complexes, including protein-nucleic acid, nucleic acid-ligand, and protein-protein interactions. As in Binding MOAD, the binding data is collected from the primary publication of the respective crystal structure. In general, PDBbind is curated in a way similar to Binding MOAD, but it has some key differences. Unlike Binding MOAD, PDBbind does not have strict controls for protein quality, such as a threshold for electron density (crystal structures with resolution as large as 4.7Å are included). PDBbind limits its entries to complexes with only one ligand in the crystal structure, and it excludes any complexes that do not have binding data (e.g., the many structures with only simple cofactors bound are excluded). PDBbind is updated annually and provides a user-friendly interface for data search, browsing, and download. It is accessible to both academic and commercial users. [url: http://www.pdbbind-cn.net/]

Catalytic Site Atlas (CSA)

The CSA is a comprehensive listing of protein residues with known catalytic function developed by Janet Thornton and hosted at the European Bioinformatics Institute [111]. The database currently contains 968 enzymes annotated based on evidence from the literature. A residue is listed only if the evidence shows a direct chemical function for that residue in the protein active site. Additionally, CSA provides over 26,000 entries whose catalytic residues are inferred by strict PSI-BLAST alignment to the 968 curated entries. Since a protein structure is required to un-ambiguously identify functional residues, all entries contain cross-references to their respective PDB entries. Extensive browsing and searching capabilities, as well as a downloadable file, are provided at the CSA website. Since CSA requires evidence of catalytic function on a per-residue basis, it represents a focused but limited view of the ligand binding site. Residues that might participate in important binding interactions but do not have a catalytic role are omitted. These excluded residues may help define the specificity of binding.[url: http://www.ebi.ac.uk/thornton-srv/databases/CSA/]

PDBeMotif (formerly MSDmotif & MSDsite)

Originally developed by Kim Henrick, PDBeMotif is part of the PDBe (PDB Europe) suite of tools available at the European Bioinformatics Institute. PDBeMotif is less of a database and more of an interface, allowing one to search the full set of the structures available in the PDB with sequence, chemical structure, or 3D sub-structure [112]. In fact, the open source PDBeMotif tool can be freely downloaded and used on any set of public or private protein structures. The search capabilities allow one to examine characteristics of binding sites of single proteins or classes of proteins grouped by structural families (CATH, Pfam), functional families (EC, TIGR), or a number of other classification methods, such as genetic families or ProSite motifs. Binding-site environmental characteristics include residue composition, conserved 3D structural features, various bond types (such as ionic, hydrogen bonds, or planar groups), and many others. They can be browsed per-structure or summarized for custom groups of proteins. PDBeMotif provides the most full-featured and fastest portal to the full variety of protein-ligand complexes present in the PDB. However, no quality control for protein structures is performed, and ligands are defined as any HETATM group, including modified amino acids that are part of the main protein chain. The burden of filtering the data to desired quality is left to user, although extensive search and filtering functionality is provided for this purpose. [url: http://www.ebi.ac.uk/pdbe-site/pdbemotif/]

FireDB

FireDB is a database enumerating residues involved in ligand binding in protein structures available from the PDB [34]. FireDB was developed by Alfonso Valencia and is hosted at the Spanish National Cancer Research Center**.** The current release of the database contains over 170,000 unique protein chains and over 13,000 accompanying molecular compounds. The ligand-binding residues are identified at several distance cutoffs from the ligand (3.5, 4.0, and 4.5Å), and functional annotations are transferred from the Catalytic Site Atlas (see above). The database also clusters protein sequences at 97% sequence identity and generates consensus lists of binding-site residue contacts. The conservation of binding sites across all proteins is also considered, and tools are provided to explore evolutionary relationships of distantly related proteins with conserved binding sites. Some ligand filtering is performed to eliminate solvents, ions, DNA, RNA, peptides, and uncommonly large ligands (where the ligand is two-thirds the size of the protein). FireDB allows NMR structures and has no criteria for excluding low-quality protein structures. [url: http://firedb.bioinfo.cnio.es/Php/FireDB.php]

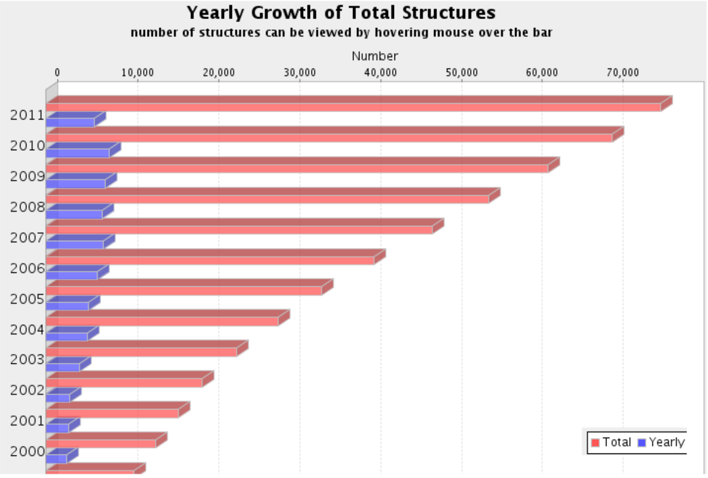
POCKETOME

The Pocketome is a collection of experimentally identified (i.e. crystal structure), small-molecule binding pockets that was developed by Ruben Abagyan. It is hosted at the Skaggs School of Pharmacy and Pharmaceutical Sciences at University of California, San Diego. Each entry in the Pocketome corresponds to a small-molecule binding site in a protein that is represented by at least two PDB entries, , has been co-crystallized with at least one drug-like small molecule, has an associated UniProt entry.. Unbound crystal structures are also included if they meet the quality criteria. The dataset was originally compiled to evaluate the PocketFinder binding-site prediction algorithm. It was also made available as a searchable online database. The database included more than 5500 bound structures and twice as many un-bound structures in its original form (not every holo structure has an apo structure), and it has been updated periodically to keep pace with the growth of the PDB. Currently, the online database contains 988 entries, which encompass more than 11,000 PDB structures. However, an objective analysis of the content is difficult because bulk download of the annotated database is not available. Multiple examples of the same binding site in different crystal structures are aligned and compared to each other to calculate pair-wise, pocket-ligand steric clashes, binding-site atom RMSD, and overall shape similarity of the pockets. For each entry, a residue contact map is provided summarizing ligand-contacting residues in all structures in the entry. Each entry can be visualized via the web browser or downloaded for local analysis using the Molsoft ActiveICM software [113]. [url: http://www.pocketome.org]

2.3 Binding MOAD Annual Update

As the number of high-quality structures annually deposited to the PDB continues to grow (Figure II-1), a thorough and efficient annual-update process is crucial to keeping Binding MOAD up-to-date with the available data. During creation of Binding MOAD, care was taken to automate as many steps of the data processing and curation as possible. Still, many steps require manual interpretation of the data, and constant improvements in Binding MOAD's primary data source – the PDB – require appropriate adjustments and corrections to the process. While the details of the Binding MOAD pipeline have been described elsewhere, the general outline along with improvements and extensions to the procedure is described below.

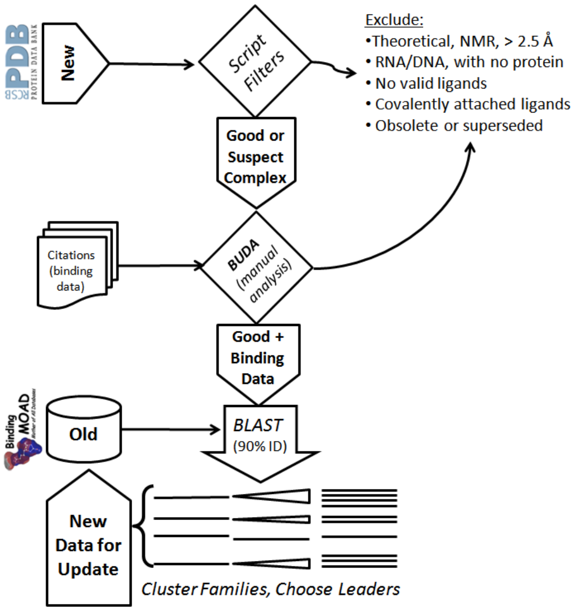
Figure II-1: Annual growth of the PDB. Figure from rscb.org



To ensure that every protein-ligand pair in Binding MOAD has an associated protein structure, the database is updated in a top-down approach, starting with all the protein structures submitted to the PDB over the previous calendar year (Figure II-2). Initial filtering of acceptable structures is performed by automated scripts, and any ambiguous cases are flagged for further manual examination. The manual-examination step also entails the look-up of any binding data available in the primary literature for the crystal structure. After the manual curation has been completed, all proteins that have not been rejected are clustered into families with existing Binding MOAD entries, and a representative leader is chosen for each family. Once the families and leaders are determined, the annual update is complete, and the production server that drives the

Binding MOAD website is loaded with the new dataset. Below is a summary of the individual steps in the above process.

Figure II-2: Annual update of the Binding MOAD database and description of the individual steps.



Annual steps to update Binding MOAD:

1. Use the PDB's list of obsolete entries to identify any existing structures in Binding MOAD that should be removed.
2. The previous version of the PDB download is compared to the new download to identify all new structures that have been added to the PDB since the last version of Binding MOAD was created. Biounit files are downloaded for the new entries.
3. Good and suspect protein-ligand complexes are identified in the new structures using our filtering scripts.
4. Any new HETs must be classified as suitable ligands or added to the suspect, partial, or reject lists.
5. Literature references are scraped as HTML from publisher websites and loaded into BUDA – a tool designed to facilitate manual look up of binding or kinetic values.
6. Sequences are added to existing classes and protein families, but regrouping all sequences from scratch may be necessary to periodically confirm our protein classes and families.
7. Each new structure is compared with the leader of its homologous protein family to determine if the new structure is a better representative of the family. Any new families are also evaluated to choose leaders.
8. Updated data is loaded into the Binding MOAD database.

2.3.1 Automatic Filtering Scripts

Remediated biounit PDB files are retrieved from the RSCB website. The use of biounits instead of mmCIF format files that were initially used to create Binding MOAD was promted by the PDB’s commitment to the format and the 2006 remediation that fixed numerous issues in the biounit files. The biounit also provides the most biologically-appropriare structure of the protein, important when annotating the protein with binding data that was likely obtained using the biologically-relevant protein form. The biounit files are processed with a series of Perl scripts that use the Bioperl library parser to interpret the PDB format data. Ligands and protein atoms are identified using a combination of SEQRES and HETATM records. Multi-part ligands – those that contain more than one HET group – are identified by their spatial proximity, and every ligand is checked for covalency by measuring the distance to the closest protein atom. A multi-part name is constructed for a multi-part ligand based on its component HET groups. Any ligand within 2.1 – 2.4 Å of the protein is flagged as suspect and examined manually. Short contacts to metals are also examined manually to differentiate coordinated ions in metal-containing enzymes from covalently bound metals, such as the iron present in heme groups. Ligands are flagged as “suspect” according to their memberships in our list of unusual HET groups (Table II-1).

Table II-1:Definition of unusual HET groups. For brevity, not all compounds are listed.

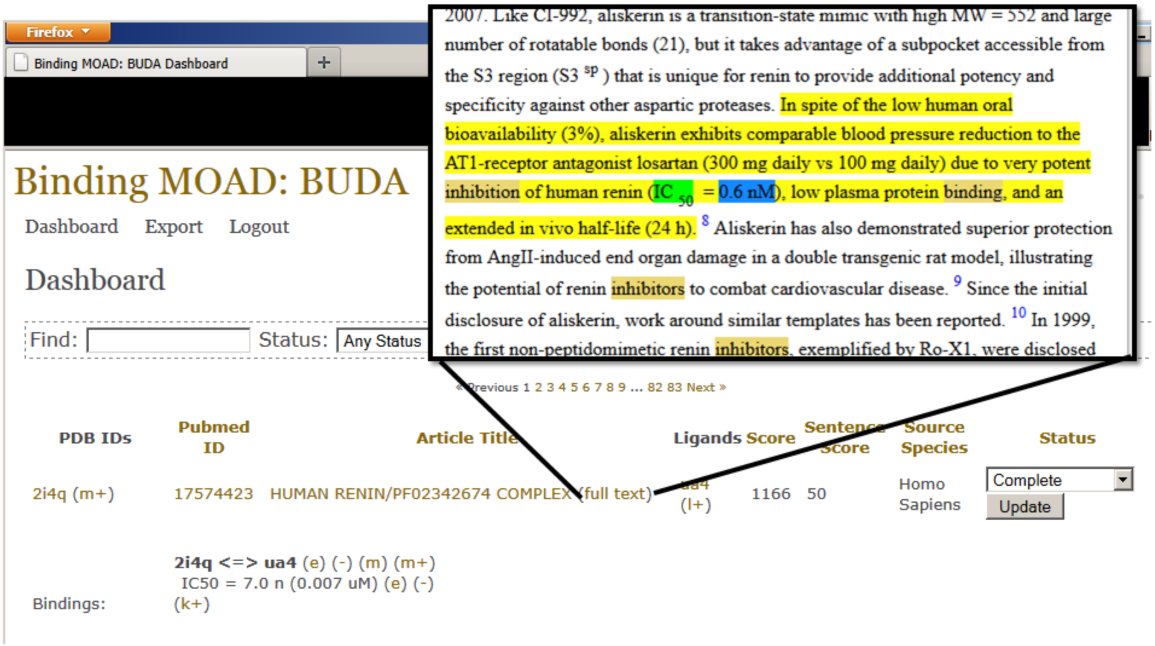
Classification	**Ligand Types (Examples)**
Suspect Ligands (111)	**Sugars** (glucose, galactose, fructose, xylose, sucrose, b-D-xylopyranose,trehalose) **Small organic molecules** (phenol, benzene, toluene, t-butyl alcohol) **Membrane components** (phosphatidylethanolamine, palmitic acid, decanoic acid) **Small metabolites** that may be buffer components (citric acid, succinate, tartaric acid)
Partial Ligands (78)	**Chemical groups** (amino group, ethyl group, butyl group, methoxy, methyl amine) **Inorganic centers** of transition state or product mimics (aluminum fluorides, beryllium fluorides, boronic acids) **Modifications to amino acids** (oxygens of oxidized CYS, phosphate group on TYR)
Rejected Ligands (511)	**Unknown or dummy groups** (UNK, DUM, unknown nucleic acid, fragment of) **Salts and buffers** (Na^+, K^+, Cl^-, PO_4^{-3} , CHAPS, TRIS, tetramethyl ammonium ion) **Solvents** (DMSO, hexane, acetone, hydrogen peroxide) **Crystal additives and detergents** (polyethylene glycol, oxtoxynol-10,dodecyl sulfate, methyl paraben, 2,3 propanediol, pentaethylene glycol, cibacron blue) **Metal complexes** that associate to the protein surface and are used for phase resolution (terpyridine platinum, bis bipyridine imidazole osmium) **Metal ions** that are part of the protein (Mg^{+2}, Zn^{+2}, Mn^{+2}, Fe^{+2}, Fe^{+3}) **Catalytic centers** that are part of the protein (4Fe-4S cluster, Ni-Fe active center) **Heme groups** (heme D, bateriochlorophyll, cobatamin, protoporphyrin IX)

2.3.2 By-Hand Curation of the Data

Literature citations for all structures passing the automatic filtering scripts are read to confirm the validity of the ligands and record the binding data, if present. Suspect ligands are also examined at this step. Suspect ligands may be valid in cases where they are actually products or reactants of an enzyme or otherwise take part in protein function. Partial ligands are molecules that cannot be a ligand on their own and are often components of multi-part ligands. In rare cases, even the ligands on the rejected ligand list may be marked as valid. This usually occurs in case of crystal additives and detergents, which are sometimes present in the structure as an enzyme reactant or target of a transport protein, as opposed to simply being a component of the crystallization medium. The information in the literature citation is normally sufficient to determine whether the ligand is a biologically relevant molecule or should be considered invalid. The reasons for retaining a suspect, partial, or rejected ligand as a valid ligand are recorded in a comment field for future reference.

We have recently incorporated a browser-based tool that uses natural language processing (NLP) to assist with locating binding data and annotating ligand validity. The tool, termed BUDA, was developed in collaboration with Torrey Path LLC (formerly Metamatics LLC), and it is now hosted locally as an integral part of the annual update process. Literature citations are given a score based on the probability of containing binding data, as determined by an NLP text-mining algorithm. Citations with higher scores are examined first. To aid the examination, words and sentences indicating the likely location of the binding constant are highlighted in an HTML version of the citation. Once a binding constant is identified, it is recorded right in the browser-based application. Ligands can also be toggled as valid or invalid, and comments supporting an annotation call can be associated with each protein-ligand pair. All information is immediately saved to a database and exported to Binding MOAD once all structures are processed. The BUDA application greatly speeds up the manual curation process by improving the search for binding data and reducing errors in formatting and spelling that are otherwise inherent in free-text entry (such as using a spreadsheet).

Figure II-3: Literature citation analysis tool (BUDA). Inset shows text highlighting that identified sentences likely to contain binding data. Data and ligand annotations are recorded in allocated fields and saved for eventual export to Binding MOAD.



2007. Like CI-992, aliskerin is a transition-state mimic with high MW = 552 and large number of rotatable bonds (21), but it takes advantage of a subpocket accessible from the S3 region (S3 ^{sp}) that is unique for renin to provide additional potency and specificity against other aspartic proteases. In spite of the low human oral bioavailability (3%), aliskerin exhibits comparable blood pressure reduction to the AT1-receptor antagonist losartan (300 mg daily vs 100 mg daily) due to very potent inhibition of human renin (IC_{50} = 0.6 nM), low plasma protein binding, and an extended in vivo half-life (24 h). [8] Aliskerin has also demonstrated superior protection from AngII-induced end organ damage in a double transgenic rat model, illustrating the potential of renin inhibitors to combat cardiovascular disease. [9] Since the initial disclosure of aliskerin, work around similar templates has been reported. [10] In 1999, the first non-peptidomimetic renin inhibitors, exemplified by Ro-X1, were disclosed

PDB IDs	Pubmed ID	Article Title	Ligands	Score	Sentence Score	Source Species	Status
2i4q (m+)	17574423	HUMAN RENIN/PF02342674 COMPLEX (full text)	ua4 (l+)	1166	50	Homo Sapiens	Complete

Bindings: **2i4q <=> ua4** (e) (-) (m) (m+)
IC50 = 7.0 n (0.007 uM) (e) (-)
(k+)

2.3.3 Grouping Proteins to Address Redundancy

Redundancy of proteins present in the PDB is addressed at two levels in Binding MOAD. The majority of the proteins deposited in the PDB are enzymes, and enzyme classification (EC) codes are used to broadly group entries into functionally similar classes. Within these classes, and among the non-enzymes, proteins are grouped into homologous protein families based on sequence similarity (Figure II-2).

For each Binding MOAD update cycle all proteins (newly filtered and those from previous years) are aligned by BLASTp [114]. A cutoff of 90% sequence identity is used to group entries into a family. Enzyme proteins that match entries in multiple EC classes are only assigned a single EC class (see [106] for clustering details) and are only present in a single family. For each family, a best representative “leader” structure is chosen. The leader is the structure with the tightest binding ligand. If a family has no entries with binding data, the following order of priorities is used for choosing a leader: 1. best resolution, 2. wild-type structure over mutants, 3. most recent deposition date, and finally 4. when all the criteria are the same, the leader is chosen based on comments in the crystallography paper.

2.3.4 Importing SMILES and Ensuring Ligand-Name Consistency

HET groups in the PDB are periodically remediated to eliminate redundant names or enforce a specific formatting for existing names. Since Binding MOAD only imports new structures during the annual update, outdated ligand HET names can accumulate in the database. As part of an effort in this thesis to examine protein-ligand binding sites, the consistency of ligand names in Binding MOAD needs to be ensured. An additional step was implemented in the 2008 Binding MOAD update to synchronize Binding MOAD ligand names with the most up-to-date HET group and to import standardized SMILES strings for identifying unique ligand structures. The Chemical Component Dictionary (CCD) [115] distributed by the PDB contains all the single-HET names and formulas present in the PDB. A current version of the CCD database (in SQL format) is downloaded, and every HET group name in Binding MOAD is checked against the database. Standard OpenEye SMILES strings for single HET groups are imported from the CCD directly into the Binding MOAD database to supplement ligand information already stored in Binding MOAD. A cross-reference list of obsolete HET codes and their superseding versions is compiled and saved separately from the database. Multi-part ligands are also checked to make sure any component HETs have not been superseded. As opposed to the SMILES, the ligands names are not corrected in the Binding MOAD database itself. This would introduce inconsistency with respect to the biounit files that accompany the Binding MOAD database. Since previous years' biounits are not re-downloaded during the annual update, they will not contain the updated HET group names. A Python script handles the SMILES import and the construction of the HET name cross-reference list.

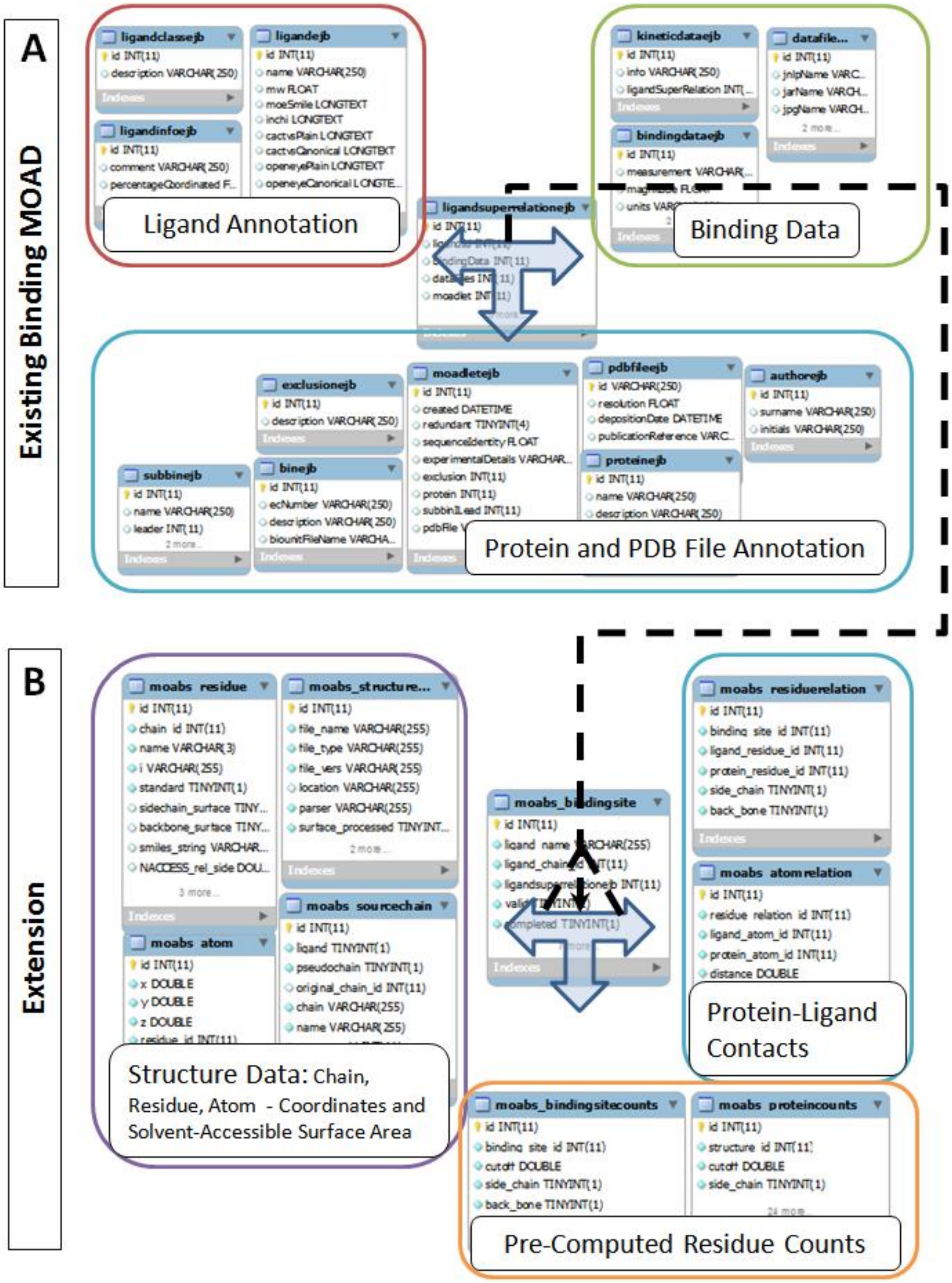
2.4 Extension of Binding MOAD for Binding Site Analysis

Currently, the relational database and object model underlying the Binding MOAD database is optimized for serving binding data on individual protein-ligand pairings and the organization of that data with respect to protein families and Enzyme Classification (EC) classes. Every unique protein-ligand pair is associated with binding data, if available, and the protein and ligand are in turn linked to additional annotation, such as the 90% sequence identity bin for the protein and the type for the ligand (valid or invalid). Associated structure data is also available in the form of the original PDB biounit containing the protein-ligand pair and a SMILES string encoding the 2D molecular structure for most ligands. Although the Binding MOAD curation pipeline performs many useful calculations on the biounit structures to identify multi-part, covalently attached, and suspect covalent ligands, no references between the ligand definition and the biounit structure are stored in the relational database. Since the biounit file may have multiple ligand locations, unconventional labeling of chains or residues, or other inconsistencies inherent in structural-data files, matching the ligand name provided by Binding MOAD to the atomic coordinates is not always a straightforward process. In order to perform the rigorous and efficient analysis of protein-ligand binding sites described in this thesis, an extension of the relational database was implemented to directly link existing Binding MOAD annotation data to structural data in the biounit files. Aside from being used for analysis of binding site propensities in this thesis, the current implementation of the extended database is undergoing internal testing and evaluation for eventual incorporation into the publicly accessible Binding MOAD website.

2.4.1 Relational Database Object Model

To comply with the current implementation of the Binding MOAD database, all extensions to the relational model were developed using MySQL Server conventions (version 5.1). The initial point of extension was the LigandSuperRelation table (Figure II-4A), which maintains a binding-pair relationship between a PDB ID and a ligand that is uniquely identified by its HET group name. The goal of the extension is to create a binding-site definition where ligand uniqueness is defined by the coordinates of the specific ligand residues in the structure file. This approach extends the protein-ligand pairs with specifics about the interacting residues and atoms. This implies that a many-to-one relationship can exist between a structure-based binding site and a LigandSuperRelation binding pair. For example, multiple sulphate (SO4) molecules (a frequent invalid ligand) can be present in a PDB biounit structure, and each of the molecules will have a unique location on the protein surface and unique set of interacting binding-site residues. Meanwhile, only one LigandSuperRelation will exist between the protein PDB ID and the ligand named SO4 in the current Binding MOAD architecture. Therefore, the extended schema introduces the BindingSite object, which represents a structurally-unique protein-ligand pairing (middle of Figure II-4B). Several tables record the residue-residue and atom-atom interactions between the ligand and protein and tie in the structural data into the binding-site definition. Finally, a set of tables models the structure-chain-residue-atom hierarchy of the PDB file and stores residue and atom identities. Additional residue-level and atom-level data, such as solvent-accessible area and atomic coordinates, are also stored. The full entity relation (ER) diagram of the database schema is presented at the end of this chapter in Figure II-7.

Figure II-4: Simplified entity relation diagram for the Binding MOAD relational database, illustrating the relationship between A) the protein, ligand, and binding-data annotations in the existing schema and B) the structure's residue, atom, protein-ligand interaction, and residue-count data in the extended schema. There is a one-to-many relationship between the central table in A (ligandsuperrelationejb) and central table in B (moabs_bindingsite). Some tables and fields omitted for clarity.



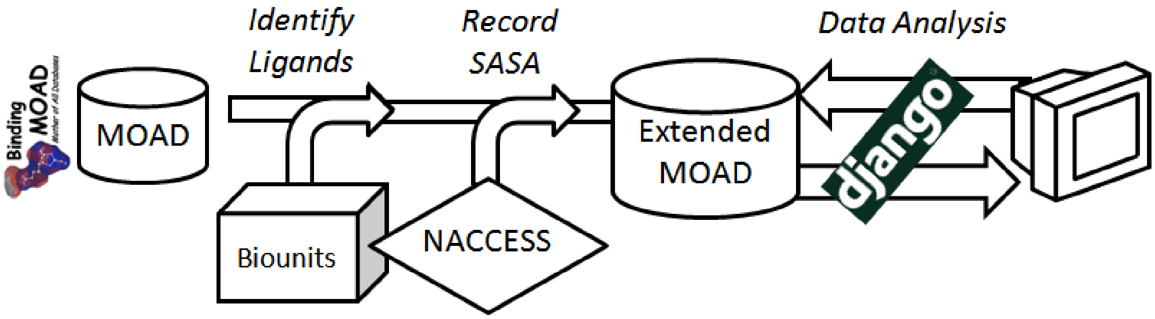
In the extended object model, all information regarding chain and residue identity is recorded for each protein in Binding MOAD as it occurs in the PDB biounit file(s). This information is stored only once for each protein, as it is independent from the binding-site definition. Primary keys are generated for all the Chain, Residue, and Atom objects for unique identification that is independent of the chain, residue, or atoms IDs of the PDB format. A BindingSite object is defined relative to this structural information, and it does not store any of the structural data itself. The BindingSite object serves to make two important conceptual relations that extend the protein-ligand pair defined by LigandSuperRelation: the first is to the chain in the biounit file that contains the ligand, and the second is to the set of protein residues that are members of the binding site. For one protein-ligand pair, there can be multiple ligand chains in the structure file that represent the ligand. In turn, for each of the ligand chains, its respective binding site residues on the protein are conditional on the interaction criteria between the ligand and protein atom. Depending on the intended data analysis, multiple criteria for defining a binding site may be used, and thus, multiple binding sites can be defined for each ligand chain, each with their own set of protein-ligand interactions. The ResidueRelation and AtomRelation objects (Protein-Ligand Contacts box in Figure II-4) store the relationships between Residue and Atom objects (Structure Data box in Figure II-4) that constitute the binding site specified by the BindingSite object. This relationship model is simple yet powerful. It keeps the primary structure data separate from the concept of a binding site, and it can be easily modified and extended with any novel binding-site definition criteria. Conversely, any update to chain or residue information in the biounit file that does not affect the atomic coordinates (a change in HET name or residue re-indexing) can be propagated to the database without the need to update the binding-site definition tables.

A major goal of this thesis work is to calculate propensities of residues in collections of binding sites. This requires the calculation of residue frequencies in binding sites and in the whole protein. The simple relational model described so far can be used to count the residues and calculate the relevant frequency and propensity statistics. However, even with improved speed of querying provided by indexing and query optimization, the retrieval of individual residues on-demand is prohibitively expensive for a large set of structures. To improve queries of residue content for multiple structures, additional tables are provided for storing pre-calculated residue counts for each protein and each BindingSite object. These ProteinCount and BindingSiteCount objects store total residue counts for each of the 20 amino acids in individual fields and a vector array of these totals in a single field. This data has to be computed ahead of time, but it results in significant speedup when querying binding-site composition. Just like there can be multiple BindingSite objects for each ligand chain, there can be many ProteinCount or BindingSiteCount objects for different counting criteria. For example, a ProteinCount object can record the residue totals for the protein surface or the residue totals for the protein as a whole. For the same BindingSite object, there might be a BindingSiteCount object with totals of residues that only have side-chain interactions with the ligand and one object with totals of residues that have backbone interactions. Propensities for a set of protein-ligand pairs of interest can quickly be computed from queries to the respective BindingSiteCount and ProteinCount objects, avoiding numerous small queries against the ResidueRelation and Residue tables. This is particularly important when analyzing the data may require inspection of multiple subsets or the assessment of certain criteria's impact upon the resulting data (i.e., what happens if a parameter changes, or how do subsets compare to each other). This speedup is important, as the latter query would result in prohibitively long wait times for a web-facing implementation like BindingMOAD.org. While further optimization of these pre-calculated residue totals is possible, the current implementation strikes a balance between usability and speed that is sufficient for current analyses.

2.4.2 Framework for Import, Processing, and Analysis of Binding Site Data

The extended version of Binding MOAD with additional binding-site data was prototyped independently of the Binding MOAD update process, and it is currently implemented as a stand-alone application with a MySQL database backend and Python-based Django web framework front-end (Figure II-5). The first challenge in creating the database of binding-site information is populating the data in a manner consistent with existing Binding MOAD conventions. A series of Python scripts were developed for loading the data into the extended relational database. The import scripts use the existing Binding MOAD SQL database to obtain the PDB biounit file name and the names of ligands present in the biounit; then it parses the biounit file (mmLib parser is used) to identify and load the ligand- and protein- residue information. The identified ligands are then extracted from the biounit files, and the un-liganded structure is processed by NACCESS [91] to calculate the solvent accessible surface area (SASA) of each residue. Then the scripts directly load the structure, binding site, and SASA data into a new database using MySQLdb Python libraries. The direct data load ensures that database object relations are kept consistent, and it eliminates extra steps of writing/loading data to/from flat files.

Figure II-5: Implementation of the extended Binding MOAD database. Scripts are used to load the extended database. A Django-powered front end implements data analysis and display functions.



As previously described, Binding MOAD uses a series of heuristics to ensure the definitions of the ligand names are as accurate as possible for multi-part and suspect ligands. However, since specific chain and residue names are not recorded, the binding site import scripts need to re-trace certain parts of the heuristic analysis to try and unambiguously match an existing ligand name to its corresponding HETATM records in the structure file. For a majority of single HET group ligands, this process is straightforward; a simple name match to a HETATM residue name is sufficient. If multiple heteroresidues with that name are found, each is recorded separately (chain names are kept if unique, or mapped to unique pseudo-chain names based on residue indices). Multi-part ligands can be problematic as the ordering of the HET groups in the ligand name might not match the order of the residues in the structure file. This is especially true for structures that have multiple-sugar ligands present. Sugar chains tend to be fragmented, and if the chain information in the biounit file does not uniquely identify the various fragments, telling the individual (un-connected) fragments apart might be near-impossible by name alone. A series of heuristics attempts to match the multi-part ligand name to the corresponding set of HET groups in the biounit file. Binding-site data is only imported for un-ambiguously identified ligands to prevent inaccurate binding-site definition or inconsistencies with the name-based ligand definition in the existing Binding MOAD database. Additional biounit quality issues, such as erroneous residue indexing or partially-missing ligand data, can also prevent un-ambiguous identification. As a result of these limitations, about 8% of protein-ligand pairs in Binding MOAD are not extended with structural binding-site data. We may see a slight decrease in this failure rate once the import of the binding sites is incorporated into the Binding MOAD update process and the heuristics for ligand identification are unified.

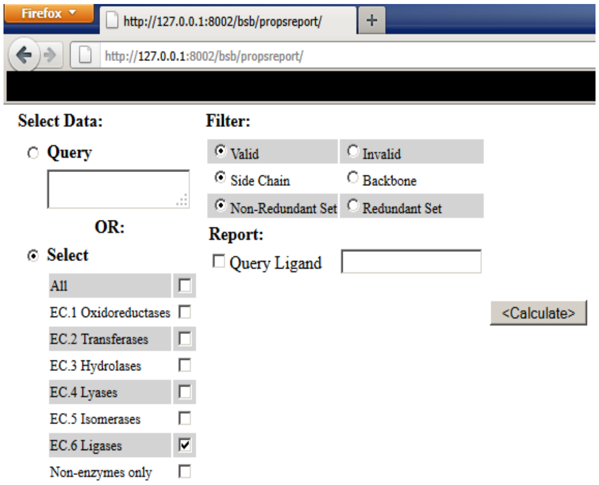
With the ligand chain and residue(s) identified, binding-site residues in the protein are defined using a distance cutoff from the atoms in the ligand chain. ResidueRelation and AtomRelation records are populated with interacting residues and atoms, respectively. For an interacting residue, all protein-ligand atom pairs within the interaction distance are noted in the AtomRelation table, along with the absolute distances to the ligand atom. Since this distance information is retained, a post-processing step can re-define binding sites with interaction cutoffs smaller than the distance used to populate the database. Currently, the atom information is stored only for the binding-site residues and the ligand molecule. Storing all the protein atoms in the database is possible, of course, but the extra data significantly increases the size of the database from several hundred megabytes to several gigabytes, with a deleterious effect on query performance and without benefit to the intended data analysis.

The identified ligand chains are also used to create an un-liganded version of the biounit file, extracting all valid and invalid ligands that are not part of the protein. The SASA of every residue in the un-liganded structure is then calculated by NACCESS [91], ignoring any heteroatoms and water molecules remaining in the structure. Absolute and relative SASA for each residue side chain and main chain is recorded in the database. The default definition of a surface residue in this thesis is $\geq 5Å^2$ of side-chain or main-chain SASA, and the surface residues are flagged as such to avoid querying the raw SASAs numbers when compiling lists of surface residues. However, as with binding site residues, storing the raw data allows for an easy re-classification of which residues are considered surface at a later time.

The Python import scripts directly load all the data into the MySQL database, at which point the binding-site data is ready for mining. The development of the data model and the optimal query methods required multiple development/testing cycles, and the Python-based Django web framework was used as a prototype front-end for processing and querying the database. Django is an open-source project designed for quick development of web-based applications, but its strength lies in providing an object-oriented database interface that is optimized for querying, filtering, and otherwise manipulating a MySQL data source. In other words, the developer does not have to write the raw SQL queries for retrieving or writing data from/to the database, but instead, can use Python classes and Django's programmer-friendly syntax to build complex and flexible queries in a fraction of the time. This simplifies application development by allowing the developer to concentrate perfecting the methodology behind the data-mining code. On the user-facing side of the application, a light-weight web server and a feature-rich webpage template system provided in Django allow for a quick mock-up of a user-friendly interface for executing queries and displaying the results of the data processing. This interface uses standard HTML fields for specifying queries and displaying raw data in report form, and it was supplemented with Google Charts components for graphical display of key results. A simple example of a query for residue propensities, and the query result, is provided in Figure II-6. Developing the binding-site analysis code as a web-based application helped guide the relational database design for eventual incorporation into the existing Binding MOAD database and web server. Moreover, it provided a convenient interface for the retrieval of the binding-site information prior to its incorporation into the Binding MOAD web server. This made the binding-site data available for internal use by any lab members, without requiring them to write their own scripts for accessing the relational database.

Figure II-6: Example of a query for calculating residue propensities for a set of binding sites (in this case all valid binding sites with side-chain contact residues in non-redundant proteins of the Ligase enzyme class) and the resulting graphical and raw-data output.

Query


Firefox
http://127.0.0.1:8002/bsb/propsreport/
+
http://127.0.0.1:8002/bsb/propsreport/

[redacted]

Select Data:

○ Query

OR:

● Select

All	☐
EC.1 Oxidoreductases	☐
EC.2 Transferases	☐
EC.3 Hydrolases	☐
EC.4 Lyases	☐
EC.5 Isomerases	☐
EC.6 Ligases	☑
Non-enzymes only	☐

Filter:

● Valid ○ Invalid

● Side Chain ○ Backbone

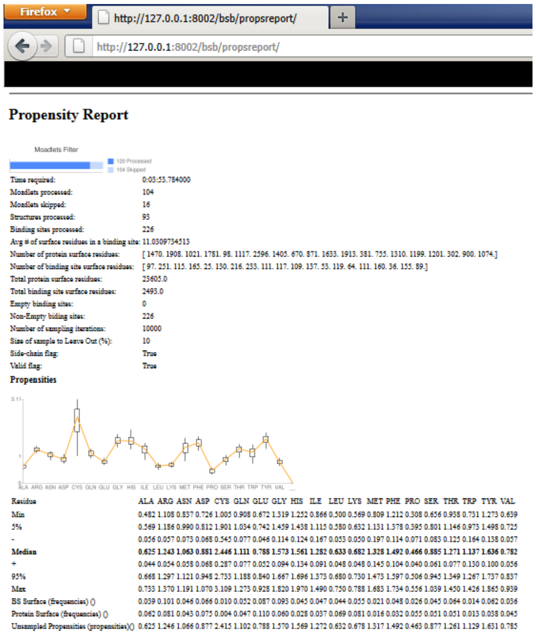
● Non-Redundant Set ○ Redundant Set

Report:

☐ Query Ligand

<Calculate>

Result


Firefox
http://127.0.0.1:8002/bsb/propsreport/
+
http://127.0.0.1:8002/bsb/propsreport/
Moadlets Filter
120 Processed
104 Skipped
3.11
1
0
ALA ARG ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL

[redacted]

Propensity Report

Time required:	0:03:53.784000
Moadlets processed:	104
Moadlets skipped:	16
Structures processed:	93
Binding sites processed:	226
Avg # of surface residues in a binding site:	11.0309734513
Number of protein surface residues:	[1470. 1908. 1021. 1781. 98. 1117. 2596. 1405. 670. 871. 1633. 1913. 381. 755. 1310. 1199. 1201. 302. 900. 1074.]
Number of binding site surface residues:	[97. 251. 115. 165. 25. 130. 216. 233. 111. 117. 109. 137. 53. 119. 64. 111. 160. 36. 155. 89.]
Total protein surface residues:	23605.0
Total binding site surface residues:	2493.0
Empty binding sites:	0
Non-Empty biding sites:	226
Number of sampling iterations:	10000
Size of sample to Leave Out (%):	10
Side-chain flag:	True
Valid flag:	True

Propensities

Residue	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
Min	0.482	1.108	0.837	0.726	1.005	0.908	0.672	1.319	1.252	0.866	0.500	0.569	0.809	1.212	0.308	0.656	0.938	0.731	1.273	0.639
5%	0.569	1.186	0.990	0.812	1.901	1.034	0.742	1.459	1.438	1.115	0.580	0.632	1.131	1.378	0.395	0.801	1.146	0.973	1.498	0.725
-	0.056	0.057	0.073	0.068	0.545	0.077	0.046	0.114	0.124	0.167	0.053	0.050	0.197	0.114	0.071	0.083	0.125	0.164	0.138	0.057
Median	**0.625**	**1.243**	**1.063**	**0.881**	**2.446**	**1.111**	**0.788**	**1.573**	**1.561**	**1.282**	**0.633**	**0.682**	**1.328**	**1.492**	**0.466**	**0.885**	**1.271**	**1.137**	**1.636**	**0.782**
+	0.044	0.054	0.058	0.068	0.287	0.077	0.052	0.094	0.134	0.091	0.048	0.048	0.145	0.104	0.040	0.061	0.077	0.130	0.100	0.056
95%	0.668	1.297	1.121	0.948	2.733	1.188	0.840	1.667	1.696	1.373	0.680	0.730	1.473	1.597	0.506	0.945	1.349	1.267	1.737	0.837
Max	0.733	1.370	1.191	1.070	3.109	1.273	0.928	1.820	1.970	1.490	0.750	0.788	1.683	1.734	0.556	1.039	1.450	1.426	1.865	0.939
BS Surface (frequencies) ()	0.039	0.101	0.046	0.066	0.010	0.052	0.087	0.093	0.045	0.047	0.044	0.055	0.021	0.048	0.026	0.045	0.064	0.014	0.062	0.036
Protein Surface (frequencies) ()	0.062	0.081	0.043	0.075	0.004	0.047	0.110	0.060	0.028	0.037	0.069	0.081	0.016	0.032	0.055	0.051	0.051	0.013	0.038	0.045
Unsampled Propensities (propensities)()	0.625	1.246	1.066	0.877	2.415	1.102	0.788	1.570	1.569	1.272	0.632	0.678	1.317	1.492	0.463	0.877	1.261	1.129	1.631	0.785

2.4.3 Optimizing Data Mining for a Web Interface

Incorporating binding-site data into the Binding MOAD database has several intended uses:

1. Retrieve a single ligand binding site and its residue content.
   a. Visualize the binding-site residues.
2. Query a group of ligand-binding sites for residue frequencies or propensities.
3. Retrieve binding sites similar in residue composition to the query.

The first use case is the simplest, and does not require extensive database querying. The use case in 1a is also straightforward, and currently, a Python script can be generated for visualization of any binding site in PyMol [116]. Use cases 2 and 3 require querying large numbers of binding sites, several thousand in worst case scenarios, and thus require appropriate optimizations. The largest speed-up in querying binding-site residues can be achieved by pre-calculating residue count totals and storing them in the database. A series of pre-processing functions are used to accomplish this after the data import has been completed and before the database is deployed for querying. As noted in Figure II-4, the extended database schema allocates several tables for pre-calculated binding-site data (ProteinCounts and BindingSiteCounts objects). The pre-processing functions iterate through all the proteins in the database and add their surface residues, storing the totals in the ProteinCounts table for each protein structure entry. Two ProteinCounts records are created for each protein, one for totals of surface residues with exposed side chains and one for residues with exposed main chains. Similarly, for each unique binding site in the database, the totals for side-chain and main-chain surface residues for the 20 standard amino acids are calculated and placed in the BindingSiteCounts table. This pre-calculated data makes it simple to retrieve the member residues of a binding site if the BindingSite primary key is known instead of executing a multi-table join query with BindingSite, ResidueRelation, and Residue tables. The binding-site similarity data for use case 3 is also pre-calculated, using Tanimoto similarity of the residue count vectors, and stored in a separate table for easy retrieval by BindingSite primary key. Additional efficiency of database queries could be achieved by optimizing the actual query commands used to retrieve the data, but Django already implements query optimization, and for our purposes, simply following the best-practice conventions of the Django query syntax is sufficient for ensuring query optimality.

After the data has been retrieved from the database, the second major bottleneck to serving the result back to the user is the calculation of actual residue frequencies and propensities. This is an issue mainly for use case 2, where residue counts of several thousand binding sites need to be added and divided thousands of times for assessment of confidence intervals. Since it is impossible to pre-calculate frequencies for every desired subset of binding sites, the propensity calculations and the sampling need to be performed on the fly. Luckily, the NumPy scientific computing library for Python provides a set of powerful vector-based numerical functions for quick calculations on massive amounts of data. NumPy functions are leveraged extensively for the propensity calculations, and they provide a major speed-up as compared to generic Python math libraries. Once the residue-count data is retrieved from the database, it is stored in a NumPy matrix, which can be easily sliced to produce sub-sampled of the data. This makes propensity calculation for sampling runs lightning-fast. In fact, the current rate-limiting step in calculating the propensities for large sets of binding sites is still the database-query step. All code dependent on the NumPy library is implemented in a separate Python module to facilitate possible future use with a non-Python web framework, such as the Java-based JBoss Application Server, the framework for the current Binding MOAD implementation.

With currently implemented optimizations, worst-case queries for residue propensity data (5000+ proteins) take between 5 and 10 minutes to execute. While this processing time is still well above the less-than-a-minute wait time that most web users expect from a website, it can be further streamlined for a production version of the Binding MOAD server. The overall goal of the optimizations developed as part of this thesis was to construct a relational schema and processing code that is consistent with current best-practices in database and code design. The current implementation of the extended Binding MOAD database allows for extremely flexible queries of the binding sites present in Binding MOAD, and it provides results within minutes. The loss of speed that comes from using a relational database instead of flat files is counterbalanced by the ease of formulating complex queries and the existence of a user-friendly front end for internal laboratory use. More importantly, the use of the relational database schema that extends Binding MOAD's current relational object model will facilitate the incorporation of the binding-site data into the existing Binding MOAD web server.

2.5 Future Directions for the Extension of Binding MOAD

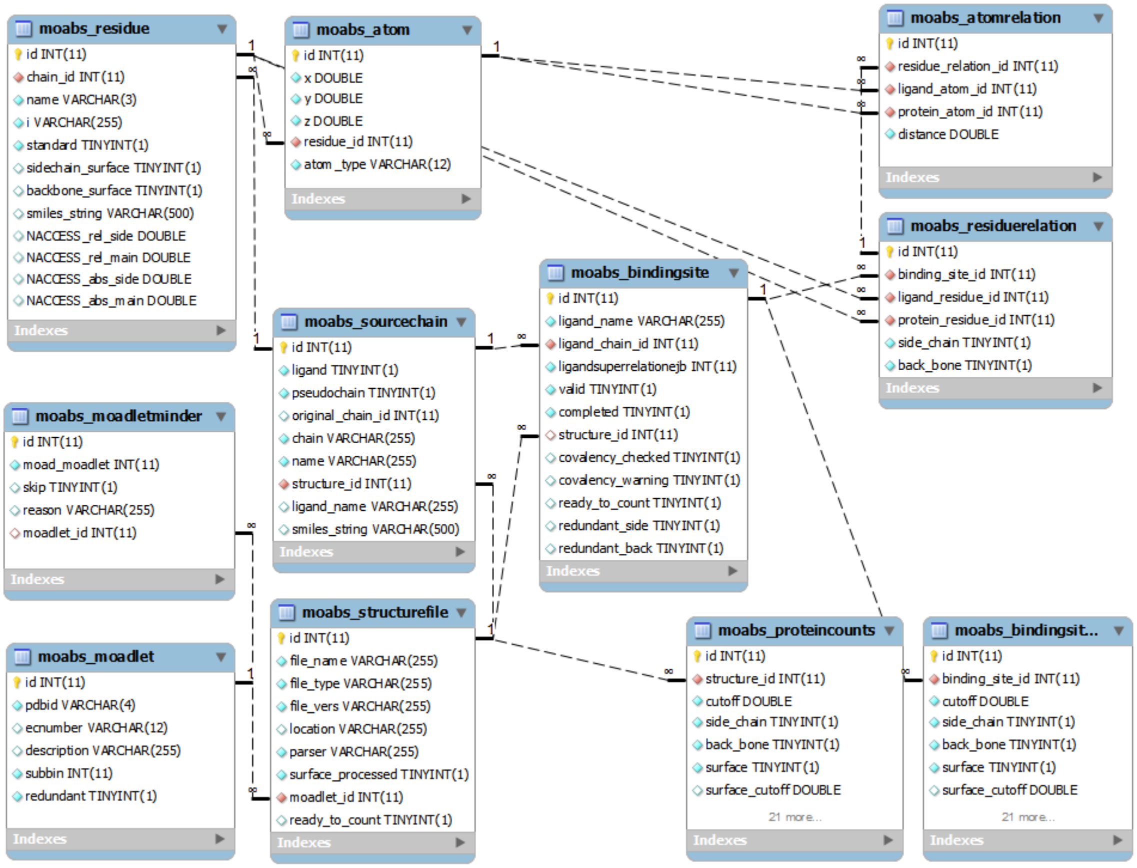
The extended relational database described above will be integrated into the Binding MOAD application server and web site before the next annual update process. The Python code used to calculate sampled residue frequencies and propensities can be re-factored into Java code compatible with the Java Beans currently powering the Binding MOAD website, or it can be used as a standalone module, providing functionality through a Java linker function. No special webpage elements are needed to display the residue count or propensity on the Binding MOAD website, unless summary graphs are desired. Such graphs can be implemented using Google Graphs components or plotting libraries such as gnuplot or matplotlib, all of which can generate plots on-demand with the speed required for a web application. Under the JBoss framework, the Binding

MOAD server is rebuilt whenever the database or component code is updated. To improve performance, any graphics for individual binding sites (or even sets of binding sites) can be pre-generated during the rebuild.

Incorporating binding-site residue data into the Binding MOAD web application will give users a wider set of tools for exploring the database. For example, a user may wish to view the residue content of binding sites in a particular protein family to ascertain the variation in the ligand-binding environment. This is similar to residue conservation information provided by the Pocketome and PDBeMotif databases, and it may be presented graphically or via raw statistics. The user would also be able to compare a binding site to others in the database using their composition, either by retrieving sites similar to a known site or similar to a desired query pattern. Comparing the composition of two sites can be reduced to a vector distance calculation, and familiar distance metrics such as Tanimoto or RMSD will provide intuitive and adjustable cutoffs for defining relevant matches.

A unique feature of the Binding MOAD binding-site set will be the ability to generate residue frequencies and propensities for an arbitrary set of protein-ligand sites. These residue statistics can be pre-computed for the protein families and EC classes that are already incorporated into Binding MOAD. They will also be available for any result set of a Binding MOAD query, which currently includes search parameters for HET groups, protein names, structure-resolution cutoffs, and presence of binding data. Additional search parameters for ligand validity and type of protein-ligand contacts will also be included for propensity queries. Sampling techniques used in this thesis will be applied to provide a user with statistical confidence of the obtained propensity values. The functionality to compare propensities from one query to the propensities from another query and test the statistical significance of the differences can also be easily implemented in the current framework. The combination of these search criteria and the ability to return summary data of residue content for any query will distinguish Binding MOAD from existing protein-ligand databases.

Figure II-7: Entity relationship diagram of the extension to the Binding MOAD relational object model. The schema is derived directly from the MySQL database. Tables that store the structural information are on the left. Tables that store the binding-site residue relationships are on the upper right. Tables that store pre-calculate residue-count data are on the lower right. The moabs_bindingsite table (middle of figure) provides a foreign key to the existing Binding MOAD schema.



CHAPTER III

Exploring the Composition of Protein-Ligand Binding Sites on a Large Scale

3.1 Introduction

Understanding general properties of protein-ligand binding sites is of great importance if we are to gain insight into the functional diversity of the proteome. One of the most fundamental properties of the receptor surface is the set of amino acids available for interactions with ligands. In many protein families, this set is well known and structurally conserved due to the functional role of the residues, and several insightful studies have summarized catalytic residue content in sets of enzymes [31] [117, 118]. However, the more general trend of amino-acid distribution within binding sites across a variety of protein and ligand types is less understood; previous studies have explored limited sets of proteins [37] or interactions of specific interest [119]. With ever-increasing numbers of protein structures available and numerous databases dedicated to protein-ligand analysis [34, 47, 88, 107, 112], a wider view of the residue composition of binding sites is now possible and necessary. Establishing general trends of binding-site composition can help develop valuable tools for identifying a protein functional site without prior information about the protein's sequence or structural homology. Such tools can be invaluable for the characterization of proteins of unknown function emerging from current structural genomics projects [120]. The recent use of binding-site composition to bolster methods for *de novo* prediction of binding sites [18, 33, 121] [122] is an encouraging example of the utility of the general binding-site composition trends.

To study the composition of ligand binding sites across the broadest set of available protein structures, we analyzed the propensity of residues in all the binding sites present in the Binding MOAD database - one of the largest sets of curated protein-ligand complexes [107]. Our analysis summarizes surface composition of binding sites of biologically relevant ligands, such as natural reactants, drugs, and co-factors. We also show how composition of binding-site surfaces varies with number of structures analyzed; this measure of statistical significance is not presented to this extent in other studies to date. Another unique aspect of this study is our examination of the binding of spurious co-crystals, such as crystallization buffers, solvents, and stray ions, which exhibit markedly different trends than the binding of functional ligands.

3.2 Methods

3.2.1 Large, Non-Redundant Binding-Site Dataset

We began by assembling a non-redundant set of 3295 protein-ligand structures, each representing a closely related protein family from the 2009 release of Binding MOAD. The non-redundant set of Binding MOAD is composed of families grouped by 90% sequence identity; the 3295 complexes embody the variation of the full set of 14,720 complexes with 41,721 binding sites. A binding site was defined as the set of protein residues which have at least one non-hydrogen atom within 4.0 Å of a ligand's non-hydrogen atom. These residue interactions were then labeled as side chain (SC) or backbone-only (BB-only) depending upon which atoms participated in the interaction. A residue classified with a BB-only interaction did not have any side-chain atoms within the interaction distance. Residues were classified as SC if the interaction was solely through the side chain or through both its side chain and backbone atoms. Glycine residues are considered a special case, and their interactions are always classified as SC regardless of the absence of a side chain. A single protein residue could have interactions with more than one ligand, in which case the residue interactions were considered independent, and the residue was included in each ligand's binding site. Since a ligand-based definition of the binding site was used, smaller ligands may not make contacts with all possible residues in a large binding site. Only the residues in contact with the ligand are part of the calculation of a site's solvent-accessible surface area (SASA).

In accordance with Binding MOAD annotation, each binding site was classified as "valid" or "invalid" depending on the biological relevance of the ligand [107]. Since all structures in Binding MOAD must contain a valid ligand, the likelihood of an invalid ligand occupying a biologically relevant site is greatly reduced. While it is still possible, the rate of such occurrence is much less than using all the structures in the Protein Data Bank (PDB) [2]. For each protein structure, multiple sites of a unique ligand were analyzed for redundancy by comparing the counts of each residue. Binding MOAD uses biounit structures, which can contain multi-meric proteins composed of identical subunits. To avoid over-representing ligand sites of multi-meric proteins, only one site was retained when multiple sites with an identical ligand and identical binding -ite residues existed in the same structure.

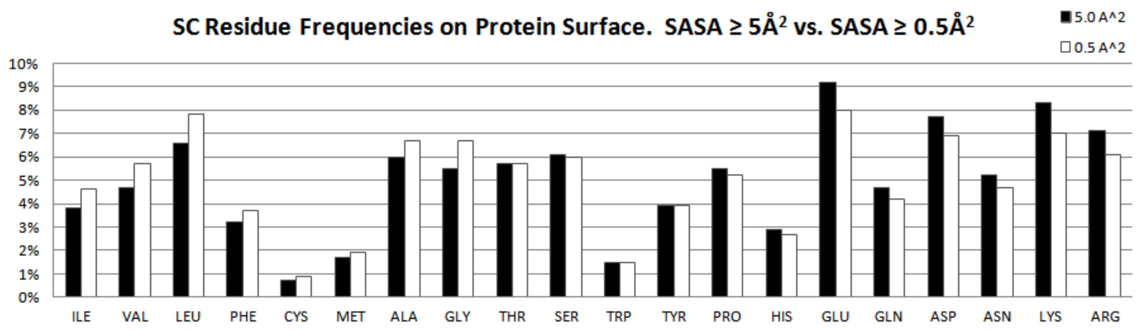
3.2.2 Surface Residue Definition

Solvent accessibility of residues was calculated using the NACCESS program [91]. NACCESS rolls a probe with the diameter of a water molecule across the entire van der Waals (vdW) surface of the protein and uses the path traced by the probe's center to calculate the SASA of each residue. It is important to note that this is different from the molecular surface area (MSA), which is the path traced by the probe's surface. Known ligands were removed from the structure before the SASA calculation. The default probe size was used, and any waters, hydrogens, or remaining hetero-residues were ignored (also default behavior for NACCESS). The NACCESS value of *abs_side* was used to define surface residues for the SC set and *abs_main* to define surface residues for the BB-only set. These report the absolute areas (in Å^2) of the residue side chain and backbone, respectively (calculated using default NACCESS atom types and vdW radii). Since NACCESS treats the Gly Cα as a side chain, the largest of the *abs_main* or *abs_side* values was used for that residue. SASA was calculated for all residues in a protein, which included any binding-site residues.

We chose to use the common standard of ≥5Å^2 SASA as the definition of a "surface" residue [123, 124]. However, we were concerned that this definition included only 84% of SC binding-site residues (data not shown), so we also examined the effect of lowering the minimum SASA cutoff to 0.5Å^2 (Table III-1) to ensure we were not omitting significant parts of the binding site. Lowering the cutoff for the surface definition increased the total number of binding-site residues (SC and BB-only) so that

98% of the residues within interaction distance of the ligand were considered “surface”. However, the respective increase in total binding-site SASA was only 0.2%, a contribution so small that it is misleading to count residues. Furthermore, the 0.5-Å^2 definition led to inappropriate frequencies for amino acids on the surface of the protein (Figure III-1). Specifically, more hydrophilic residues such as Arg, Asp, Lys, and Glu have the highest surface frequencies with the 5-Å^2 cutoff (>7%), which is in keeping with other studies [30]. Although the relatively hydrophobic Leu had high frequencies with both definitions, it is not appropriate that counting many small-SASA contributions (at 0.5-Å^2 cutoff) should make it more frequent (7.8%) than Arg (6.1%) or Lys (7%). Including the inconsequential contributions of small-SASA residues when counting residue frequencies and propensities simply leads to erroneous conclusions.

Figure III-1: Frequencies of solvent-accessible SC residues on the protein surface with a cutoff of SASA≥5Å2 (black bars) and SASA≥0.5Å2 (white bars). Residues are sorted by decreasing hydrophobicity.



3.2.3 Residue Propensity Calculation

In accordance with previous studies, we used residue propensity as a measure of residue over-representation to explore the binding-site composition [31] [117] [37] [121]. The cumulative propensity P_i for each amino acid i = *Ala, Arg, Cys ...etc.* was calculated by taking the ratio of the frequency of the amino acid in binding-sites F_i^{BS} and its frequency on the protein surface F_i^{PS}. The binding-site frequency was obtained by summing across the surfaces of all binding sites $s = 1...S$ in a binding-site class (SC or BB-only). The protein frequency F_i^{PS} was obtained by summing up the occurrence of the amino acid across the surfaces of all proteins $p = 1...P$, where P= 3295 in our case.

Equation 1: Propensity calculation.

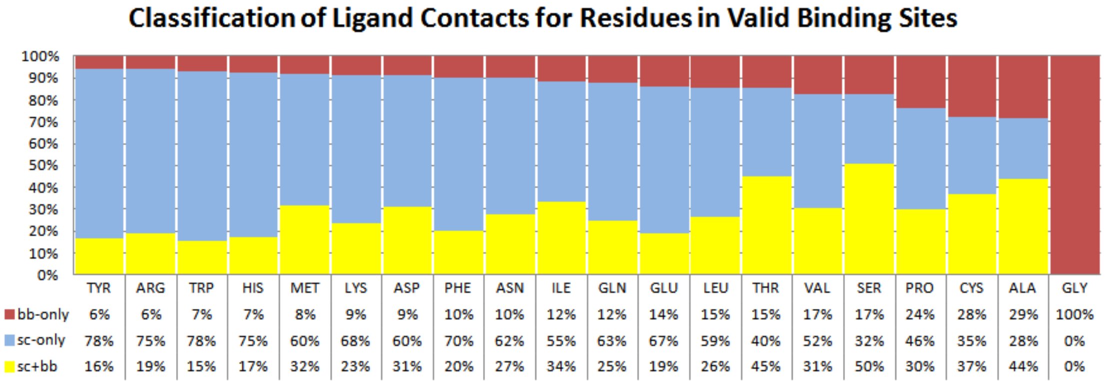
$$P_i = \frac{F_i^{BS}}{F_i^{PS}} \quad \text{where} \quad F_i^{BS} = \frac{\sum_s N_i^s}{\sum_s \sum_i^{20} N_i^s} \quad \text{and} \quad F_i^{PS} = \frac{\sum_p N_i^p}{\sum_p \sum_i^{20} N_i^p}$$

The propensities were calculated separately for valid versus invalid binding sites, and SC versus BB-only sets. Propensities greater than 1.0 show over-representation of a residue in the binding sites relative to the entire protein surface, and values less than 1.0 show underrepresentation. Since propensity is a ratio of ratios and unit changes in its value represent fold changes in frequency, we present the propensity values on log-scaled axes.

Note that the residue counts were summed across the set of structures or binding sites before division. This is necessary because calculating a propensity value for a single protein may result in division-by-zero errors when rare residues, such as cysteine, are absent on the protein surface. Per-protein propensities for rare residues can also result in extremely large propensity values due to division by a small protein-surface frequency, making summary results harder to interpret. Moreover, since the average size of a binding site is around 11 residues, many amino-acids are not represented in all sites and lead to zero per-protein propensities. In calculations of propensities for a set of binding sites, only proteins that contained at least one site of that type (SC or BB-only, valid or invalid) were included in the calculations.

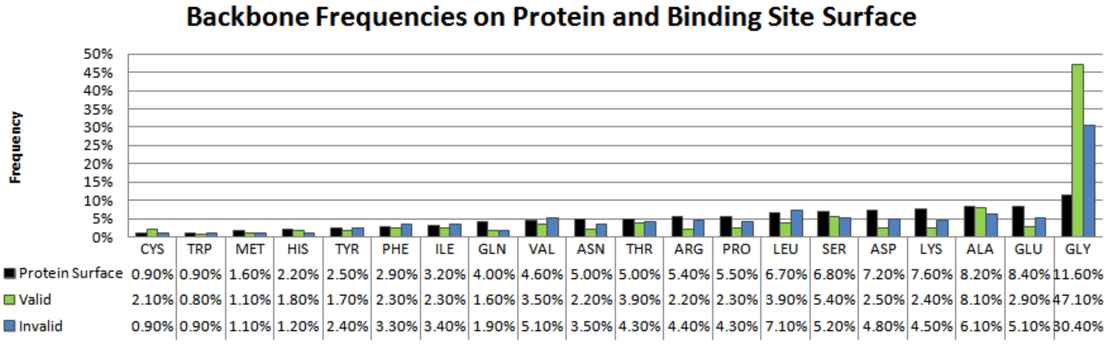
The BB-only interactions are relatively rare (Table III-1), and they are dominated by glycine (Figure III-2 and Figure III-3). Most residues with BB-only contacts to the ligand point their side chains ‘away’ from the ligand; otherwise, a side-chain atom would likely be within the interaction distance, and the residue would be classified as having SC contacts. Additionally, since BB-only contacts represent equivalent atom types from residue to residue, they are not expected to provide diverse interaction environments based on residue type. For all these reasons, we focus our results and discussion on residues in the SC category.

Figure III-2: Relative frequency of sc-only, BB-only or both (SC+BB) interactions per residue. The residues with "SC" interactions in our analysis combine the sc-only and "SC+BB" contacts (blue+yellow). Residues ordered by increasing BB-only frequency. Here, all Gly interactions are shown as BB-only to show its overall contribution to BB-only contacts.



	TYR	ARG	TRP	HIS	MET	LYS	ASP	PHE	ASN	ILE	GLN	GLU	LEU	THR	VAL	SER	PRO	CYS	ALA	GLY
🟥 bb-only	6%	6%	7%	7%	8%	9%	9%	10%	10%	12%	12%	14%	15%	15%	17%	17%	24%	28%	29%	100%
🟦 sc-only	78%	75%	78%	75%	60%	68%	60%	70%	62%	55%	63%	67%	59%	40%	52%	32%	46%	35%	28%	0%
🟨 sc+bb	16%	19%	15%	17%	32%	23%	31%	20%	27%	34%	25%	19%	26%	45%	31%	50%	30%	37%	44%	0%

Figure III-3: Frequencies of BB-only residue contacts on binding site, sorted by increasing frequency of on the protein surface. Surface residues with ≥5Å^2 backbone SASA are shown. Gly interactions are shown as BB-only to stress that it constitutes the vast majority of such contacts.



	CYS	TRP	MET	HIS	TYR	PHE	ILE	GLN	VAL	ASN	THR	ARG	PRO	LEU	SER	ASP	LYS	ALA	GLU	GLY
■ Protein Surface	0.90%	0.90%	1.60%	2.20%	2.50%	2.90%	3.20%	4.00%	4.60%	5.00%	5.00%	5.40%	5.50%	6.70%	6.80%	7.20%	7.60%	8.20%	8.40%	11.60%
🟩 Valid	2.10%	0.80%	1.10%	1.80%	1.70%	2.30%	2.30%	1.60%	3.50%	2.20%	3.90%	2.20%	2.30%	3.90%	5.40%	2.50%	2.40%	8.10%	2.90%	47.10%
🟦 Invalid	0.90%	0.90%	1.10%	1.20%	2.40%	3.30%	3.40%	1.90%	5.10%	3.50%	4.30%	4.40%	4.30%	7.10%	5.20%	4.80%	4.50%	6.10%	5.10%	30.40%

3.3 Results and Discussion

The set of 3295 structures yielded 7712 valid binding sites and 4909 invalid binding sites (Table III-1), which together represent a comprehensive set of protein-ligand variety present in the PDB. After taking into account site redundancy and eliminating incredibly small binding sites (those that could not accommodate a solvent probe atom and thus did not have "exposed" residues), there were 5562 valid and 3552 invalid sites. Roughly a third the 3295 structures had invalid binding sites in addition to one or more valid site, consistent with the higher number of invalid ligands per structure than valid in the structures where invalid ligands do occur. On average, valid binding sites were 6 times larger in terms of number of residues and 2 times larger by surface area that invalid ones. This is expected because valid ligands in our set tend to be larger and more buried than invalid ones. However, it means that the number of binding-site residues available for frequency and propensity calculations is different between valid and invalid sites.

Table III-1: Summary data of structures from the 2009 Binding MOAD release used in the propensity calculations. Enzyme class memberships determined based on EC annotations from the PDB. All SASA areas calculated by NACCESS.

Structures		Sites		SASA Cutoff (Å^2)	Non-Redundant Sites	Residues Per Site Avg. (Median)	SASA per site Avg. (Median) Å^2	Protein:Site # Residues
Valid	3295	7712	SC	**5.0**	5562	11.4 (11)	433 (399)	10:1
				0.5	5514	13.2 (12)	441 (406)	10:1
			BB-only	**5.0**	3213	2.4 (2)	35 (28)	39:1
				0.5	3943	3.7 (3)	34 (26)	37:1
Invalid		4909	SC	**5.0**	3461	3.6 (3)	194 (178)	50:1
				0.5	3581	4.1 (4)	195 (178)	51:1
			BB-only	**5.0**	1358	1.6 (1)	25 (21)	143:1
				0.5	1739	1.9(2)	22 (16)	165:1
Valid Only Enzymes	2354	6063	SC	**5.0**	4301	11.7 (11)	434 (399)	10:1
Valid Non-Enzymes	835	1597	SC	**5.0**	1261	10.3 (10)	431 (401)	11:1

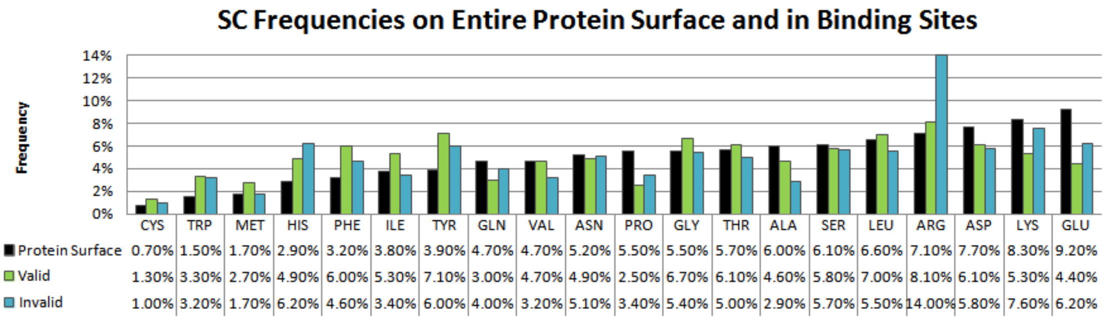
3.3.1 Residue Frequencies and Propensities

Most proteins exist in aqueous environments, such as that of a cell. Therefore, it is generally accepted that the solvated outer surface of the protein is composed of amino acids that tend to be hydrophilic in nature. Conversely, the core of the protein is more hydrophobic, a factor that contributes to the proper folding and stability of proteins [125, 126]. For example, hydrophobic residues tend to bury larger areas of their side chains upon protein folding than hydrophilic ones [29]. However, the composition of the solvent-exposed protein surface is not uniformly hydrophilic in nature [29], and the correlation between residue hydrophobicity and solvent-exposure is limited [30]. Since binding sites are a part of a protein’s surface, the comparative analysis of binding site composition must be performed with respect to the composition of the entire surface.

In our analysis, charged and polar residues make up the largest portion of protein surfaces (black bars in Figure III-4A), but surprisingly, Ala and Leu are more prevalent than the more hydrophilic Thr and Ser. All four of these residues are frequent in sequence. Less-frequent hydrophobic residues such as Met, Phe, Trp, and Cys have low surface frequencies. If we relax the surface definition to include less-solvent-accessible residues, (Figure III-1) very hydrophobic amino acids like Ile, Val, and Leu increase in their relative surface frequency. However, as discussed previously, their contribution in terms of fraction of overall surface area would be miniscule. Gly, which is common in protein sequence, has a surface frequency comparable to the more hydrophilic Asn and Pro.

Figure III-4: A) Frequencies of solvent-accessible side chains on the protein surface and in binding sites with SASA cutoff ≥ 5Å^2. B) Median propensity of residues in ligand binding sites of valid and invalid ligands, analyzed across all proteins. Residues in A and B are ordered by increasing frequency on surface. C) Ratio of residue propensity for valid versus invalid binding sites. Residues ordered by decreasing ratio. Error bars in B and C indicate 95th percentiles of 10,000 leave-10%-out samples.

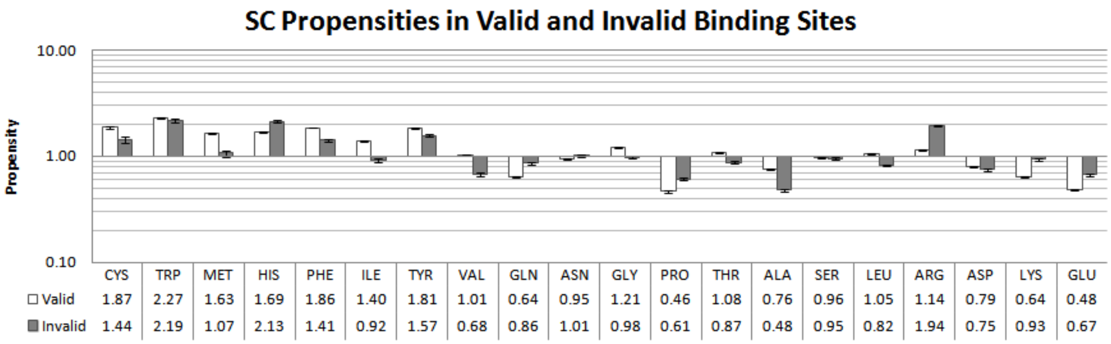
A



SC Frequencies on Entire Protein Surface and in Binding Sites

	CYS	TRP	MET	HIS	PHE	ILE	TYR	GLN	VAL	ASN	PRO	GLY	THR	ALA	SER	LEU	ARG	ASP	LYS	GLU
■ Protein Surface	0.70%	1.50%	1.70%	2.90%	3.20%	3.80%	3.90%	4.70%	4.70%	5.20%	5.50%	5.50%	5.70%	6.00%	6.10%	6.60%	7.10%	7.70%	8.30%	9.20%
🟩 Valid	1.30%	3.30%	2.70%	4.90%	6.00%	5.30%	7.10%	3.00%	4.70%	4.90%	2.50%	6.70%	6.10%	4.60%	5.80%	7.00%	8.10%	6.10%	5.30%	4.40%
🟦 Invalid	1.00%	3.20%	1.70%	6.20%	4.60%	3.40%	6.00%	4.00%	3.20%	5.10%	3.40%	5.40%	5.00%	2.90%	5.70%	5.50%	14.00%	5.80%	7.60%	6.20%

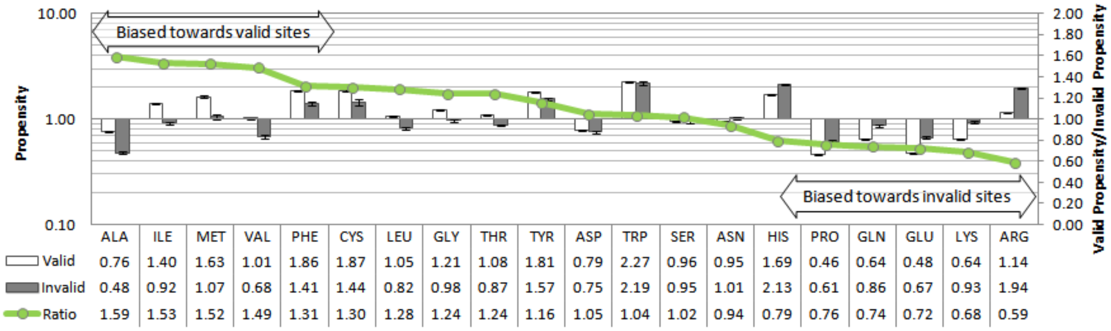
B



SC Propensities in Valid and Invalid Binding Sites

	CYS	TRP	MET	HIS	PHE	ILE	TYR	VAL	GLN	ASN	GLY	PRO	THR	ALA	SER	LEU	ARG	ASP	LYS	GLU
□ Valid	1.87	2.27	1.63	1.69	1.86	1.40	1.81	1.01	0.64	0.95	1.21	0.46	1.08	0.76	0.96	1.05	1.14	0.79	0.64	0.48
■ Invalid	1.44	2.19	1.07	2.13	1.41	0.92	1.57	0.68	0.86	1.01	0.98	0.61	0.87	0.48	0.95	0.82	1.94	0.75	0.93	0.67

C



	ALA	ILE	MET	VAL	PHE	CYS	LEU	GLY	THR	TYR	ASP	TRP	SER	ASN	HIS	PRO	GLN	GLU	LYS	ARG
□ Valid	0.76	1.40	1.63	1.01	1.86	1.87	1.05	1.21	1.08	1.81	0.79	2.27	0.96	0.95	1.69	0.46	0.64	0.48	0.64	1.14
■ Invalid	0.48	0.92	1.07	0.68	1.41	1.44	0.82	0.98	0.87	1.57	0.75	2.19	0.95	1.01	2.13	0.61	0.86	0.67	0.93	1.94
● Ratio	1.59	1.53	1.52	1.49	1.31	1.30	1.28	1.24	1.24	1.16	1.05	1.04	1.02	0.94	0.79	0.76	0.74	0.72	0.68	0.59

Residue propensities presented in Figure III-4 B and C, present the propensity of residues to appear in protein surface regions involved in ligand binding. Pro, Glu, Gln, Lys, and Ala disfavor binding sites (propensities of 0.46 – 0.76). Arg, Thr, Val, Leu, Ser, and Asn have propensities within ±0.2 of 1.0, showing that these are relatively unbiased in their contributions to binding sites versus the rest of the protein surface (Figure III-4B). Though Arg, Leu, and Asp have the first, third, and fourth largest contributions to binding sites (Figure III-4A) their relative propensities are ~1 because of their equally high prevalence on the entire protein surface. Larger propensities for binding sites occur when a residue is frequently observed in binding sites, but is rare on the general surface. Cys, Trp, Met, His, Phe, Ile, and Tyr all have low protein surface frequencies (left side of Figure III-4A) and show propensities of ≥ 1.4 (left side of Figure III-4B). Tyr and Phe are excellent examples. They are the second and seventh most common resides in binding sites, respectively, but they are rare on the protein surface. These residues are bulky and aromatic, so their exposure to solvent is rather unfavorable. It is reasonable that evolution is judicious in their use, placing them where they are most needed for a functional role, such as conservation in binding sites [117] [119] [30]. Trp also has a high propensity for binding sites and similar physical properties, but its exceptional propensity actually reflects its rarity on the protein surface (< 2% of all SC contacts). The same pattern is seen for Cys, which is even rarer on the surface (< 1% of SC contacts). Gly is notable because backbones are rare on protein surfaces (about 17% of the total protein surface area), but when they are present, they are overwhelmingly Gly. Gly backbones account for 13% of all backbone protein surface area (data not shown) and they are highly biased to be located in binding sites. Gly backbones account for ~50% of BB-only interactions in valid binding sites.

However, when considered along with other SC interactions, as it is in our analysis, Gly does not show a large propensity for binding-site regions. Overall, our propensities for valid binding sites agree well with previously published propensities from a set of ~35,000 redundant ligand-binding sites [32] (R^2=0.81) and those from a smaller set of 41 drug-binding sites [121] (R^2=0.79). Propensities for invalid sites were less well correlated with these data (R^2=0.27 and R^2=0.61, respectively).

3.3.2 Comparison of Frequencies and Propensities in Invalid versus Valid Sites

A unique aspect of this study is our ability to compare the binding-site interaction patterns for valid ligands to those in sites of spurious additives. This provides a type of "experimental control" which is usually not possible in analyses of binding-site databases. The issue at hand is not necessarily the recognition of additives themselves, but instead, with how valid and invalid binding differs. Figure III-4C demonstrates the propensities for valid and invalid binding sites, ordered by the ratio between of the two. This data emphasizes our caution in over-interpreting the high propensities of Cys and Trp. They do not show any significant bias for valid ligands. One could argue that Trp, Cys, or any other residue may be inherently "sticky" for small molecules, so they are meaningful for biological insights regardless of the presence of valid versus invalid ligands. However, we find that there are residues which show a significant bias between the classes. This significance was confirmed by randomly shuffling valid and invalid labels 1000 times (maintaining their relative proportion) and re-calculating the propensities and ratios each time. All residues had an average ratio of 1 across the shuffled sets. The maximum and minimum of the shuffled ratios was 1.2 and 0.8 respectively, both for Cys, with all other residues having considerably narrower minimum and maximum values (data not shown). We therefore consider propensity ratios >1.2 and < 0.8 as significant trends.

Ala, Ile, Met, and Val are the most biased toward biologically relevant binding sites over indiscriminant associations (ratio > 1.4), followed by a second tier of Phe, Cys, Leu, Gly, and Thr (ratio > 1.2). Conversely, His, Pro, Gln, Glu, Lys, and Arg show a bias towards invalid binding sites (ratio < 0.8), although all but His and Arg have propensity for the surface rather than binding sites. Considering Arg has among the highest catalytic propensities [31], it should be present in many valid binding sites, but we do not see strong correlations between binding site propensity (valid or invalid) and catalytic propensity (data not shown) or large differences in propensity values when enzymes are considered separately from non-enzymes (Figure III-7). Instead, looking at the distribution of Arg interactions in binding sites of invalid ligands (Table III-3) demonstrates that they make up most SC interactions in 11 of the top 20 ligand sites and are present at high rates (> 15% of SC interactions) in sites of small, charged molecules, such as sulfate (SO4), phosphate (PO4), acetate (ACY), and chloride (CL) ions. They are also especially frequent in citrate (CIT) sites, which appear on both valid and invalid lists, depending on the function of the bound protein. Of the residues that show valid to invalid ratios of > 1.2, only Ile, Met, Phe, and Cys show a propensity for binding sites versus the protein surface.

Table III-2: **Composition of binding sites with respect to bound ligand for the top 20 valid ligands. Ligand listed in decreasing fraction of 5562 binding sites. Most frequently interacting residue for each ligand is in bold.**

HET	Ligand%	ALA%	ARG%	ASN%	ASP%	CYS%	GLN%	GLU%	GLY%	HIS%	ILE%	LEU%	LYS%	MET%	PHE%	PRO%	SER%	THR%	TRP%	TYR%	VAL%
NAD	4.49	5.90	4.75	6.12	7.01	1.56	2.14	3.43	7.49	3.60	**8.33**	6.82	4.01	1.87	4.58	4.51	5.98	7.94	1.44	4.80	7.73
FAD	3.90	6.77	7.09	4.02	4.47	1.91	3.51	4.28	7.33	4.72	6.79	6.21	3.98	1.42	4.65	4.00	6.91	**7.79**	3.42	6.28	4.44
ADP	3.09	4.48	10.46	5.43	5.37	0.50	2.35	3.36	**11.42**	2.80	4.98	5.20	9.96	1.85	4.14	2.74	5.76	8.67	0.78	5.04	4.70
NAP	2.97	6.37	8.97	5.68	4.02	0.59	2.25	2.01	**9.93**	3.01	6.78	5.99	5.64	2.01	2.25	4.12	8.76	8.65	0.97	5.78	6.23
FMN	2.34	5.09	**10.99**	7.17	2.43	1.39	3.88	2.14	9.14	5.73	4.34	4.63	4.51	3.18	3.30	2.89	8.21	6.94	3.18	6.54	4.34
ATP	1.80	2.76	**12.20**	4.26	6.27	0.17	2.42	7.52	10.78	2.26	4.43	5.51	12.03	1.92	5.43	0.75	5.35	8.02	1.42	2.26	4.26
GDP	1.73	3.17	4.39	3.98	11.44	2.96	1.74	3.17	8.27	1.43	1.94	8.27	**19.10**	0.31	4.60	1.63	8.17	10.52	-	1.94	2.96
GLC	1.55	3.95	9.65	6.58	**12.94**	0.22	6.14	7.46	3.07	7.46	2.41	1.10	2.63	2.19	7.46	0.88	1.75	1.54	11.40	10.53	0.66
NDP	1.37	6.19	9.32	4.66	3.83	1.18	2.30	2.85	9.32	2.85	5.29	5.85	5.78	2.64	1.32	2.85	**10.44**	8.35	1.74	7.38	5.85
SAH	1.20	5.07	2.97	3.21	10.51	1.85	2.10	4.45	**11.50**	2.35	5.07	8.16	1.98	4.45	7.91	2.97	6.06	3.83	4.20	7.29	4.08
ANP	1.10	4.90	7.48	6.62	7.23	-	3.43	4.04	**10.54**	1.84	6.37	4.78	9.56	2.21	4.53	1.84	5.51	7.97	0.98	4.29	5.88
COA	0.97	8.85	7.51	3.35	2.95	0.80	4.29	0.94	7.24	4.29	4.56	8.45	**8.98**	4.29	6.84	2.55	6.43	4.16	1.88	5.36	6.30
NAG	0.81	2.34	6.54	**19.16**	9.35	3.74	4.21	3.74	4.21	1.40	2.80	4.21	2.34	1.87	3.27	1.40	2.34	5.61	14.49	4.67	2.34
CIT	0.79	3.04	**16.22**	7.77	4.73	0.34	2.03	3.04	6.76	11.15	4.73	3.72	6.42	2.03	2.70	3.38	7.77	4.73	2.03	5.74	1.69
AMP	0.77	4.48	**10.70**	2.74	5.72	1.74	3.73	5.97	6.97	5.97	5.97	4.98	5.97	1.49	6.72	1.74	5.97	7.71	1.00	6.47	3.98
NAI	0.76	7.79	3.89	6.17	7.38	0.13	2.28	2.55	8.72	2.15	**9.40**	8.99	4.83	2.68	1.88	3.49	7.38	6.31	0.67	4.30	8.99
MAN	0.72	5.91	-	**18.72**	16.75	-	9.36	1.97	5.91	2.46	-	5.42	3.45	-	1.48	2.46	1.97	3.94	5.42	12.32	2.46
SAM	0.67	5.20	4.98	3.62	**11.09**	0.45	3.85	6.11	8.82	5.43	5.20	7.24	2.71	2.26	7.92	4.07	4.98	3.85	2.04	7.01	3.17
GNP	0.65	4.22	0.84	2.95	8.44	1.90	2.11	1.27	12.66	0.84	1.27	8.02	**18.78**	0.42	5.49	3.16	8.86	14.14	-	3.16	1.48

Table III-3: Composition of binding sites with respect to bound ligand for the top 20 invalid ligands. Ligands listed in decreasing fraction of 3461 binding sites. Most frequently interacting residue for each ligand is in bold.

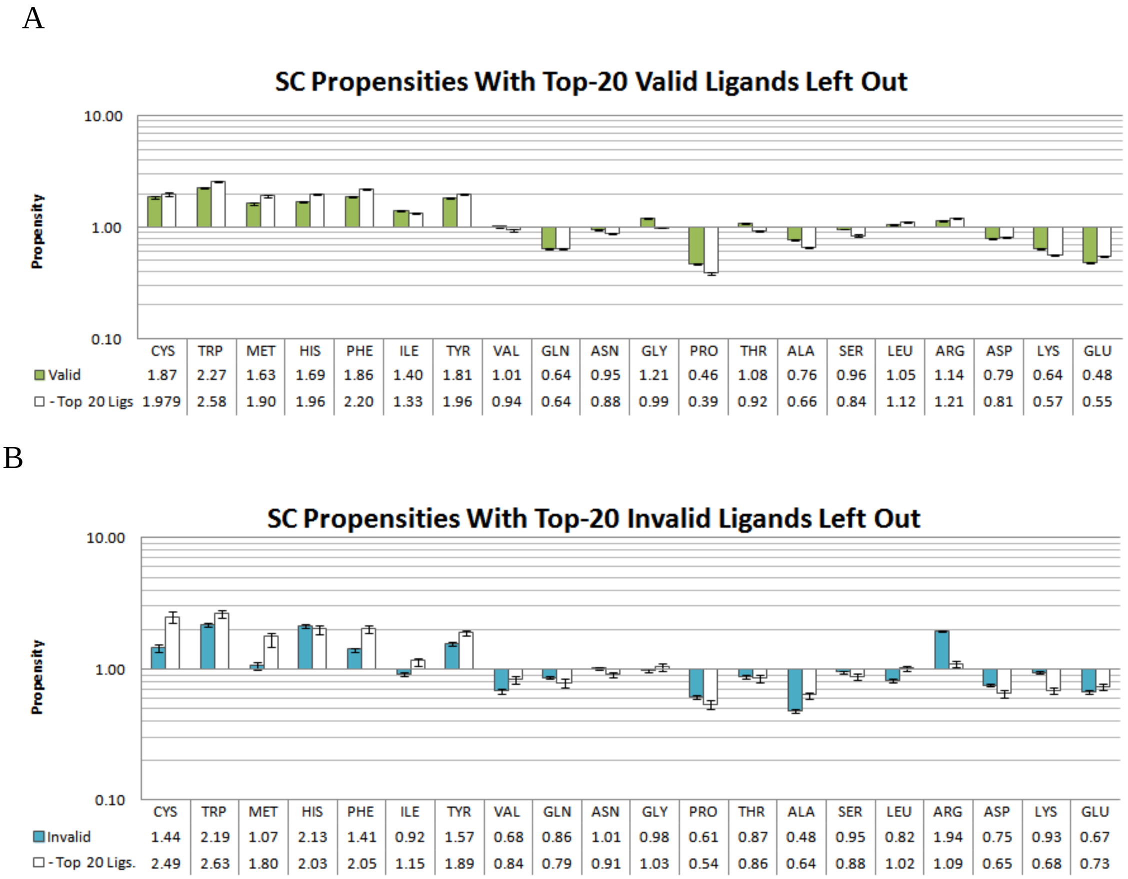
HET	Ligand%	ALA%	ARG%	ASN%	ASP%	CYS%	GLN%	GLU%	GLY%	HIS%	ILE%	LEU%	LYS%	MET%	PHE%	PRO%	SER%	THR%	TRP%	TYR%	VAL%
SO4	26.09	2.22	**24.72**	4.85	4.02	0.69	4.05	5.37	4.88	8.03	1.35	2.74	10.42	1.28	1.84	2.98	7.24	5.89	1.87	4.05	1.52
GOL	16.93	3.03	**11.63**	4.95	7.53	0.31	4.55	7.89	5.17	4.55	4.10	5.35	7.35	1.29	4.95	3.39	4.95	4.68	3.88	6.68	3.79
EDO	9.91	3.37	**10.50**	6.50	6.35	0.78	5.09	6.50	4.23	3.68	5.25	6.58	6.74	1.18	5.88	4.08	4.23	4.86	4.47	6.58	3.13
CL	6.79	2.42	**15.80**	7.62	3.90	1.67	2.79	3.16	6.69	10.04	1.86	5.39	10.04	1.12	2.79	4.83	5.39	5.39	1.86	4.09	3.16
PO4	5.46	2.22	**18.03**	4.30	7.21	0.97	3.74	6.38	9.71	9.02	0.97	1.53	10.68	0.69	2.36	1.80	8.74	4.58	0.69	4.58	1.80
ACT	3.24	1.81	**13.18**	3.10	3.62	1.55	4.39	6.20	2.58	9.30	4.65	6.20	8.79	2.07	5.94	1.81	5.94	3.36	2.58	7.75	5.17
MPD	2.25	3.26	8.31	5.34	8.31	-	3.56	6.53	4.15	2.67	5.04	8.31	3.26	2.67	6.82	6.53	4.75	3.26	3.56	**9.50**	4.15
EGL	1.85	1.55	**11.92**	5.70	7.25	-	4.15	5.18	1.55	6.22	2.59	8.81	7.77	1.55	6.22	6.74	2.59	7.25	1.55	6.74	4.66
FMT	1.56	3.80	**13.92**	8.23	11.39	-	2.53	3.80	5.70	5.06	1.90	3.16	8.86	1.27	0.63	2.53	5.70	9.49	4.43	3.80	3.80
TRS	1.16	2.11	5.79	6.32	**9.47**	1.05	5.79	3.16	8.95	4.21	4.74	7.37	5.79	1.58	2.63	4.21	2.11	4.74	6.32	6.84	6.84
ACY	1.10	1.87	**18.69**	5.61	9.35	-	1.87	2.80	6.54	3.74	2.80	3.74	8.41	2.80	5.61	3.74	8.41	6.54	0.93	5.61	0.93
PEG	0.98	7.08	9.73	2.65	10.62	0.88	1.77	8.85	7.96	4.42	-	**12.39**	3.54	1.77	2.65	4.42	4.42	2.65	3.54	7.08	3.54
IPA	0.81	3.23	**11.83**	5.38	3.23	3.23	1.08	-	2.15	6.45	7.53	**11.83**	-	4.30	8.60	3.23	8.60	2.15	2.15	7.53	7.53
BOG	0.78	1.37	6.85	1.37	2.05	-	0.68	4.11	3.42	2.74	11.64	**19.86**	3.42	2.05	13.70	3.42	2.05	3.42	5.48	4.79	7.53
IOD	0.75	2.13	**12.77**	4.26	2.13	2.13	2.13	-	2.13	4.26	2.13	6.38	**12.77**	2.13	2.13	6.38	8.51	10.64	8.51	6.38	2.13
EOH	0.64	3.45	8.62	5.17	6.90	5.17	5.17	5.17	6.90	6.90	5.17	10.34	-	-	1.72	-	5.17	5.17	5.17	**12.07**	1.72
BR	0.58	-	6.98	6.98	-	-	6.98	6.98	4.65	6.98	9.30	2.33	4.65	2.33	9.30	**18.60**	-	2.33	-	6.98	4.65
MES	0.55	4.71	9.41	4.71	3.53	-	5.88	8.24	2.35	4.71	2.35	**10.59**	5.88	1.18	5.88	7.06	8.24	2.35	5.88	4.71	2.35
MG	0.52	-	10.00	6.67	**53.33**	-	3.33	20.00	-	3.33	-	-	-	-	-	-	-	3.33	-	-	-

In solution, all charged side chains may be expected to attract small polar ligands classified as invalid in our dataset. However, we see higher frequencies for positively charged residues (Arg, Lys) than for negatively charged ones (Glu, Asp) in invalid binding sites. It is unusual that Glu and Asp are under-represented in invalid binding sites because an abundance of positively charged ions are present in buffers just like negative ions. Asp and Glu are indeed frequent in Mg^{+2} sites, where they comprise 22 of 30 residues across 18 sites. However, the binding of positive ions is not observed often in our dataset; Mg^{+2}, Na^{+}, and Ca^{+2}, and are 20th and 23rd and 26th highest occurring invalid ligands by frequency, and together they represent less than 0.8% of all invalid binding sites. This is in contrast to Cl^{-}, I^{-}, and Br^{-}, which all make the top 20 list and comprise ~8% of invalid sites (Table III-3). The higher desolvation cost of a positive ion might make such binding interaction less frequent, and, thus, less likely to appear in protein crystal structures (outside of functional active sites where they frequently appear as co-factors).

3.3.3 Assessment of Ligand Bias on Propensity Values

There is a significant bias in the PDB among the valid ligands (abundance of nucleosides) and invalid ones (common buffer molecules). To measure the bias introduced by preponderance of such ligands, we recalculated propensities while leaving out any binding sites containing the most frequent 20 ligands. Omission of the most frequent valid ligands (~32% of the set) slightly raised propensities of Trp, Phe, His, Met, and Glu and lowered those of Ser, Ala, and Pro (Figure III-5A). However, the omission had little effect overall. In contrast, propensities for invalid binding sites were significantly affected by the removal of the 20 most frequent invalids, which account for about 82% of invalid sites (Figure III-5B). The propensities for Trp, Phe, Met, and Tyr rose sharply while propensities for Arg and Lys fell, indicating a respective increase and decrease in frequencies of these residues in the remaining binding sites (protein surface frequencies remained the same, data not shown).

Figure III-5: Propensities of residue SC interactions in valid sites, with and without the top 20 ligands by frequency. A) Propensities in valid sites. B) Propensities in invalid sites. The error bars represent 95th percentile bounds based on leave-10%-out clustering within each set. Residues are ordered by increasing frequency on protein surface.



	CYS	TRP	MET	HIS	PHE	ILE	TYR	VAL	GLN	ASN	GLY	PRO	THR	ALA	SER	LEU	ARG	ASP	LYS	GLU
🟩 Valid	1.87	2.27	1.63	1.69	1.86	1.40	1.81	1.01	0.64	0.95	1.21	0.46	1.08	0.76	0.96	1.05	1.14	0.79	0.64	0.48
□ - Top 20 Ligs	1.979	2.58	1.90	1.96	2.20	1.33	1.96	0.94	0.64	0.88	0.99	0.39	0.92	0.66	0.84	1.12	1.21	0.81	0.57	0.55

	CYS	TRP	MET	HIS	PHE	ILE	TYR	VAL	GLN	ASN	GLY	PRO	THR	ALA	SER	LEU	ARG	ASP	LYS	GLU
🟦 Invalid	1.44	2.19	1.07	2.13	1.41	0.92	1.57	0.68	0.86	1.01	0.98	0.61	0.87	0.48	0.95	0.82	1.94	0.75	0.93	0.67
□ - Top 20 Ligs.	2.49	2.63	1.80	2.03	2.05	1.15	1.89	0.84	0.79	0.91	1.03	0.54	0.86	0.64	0.88	1.02	1.09	0.65	0.68	0.73

These changes highlight the dependence of the propensities upon the size of the dataset and the variety of ligands it contains. While the propensities calculated for valid binding sites represent a broad array of ligands, invalid propensities are dominated by interactions that are made to the most frequent ligands, namely – sulfate, glycerol, ethylene glycol, and phosphate. This bias is inherent in protein crystallographic data and should be kept in mind when performing broad statistical analysis of residue interactions. Moreover, the large changes in propensities for the reduced set of invalid binding sites are hard to interpret, since subsets of such small size (352 structures remained) have large variation in the leave-10%-out cross-validation. As we see in Table III-4, random subsets from such a small set result in high standard deviations, even if all ligands are allowed, indicating that a sufficiently large set of sites has not been sampled to produce confident propensity estimates. This exposes a caveat of any frequency- or propensity-based protein analysis with small sets of proteins: variation of binding-site frequencies in small sets of structures can have large effects on propensities (see below). Such comparison should only be done in the context of overall residue frequencies and with the knowledge of the uncertainty inherent to the small dataset.

3.3.4 Influence of the Size of the Datasets on Propensity Confidence

To assess the statistical significance of the data, propensity calculations for each set of binding sites were carried out 10,000 times, each time leaving out a random 10% of the proteins (i.e., retaining ~3000 structures at random). For each residue, the median of the 10,000 propensity values is reported, and the 95th percentile bounds are used for the error bars. To assess the dependence upon the size of the dataset, a separate series of calculations were conducted using the procedure above. Progressively larger sets of proteins were randomly chosen from the set of 3295 structures, and propensities were calculated for that set without additional leave-10%-out sampling. The set size was incremented in intervals of 1% of the full structure set and 100 samples were taken at each percentage points, resulting in a total of 10,000 values. Frequencies and propensities were calculated for each sample (Figure III-6). Additionally, propensity medians, standard deviation, and 95th percentiles for six representative residues were calculated from 10,000 random samples at four different set sizes: 100, 500, 1000, and 2000 structures (Table III-4).

The variation in SC frequencies and propensities were thus assessed by sampling random sets of varying numbers of structures (Figure III-6) 100 times each. For clarity, we focused on 6 representative residues: Lys and Glu as the most frequent on protein surface, Val and Asn as moderately frequent, and Cys and Trp as the least frequent. The protein surface contains the most residues by number, and the residue frequencies converge to within ±0.5% variation once ~ 500 or more structures are sampled (Figure III-6A). The binding sites are much smaller than the protein surface, so a larger number of structures are needed to achieve convergence of ±0.5% variation: ~1500 structures for valid sites (Figure III-6B) and ~2500 structures for invalid sites (Figure III-6C). The propensity values fluctuate in proportion to the frequencies (Figure III-6D) and converge around ~1000 structures in a dataset. Standard deviations of propensities for Lys and Glu in valid and invalid binding sites are below 0.1, even in subsets as small as 500 structures (Table III-4). The propensities of rare residues do not converge to such small standard deviation until sets as large as 2000 structures are sampled, especially in the case of propensities for invalid sites. Convergence to mean values of the underlying population is guaranteed as the sample set size approaches the size of the full set; however, the rate of this convergence indicates whether relatively small subsets sufficiently sample the full population means. When constructing a dataset for computing propensities, a balance is required between eliminating redundant or poor quality structures and maintaining a sufficient set size. Based on our results, a set of at least 1000 structures is required to confidently measure general binding-site propensities for valid ligands and 2500 for invalid ligands. Of course, this figure is based on a random and non-redundant protein set. Frequencies and propensities for a set of related proteins (for example, those from the same structural fold family) may show such convergence with fewer structures. We recommend that any propensities calculated on a limited set of structures should be assessed by comparison to the best-available general propensities (such as ones presented here) and by taking into account the variation in random subsets of similar size.

Figure III-6: A) Protein surface, B) valid binding-site, and C) invalid binding-site frequencies, and D) valid binding-site propensities of six residues. Values for subsets of the protein structure set, from 1% to 99% of the full set, are shown with 100 samples at each percent point.

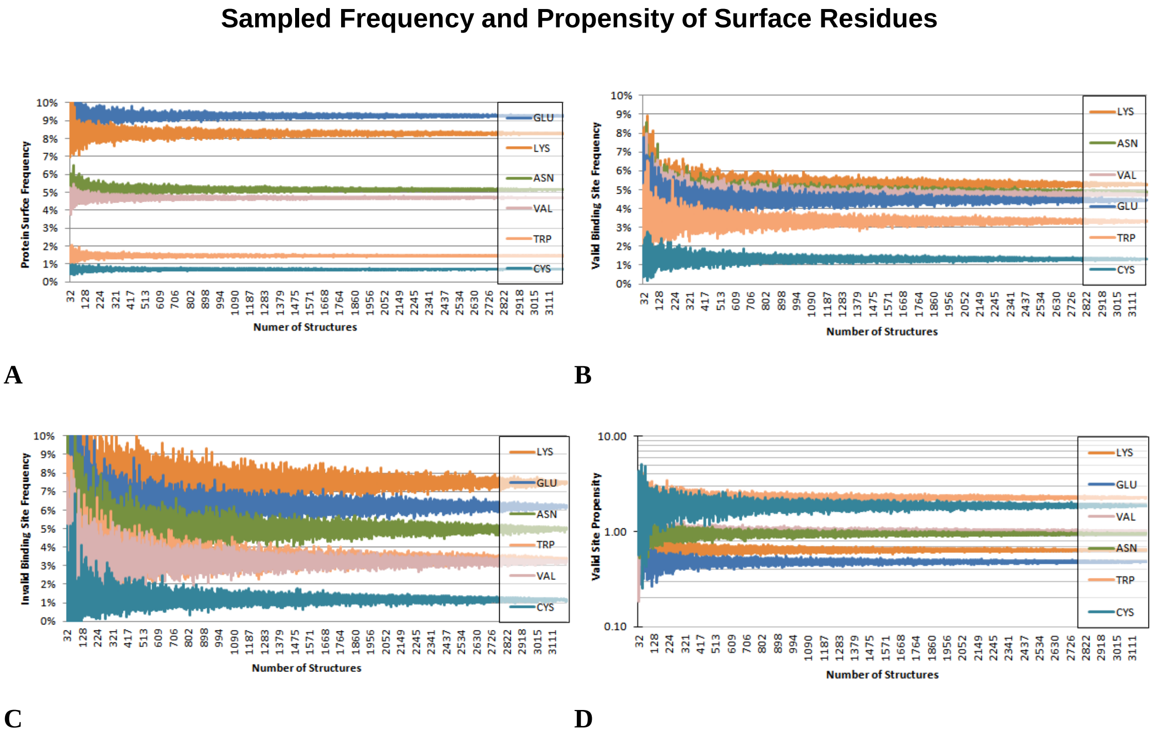


Table III-4: Propensity median, standard deviation, and 95th percentile bounds for 6 representative residues in sampled subsets of protein structures. All values based on 10,000 random samples from the full protein set.

Propensity median and **standard deviation** (with 95th percentile bounds) of Representative Residues

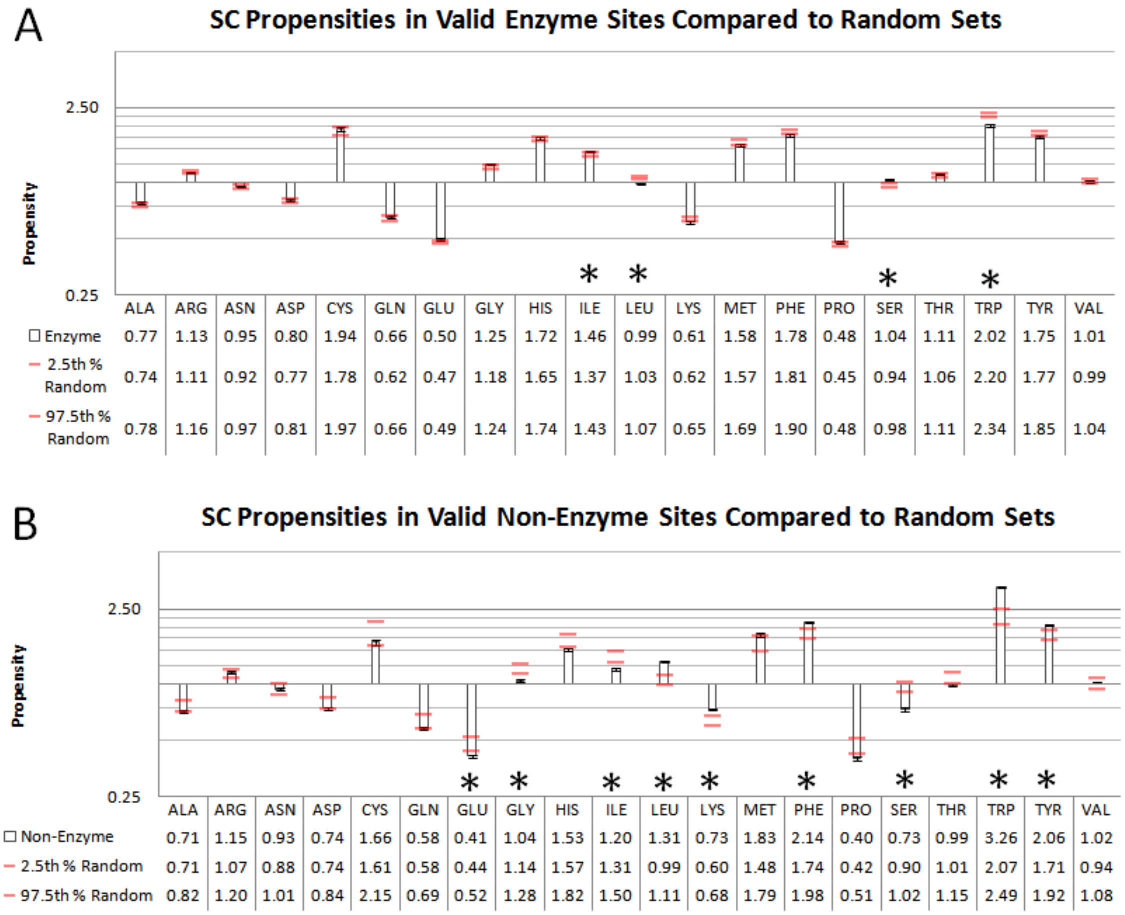
		100 Structures		500 Structures		1000 Structures		2000 Structures	
		Valid	Invalid	Valid	Invalid	Valid	Invalid	Valid	Invalid
Frequent	LYS	0.64, **0.08** (0.48/0.82)	0.91, **0.21** (0.55/1.38)	0.64, **0.03** (0.57/0.71)	0.91, **0.08** (0.75/1.09)	0.64, **0.02** (0.60/0.68)	0.91, **0.06** (0.81/1.03)	0.64, **0.01** (0.62/0.66)	0.91, **0.03** (0.86/0.97)
	GLU	0.48, **0.07** (0.34/0.62)	0.67, **0.16** (0.38/1.01)	0.48, **0.03** (0.42/0.54)	0.66, **0.07** (0.54/0.80)	0.48, **0.02** (0.44/0.52)	0.67, **0.04** (0.58/0.76)	0.48, **0.01** (0.46/0.50)	0.67, **0.02** (0.62/0.71)
Moderate	VAL	1.01, **0.14** (0.75/1.30)	0.66, **0.24** (0.26/1.19)	1.01, **0.06** (0.90/1.12)	0.68, **0.10** (0.49/0.88)	1.01, **0.04** (0.94/1.08)	0.68, **0.06** (0.55/0.81)	1.01, **0.02** (0.97/1.05)	0.68, **0.03** (0.61/0.75)
	ASN	0.94, **0.13** (0.71/1.22)	0.96, **0.26** (0.52/1.57)	0.95, **0.05** (0.85/1.05)	0.99, **0.11** (0.79/1.22)	0.95, **0.03** (0.88/1.01)	0.99, **0.07** (0.86/1.13)	0.95, **0.02** (0.91/0.98)	0.99, **0.04** (0.92/1.06)
Infrequent	CYS	1.86, **0.52** (0.96/3.00)	1.46, **1.13** (0.00/4.52)	1.87, **0.21** (1.48/2.29)	1.62, **0.47** (0.89/2.71)	1.87, **0.14** (1.61/2.15)	1.66, **0.30** (1.13/2.29)	1.87, **0.07** (1.73/2.01)	1.67, **0.16** (1.35/1.97)
	TRP	2.26, **0.41** (1.53/3.13)	2.20, **0.84** (0.77/4.06)	2.27, **0.16** (1.96/2.60)	2.28, **0.34** (1.64/2.98)	2.27, **0.11** (2.07/2.49)	2.29, **0.22** (1.87/2.72)	2.27, **0.06** (2.16/2.38)	2.29, **0.12** (2.06/2.52)

As an example, we looked at the propensities in enzyme and non-enzyme valid ligand binding sites, which have been previously shown to differ in their ligand efficiencies [22]. Figure III-7 shows the propensities along with red lines indicating the 95th percentile bounds of valid propensities from random sets of structures sampled 10,000 times from the full dataset (as presented in Table III-4). For enzymes, sets of 2500 structures were sampled, while for the smaller non-enzyme set only 1000 structures were sampled. The leave-10%-out sampling used during the propensity calculations provides a measure of stability for the propensity values. In contrast, the sampling of random structures provides a bound for propensity values that can be expected by chance. Therefore, for enzyme or non-enzyme propensities to be considered different from the general (randomly observed) valid binding-site propensities, their ±95th percentile range must be outside the 95th percentile range of propensities obtained from random structure sets of the same size. The asterisks in Figure III-7 mark residues that fulfill this criterion. This is the strictest-possible criterion, because only minimal overlaps of the median distributions can still be considered statistically significant. The average values of random sampling will be enzyme-biased because Binding MOAD and the PDB are themselves enzyme-biased. Therefore, exceptional propensity trends for non-enzyme may be more likely.

The set of enzyme structures makes up more than two-thirds of the structure set used to compute propensities in this study. Binding-site propensities computed on this number of structures are very close to general propensity trends seen across all valid binding sites. Accordingly, the variation of propensities in corresponding random samples is very low. In enzyme binding sites, Ile and Ser have median propensities higher than random, and Leu and Trp lower ones. The set of non-enzymes has nine residues that have propensities significantly different than those seen at random. Leu, Lys, Phe, Trp, and Tyr have significantly higher binding-site propensities than those seen in sets of random structures, and Glu, Gly, Ile, and Ser have lower-than-random propensities. In a recent study comparing residue composition of enzyme and non-enzyme sites, Leu, Met, Trp and Tyr have been shown to have much higher frequencies in binding sites of high-affinity non-enzyme proteins than in enzyme high-affinity binding sites [22]. Combined with our propensity observations, the presence of Leu, Trp, and Tyr residues in binding sites without enzymatic function may be a distinguishing trend. Although Met propensity is higher in non-enzyme sites, it is within random sampling error. The Carlson et al. study also observed relatively low non-enzyme binding-site frequencies for Val, Ile, Asp, and Gly. Our propensities for Ile and Gly are consistent with their findings, but Asp has no propensity trend among enzymes versus non-enzymes, aside from its low propensity for binding sites in general. The elevated propensity of Lys and Phe and lower propensities for Glu and Ser for non-enzyme sites are unique trends observed in the current study.

As smaller sets of structures are used for calculating propensity values, there is a greater chance of seeing values that deviate from general binding-site propensity trends. However, the 95th percentile margins of error from randomly sampled sets of similar size will also change, becoming wider, especially for less-frequent residues. Therefore, it is important to conduct comparisons to randomly-sampled propensity values as suggested herein, to distinguish set-specific trends from the overall propensity trends in the currently available data.

Figure III-7: Propensities in A) enzyme and B) non-enzyme valid binding sites The black error bars represent 95th percentile bounds based on leave-10%-out clustering. For context, red lines represent 95th percentile bounds of propensities from 10,000 random samples of A) 2500 random, diverse proteins and B) 1000 random, diverse proteins (as seen in Table III-4). Stars indicate residues whose median propensity value (±leave-10%-out 95th percentile error) falls outside of the 95th percentiles of the randomly-sampled propensities.



	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
□ Enzyme	0.77	1.13	0.95	0.80	1.94	0.66	0.50	1.25	1.72	1.46	0.99	0.61	1.58	1.78	0.48	1.04	1.11	2.02	1.75	1.01
− 2.5th % Random	0.74	1.11	0.92	0.77	1.78	0.62	0.47	1.18	1.65	1.37	1.03	0.62	1.57	1.81	0.45	0.94	1.06	2.20	1.77	0.99
− 97.5th % Random	0.78	1.16	0.97	0.81	1.97	0.66	0.49	1.24	1.74	1.43	1.07	0.65	1.69	1.90	0.48	0.98	1.11	2.34	1.85	1.04

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
□ Non-Enzyme	0.71	1.15	0.93	0.74	1.66	0.58	0.41	1.04	1.53	1.20	1.31	0.73	1.83	2.14	0.40	0.73	0.99	3.26	2.06	1.02
− 2.5th % Random	0.71	1.07	0.88	0.74	1.61	0.58	0.44	1.14	1.57	1.31	0.99	0.60	1.48	1.74	0.42	0.90	1.01	2.07	1.71	0.94
− 97.5th % Random	0.82	1.20	1.01	0.84	2.15	0.69	0.52	1.28	1.82	1.50	1.11	0.68	1.79	1.98	0.51	1.02	1.15	2.49	1.92	1.08

3.4 Conclusion

Our study highlights the differences in side-chain interactions with valid and invalid ligands and the frequency of residues taking part in these interactions, in contrast to the surface composition of the whole protein. Most importantly, the relative propensity of valid versus invalid binding sites should help to improve methods for identifying binding sites in proteins of unknown functions and other proteomic methods where understanding of general composition of protein-ligand binding sites is required. Better understanding of these interactions, and how they differ across binding sites, can help focus statistical analysis of broad sets of protein surfaces toward the most biologically relevant ligands. It also exposes the variation in residue frequencies on the protein and binding-site surfaces in randomly chosen sets of proteins. Given how this variation can affect the interpretation of frequency- and propensity-based analysis of protein surfaces, we recommend that at least 1000 diverse protein complexes are needed for significant general conclusions for biologically relevant valid binding sites. When calculating propensities for smaller sets of structures, such as proteins of a functional family or similar ligand-binding sites, it is important to compare them to those of randomly sampled sets of structures. This can help determine how significant the trends are with respect to the variety of protein-ligand sites currently available in databases such as BindingMOAD.

CHAPTER IV

Propensity-Based Scores to Improve a Binding-Site Prediction Algorithm

4.1 Introduction:

Proteins are the workhorses of the biochemical world, and they perform their functions through interactions with macromolecules (e.g. other proteins, DNA, RNA) or small molecule ligands (e.g. reactants, co-factors, or signaling molecules). The bewildering variety of protein functions are all mediated through contact between the protein surface and the interacting molecule, giving rise to an equivalent variety of protein interaction sites, each specialized to bind certain ligands and perform a certain function. Understanding how the protein binding surfaces allow for specificity and selectivity of their binding partners is crucial for such endeavors as characterizing protein function and developing drugs to target specific protein binding events and modulate the protein function. Identification of these binding sites on the protein is a critical goal of SBDD. Due to the increasing amount of un-characterized protein structures deposited in structure databases, such as those coming from large structural genomics projects [127], the need for tools to effectively predict un-known binding sites is pressing [120].

A large fraction of drugs are developed to target protein binding sites that bind small molecules. Those sites are usually smaller than protein-protein binding sites, and therefore present a smaller target area for the rational design of drug molecules with good pharmacokinetic properties. Current SBDD methods rely on existing protein structures to inform the drug design process. The growing number of structures of proteins co-crystallized with small molecule ligands in the PDB [2] has motivated the rise in systematic study of the common features of the surface responsible for molecular interaction between proteins and ligands [13] [128] [10] [129]. Ligands bind to their binding sites in lieu of the rest of the protein for several reasons: including geometric complementarity, physicochemical complementarity, and other energetically favorable effects. None of these factors alone can account for the overall specificity or selectivity of the binding site [13] [130]. However, in general, we can expect the site of ligand binding to be different in composition and/or geometry from the rest of the protein in order to allow the interactions necessary for diffusion-mediated ligand binding.

The differences between protein-ligand sites and the rest of the protein have been exploited to construct many algorithms for the prediction of binding sites. These can be classified into three broad categories – those utilizing the protein geometry for their predictions, those utilizing energy-based descriptions of the protein surface, and those using knowledge-based methods to compare protein surface features to known binding sites (reviewed in [3] [36] [11]). The knowledge-based methods can perform well when the protein of interest has homologs or orthologs in existing databases or a similar ligand-binding mode can be identified, but they do not encompass information that can be generalized across many proteins. Ligand cross-reactivity with several proteins, for example, can occur without significant similarity in their binding sites. The geometric methods rely on the general observation that the largest and/or deepest cavity on the protein surface is one that is likely to bind a ligand [48] [53]. The energy-based methods use known physicochemical forces of atom-atom interactions, such as vdW interactions, hydrophobicity, or hydrogen-bonding potentials of the protein surface atoms, usually by measuring those forces with respect to probes or molecule fragments [58] [43]. Current state-of-the-art methods achieve success rates of up to 95% on holo structures, and most perform worse when tested on apo structures of the corresponding proteins[57]. However, the test sets for these methods often contain many redundant protein structures, which may inflate the performance data.

In recent years, prediction methods have started to combine measures of geometry with energy-based or genome-based information to increase their predictive power [42] [131] [132] [133]. Several methods measure evolutionary conservation by using multiple alignments in a family of proteins to identify residues that appear in the binding sites of multiple family members. This approach is especially powerful in selectively predicting the functional class of binding sites, especially if only sub-sequences of ligand-binding residues are aligned [134]. However, it is dependent on the size of the available protein families, and does not capture cases of convergent evolution, where seemingly un-related proteins evolve binding sites that bind a similar set of ligands. Alignment-independent analysis of binding-site residue conservation also shows preferences of certain residues for binding surfaces as opposed to the rest of the ligand surface [32, 33] . Similarly, an analysis of atom triplet propensities for binding surfaces showed they can be used to predict protein-ligand sites by delineating protein surfaces that contain atom triplets biased towards binding sites [122]. However, to our knowledge, none of the methods have utilized general trends of residue composition to help improve their predictions.

In this study, we obtain the propensities of amino acids within ligand binding sites in a large set of non-redundant protein-ligand complexes available in the Binding MOAD database. We then use the propensity measurements to improve the rank-scoring performance of a binding-site prediction algorithm. Since residue composition of the binding site reflects the atom types available for energetic interactions with the ligand, we choose a binding site prediction method that utilizes protein surface geometry, instead of potential surface interaction energy, for predicting potential binding pockets. The geometric properties of the protein surface captured by such a method will be complementary to our residue propensity information. Specifically, we employ the SiteFinder method implemented in the Molecular Operating Environment (MOE) software suite. Despite rough estimates of hydrophobicity in its cavity selection algorithm, SiteFinder is still considered a primarily geometry-based method [57]. SiteFinder locates a series of large concave pockets on the protein surface, and ranks these pockets according to a scoring scheme that incorporates size and hydrophobicity of the pocket.

A recently conducted study used the pockets predicted by SiteFinder predictions and a rank-score based on a protein-ligand binding index (similar to residue propensities) to predict known drug-binding pockets in a set of crystal structures [121] and homology models [135]. Our study distinguishes itself in the quality, size and diversity of the training set used to calculate the residue propensities, the assessment of residue pair propensity values in addition to single residues, and the use of a consensus score to combine propensity-based scores with those of SiteFinder. While Soga et al.

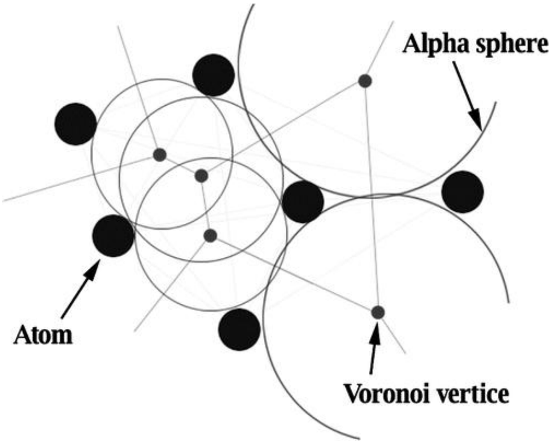
demonstrate that a propensity-based score can be used to successfully rank-order predicted sites, we perform a more thorough comparison of various scoring schemes and reveal the factors behind the high success rates of propensity-based scores. We also demonstrate certain cases where a consensus score (combination of SiteFinder and propensity-based score) may be more helpful in ranking successful SiteFinder prediction highly.

4.2 Methods:

4.2.1 SiteFinder

SiteFinder is implemented in the MOE software package from the Chemical Computing Group (CCG) [56] and belongs to the class of binding site identification algorithms that rely primarily on the geometry of the protein surface to predict ligand-binding cavities. To identify a potential predicted site, SiteFinder first uses Delaunay triangulation [52] to obtain a set of vertices that each correspond to plane of a Voronoi tessellation of the protein (a dual graph of the Voronoi tesselation) [51]. A set of spheres with varied radii, called the alpha-spheres (Figure IV-1), are then associated with the triangulated points, such that each alpha-sphere touches three protein atoms and has no internal atoms [54]. SiteFinder prunes this set of α-spheres based on their size and solvent exposure, eliminating those that are too buried or too exposed, and then labels the α-spheres as hydrophobic or hydrophilic based on the local atom environment near the sphere. The α-spheres are clustered into cavities, which consist of more than one α-sphere, and contain at least one hydrophobic sphere. The relative rank of these predicted cavities is determined by a score representing the number of hydrophobic atoms in contact with the α-spheres, which normally scales with cavity size. Using this algorithm, SiteFinder locates cavities of various shapes and sizes with a dense hydrophobic character.

Figure IV-1: Example of an alpha sphere. The alpha sphere is displayed in red, and the three contacted atoms (2D) are in black. Figure taken from Schmidtke et al. [57]



4.2.2 Running SiteFinder with Optimized Parameters:

One of the pitfalls of the ‘out-of-the-box’ SiteFinder algorithm is the relatively large size of its predicted binding pockets [57]. This can result in deceptively large success rates, as the large swaths of protein surface covered by the predicted pockets have a greater chance of including a known site. However, SiteFinder remains one of the top current geometric prediction algorithms, and its implementation in MOE makes it relatively simple to use on a large set of structures. Schmidtke et al. recently performed a systematic scan of SiteFinder parameters to find the optimal balance between predicted site size and prediction success [57]. We therefore use SiteFinder with optimized parameters of *da_dist* = 4.0, *connect_dist* = 4.6 Å, and *site_minrad* = 1.8 Å. SiteFinder was executed with a modified version of a *batch_sitefinder* SVL script distributed by CCG, and the member residues of the predicted binding sites were output to flat files for further processing. The member residues were identified by SiteFinder as those in the immediate vicinity of the binding-site spheres delineating a predicted pocket.

4.2.3 Prediction Set (structures from 2003 and earlier):

The evaluation is comprised of binding sites from 1080 non-redundant proteins from the Binding MOAD database that have been published in the PDB prior to 2003. We chose to use this date cutoff to approximate the set used by An et al. for their analysis of the “pocketome” [13]. Our set of proteins is smaller than the 5000+

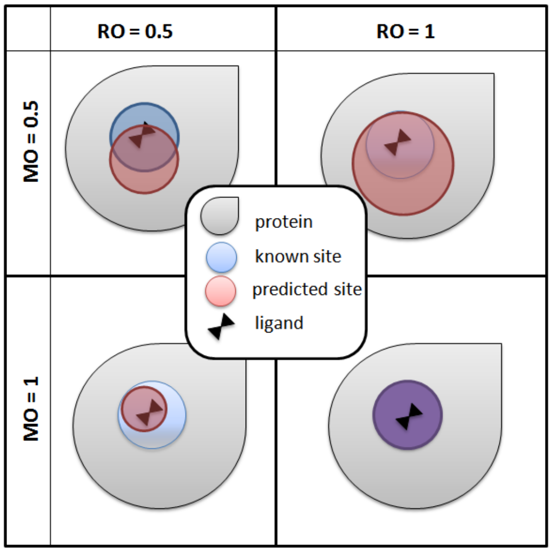
structures in the pocketome because we limit it to proteins that appear in Binding MOAD, and we have removed redundancy. The non-redundant set is grouped by 90% sequence identity. Any ligands listed in Binding MOAD are considered to represent “known” binding sites. Ligands were also classified as “valid” or “invalid” according to Binding MOAD annotations. Valid ligands include any biologically relevant co-crystallized molecules known to play a role in the protein function. These include a wide variety of natural products, as well as drugs and other human-made molecules. Conversely, any “opportunistic” co-crystals not directly involved in protein function, such as buffers, solvents or other crystallographic additives, were classified as invalid. Only valid ligand binding sites were considered as known sites. To ensure that we match any and all possible correct predictions, any redundant ligand sites within a protein were considered. Known binding sites were defined by residues with at least one side-chain atom within 4.0 Å of the ligand. If more than one ligand was present within the cutoff distance from a residue, it was included in the binding site definition of both ligands.

Before processing with SiteFinder, the biounit files for these structures were obtained from Binding MOAD and processed to remove any HETATM records corresponding to the known ligands. Solvent atoms and salts are ignored by SiteFinder. Other HETATM records in the biounit, such as modified residues were retained for completeness. The stripped proteins were loaded into MOE and processed by SiteFinder with optimized parameters noted above.

4.2.4 Determining a correctly predicted binding site:

The SiteFinder algorithm explicitly outputs the residues that belong to the surface patches outlined by the alpha spheres in the predicted cavity. The identities of the residues in the predicted and known sites were used to determine the quality and success of the prediction. Since the binding-site prediction and the known-site extraction were performed on the same biounit, and our definition used for a binding site is residue-based, the identity of the residues and their parent protein chain in the predicted and known sites is sufficient to determine correspondence. For each protein, residues in each predicted site were matched to those in known binding sites. For each predicted/known site pair the extent of the match was quantified as follows. A relative overlap (RO) was defined as N_m/N_k, where N_m is the number of matched residues, and N_k is the number of residues in the known site. Predicting all the residues of a known site would thus result in an RO of 1, and an RO of 0 would signify no match (Figure IV-2). A large predicted site will thus have a higher probability of matching a known pocket. Since a predicted pocket that covers the entire protein surface and achieves an RO of 1 is not very valuable in a real-world application, we also compute the mutual overlap (MO) of the predicted and know sites. MO is defined as N_m/N_p, where N_p is the size of the predicted site, and MO values closer to 1 indicate a more accurate prediction, namely one that contains only residues of the known binding site. Schmidtke et al. recently showed that SiteFinder is one of the methods prone to generate large predicted sites with a relatively low MO [136], and that this criterion is valuable for assessing the quality of a prediction.

Figure IV-2: Graphical illustration of the RO and MO criteria used for assessing a successful binding-site prediction.



Each predicted pocket in a protein is evaluated by calculating an RO and MO with respect to each of the known sites in the protein. A given predicted site with an RO > 0.5 is considered a correct prediction. If the predicted pocket contains matches to more than one known site, as can be the case when there are two known binding sites close to each other in a single large cavity, we choose the site with the largest MO as the best match. Since SiteFinder sometimes generates several dozen predicted sites per protein, there is often more than one correctly predicted site for protein structures that contain more than one ligand. Therefore, for certain parts of the analysis we limit ourselves to structures with only one bound ligand.

4.2.5 Propensity Set (structures between 2004-2009)

Residue propensities used for scoring the predicted binding sites were independently calculated from a non-redundant subset of the Binding MOAD database that included 2123 structures released in the PDB between 2004 and 2009. This set is completely independent of the prediction set described above. To avoid biasing propensities on multimers, we omit redundant sites within a structure, i.e., those that have the same ligand and same binding site residues, such as equivalent sites in a multi-meric protein. Surface-based propensities of individual residue side chains were calculated via the following propensity formula (also described in Chapter 2):

$$P_i = \frac{F_i^{BS}}{F_i^{PS}} \quad \text{where } F_i^{BS} = \frac{\sum_s N_i^s}{\sum_s \sum_i^{20} N_i^s} \quad \text{and } F_i^{PS} = \frac{\sum_p N_i^p}{\sum_p \sum_i^{20} N_i^p}$$

Where P_i is the propensity of amino acid i ($I = Ala, Arg...$) and F_i^{BS} and F_i^{PS} are the surface frequencies of the amino acid i in the binding sites or protein surface. The frequencies are calculated by summing over all occurrences of an amino acid in binding sites (N_i^s) or on the protein surface (N_i^p) in a set of structures.

Propensity of a pair of co-occurring residues to appear in binding sites is calculated in a similar manner [137]. The propensity of two co-occurring amino acids i and j is:

$$P_{ij} = \frac{F_{i,j}^{BS}}{F_{i,j}^{PS}}$$

where $F_{i,j}^{BS}$ and $F_{i,j}^{PS}$are concurrence frequencies of two amino acids in the binding sites or the protein surfaces respectively. The binding site concurrence frequency $F_{i,j}^{BS}$ is determined by summing the occurrence of each pair of amino acids over the set of binding sites, and dividing by the total number of amino-acid pairs in that set:

$$F_{i,j}^{BS} = \frac{\sum_s N_{ij}^S}{\sum_s \sum_i^{20} \sum_j^{20} N_{ij}^S}$$

If n_x is the number of residue of type *x* in a binding site *s*, then the concurrence of residues i and j in a binding site $N_{ij}^{s} = \ n_i \times n_j$ if $i \neq j$, and $\binom{n}{2}$ if $i = j$. The frequency of protein surface concurrence was calculated assuming that the occurrences of amino acids *i* and *j* are independent:

$$F_{i,j}^{PS} = W_{i,j} \times \ F_i^{PS} \ \times \ F_j^{PS} \text{ with } W_{i,j} = \begin{cases} 1(i = j) \\ 2\ (i \neq j) \end{cases}$$

Propensities were calculated separately by ligand class, resulting in four propensity sets: Valid or PairValid – based on residue frequencies in biologically relevant valid binding sites, and Invalid or PairInvalid – based on residue frequencies in spurious binding sites. The ligand class was based on annotation from Binding MOAD. Care was taken to ensure that only reasonable invalid ligands were retained for the propensity calculation. For example, covalently attached ligands, heme groups and unknown ligands (UNK, UNX, etc) were excluded from the invalid set. Metal ions that were not covalently attached (between 2 and 3 Å from the protein surface) were examined manually, and those involved in ligand coordination were excluded. All retained invalid ligands are thus known molecules that are not covalently attached, and to our knowledge, are not involved in facilitating the biological functionality of the protein. The valid ligand set was not filtered for specific ligand classes and includes many ligand types, from natural small molecules to drugs and small peptides.

Propensity values of the Protein Ligand Binding (PLB) index were obtained from Soga et al. for comparative analysis [121].

4.2.6 Calculation of Raw and Consensus Scores

A series of propensity-based scores were calculated for each pocket predicted by SiteFinder for comparison to, and combination with, the default SiteFinder hydrophobicity score. Single and residue-pair propensities from valid and invalid sites constitute a set of four scores. A score based on the PLB index and one based on size were also calculated. For each predicted site, scores based on single-residue propensities were calculated using the equations below, where p_i is the propensity of residue of type *i* (20 different types of amino acids) and n is the number of residues of that type present in the predicted pocket.

$$score = \sum_i^{20} P_i N_i \text{ (or } log \quad score = \sum_i^{20} \log(P_i) N_i \text{)}$$

For calculation of pair propensities, the equation is extended to:

$$pair\ score = \sum_i^{20} \sum_j^{20} P_{i,j} N_{i,j} \text{ (or } pair\ log\ score = \sum_i^{20} \sum_j^{20} \log(P_{i,j}) Nn_{i,j} \text{)}$$

where N_{ij} was defined as $n_i \times n_j$ when $i \neq j$ and $\binom{n}{2}$ when $i = j$, and P_{ij} is the propensity of i and j calculated previously. Log-based scores were also explored to emphasize the differences in propensity values and avoid additive scores that scaled with the size of the score binding site. We do not use log scores for our analysis unless otherwise stated.

To calculate a consensus or combination score the propensity scores and the SiteFinder scores were first converted to normalized scores (z-scores) according to the standard formula,

$$z\ score = \frac{x-\mu}{\sigma}$$

where x is the score for a given site, μ is the mean score for all sites in the protein, and σ is the standard deviation for all score in the protein. A simple addition of the SiteFinder z-score and a propensity z-score gave the combined score for a predicted site. For each protein, the relative ranks for all the predicted sites were calculated by sorting the z-scores from highest to lowest. Rank for each score type was calculated separately. In all, 10 different rankings were calculated in addition to the default SiteFinder score: Size-based, PLB-based, Valid, Invalid, PairValid, PairInvalid, CombinedValid, CombinedInvalid, CombinedPairValid and CombinedPairInvalid rank. A rank based on the PLB index from Soga et al. was calculated, using propensity values obtained from the publication.

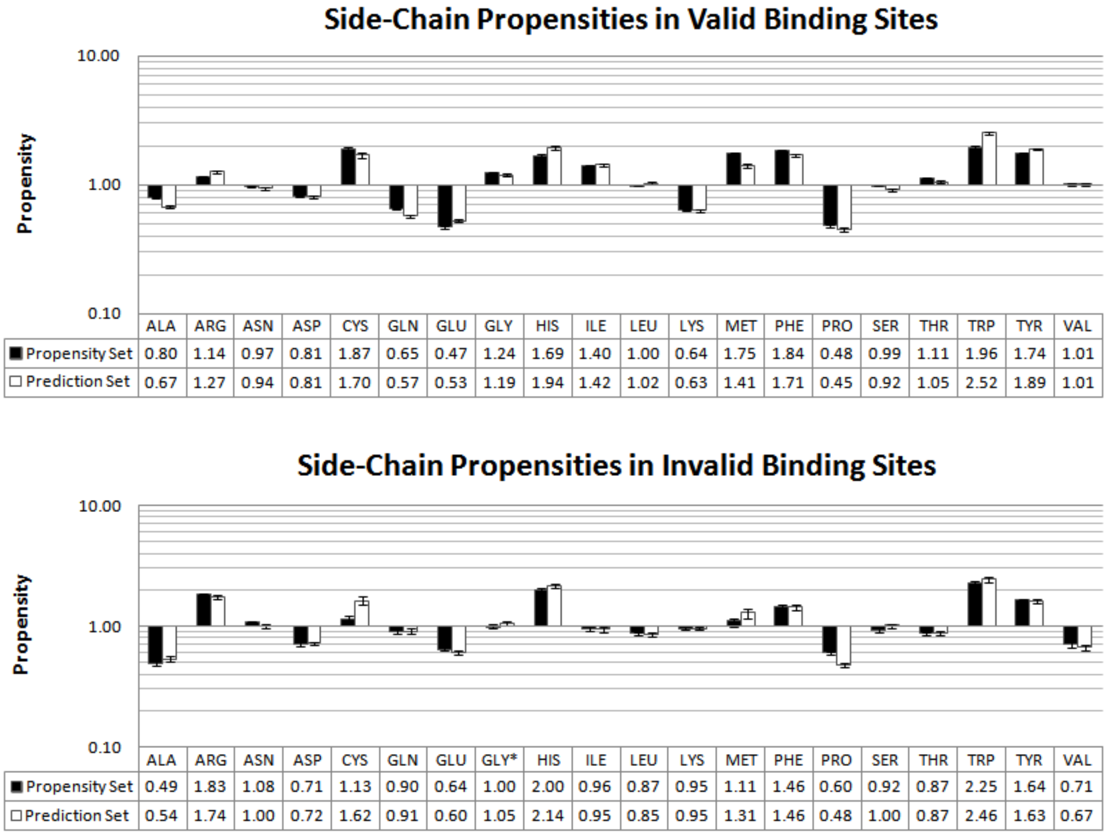
4.3 Results

4.3.1 Propensities

The propensity set contained 2123 structures, all deposited from 2004-2009 in the PDB. This set includes 2471 valid binding sites, and 1084 invalid binding sites. Propensities for the 20 standard amino acids were calculated for the structures from 2003 and earlier, and the structures from 2004-2009, simply to show that one set is not grossly different from the other (Figure IV-3). However, only the propensities from the

2004-2009 propensity set were used to calculate scores of the site-prediction because it is inappropriate to train data to reproduce the answer one wishes to find. The propensities for valid binding sites were largely comparable across the two sets, with some small differences in the rarely occurring residues, such as Trp, His, Cys, and Met. The residues with the largest valid binding site propensities - those over-represented in the valid binding sites – are Trp, Cys, His, Met, Phe, and Tyr. Conversely, Ala, Gln, Glu, Lys, and Pro have low propensities, indicating their under-representation in the valid ligand sites relative to the rest of the protein surface. The propensities derived from the propensity set showed some differences between the valid and the invalid binding site sets. The largest relative difference was observed in Ala, Arg, Asn, Trp, Tyr, and Val. The former three residues have higher propensity in invalid binding sites, while the latter three have higher propensities in the valid binding sites. The more hydrophobic Ala, Trp, and Val, have binding site propensities above 1, while Arg, Asn, and Tyr have propensities near or below 1 in both types of binding sites. No single residue shows a change from a high binding-site propensity (>>1) to a very low binding site propensity (<<1), or vice versa, when valid sites are compared to invalid (Ile shows the strongest such trend). This indicates that despite variability of residues present on the binding surfaces of a protein, the variation in residue frequencies between a ligand-binding surface and a non-binding surface is greater than that between the valid versus invalid ligand-binding surface. Despite this trend, we still hypothesize that valid binding-site residue propensities can be more successful than those of invalid binding sites in scoring binding site predictions. After all, it is important to locate biologically relevant binding sites and reduce the chance of highly ranking surface pockets that are closer in character to a non-functional binding region for invalid ligands. Pair-wise residue propensities derived from the propensity set are presented in Figure IV-4. The trends in pair-wise propensities follow some of the trends of single propensities, especially for infrequently occurring residues.

Figure IV-3: Residue propensities derived from the prediction set (white) and propensity set (white) of valid ligand binding sites A) and invalid ligand binding sites B). Error bars represent 95% quartiles of leave-10%-out cross-validation sampling.



	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
■ Propensity Set	0.80	1.14	0.97	0.81	1.87	0.65	0.47	1.24	1.69	1.40	1.00	0.64	1.75	1.84	0.48	0.99	1.11	1.96	1.74	1.01
□ Prediction Set	0.67	1.27	0.94	0.81	1.70	0.57	0.53	1.19	1.94	1.42	1.02	0.63	1.41	1.71	0.45	0.92	1.05	2.52	1.89	1.01

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY*	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
■ Propensity Set	0.49	1.83	1.08	0.71	1.13	0.90	0.64	1.00	2.00	0.96	0.87	0.95	1.11	1.46	0.60	0.92	0.87	2.25	1.64	0.71
□ Prediction Set	0.54	1.74	1.00	0.72	1.62	0.91	0.60	1.05	2.14	0.95	0.85	0.95	1.31	1.46	0.48	1.00	0.87	2.46	1.63	0.67

Figure IV-4: Residue propensities derived from the propensity set of valid ligand binding sites A) and invalid ligand binding sites B). Higher values highlighted in red, lower values in blue.

A	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	1.51	0.70	0.71	0.56	1.34	0.46	0.28	1.10	1.00	1.22	0.90	0.46	1.23	1.22	0.37	0.70	0.83	1.21	1.27	0.93
ARG		2.07	0.95	0.76	1.57	0.67	0.51	1.15	1.69	1.30	0.90	0.67	1.54	1.69	0.42	1.01	1.11	1.94	1.94	0.96
ASN			2.47	0.67	1.49	0.58	0.42	1.09	1.41	1.28	0.79	0.58	1.41	1.47	0.43	0.95	0.96	1.88	1.77	0.88
ASP				1.33	1.29	0.43	0.30	1.05	1.03	1.15	0.67	0.54	0.94	1.12	0.35	0.71	0.83	1.42	1.28	0.71
CYS					35.13	1.00	0.74	1.99	2.84	2.59	2.05	1.10	3.24	3.27	0.87	1.83	2.22	3.76	3.05	1.87
GLN						1.73	0.27	0.74	0.90	0.82	0.64	0.37	1.03	1.07	0.25	0.63	0.69	1.40	1.21	0.60
GLU							0.65	0.50	0.80	0.60	0.39	0.28	0.65	0.81	0.17	0.39	0.43	1.22	0.88	0.41
GLY								3.07	1.54	2.10	1.17	0.85	1.73	1.66	0.68	1.28	1.49	1.73	1.84	1.31
HIS									7.36	1.89	1.43	0.77	2.29	2.67	0.59	1.35	1.40	3.45	2.95	1.30
ILE										5.08	1.59	0.77	2.53	2.62	0.78	1.35	1.55	2.18	2.36	1.71
LEU											2.06	0.61	2.01	1.99	0.45	0.96	1.07	1.83	1.79	1.28
LYS												1.02	0.81	0.92	0.26	0.67	0.80	1.01	1.08	0.59
MET													13.38	3.73	0.62	1.37	1.57	3.15	2.74	1.79
PHE														7.45	0.75	1.49	1.58	3.89	3.23	1.88
PRO															1.01	0.45	0.50	0.75	0.71	0.54
SER																2.11	1.05	1.93	1.74	0.95
THR																	2.42	1.79	1.82	1.11
TRP																		19.16	5.41	1.76
TYR																			7.55	1.85
VAL																				2.83

B	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	2.54	0.83	0.41	0.21	1.60	0.37	0.24	0.67	0.80	0.85	0.86	0.25	0.84	1.34	0.32	0.55	0.64	1.67	0.97	0.61
ARG		6.68	0.91	0.67	2.25	0.87	0.55	1.00	2.39	1.01	1.01	0.79	1.19	1.74	0.44	1.20	0.94	2.44	1.85	0.75
ASN			5.38	0.32	1.51	0.52	0.33	0.60	1.13	0.67	0.60	0.52	1.09	0.88	0.38	0.52	0.60	1.41	1.18	0.43
ASP				2.68	0.54	0.25	0.28	0.46	0.77	0.30	0.32	0.48	0.41	0.48	0.25	0.44	0.40	0.80	0.76	0.31
CYS					89.28	2.00	1.13	1.23	3.51	2.44	2.25	0.96	4.24	4.68	1.28	1.38	2.11	8.53	2.48	1.68
GLN						4.94	0.28	0.39	1.41	0.76	0.70	0.36	0.77	1.27	0.33	0.52	0.50	1.68	0.97	0.55
GLU							1.92	0.47	0.64	0.49	0.41	0.32	0.73	0.69	0.20	0.30	0.29	1.30	0.69	0.32
GLY								5.44	1.42	0.78	0.74	0.68	1.17	1.28	0.45	0.75	0.58	1.37	1.41	0.59
HIS									21.93	1.84	1.98	1.16	2.04	3.00	0.73	1.47	1.43	3.33	2.85	1.08
ILE										7.66	1.71	0.50	2.72	2.98	0.47	0.81	0.92	3.18	1.89	1.33
LEU											3.63	0.44	1.72	2.31	0.48	0.73	0.81	2.28	1.63	0.85
LYS												3.01	0.62	0.76	0.26	0.49	0.42	0.96	0.93	0.31
MET													19.97	4.14	0.65	1.10	0.87	6.04	3.51	1.72
PHE														13.84	0.90	1.49	1.47	5.27	3.45	1.77
PRO															2.61	0.30	0.31	0.60	0.63	0.32
SER																4.71	0.79	1.64	0.97	0.63

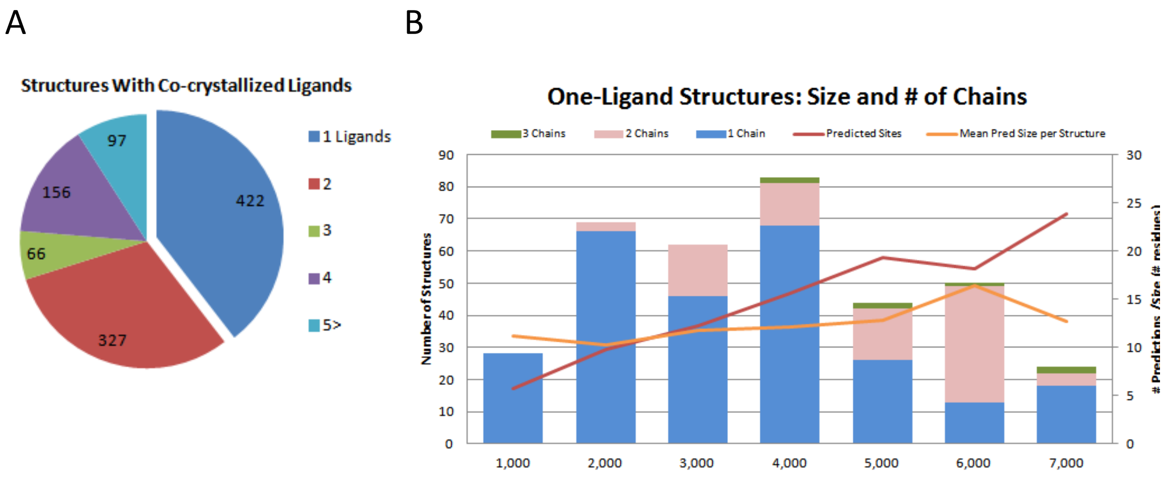
THR	4.28	1.51	0.91	0.46
TRP		43.10	4.35	2.24
TYR			11.39	1.34
VAL				4.13

4.3.2 SiteFinder Prediction Success

The SiteFinder algorithm predicted 29,157 pockets in the 1080 structures from 2003 and earlier. An average of ~27 predictions per protein were found, with an average prediction size of 20 residues (median size 10 residues). Of the 1080 structures, 1072 (99%) contained one or more correct (RO > 0.5) predictions, and 1850 of the known binding sites were matched by at least one prediction. If the rank of the correctly predicted sites is not considered, this represents a 75% overall success rate (1850 /2471); 1317 of these predictions were ranked among the top 3 in their respective protein by SiteFinder - a 53% top-3 success rate overall. The SiteFinder top-ranked prediction success is 67% with the limitation that only one correct prediction per protein is required even if there are multiple known ligand sites a protein. This optimistic rate climbs to 89% if the top-3 ranked predictions are considered.

Over a quarter of the structures in the prediction set had more than 3 bound ligands (Figure IV-5A), so 100% success in top-3 ranked sites is impossible in this dataset. We thus limited our further comparative analysis to the 422 proteins with only one co-crystallized ligand (Figure IV-5B). This limits the assumptions about the correct number of expected true-positive results when calculating a success rates, and it simplifies the interpretation. SiteFinder correctly identified 60.4% of the known ligand sites as a top-ranked prediction (Table IV-1), with an average RO of 0.99 and average MO of 0.30. The success rate in the top-3 ranked sites was 87.9%, picking up an additional 100 correctly-identified predictions. These success rates are lower than those determined for SiteFinder by Schmidtke et al. in their prediction exercise with the pocketome dataset. In that study, 77% and 96% predictions ranked in the top-1 and top-3, respectively. The pocketome set is much larger (> 5000 structures) and includes redundant proteins, which may result in higher success rates. The relative improvement in performance between the top-ranked and the top-3 ranks is comparable to their results.

Figure IV-5: A) Fractions of structures in the full prediction set with respect to the number of co-crystallized ligands. B) Structures with one co-crystallized ligand, binned by heavy atom count and showing proportions of structures with various number of protein chains within bin.



4.3.3 Relative Success of Raw Scores

We compared the default SiteFinder rank score to ranks based on the size of the predicted site, the Protein-Ligand Binding (PLB) Index score of Soga et al. [121], and our propensity-based scores. We also compared to various combinations with the SiteFinder score. As can be seen from Table IV-1, all scores had similar success among their highest ranked sites. The highest success rate for the top-3 ranked sites was achieved by the z-score combination of SiteFinder and Valid propensity scores (91.3%). In fact, all scores performed better than the 87.9% top-3 success rate of the SiteFinder score alone, although the margin of the improvements was slim. Size alone, as measured by the number of residues in the predicted site, performed as well as any of the scoring schemes. It even out-performed the SiteFinder score in both top-ranked and top-3 ranked predictions. Many site-prediction methods like SiteFinder are of course based on the premise that the largest cavity on the protein is the true binding site, and due to the additive nature of the scoring metrics, additive scores have a high correlation with the size of the predicted site. Still, it is important to have a metric that performs better than a simple residue count.

Table IV-1: Success rate of predictions in proteins with one co-crystallized ligand, and success rate for the same proteins if binned by protein size.

	Success for One-ligand Proteins			Success Rates for One-ligand Proteins Binned by Size (# atoms)						
Scores	**Top rank**	**Top 3**		**1K**	**2K**	**3K**	**4K**	**5K**	**6K**	**7K**
SiteFinder Score	60.4%	87.9%	**Top Rank**	78.6%	79.7%	69.4%	71.1%	40.9%	28.0%	50.0%
			Top 3	92.9	91.3	88.7	95.2	81.8	86.0	79.2
Size	59.2	88.8								
PLB Index[135]	62.4	91.1								
Propensity Scores:										
Valid	62.6	90.9	**Top Rank**	82.1	79.7	74.2	73.5	52.3	24.0	54.2
			Top 3	96.4	91.3	91.9	94.0	93.2	90.0	95.8
Invalid	62.0	91.1								
PairValid	63.1	91.1								
PairInvalid	62.0	90.6								
Combination Scores:										
SiteFinder & Valid	62.7	90.9	**Top Rank**	82.1	79.7	74.2	75.9	47.7	24.0	54.2
			Top 3	92.9	92.8	91.9	94.0	93.2	90.0	95.8
SiteFinder & Invalid	62.9	90.6								
SiteFinder & Pair Valid	62.7	91.3								
SiteFinder & Pair Invalid	62.9	91.1								
SiteFinder & log Valid	--	--	**Top Rank**	67.9	69.6	61.3	67.5	65.9	82.0	62.5
			Top 3	96.4	87.0	88.7	90.4	84.1	90.0	75.0
SiteFinder & log Invalid			**Top Rank**	71.4	56.5	51.6	54.2	47.7	80.0	37.5
			Top 3	92.9	81.2	75.8	73.5	68.2	88.0	62.5

We expected scores based on Valid propensities to perform better than those based on Invalid propensities, since all known ligands sites in the prediction set are valid biologically relevant co-crystals, and their sites should be preferentially weighted by a score derived from such ligand sites. However, the Valid score does no better than the Invalid score, and it does not constitute a significant advantage. A Pearson correlation of 1 between the two scores confirms their strong linear relationship (Table IV-2). The PLB index is also highly correlated with our propensity-based scores, and it has almost identical success rates. Any differences in actual values of residue propensities was inconsequential with respect to prediction success, and all three of the above-mentioned scores have Pearson correlations of 1 with predicted binding-site size. Scores based on residue-pair propensities are less correlated with other scores because, like the number of residue pairs, the score grows exponentially with the size of the predicted site. However, the different score distribution does not translate into a difference in prediction success, and the ranks of the pair-wise scores are highly correlated to the single-residue scores and place almost exactly the same number of correctly matched predictions among their top ranks (Table IV-1).Due to the high correlation between the Valid scores and other propensity-based scores, we focus the following results and discussion on the SiteFinder, size, and Valid scoring schemes, addressing the remaining scores as needed when differences are observed.

Table IV-2: Pearson correlations among the various scoring schemes.

Pearson Correlation of Propensities	Size	PLB	Valid	Invalid	PairValid	PairInvalid
SitFinder	0.98	0.98	0.98	0.98	0.80	0.80
Size		1.00	1.00	1.00	0.77	0.78
PLB			1.00	1.00	0.75	0.76
Valid				1.00	0.76	0.76
Invalid					0.76	0.76
PairValid						1.00
PairInvalid						

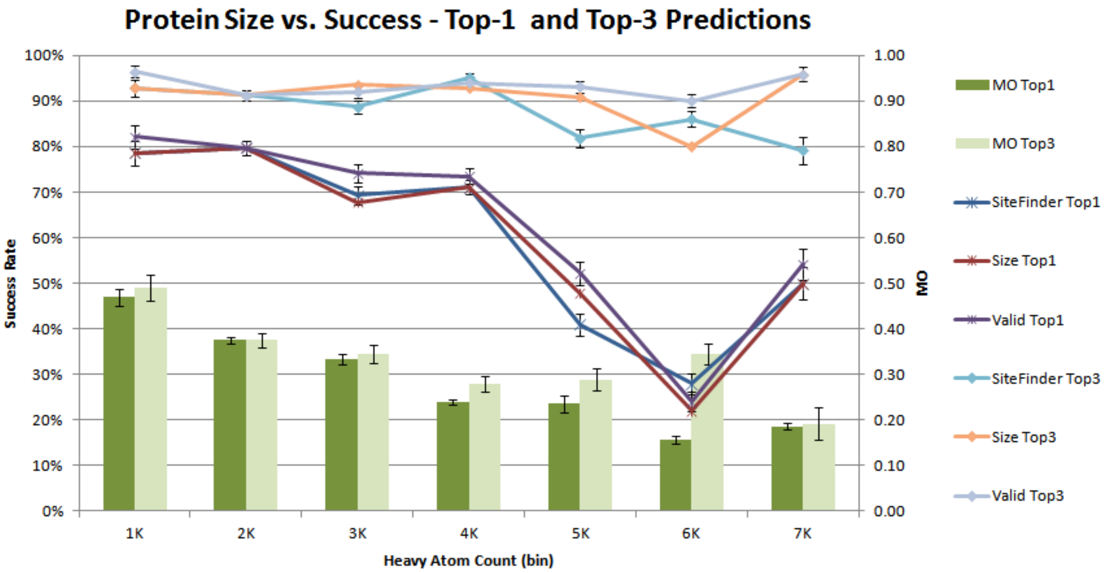
4.3.4 Prediction Success and Protein Size

Size of the predicted binding sites dominates the prediction success in our set of single-site proteins. The matching prediction size largely does not correspond to the actual size of the known binding site, indicating poor precision (Pearson correlation = 0.14). It is only somewhat correlated with the protein size (Pearson correlation = 0.56). SiteFinder generates anywhere from a few to several dozen predicted sites for every protein. Overall, the median size of these predicted sites is consistently ~10±2 residues. However, the number of predicted sites per protein *is* proportional to protein size (Pearson correlation = 0.89), and the *mean* size of these prediction also increases with increasing protein size. Due to the dominating effect of size of the predicted site size on the success rate, all additive scoring schemes will agree on the largest of the predictions as the top-ranked one, when a protein contains a few extremely large predictions followed by a number of much smaller pocket candidates. In an additive score, where each residue is multiplied by its respective propensity, size can be thought of as a score with propensity of 1 for each residue. The dominating effect of size implies that in most cases, the contribution to a propensity-based score by residue propensity values that do not equal 1 is not large enough to counteract the difference in size between the largest predictions and their smaller, but possibly relevant, alternatives. Because predicted pocket size is such a good indicator of a correct prediction for a majority of cases, any attempt to normalize a propensity-based score to remove the size effect would significantly lower success rate. After all, there are a larger number of smaller predicted pockets for a protein than big ones, and most of them will not overlap with a known site.

Protein size is well known for any system where the binding site is unknown. Normally only sequence and structure information are available for such a system. Barring any sequence or structural similarity to known proteins, which can shed a light on the potential protein function, the shape and size of the protein are two factors that may indicate the success of a binding-site prediction algorithm. Protein size is also the one factor that correlates with number of predictions per protein, their average size, and the MO values of the matches (Figure IV-5B). To better understand the cases where protein size might have a varied influence on the success of SiteFinder, we chose to look at a breakdown of our one-ligand protein set by their weight. We binned proteins by size ranging from 1,000 to 8,000 heavy atoms (Figure IV-5B). There were 360 proteins in this range, with the remaining 62 proteins having > 8000 atoms. We only looked at the seven sets of smallest proteins (Figure IV-6), because there were too few proteins per bin to establish accurate success rates for proteins with > 8000 atoms (Figure IV-8). To better appreciate the differences in success rates within and across bins, we sampled the set of 422 proteins 1000 times, each time discarding a random 10% of the data and repeating the binning and success-rate calculations. The error bars in Figure IV-6 indicate the standard deviation of these samples for the SiteFinder and Valid scores. Valid scores improved the success of the top-ranked predictions in the 3000 and 5000 atom bins by 4.8% and 11.4%, respectively. These differences are greater in magnitude than the sum of the standard deviation from sampling. For the top-

3 ranks, Valid scores significantly improved on SiteFinder in the 3000, 5000 and 7000 atom bins by 11.4%, 4%, and 16.7%, respectively.

Figure IV-6: Prediction success of various scores as a function of protein size. Structures with one ligand are binned by number of heavy atoms. Success rates are calculated separately for each bin. Median MO values for prediction matches in each bin are also shown. Error bars show standard deviation for SiteFinder and Valid scores in 1000 rounds of leave-10%-out sampling of the full set of 422 proteins.



A clear trend among top-ranked predictions is the lower success rate for larger proteins. Performance of the SiteFinder score dropped off starting at proteins of 4000 atoms in size, and a sharp dip was observed for all scores in the 6000 atom bin (6000-7000 atoms). Drop-off in performance with size was not evident when the top-3 ranked predictions were considered. The top-3 ranks recapitulate the lack of performance seen in the top-ranked predictions, and we can surmise that SiteFinder generates very large and thus, highly-ranked sites for large proteins. While the largest prediction may not necessarily be correct, the second or third largest likely is. This is consistent with the trends for larger and more numerous SiteFinder predictions with increasing protein size. It may also reflect the multi-meric composition of proteins in the higher size bins, where two or more chains may have reasonable binding sites, but only one contains a ligand bound in the crystal structure (these would be classified as having only one known site in our data set).

It is curious that proteins in the 6000 bin were poor targets for prediction, especially among the top-ranked predicted sites. This poor performance can be attributed to a disproportionate number of antibodies present in this bin. Of the 50 structures in the bin, 31 are antibody variants. This is 70% of all antibodies present in the 422 structures of the one-ligand set. Although 28 of the 31 (41 of the 45 in the full set) antibody proteins in the bin have a matching prediction within the top-3 ranks, only 14 of them are ranked as the top prediction. Antibody complexes are composed of a “light” chain and a “heavy” chain. They have a large central cavity at the junction of these chains. This partially hydrophobic cavity is detected by SiteFinder as the top-ranked pocket in lieu of the antigen-binding sites at the “ends” of the immunoglobulin domains that usually contain the co-crystallized ligand. Although we use a non-redundant protein set to eliminate multiple proteins with similar structures, the antibody family has high sequence variability with relatively similar tertiary structure, and thus it introduces some bias into the dataset.

4.3.5 Relative Success of Consensus Scores

We examined the effectiveness of consensus scoring by adding the z-scores of a propensity-based score to the SiteFinder score for each predicted site and then re-ranking the predictions by the combined z-scores. Due to the previously-mentioned influence of predicted site size on the success rate, combination scores did not perform much better than the individual scores alone (ComboValid in Figure IV-7). Therefore, we examined the combination of the SiteFinder score with a non-additive, propensity-based scoring scheme in which the number of residues is multiplied by the log of the propensity of that residue, instead of the raw propensity value (see Methods). Propensities of < 1.0 will generate negative log values, so the final score will not always be additive with respect to the number of residues in a site. Using the log version of a propensity-based score on its own results in very poor performance because some large sites may have low z-scores in contrast to smaller sites (data not shown). However, in combination with the size-dependent SiteFinder score, the log score may provide complimentary information that reflects the composition of the binding site rather than its size. In Figure IV-7, we see that the combination of SiteFinder score with a log Valid propensity score (ComboLogValid) has high top-rank success rates in proteins over 5000 atoms in size, improving the success rate of SiteFinder alone by 25%, 54%, and 12.5% in the three largest protein bins (Table IV-1 and Figure IV-7). The advantage of the consensus score is not observed in the success rates of the top-3 ranked predictions. This indicates that log-based scores can re-rank a slightly smaller correct pocket from second or third place to top rank, but they are less effective at re-ranking smaller correct predictions that may be present further down in the prediction list. With size of the predicted site having a diminished influence, consensus scores based on Valid propensities significantly outperform those based on Invalid propensity values (Table IV-1). The consensus score of SiteFinder and the log of Invalid propensities has top-rank success rates 10% lower on average than those of the Valid log-based consensus score. Consensus scores calculated for log values of PairValid and PairInvalid propensities still retain some size dependence, and they perform with intermediate success between that of SiteFinder alone and consensus scores based on log values of single-residue propensities. Unfortunately, using any consensus score also reduces the success rate of the SiteFinder score for structures smaller than 5000 atoms in size, both in the top-ranked, and the top-3 ranked predictions.

Figure IV-7: Prediction success using consensus scores, reported as a function of protein size. Structures with one ligand are binned by number of heavy atoms. Success rates are calculated separately for each bin. Error bars show standard deviation for SiteFinder scores in 1000 rounds of leave-10%-out sampling of the full set of 422 proteins.

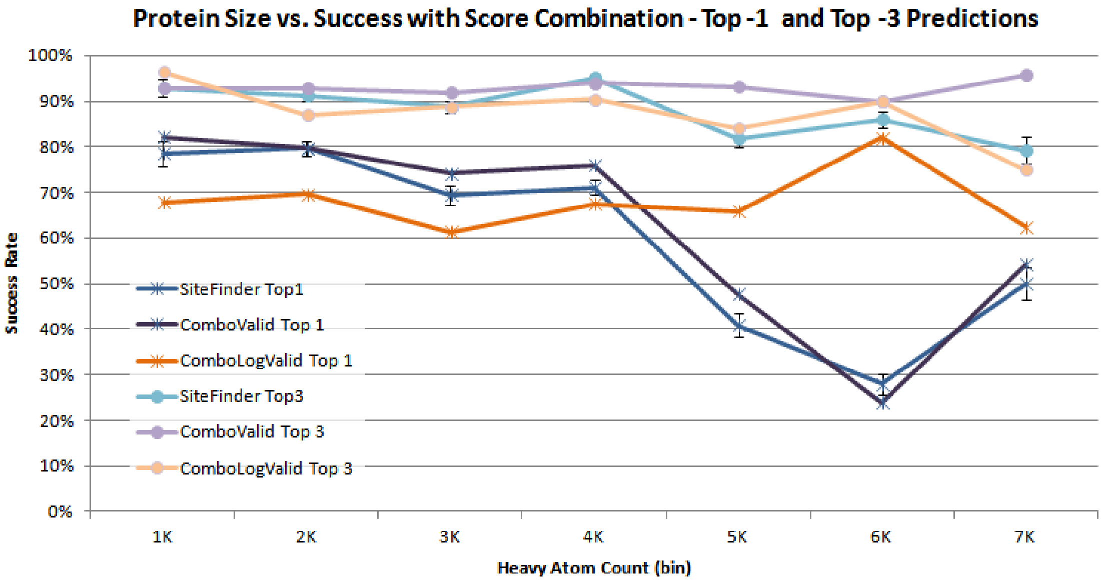
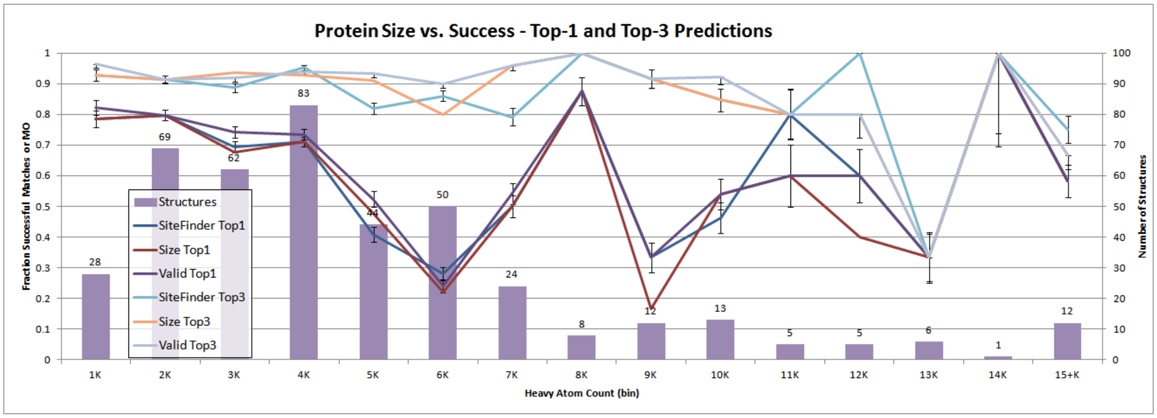


Figure IV-8: Prediction success, reported as a function of protein size. Structures with one ligand are binned by number of heavy atoms. Success rates are calculated separately for each bin. Number of structures in each bin is also shown. Error bars show standard deviation for SiteFinder and Valid scores in 1000 rounds of leave-10%-out sampling of the full set of 422 proteins.



4.4 Conclusion

We tested the success of the SiteFinder algorithm to predict binding sites over a set 1080 diverse proteins, and a subset of 422 proteins with only one ligand-binding site. The default SiteFinder score were compared to size-based scores based on residue propensities derived from an independent training set of thousands of diverse protein-ligand complexes. Propensity-based scores were found to rank binding-site predictions as well as, or better than, the SiteFinder score. However, we found that the size of the predicted binding sites alone can account for most of the successful matches, and the success of any additive rank score that scales with prediction size will be driven by the size component. This accounts for high success rates of propensity based scores presented in this study and previous publications [121] [135]. We found that the top-rank success rate of SiteFinder predictions drops off as the size of the protein increases, and certain protein classes, such as antibody proteins, are especially challenging for this prediction method. The propensity scores perform better in placing correctly predicted sites at the top for proteins over 5000 atoms in size, and a consensus score combining the additive SiteFinder score and a non-additive propensity score can rescue poor success rates. While the confirmation of overall success of SiteFinder on a large, diverse set of proteins is encouraging, the dominance of predicted site size in successfully ranking correct predictions indicates room for improvement in scoring the potential pockets generated by geometry-based binding site prediction methods.

CHAPTER V

Can Scoring Functions Distinguish Biologically Relevant Binding from Irrelevant, Opportunistic Binding?

5.1 Introduction

Databases of protein-ligand structures are important tools for studying principles of molecular recognition. Our resource, Binding MOAD, has helped yield new insight into protein-ligand binding, ranging from fundamental differences between enzymes and non-enzymes[22] to residue propensities in binding sites. Binding MOAD is valuable not only for generating new knowledge, but also as a yardstick for testing current understanding and hypotheses. It can serve as gold standard collection of known protein-ligand structures for evaluating and improving algorithms used in molecular docking. Here, we demonstrate its use in creating a new test set to evaluate scoring functions, one that poses a new question, "can scoring functions distinguish biologically relevant binding from irrelevant, opportunistic binding across diverse proteins?"

Molecular docking has become an increasingly important computational tool in modern structure-based drug design (see refs [138], [139], [140], [141], [142], [143] for review). Given the three-dimensional structure of a protein, the process starts with sampling possible ligand orientations and conformations (referred to as modes) at the selected site of the protein target and then ranks these modes according to their scores calculated with a scoring function. The development of accurate scoring functions to evaluate putative modes is a critical and challenging element in molecular docking. For years, different scoring functions have been developed that boast different computational speeds and accuracy. Roughly, these scoring functions can be grouped into three categories according to their derivation: force-field based, empirical, and knowledge based.

Force-field-based scoring functions are fully or partially evaluated on a set of force-field parameters derived from both experimental work and quantum mechanical calculations to describe the interactions among atoms.[144] [98] When considering the multitude of ways that explicit water molecules can compliment binding (see ref [145] and references therein), conformational sampling and force-field based scoring functions are computationally too expensive to be used in virtual database screening. As an alternative, the solvent effect can be implicitly considered by the Poisson-Boltzmann model (e.g., [146], [147], [148],[149]) or generalized-Born model (see [150] for review) in post-docking scoring (e.g.,[151], [152], [153]). The most simplified method for modeling the solvent effect is to use a distance-dependent dielectric constant to calculate the electrostatic interaction energy term [154], which can be directly used to speed the docking process at the expense of accuracy.

A second category is empirical scoring functions whose parameters are derived by reproducing the binding affinities of a training set of protein-ligand complexes with known three-dimensional structures (e.g., [155], [156], [157], [158], [159]). Compared to force-field-based scoring functions, empirical scoring functions score protein-ligand complexes quicker because of their relatively simple energy terms. The generality of an empirical scoring function is typically restricted by the composition of its training set.

The third kind of scoring functions are the knowledge-based scoring functions [160], [161], in which an inverse Boltzmann relationship is used to determine pair-wise energy potentials, directly converted from the occurrence frequencies between atom pairs in a database of protein-ligand structures[25, 162-168]. The derived pair potentials try to embody all the effects that govern ligand binding such as electrostatic interactions, van der Waals interactions, hydrophobic effect, desolvation penalties, etc. Knowledge-based scoring functions have a good balance between accuracy and speed. Compared to empirical scoring functions, knowledge-base scoring functions can be more general as a result of larger and more diverse training sets of protein-ligand structures available from the Protein Data Bank (PDB) because any structure can be used even if binding affinity data is unknown [2]. The pair-potential feature of the knowledge-based scoring functions also makes the scoring process as fast as the empirical scoring functions.

Currently, there are three common criteria that are used to evaluate a scoring function.[167] The first criterion is binding-mode prediction, how closely a predicted ligand-binding mode resembles the experimental structure. The second criterion is binding-affinity prediction, whether or not the scoring function can rank order compounds by affinity or reproduce the experimentally determined binding data. The third criterion is enrichment in virtual database screening, whether or not the true inhibitors/binders can be ranked at the top of a large database of ligands according to their binding scores for a protein target. Most current scoring functions perform satisfactorily in one or two criteria [102]; however, it is challenging for a scoring function to perform well in all three. [167]

One common feature for the three above criteria is that they are designed to evaluate a scoring function on a single protein-ligand complex or a specific binding site, target without considering the biological types of the bound ligand. With the rapid development of proteomics projects, more and more protein-ligand structures are being determined experimentally and deposited in the PDB. It is noticeable that many bound ligands in the PDB are biologically irrelevant; typical examples include additive molecules such as detergents for crystallization purposes or buffer molecules. The presence of these molecules bound to protein surfaces usually results from their high concentrations rather than from tight binding interactions (a case referred to as "opportunistic binders" or "invalid ligands") [106, 108]. Whether a scoring function is able to discern invalid ligands from weakly-bound, biologically relevant ligands is a new criterion proposed in the present work. It is desirable to extend scoring functions to evaluate protein binding sites for the determination of function or druggability of a pocket. This goal requires scoring functions to be able to discern biologically relevant binding events from opportunistic ones over a wide range of proteins. This can be particularly challenging if the biologically relevant binding is weak. An important counter issue is "appropriate failures" when additives in a *valid binding* site should score well if they are chemically similar to the biologically relevant ligand.

In this work, a diverse benchmark of valid and invalid protein-ligand complexes from the PDB is presented. Four different scoring functions, representing different categories, are used to test this new benchmark. The influence of including entropic penalties for rotatable bonds in ligands was also examined.

5.2 Methods

We selected four scoring functions to test our new benchmark. They include the empirical scoring function X-Score [100], a force field-based scoring function in DOCK 4.0 [154] [95], a semi-empirical force-field-based scoring function in AutoDock 4.0 [96, 169], and the knowledge-based scoring function ITScore in MDock [167] [170].

X-Score

The empirical scoring function X-Score includes three individual scoring functions of HSScore, HPScore, and HMScore [100]. The van der Waals energy term (VDW) is calculated by a Lennard-Jones 8-4 potential, the hydrogen-bonding term (HB) is obtained from the hydrogen bonds between protein and ligand, and the rotor term (RT) stands for the number of effective rotatable bonds in the ligand molecule. The HS, HP, and HM terms calculate the buried, hydrophobic molecular surface of the ligand, the pair-wise hydrophobic atom-contact potential, and the microscopic match of hydrophobic ligand atoms to the binding pocket, respectively. The coefficients in the scoring functions were obtained by fitting the binding affinities of 200 protein-ligand complexes with known structures [100]. In the present study, we used the average of the scores from the three scoring functions in Eq. (1) to represent the X-Score of a protein-ligand complex.

$$\begin{aligned} \mathrm{HSScore} &= C_{\mathrm{VDW},1}\cdot \mathrm{VDW}+C_{\mathrm{H\text{-}bond},1}\cdot \mathrm{HB}+C_{\mathrm{hydrophobic},1}\cdot \mathrm{HS}+C_{\mathrm{rotor},1}\cdot N_{\mathrm{tor}}+C_{0,1} \\ \mathrm{HPScore} &= C_{\mathrm{VDW},2}\cdot \mathrm{VDW}+C_{\mathrm{H\text{-}bond},2}\cdot \mathrm{HB}+C_{\mathrm{hydrophobic},2}\cdot \mathrm{HP}+C_{\mathrm{rotor},2}\cdot N_{\mathrm{tor}}+C_{0,2} \qquad (1) \\ \mathrm{HMScore} &= C_{\mathrm{VDW},3}\cdot \mathrm{VDW}+C_{\mathrm{H\text{-}bond},3}\cdot \mathrm{HB}+C_{\mathrm{hydrophobic},3}\cdot \mathrm{HM}+C_{\mathrm{rotor},3}\cdot N_{\mathrm{tor}}+C_{0,3} \end{aligned}$$

AutoDock

The scoring function in AutoDock 4.0 is a semi-empirical, force-field based scoring function which includes five energy terms [96, 169]

$$\begin{aligned}\Delta G = {} & W_{vdw} \cdot \sum_{i,j} \left(\frac{A_{i,j}}{r_{ij}^{12}} - \frac{B_{i,j}}{r_{ij}^{6}} \right) \\ & + W_{elec} \cdot \sum_{i,j} \left(\frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \right) \\ & + W_{hbond} \cdot \sum_{i,j} E(t) \left(\frac{C_{i,j}}{r_{ij}^{12}} - \frac{D_{i,j}}{r_{ij}^{6}} \right) \\ & + W_{tor} \cdot N_{tor} \\ & + W_{sol} \cdot \sum_{i,j} \left(S_i V_j + S_j V_i \right) e^{\left(-r_{ij}^2 / 2\sigma^2 \right)} \end{aligned} \tag{2}$$

where the first two energy terms are classic VDW and electrostatic interactions and have the same forms as the force-field scoring function in DOCK 4.0. [95, 154] The third term stands for the contribution from hydrogen bonds between protein and ligand. The fourth term considers the loss of torsional entropy of a ligand upon binding in which N_{tor} is the number of rotatable bonds in the molecule. The last term describes the solvation effect. The weighting coefficients for the five energy terms were obtained by fitting the known binding constants of 188 protein-ligand complexes [169].

DOCK

The scoring function in DOCK 4.0 [95] represents a typical force-field-based scoring function whose energy parameters are taken from the Amber force field [154]. This scoring function includes VDW and electrostatic interaction energy components

$$E = \sum_i \sum_j \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \right) \tag{3}$$

where r_{ij} stands for the distance of protein atom i and ligand atom j, A_{ij} and B_{ij} are the VDW parameters, and q_i and q_j are the atomic charges. The effect of solvent is implicitly considered by using a distance-dependent dielectric constant $\varepsilon(r_{ij})$.

ITScore

ITScore is an iterative knowledge-based scoring function developed using a training database of 781 protein-ligand complexes structures from the PDB [167], representing a set of effective pair potentials resulting from the overall effects of all binding factors.

The binding score is calculated by summing all the atomic pairs between protein atom i and ligand atom j as

$$E_{ITScore} = \sum_{i,j} u_{ij}(r) \tag{4}$$

where r is the distance between the atom pair i and j. The effective pair potentials $u_{ij}(r)$ are iteratively derived until they can discriminate the native structures from decoys for 99% of the protein-ligand complexes in the training set. The ITScore scoring function has been implemented in MDock, a program for docking against an ensemble of protein structures. [170]

5.2.1 Scoring and Adding Torsional Entropic Penalties

X-Score and AutoDock contain terms that penalize a score for each rotatable bond in a ligand on the basis that restricting each torsion carries an entropic cost. Similar terms are not included in DOCK or ITScore. The additive nature of both DOCK and ITScore inherently bias large ligands to score well (a well-known limitation of many scoring methods). This caveat can be particularly problematic in our study because several additives in the invalid set are large detergents.

To investigate the effect of incorporating penalties for torsional entropy of the ligands, we calculated two set of binding scores with and without a torsional ligand terms. To remove the torsional term from X-Score and AutoDock, we simply set the coefficients C_{rotor} and W_{tor} to zero in Eqs. (1) and (2). These are referred to as X-Score-tor and AutoDock-tor, respectively. To add a torsional term to DOCK and ITScore in a straight forward way, we added w_{tor} and N_{tor} to Eqs. (3) and (4) where N_{tor} is calculated by X-Score (its count of rotatable bound count) and w_{tor} (a scaling factor for the torsional penalty) was simply set to 1 for this work. These are referred to as DOCK+tor and ITScore+tor. This was chosen for simplicity. Furthermore, we did not wish to unfairly bias the performance of DOCK and ITScore by explicitly fitting new parameters for this purpose.

5.2.2 Receiver Operating Characteristic Curves

The performance of the scoring functions was evaluated by comparing the rank ordering of real ligands (true positive valids) versus invalids (false positives) with the receiver operator characteristic (ROC) curves. A perfect scoring method would rank order all valids before invalids, achieving a curve that starts at the origin (0,0), goes straight up to the left-hand corner of the ROC plot (1,0), and then across (1,1). A scoring method with no predictive power would equally rank valids and invalids, achieving a line with slope = 1 starting at the origin (0,0) and connecting to (1,1). Area under the curve (AUC) provides a quantitative measure for comparison for ROC curves. A perfect scoring function would have a ROC curve with an AUC of 1.0, while the poorer scoring function described above would score an AUC of 0.5, no better than random assignment. The ROCS package in R was used for calculation of the ROC curves.

5.2.3 The Test Set: Valids and Invalids.

The test set contains crystal structures of protein-ligand that are classified as valid or invalid based on biological relevance of the ligand. The subset of 177 valid hits contains complexes where the ligand is a biologically relevant molecule such as a natural substrate or a known inhibitor (Table V-1). The subset of 71 invalids contains complexes where the ligand is an extraneous molecule such as a buffer, detergent, or solvent (Table V-2). It is important to note that the invalids in this study are not computationally derived binding modes like those often used to evaluate docking and scoring; rather the invalids are obtained from existing crystal structures that represent non-functional binding events. Both types of complexes are observed protein-ligand binding events, but the ligands in the structures chosen as invalids are instances of opportunistic binding, often occurring as a result of over-abundance of the ligand in the crystallization medium. All structural data for the valid set was obtained from the Binding MOAD biounit files. Structures for the invalid complexes not available in Binding MOAD (due to absence of valid ligands) were obtained from the PDB.

Table V-1: Set of valid complexes.

PDB	HET	Affinity	PDB	HET	Affinity	PDB	HET	Affinity
1A5B	IGP		1OIT	HDT	IC50 < 0.0030 uM	2BU2	TF1	
1A6V	NPC		1OUW	MLT		2BVE	PH5	IC50=215 uM
1AI4	HAA	ki= 3.3 mM	1P0B	PQ0		2C25	SIA	
1AJN	AAN	ki=2.32 mM	1P0M	CHT		2CA8	GSW	kd=22 uM
1B42	M1A	kd=97.9 uM	1P28	HBS	kd=3.8 uM	2CAQ	GSW	kd=285 uM
1B74	DGN	Ki=50 mM	1P6B	EBP		2CBU	CTS	kd=2.1 uM
1BEU	IPL		1PA9	CSN	ki=25 uM	2CCG	TMP	
1BGG	GCO		1PZP	FTA	ki=480 uM	2CER	PGI	ki=0.6 nM
1BR6	PT1	ki = 0.6 mM	1Q8U	H52	kd = 1.1 uM	2D5C	SKM	
1C7R	PA5		1QA0	270		2DDQ	HRB	

1C9C	PP3		1QIZ	RCO		2E7F	C2F	kd=0.79 uM
1CA7	ENO		1R4S	MUA		2ENB	PTP	
1CEA	ACA	kd~11 uM	1RDY	F6P		2EXM	ZIP	
1CEB	AMH	ki~1 uM	1RGK	2AM	kd=49 uM	2FA1	BDF	
1D1Q	4NP		1ROB	C2P		2FZG	EOB	IC50=86 uM
1D8C	SOR		1SD3	SYM		2G5F	PYR	
1DL7	NCH	kd=0.32 uM	1SUX	BTS		2G7Q	AHL	
1DPM	EBP		1TH6	OIN		2GA4	1PS	
1E1X	NW1	ki=1.3 uM	1TVP	CBI		2GN2	C5P	
1E3V	DXC	kd = 45.74 uM	1U0G	E4P		2GNH	H52	ki=149 nM
1E4N	HBO		1U1W	3HA	kd=1.4 uM	2GZ8	F3F	IC50=3 uM
1ENU	APZ	ki=8.3 mM	1U2O	NEC		2HAI	PFI	IC50=0.53 uM
1FDS	EST		1UF7	CDV		2HDQ	C21	ki=40 mM
1G86	NEQ		1UTM	PEA	kd=0.971 uM	2HDR	4A3	Ki=19 mM
1GII	1PU		1UTO	PEA	kd=5 mM	2HDU	F12	ki=5 mM
1GW9	LXC		1UUX	PPI		2HHA	3TP	IC50=0.122uM
1H1H	A2P	ki~6.5 uM	1UWC	FER		2HUI	GLV	
1H7F	C5P		1UYQ	NFG		2I5X	UA5	
1HG2	IP2		1UZU	INR	ki=13.8 uM	2I6P	4NP	
1IEX	TCB		1V2G	OCA		2IUQ	TSS	
1IG0	VIB		1V48	HA1	ki=16 nM	2J75	NOY	
1IS4	LAT		1VJ5	CIU	IC50=0.12 uM	2J7B	NTZ	ki=174 nM
1JGM	PEL		1VYQ	DUX	ki=4.98 uM	2J7C	IDE	ki=10.7 nM
1JVU	C2P		1VYZ	N5B	IC50=290 nM	2J7D	GI1	ki=160 nM
1K1P	MEL		1W3O	PYR		2J7G	GI4	ki=136 nM
1KCC	GTR		1W4Q	UMF	ki=5.89 uM	2O9R	TCB	ki=21 mM
1L8B	MGP	kd=0.14 uM	1W61	PYC		2OU0	MR3	kd=592 uM
1LO0	BC1		1W6F	ISZ		2OVD	DAO	
1LP6	C5P		1W8M	E1P	ki=25 mM	2OVW	CBI	
1LPD	ADE		1W8O	LBT		2OWZ	F6P	
1LT6	GAA		1WMA	AB3	IC50=788 nM	2OYM	MNI	ki=10 uM
1MEN	GAR		1X2B	STX		2PBW	DOQ	ki=5.56 nM
1MFI	FHC	ki=2.6 uM	1XFG	HGA		2PLK	P3D	
1MMY	G6D		1XUA	HHA		2Q89	6CS	kd=0.5 uM
1MOO	4MZ		1Y20	1AC	ki=4.8 uM	2R5N	RP5	kd=700 uM
1MRD	IDP		1YC4	43P	kd=0.28 uM	2UW3	GVG	ic50=80 uM
1MSA	MMA		1YL1	ETF		2Z26	DOR	
1ND0	DP4		1YQC	GLV		3FIT	FRU	
1NHK	CMP		1ZC9	PMP	kd=0.6 mM	3KIV	ACA	kd=20 uM
1NJR	XYL		1ZI3	NLC		3LKF	POC	
1NO6	794		1ZO8	SNO		3MCT	3MC	kd=85.5 uM
1NXJ	GLV		1ZR8	AJM		3PAX	3MB	IC50=10 uM
1O0O	A2P	ki=8 uM	1ZWP	NIM		3SLI	SKD	
1O4N	OXD	IC50 > 40 mM	2ACK	EDR		43CA	NPO	kd<1 uM
1O71	POC		2AJV	COC		4ERK	OLO	ic50=27 uM
1O9O	MLM		2B56	U5P		4RHN	RIB	
1OBA	CHT	kd = 3.6 mM	2BKL	ZAH		4SLI	CNP	
1OFZ	AFL		2BKV	PGA		5EUG	URA	
1OI6	TMP		2BS7	CBS		5GPB	GPM	

Table V-2: Set of invalid complexes

PDB	HET	PDB	HET	PDB	HET
1APM	OCT	1S2U	PEG	2G4Y	TAR
1D2M	SO4	1SHV	MA4	2GFC	OCT
1D5R	TLA	1T7L	MRY	2GON	FLC
1D6F	B3P	1T7V	P6G	2GW5	IPA
1D6J	TLA	1TAQ	BGL	2HI9	GOL
1EM2	TAR	1TP7	DMX	2I3A	BTB
1FZV	MPD	1TTO	TRS	2I5P	GLC
1H53	PO4	1U3A	PE5	2INU	2PO
1HH8	FLC	1URM	BEZ	2J8X	URE
1HVV	TAR	1XEZ	BOG	2NR9	PA6
1IZ2	SUM	1YBK	BEQ	2NUD	TRS
1JJ0	SUC	1YXS	IMD	2O02	BEZ
1JLU	OCT	1ZDY	T3A	2OQA	PG4
1KWN	TAR	1ZGN	MES	2OR7	ACT
1LIH	PHN	1ZR3	MES	2P4B	BOG
1M27	FLC	2A0Q	NDG	2PGB	NAG
1MRZ	CIT	2APV	MLA	2Q9H	ACY
1N2F	DTT	2B4P	MLI	2RA5	SRT
1OLL	EDO	2BEX	GOL	2SHP	CAT
1PK3	BME	2C56	SUC	2Z9J	DMS
1PPA	ANL	2CJP	PG4	8CHO	P4C
1Q61	MG8	2E50	TRE		
1QST	EPE	2FAF	EPE		
1RJM	EP1	2FUF	FLC		
1RTV	SRT	2G47	DIO		

The valid set was chosen from Binding MOAD, a database of high-quality, protein-ligand structures with annotated ligands [106]. Structures were filtered to have the following criteria: contain as few ligands as possible (preferably only the valid ligand of interest), have no other ligands or co-factor molecules anywhere within 12Å of the ligand of interest, and be crystallized in the biologically relevant pH range of 6-8. All structures in Binding MOAD have at least 2.5Å resolution and no covalently attached ligands.

The invalid set was chosen by filtering the PDB for protein-ligand structures that met the same structural quality criteria as the valids, but *did not* meet the validity criteria as defined in Binding MOAD [106]. Ions, nucleic acids, and multi-part ligands were not considered as candidate ligands in either set. The two sets were further refined by clustering based on chemical similarity and then selecting pairs of similar ligands that were in different sets (valid vs. invalid). Preferentially, only one such pair from every cluster was chosen in order to maintain variety of ligands in the test set. Clustering was performed using the QuaSAR function in MOE, based on physicochemical descriptors – solubility (logS), hydrophobicity (SlogP), weight, and buried surface area (BSA). Solubility, hydrophobicity, weight and primary components (PCA) were calculated with MOE [56]; buried surface area was calculated with NACCESS [91]. Affinity data are available for 38% of the valid hits.

In order to ensure a fair distribution of valids and invalids across chemical space, the final sets were reviewed for even distribution of logS, SlogP, weight, and BSA values. The histograms in Figure V-1 through Figure V-4 compare the valid and invalid ligands based on these criteria. Since BSA is proportional to the number of protein-ligand contacts, and ultimately to the docking score, it was of great importance to select complexes with similar degrees of ligand burial in both valid and invalid subsets. Figure V-1 shows that the distribution of BSA between the two sets is slightly un-even, with valid ligands on average being more buried. Still ~50% of the invalids have BSA above 200 Å where ~75% of the valid hits lie. Also, both valid and invalid examples can be found throughout the BSA range. An exception is a highly buried, invalid polyethylene glycol molecule (1U3A) which was selected because it clustered with a valid ligand based on other descriptors. Such trade-offs were essential to maintain comparable distributions of the various descriptors and minimize any selection bias inherent in construction of such test sets. The ROC curve of the various descriptors in Figure V-5 shows the relative classification power of the various descriptors. BSA, weight, and the first principal component (PCA1) are the most discriminating. Each has an AUC around 0.65. While higher than the unbiased AUC = 0.5 that we had aimed for, the set is much more evenly distributes than the full set of valid hits and invalids in MOAD (data not shown).

Figure V-1: Buried Surface Area (BSA) distribution for valids and invalids.

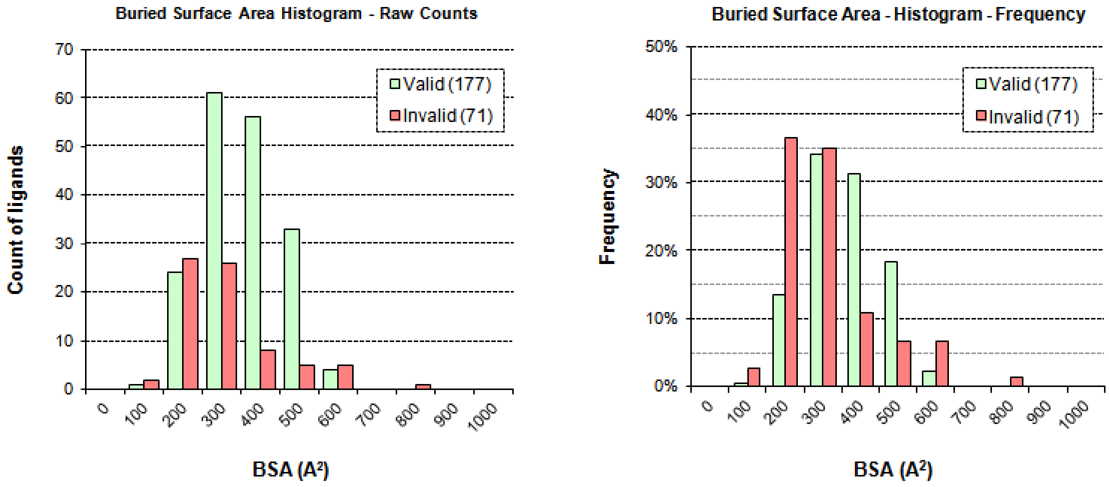


Figure V-2: Size distribution for valids and invalids

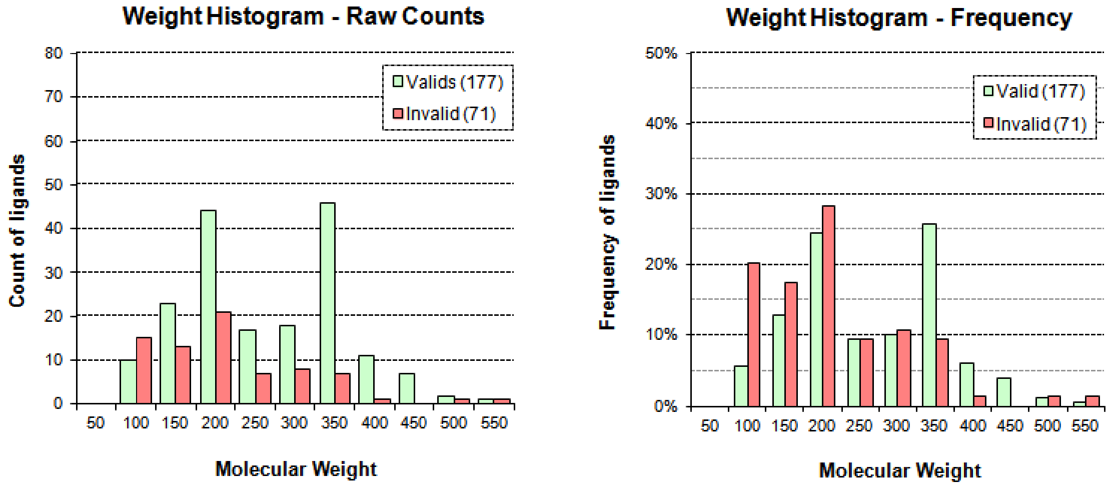


Figure V-3: Solubility (logS) distribution for valids and invalids

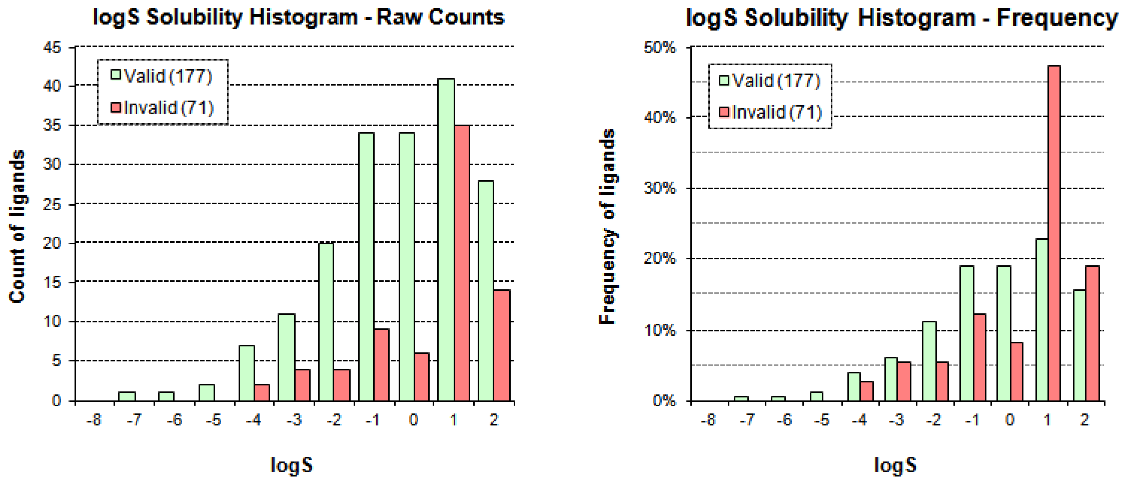
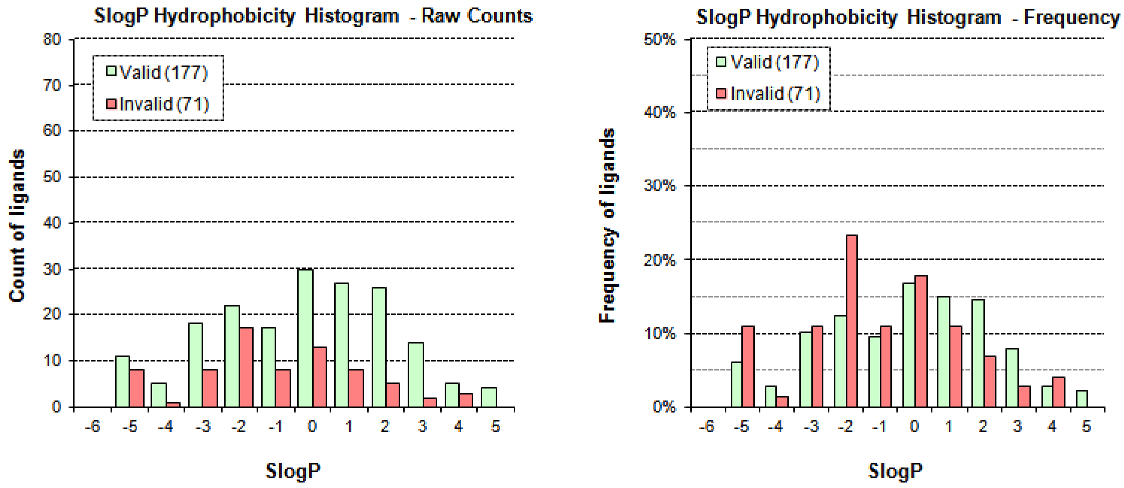


Figure V-4: Hydrophobicity (SlogP) distribution for valids and invalids.



5.2.4 Preparing and Scoring the Complexes

The scoring of all the complexes was done with ligands in their original crystallographic coordinates to avoid any bias from the various docking routines, which makes all differences in the ROC plots come solely from the scoring of the exact same poses. The complexes were prepared using Chimera software from UCSF [171]by removing water molecules and metal ions from the structures. Hydrogens and charges were added to both the protein and ligand of the complex with the former assigned

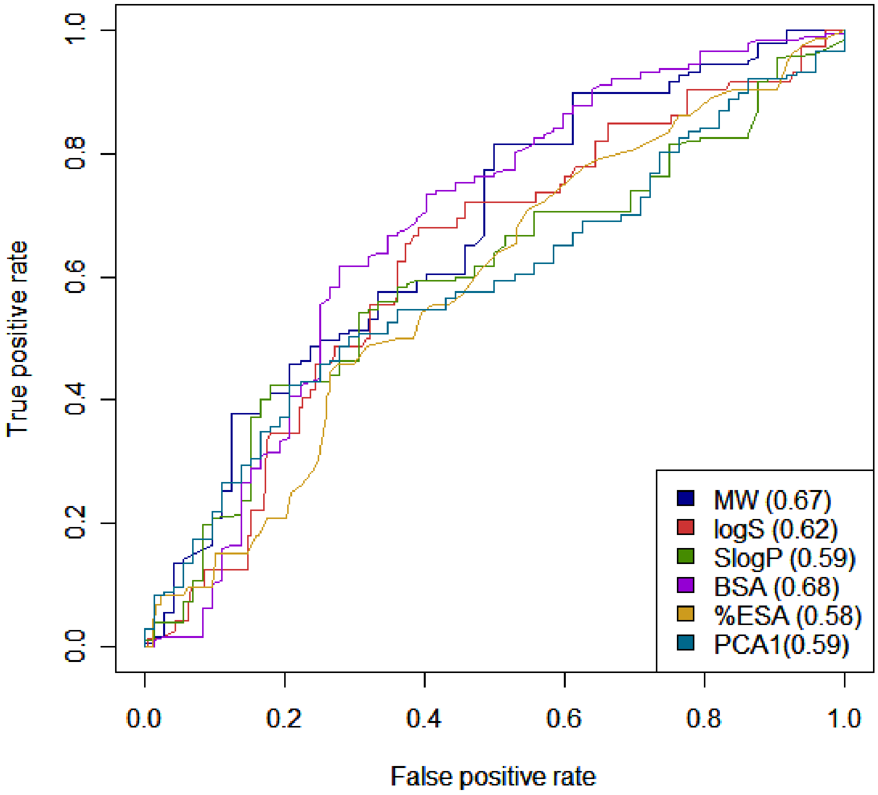
Amber charges and the latter Gasteiger [172]. Once prepared, the binding energy score for each complex was calculated using X-Score, DOCK 4.0, MDock and AutoDock 4.0.

It is important to note that the scoring with X-Score, DOCK, and ITScore were performed "blindly" as part of collaboration with the Zou lab. The Carlson group provided the Zou lab with the full list of 248 complexes, without noting which were valids and which were invalids. The Zou group scored the full list with and without torsional entropy penalties for the ligands (X-Score, DOCK+tor, ITScore+tor, X-Score-tor, DOCK, ITscore). Classification and the resulting ROC plots were calculated by the Carlson lab to reveal the performance. Scoring with AutoDock (and AutoDock-tor) was performed by the Carlson lab.

5.3 Results and Discussion

The test set is composed of 177 valid valids and 71 invalids. Although there is no ligand that is present in both sets, the molecules in the two sets are nevertheless very similar chemically. As Figure V-5 shows, the sets were chosen so that the two classes could not be easily distinguished by their molecular weight, logS, SlogP, BSA, or the PCA1 of these properties. Weight and BSA were top classifiers with AUCs of 0.67 and 0.68 respectively. Bias between weight and affinity is known, but these AUCs are still significantly lower than the AUC of the scoring functions (0.73-0.84 in Figure V-6 - Figure V-9). Most of the biologically relevant ligands in protein-ligand crystal structures are well buried [173], and tight-binding ligands have, on average, more BSA than weakly bound ligands.[174] Thus, similar BSA between the two sets is especially important as it is directly proportional to the number of contacts that contribute to the scores. Even despite our effort to achieve similar BSA distributions across the two sets, this descriptor remains the strongest naïve discriminator between the valids and invalids.

Figure V-5: ROC plot showing performance of physicochemical descriptors as classifiers: Molecular Weight (MW), logS, SlogP, Buried Surface Area (BSA), percent Exposed Surface Area (%ESA), and the principal component (PCA1). Areas under the curve (AUC) are noted in parenthesis in the legend.



5.3.1 Analysis of Scoring Functions

We were delighted that X-Score, AutoDock, DOCK, and ITScore all performed well, preferentially distinguishing valids over invalids and at a higher rate than MW or BSA (Table V-3 and Figure V-10). This is very promising for extending current scoring functions to new uses in structural proteomics like predicting druggability or function of a protein. ITScore with a torsional entropy penalty included, performed best, especially in the all-important, low-false-positive segment at the beginning of the curve. BSA significantly underperforms for more than half of the valid ligands (<0.5 true positives), its presence in the ROC plot of Figure V-10 emphasizes that all the scoring functions are out-performing a mere count of contacts in evaluating the ligands. The performance of the scoring functions is even more encouraging if one considers that they are optimized to provide a measure of protein-ligand affinity, whereas our valids and invalids were chosen on the basis of biological validity. Not all of the valid ligands are tight binders. An underlying assumption of this study is that the ligands classified in Binding MOAD as invalid are weaker binders. Though this assumption is quite reasonable, a direct comparison between experimental affinity values and the calculated scores for valid versus invalid ligands is currently impossible because affinities are typically not measured for these chance binding events. Therefore, it is encouraging to see that a measure of affinity can differentiate well between valid- and invalid-ligand binding.

Table V-3: Number of invalids in the top ranked results of the scoring functions.

	AUC	*Top 10% ranked*	*Top 25% ranked*	Top 50% ranked
AutoDock (w/tor)	0.78	0 out of 25	2 out of 62	11 out of 125
DOCK (+tor)	0.79	0 out of 25	3 out of 62	6 out of 125
ITScore (+tor)	0.84	0 out of 25	0 out of 62	6 out of 125
X-Score (w/tor)	0.81	0 out of 25	3 out of 62	13 out of 125
Cumulative Rank		0 out of 25	0 out of 62	10 out of 125

Figure V-6: ROC plot of ITScore performance

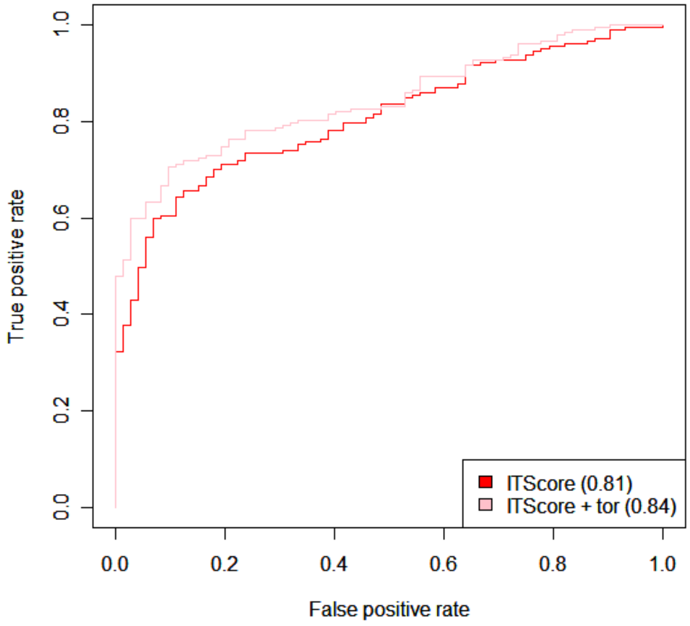


Figure V-7: ROC plot of DOCK4 performance

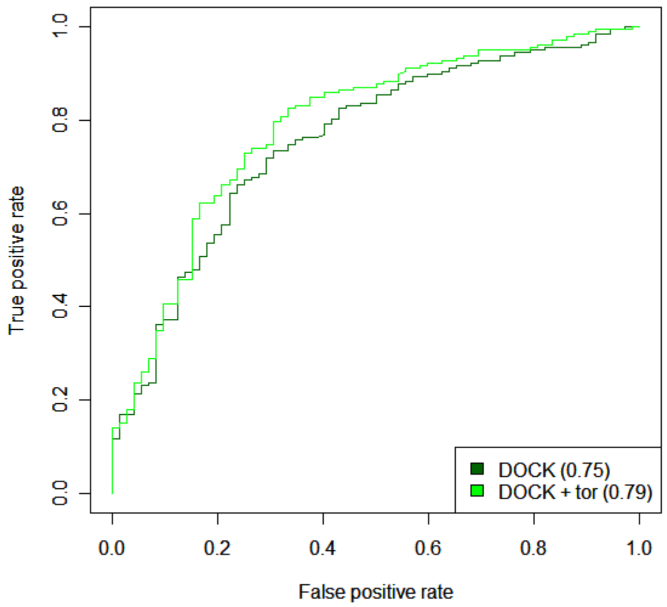


Figure V-8 – ROC plot of X-Score performance

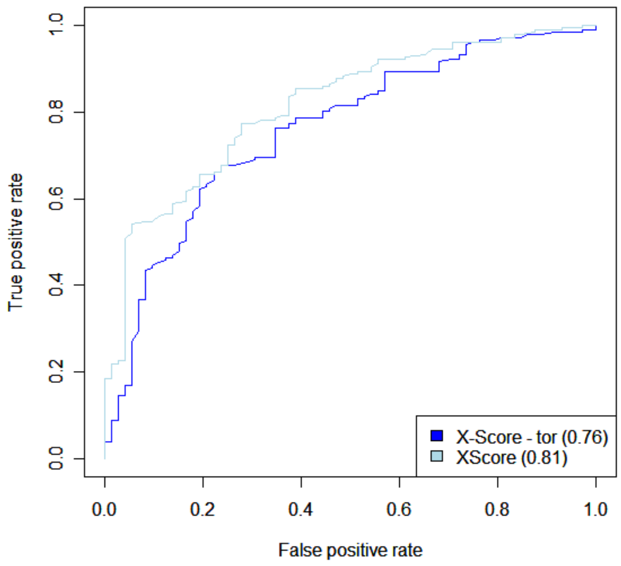


Figure V-9 - ROC plot of AutoDock4 performance

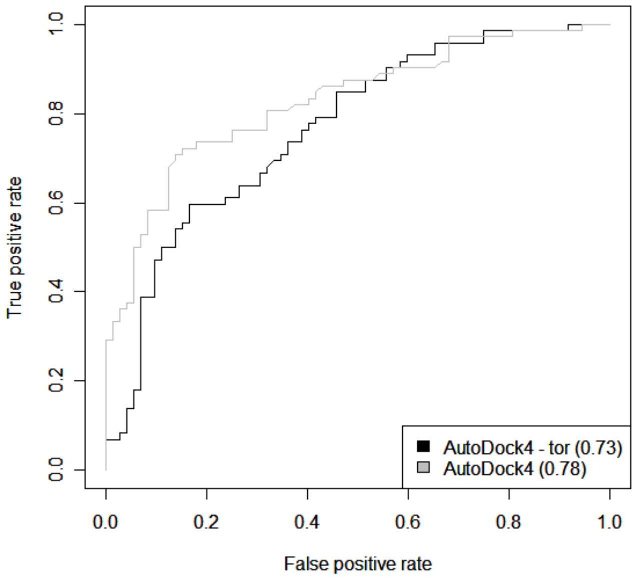
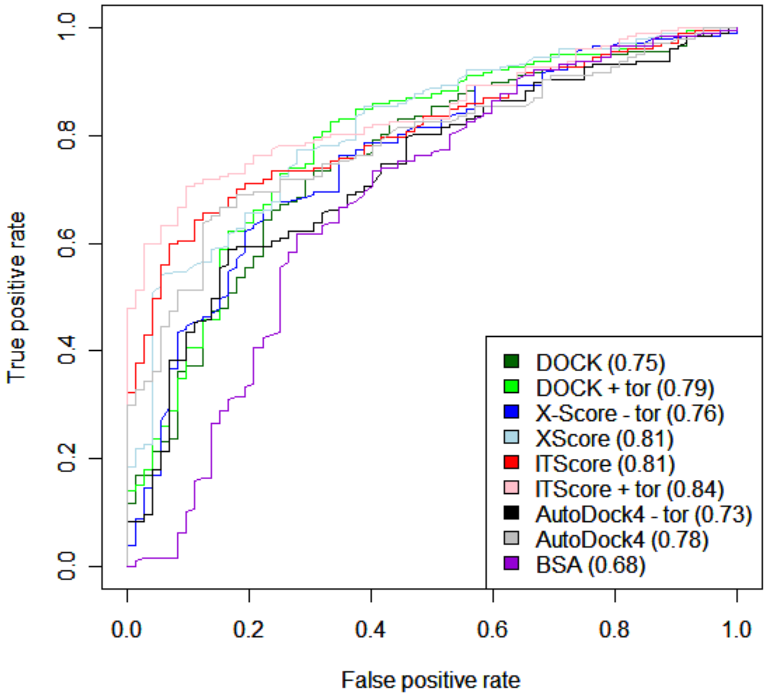


Figure V-10 - ROC plot of Scoring Function Performance



5.3.2 Torsional Entropy

Although valids were ranked over invalids whether or not torsional entropy terms were included, it is important to note that adding the penalty significantly increased the true positive rate in the highest ranked ligands (Figure V-6 – Figure V-9). This improvement was seen for all scoring functions. Including the torsional term in ITScore and DOCK improved performance (AUC increase of .03 and .04 respectively), and removing the term from X-Score and AutoDock degraded performance (AUC decrease of .05 for both). Of course removing a term from a scoring function should degrade its performance, but it was interesting to see that the torsional term resulted in relatively similar contributions in performance for all the scoring functions in terms of AUC. Furthermore the contribution of the torsional term seems to be essential for improved performance in the most critical region of the ROC plot – the lower left corner where the highest-ranked compounds are represented.

A closer look at the scores by ITScore with and without torsional penalty shows that the largest changes in score occurred for several of the invalid complexes (1U3A and 1T7V), ranking them much lower when the term was included (59 vs. 167 and 70 vs. 140 respectively). The ligands in both of these molecules are polyethylene glycols, long chain-like molecules that are not biologically important for these proteins. These ligands are large and extremely flexible, with 23 and 17 rotatable bonds, respectively. The ligands' large size results in a good contact score, even with modest burial, and the penalty term is essential to account for the high conformational entropy loss for the ligand when the torsions become restricted.

The method for including a torsional penalty to ITScore and DOCK was somewhat naïve for this study. It is reasonable to assume that performance would be further improved by properly fitting these terms into DOCK and ITScore functions. While knowledge potentials like ITScore aim to represent all physical contributions to binding, they are still restricted by any limitations of their training set. ITScore's training set is much larger than those of other functions, but torsional entropic penalties of the ligand will not be well represented unless the set includes ligands with many rotatable bonds. Pair-wise potentials are iteratively trained by identifying native poses over incorrect poses, but docking ligands with many rotatable bonds is inherently difficult because of their large conformational space. This predicament has severely restricted the ability of pair-wise potentials to account for the torsional penalty well at this time. In order to improve performance, the most appropriate approach will be to iteratively fit a new term as a corrective measure when training the pair-potential.

5.3.3 Top-Scoring Complexes

A set of valids was consistently ranked highly by at least several of the scoring functions (Table V-4). Although several valids were ranked in the top ten by 3 out of the 4 scoring functions, most non-unique top valids were highly ranked by 2 out of 4 functions. These common top-valids were most similar between functions that included the torsional penalty in their original formulation (AutoDock and X-Score) or had the penalty added (ITScore+tor and DOCK+tor). This might indicate certain scoring bias for certain ligands in our implementation of the torsional penalty for ITScore+tor and DOCK+tor. However, such bias does not undermine the usefulness of the torsional penalty in separating valids from invalids.

The common top-ranked valids range from enzymes such as hydrolases (2CER, 1VJ5, 1NO6, 1K1P), transferases (1C9C), and kinases (1OIT) to RNA binding proteins (1L8B) and immune system proteins (1LO0). Most of the valid ligands are either inhibitors, reaction intermediates designed to study enzyme function, or the enzyme substrates (EST in 1FDS). In one case, the ligand is a modified cofactor (PP3 in 1C9C) bound in the cofactor binding site. Most importantly, each of the valids is a biologically relevant ligand that makes important contacts with the protein in the active site.

One valid ranked highly by three of the four functions is the phenethyl-substituted glucoimidazole (PGI), a transition-state mimic and a potent inhibitor of beta-glycosidase (2CER) with a Ki of 0.6 nM [175]. The multiple hydroxyl groups of the ligand form numerous bonds within the active site to residues Glu 432, Gln 18, Trp 433, His 150, Asn 205 and Glu 387. However, no interactions are seen for the phenethyl group. The structure’s authors argue that entropic contributions, perhaps attributable to desolvation, are responsible for the high affinity of this inhibitor. The inhibitor is one of the strongest known for this family of proteins.

Another valid ranked in the top 10 by the majority of the functions is CIU, a potent alkylurea inhibitor of human epoxide hydrolase, bound to its target (1VJ5) with an IC_{50} value of 0.12 μM[176]. The urea oxygen forms stabilizing hydrogen bonds to Tyr 381 and Tyr 465 and the NH group form bonds with Asp 333. Additional favorable VDW interactions utilize a smattering of residues across the active site.

Structure 1C9C contains an aspartate aminotransferase complexed with a substrate analogue (pyridocal 5'-phosphate linked with an alanine, PP3) that was used to study active site-loop motion. It was ranked as a top valid by both ITScore+tor and DOCK4+tor. Highly buried (%ESA ~ 6%) and armed with phosphate and carboxylate groups, the ligand forms favorable ionic interactions with Arg 266 and Arg 386, as well as polar interactions with Asn 194, Asp 222, and several other active site residues [177]. The ligand also has relatively few rotatable bonds, reducing the torsional entropy. Curiously, X-Score and AutoDock ranked this ligand 17th and 53rd, respectively.

1OIT is an X-ray structure containing a selective inhibitor of cyclin-dependent kinase 2 (CDK2), identified through high-throughput screening and structure-activity-based optimization. The ligand, a 2-anilino imidazo[1,2-a]-pyridine, is one of the stronger binders in the valid set, with an IC_{50} of < 3 nM. The pyrimidine core of the molecule forms stabilizing hydrogen bonds to the protein backbone, while the hydrophobic anilino ring lies in a hydrophobic path of residues whose desolvation contributes to the high binding constant and likely to the CDK2 specificity of this inhibitor [178]. Additional bonds with Asp 86 further stabilize the compound. In this instance, the protein-ligand complex was ranked in the top-10 by X-Score and AutoDock, while ITScore+tor and DOCK4+tor ranked the complex lower as 31st and 11th, respectively.

Despite some correlation between AutoDock and X-Score ranks ($R^2 = 0.7$), the top rankings still vary significantly between scoring functions. An mRNA-5'-cap analogue inhibitor (7-methyl GTP) bound to a RNA-binding translation initiation protein (1L8B K_d = 0.14 μM) [179] was ranked 4th and 8th by ITScore+tor and AutoDock, respectively. Meanwhile, DOCK4+tor and X-Score ranked it 16th and 99th, respectively. The inhibitor's high potency is attributed to favorable stacking of the guanine moiety and extensive hydrogen bonding through the phosphate groups. The torsional entropy penalties for the ligand did not significantly affect the rankings.

Assessing the well-scored invalids can be an enlightening exercise because it can reveal caveats in a given scoring function. Most importantly, we expect that opportunistic invalids should score well if they are similar to a valid ligand and they are bound in the true active site. These are "acceptable failures" where a good score and rank is desirable, despite the lack of clear functional significance for the binding. Of the 71 invalids, 15 were observed to bind in active sites, and 3 of those appeared in the top-ranked invalid lists (Table V-5).

Overall, each of the scoring functions ranked a similar list of invalids highly (Table V-5) but not the rank itself or the order of ranking. Chemically the top-scoring invalids include a variety of molecules, such as common detergents (N-octanoyl-sucrose in 1IZ2, n-octane in 1APM), buffers (tartarate in 1D5R, DTT in 1N2F, HEPES in 1RJM), other small organic compounds. Most of the invalids were bound to enzymes, although not all.

Three of the four scoring functions ranked 1D5R and 1IZ2 as high-scoring invalids. Structure 1D5R has a buffer tartrate molecule bound in an active site of a PTEN tumor suppressor. The tartrate makes many similar contacts to those expected for the natural substrate – inositol (1,3,4,5)-tetrakisphosphate, but it has no known biological activity with respect to PTEN phosphatase function [180]. Meanwhile, 1IZ2 shows an N-octanyl-sucrose (SUM) detergent additive bound to an alpha1-antitrypsin. The detergent molecule's hydrophobic tail is partially inserted into a protein cavity away from the known active site, and a quarter of the ligand surface area remains exposed to solvent.

Another interesting structure is 1URM, ranked highly amongst invalids by ITScore+tor and DOCK4+tor. The structure contains a benzoate molecule which unexpectedly binds in the active site of the human peroxiredoxin 5 enzyme, making bonds with a critical mutated serine residue (mutated from the catalytic peroxidatic cysteine) and multiple hydrophobic contacts through its aromatic ring. The origin of the benzoate in the structure is unknown and no known biological function has been associated with its binding [181]. However, the molecule was also seen to bind in the same pocket of wild-type and homologous crystal structure of the protein, indicating that its binding mode might be more than serendipity.

In 1RJM, ranked as the top invalid by DOCK4+tor, the ligand is far from an active site, wedged in a hole formed by trimer of MenB chains, a lyase from *Mycobacterium tuberculosis*. The ligand's negatively charged sulfonic-acid head group forms interactions with a charged Arg side chain via some well ordered water molecules. Highly ranked by X-Score and AutoDock4, 1SHV is a structure of an SHV-1 beta-lactamase, complexed with a detergent bearing a maltose moiety. The hydrophobic tail of the Cymal-6 detergent is wedged between two hydrophobic α-helices on the "top" of the protein, while the disaccharide group is hydrogen bonded across the crystal packing interface, facilitating intermolecular aggregation. The ligand scores well despite having the maltose moiety (which accounts for ~30% of the surface area) exposed to solvent in the monomer structure, and it does not appear to have an effect on the catalytic activity of the enzyme [182]. Also scored highly by X-Score and AutoDock4 is 2SHP, an SHP-2 tyrosine phosphatase with a detergent-like ligand - CAT. The CAT's tail is nicely buried between a trio of hydrophobic α-helices, just shy of the catalytic site but it does not appear to form strong bonds with the surrounding residues.

Although several other invalids are found in the active sites of proteins (1QST, 2OR7, 1T7V, 1TTO et al.) they aren't ranked especially high among the invalids by any of the scoring functions. Some of the invalids are identified by modeling molecules into unaccounted electron density based on size, shape, and chemical environment. Usually the molecule is readily identified as a component of the crystallization matrix. For example, a molecule of polyethylene glycol was modeled to fit the density in the δ5-3-ketosteroid isomerase (8CHO) [183]. Other cases exist where the ligand was not from the crystallization matrix. Analine (ANL) is seen to bind near a proposed phospholipids substrate site in the crystal structure of phospholipase A2 (1PPA). While its presence is confirmed by the electron density, its origin is remains a mystery [184].

Table V-4: Top scoring valids. Unique valids in italics, those scored in the top 10 by 2 or more functions in plain text.

ITScore + tor

PDB	Ligand	Rank	%ESA
1C9C	PP3	1	5.9%
2CER	PGI	2	14.0%
2J7C	*IDE*	3	14.1%
1L8B	MGP	4	36.9%
1IEX	*TCB*	5	11.5%
2J7D	*GI1*	6	12.4%
1VJ5	CIU	7	6.8%
2J7B	*NTZ*	8	7.3%
1BR6	*PT1*	9	20.0%
2J7G	*GI4*	10	12.4%

DOCK4 + tor

PDB	Ligand	Rank	%ESA
1C9C	PP3	1	5.9%
1HG2	*IP2*	2	69.1%
1D1Q	*4NP*	3	12.2%
2PBW	*DOQ*	4	10.7%
1NO6	794	5	23.2%
1H1H	*A2P*	6	60.3%
1V48	*HA1*	7	5.0%
2CCG	*TMP*	8	21.2%
5GPB	*GPM*	9	16.1%
2FZG	*EOB*	10	7.1%

AutoDock4

PDB	Ligand	Rank	%ESA
1LO0	BC1	1	12.2%
2HHA	*3TP*	2	25.8%
2CER	PGI	3	14.0%
2OYM	*MNI*	4	10.7%
1Q8U	*H52*	5	61.4%
1OIT	HDT	6	25.2%
1FDS	EST	7	11.8%
1L8B	MGP	8	36.9%
1VJ5	CIU	9	6.8%
1K1P	MEL	10	24.7%

X-Score

PDB	Ligand	Rank	%ESA
1VYQ	*DUX*	1	27.0%
1VJ5	CIU	2	6.8%
1FDS	EST	3	11.8%
1OIT	HDT	4	25.2%
1GII	*1PU*	5	10.6%
1K1P	MEL	6	24.7%
2CER	PGI	7	14.0%
1LO0	BC1	8	12.2%
1NO6	794	9	23.2%
2BU2	*TF1*	10	28.1%

Table V-5: Top scoring invalids. Unique valids in italics, those ranked in the top 10 by two or more functions in plain text. Invalids found in known binding sites are marked with an asterisk

ITScore + tor

PDB	Ligand	Rank	%ESA
1M27	FLC	88	24.5%
1RTV	SRT	95	10.6%
2RA5	SRT	111	20.8%
1D5R	TLA*	112	5.0%
1URM	BEZ*	119	18.8%
1IZ2	SUM	120	25.9%
2INU	2PO	127	36.2%
2O02	BEZ	135	31.7%
1ZDY	T3A	137	18.9%
2OK6	BEZ	139	0.2%

DOCK4 + tor

PDB	Ligand	Rank	%ESA
1RJM	EP1	26	3.6%
1D5R	TLA*	29	5.0%
2RA5	SRT	35	20.8%
2B4P	MLI*	46	8.3%
1MRZ	CIT	51	34.5%
1M27	FLC	57	24.5%
1ZGN	MES	69	45.0%
1HH8	FLC	81	19.7%
1URM	BEZ*	82	18.8%
2OK6	BEZ	92	0.2%

AutoDock4

PDB	Ligand	Rank	%ESA
2SHP	CAT	54	10.4%
1SHV	MA4	60	34.1%
2OK6	BEZ	64	0.2%
1IZ2	SUM	68	25.9%
1D5R	TLA*	85	5.0%
2E50	TRE	91	26.5%
1JJ0	SUC*	98	30.0%
1LIH	PHN	100	37.2%

X-Score

PDB	Ligand	Rank	%ESA
1SHV	MA4	35	34.1%
2SHP	CAT	42	10.4%
1IZ2	SUM	44	25.9%
1RJM	EP1	95	3.6%
1Q61	MG8	103	10.4%
2OK6	BEZ	104	0.2%
1APM	OCT	105	17.8%
2P4B	BOG	106	19.7%

1PPA	ANL	101	5.2%	*1JLU*	OCT	109	17.2%
1N2F	DTT	124	16.7%	*2GFC*	OCT	111	15.5%

5.4 Conclusion

This study puts forth a new criterion for evaluating scoring functions: the ability to discern between opportunistic binding (invalids) and biologically important ligands (valids). Accordingly, a new test set is presented, which contains 177 valid and 71 invalid structures. The invalid and valid structure sets show similar distributions of physicochemical properties such as molecular weight, hydrophobicity, solubility and BSA. For different scoring functions - representative of knowledge-based, force-field, and empirical methods - are used to evaluate the test set with respect to valid/invalid discernment. The results show that the four scoring functions are able to tease out invalids from valids with a significant success rate, achieving ROC AUC scores of 0.84, 0.81, 0.79 and 0.78 for ITScore+tor, X-Score, DOCK4+tor, and AutoDock4, respectively. Additionally the approximation of ligand torsional entropy in the scoring functions was shown to have an important contribution for successful ranking of the valids versus invalids among the protein-ligand complexes.

This test set has potential to help improve algorithms used in molecular docking by providing a different measure for docking success. Its further development will be essential to extending scoring functions to new purposes, like cross docking identifying unknown protein function, or estimating the druggability of a surface pocket.

CHAPTER VI

Overcoming Sequence Misalignments with Weighted Structural Superposition

6.1 Introduction

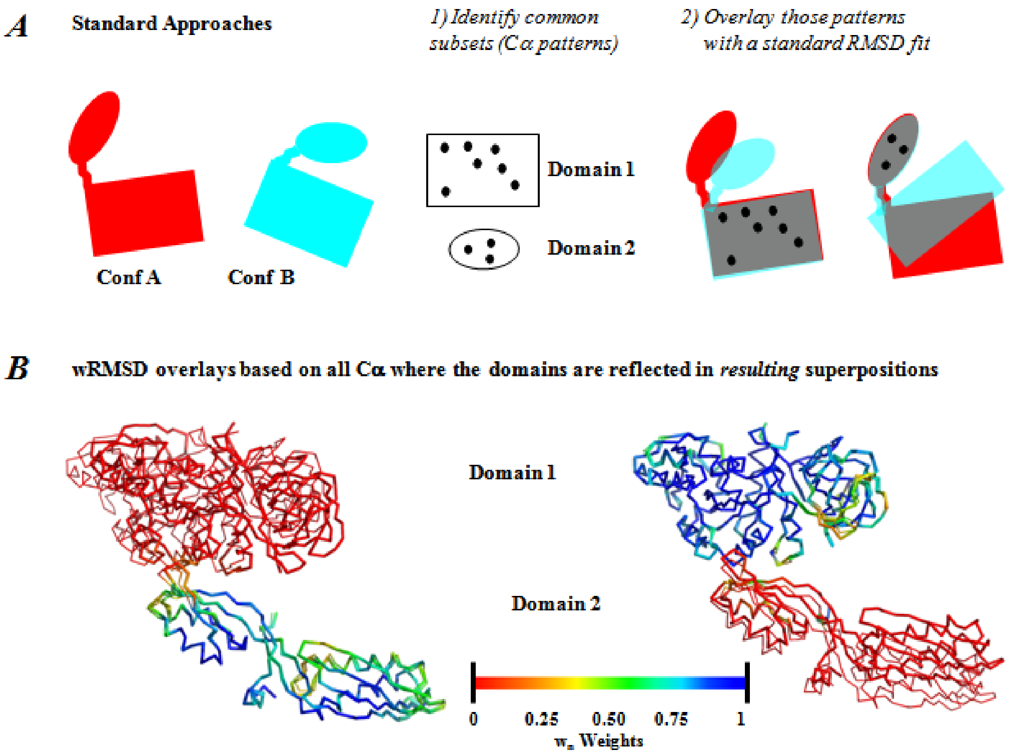
Evolutionarily related proteins generally retain a tertiary fold that is more conserved than the amino acid sequence [185, 186]. Structure is related to function; hence, proteins with similar structures may also share a common biological activity [187]. As a result, the identification of a homolog is a very useful means to infer the function and/or predict the structure of an uncharacterized protein. Many databases exist that classify proteins into families by their structures, including but not limited to SCOP [188], CATH [189], DaliDB [190], PASS2 [191], MMDB [192], ASTRAL [193], HOMSTRAD [194], and LPFC [195]. A review from Orengo and Thorton provides a very thorough discussion of protein evolution from a structural standpoint [196], and another recent review stresses that the classification in an evolutionary context is still an open problem [197].

An appropriate structural superposition provides a means to compare the similarity or dissimilarity between protein structures. However, in order to perform a structural comparison, the corresponding residues (atom pairs) between the proteins must be determined. This task can be accomplished 1) in a sequence-dependent manner using an initial sequence alignment or 2) solely through structural information in a sequence-independent manner. Sequence-based techniques can miss similarity between homologous proteins with intermediate to low sequence identity (twilight zone). Fold-based methods can identify structural similarity, even between homologs with divergent sequences, but they may be misleading in the case of flexible proteins. A technique that combines the two approaches and overcomes limitations caused by the protein flexibility would be an ideal choice for superimposing homologs. In 2005, a thorough evaluation [198] of six structural comparison techniques – SSAP [199], STRUCTAL

[200, 201], DALI [202], LSQMAN [203], CE [204], and SSM [205] - demonstrated many strengths and limitations of current approaches and the caveats of metrics used for evaluation and comparison. A more recent review with similar scope indicated room for improvement of alignments, especially in proteins with extensive conformational variability or structural repetitions [206]. Additional reviews of the field call for combining techniques and using consensus across several methods to best define a structural comparison [185, 207, 208]. Here, we present a structural alignment method that accounts for protein flexibility and utilizes a superposition-driven approach to capture structural similarity in a more systematic and intuitive way.

Previously, we introduced a superposition technique that overcomes the limitations of protein flexibility [7] by implementing a Gaussian-weighting term into the RMSD-fit algorithm determined by Kabsch [209]. The calculated weight is directly related to the distance between two atoms in space. Consequently, atom pairs in close proximity have a greater weighting than those further apart, biasing the superposition toward the regions that remain relatively rigid between conformations. Our method is the reverse of techniques used for the last 20 years, which perform two steps: 1) identify related subsets of Cα and 2) overlay those related subsets by a standard RMSD fit (sRMSD; Figure VI-1). Using our technique, the overlay defines the domains, rather than the domains defining the overlay. The resulting weights identify the domains. As illustrated in Fig. 1B, the backbones of two protein conformations are well superimposed in the blue, high-weight regions but can be seen separately in the red, low-weight regions. Each solution is based on a unique domain of the protein and each is an equally valid overlay. Complete mathematical details of the weighted RMSD (wRMSD) procedure can be found in our original work [7], and an abbreviated presentation is provided in the Appendix.

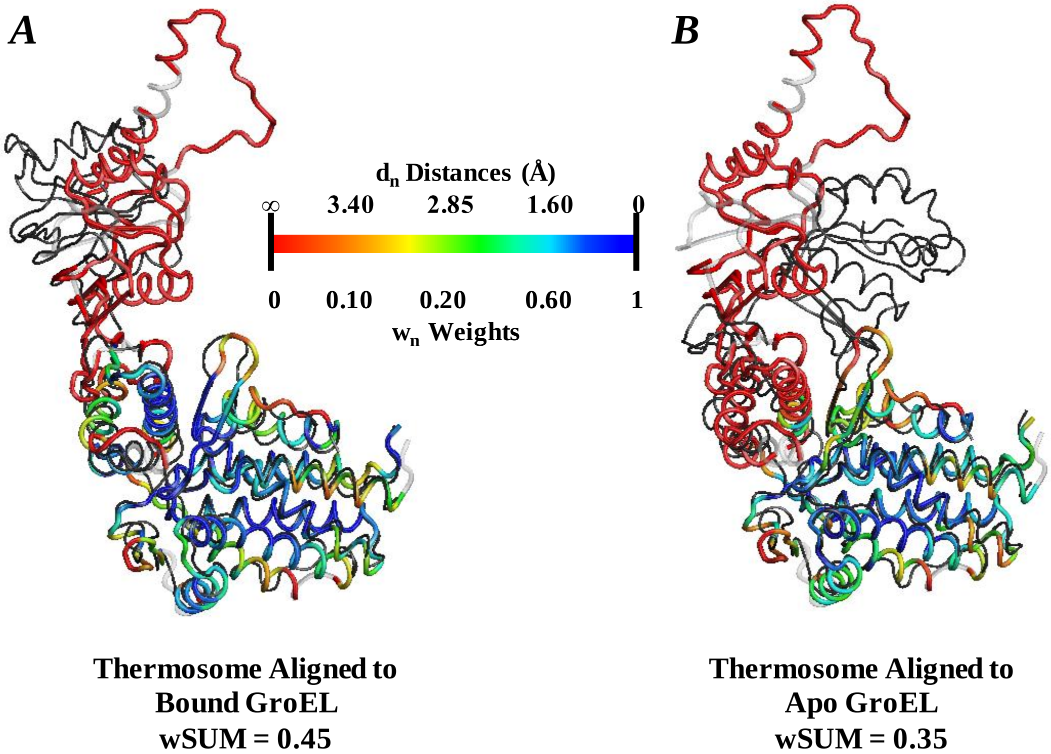
Figure VI-1 : A) Most methods for superimposing flexible proteins are based on two steps, which involve determining a subset of related atoms and overlaying the subset using a standard RMSD-fit procedure. Each technique differs in the way that it identifies the related subsets, but in the superposition step, all of the techniques designate each Cα as "in" or "out" of the calculated fit. B) Our weighted superposition is based on all Cα. Multiple solutions can be found where the domains are reflected in the resulting weights and superpositions. Blue and green regions have high weights, align well, and define a domain. Red regions have low weights and poor agreement in the overlay.


A
Standard Approaches
1) Identify common subsets (Cα patterns)
2) Overlay those patterns with a standard RMSD fit
Domain 1
Domain 2
Conf A
Conf B
B
wRMSD overlays based on all Cα where the domains are reflected in resulting superpositions
Domain 1
Domain 2
0
0.25
0.50
0.75
1
w_n Weights

In this work, we have coupled our wRMSD technique with basic sequence alignment algorithms from the EMBOSS package [210] to provide initial alignment of homologous sequences. In our previous work, we were able to show an improved superposition of two conformations of the chaperonin protein GroEL, which undergoes a large conformational change between the bound and apo forms (PDB [2] codes 1AON [211] and 1OEL [212], respectively). In Figure VI-2, we use this system again to demonstrate the potential difficulties of fitting homologous, flexible proteins. With our technique, either conformation of GroEL can be appropriately superimposed to the bound form of its archaeal homolog, the thermosome (1A6E [213]). The easier case of fitting the two bound conformations is shown in Figure VI-2A, and Figure VI-2B shows the more difficult comparison of the bound form of the thermosome to the apo form of

GroEL. The superpositions are colored by weight of the aligned pairs of Cα atoms, with higher weights indicating closer proximity, and higher average weight (wSUM) indicating stronger structural similarity. Fold-based techniques can identify the homologous regions from the similar bound conformation in regions of very low sequence similarity, but in cases where the structures of homologs are only available in alternate conformations, those same techniques may have difficulty. Our method is able to overcome errors from the initial atom pairing due to low sequence identity and large conformational differences by only weighting regions of the protein in good structural agreement.

Figure VI-2: Chaperonin family (20.8% ID). Most techniques would readily identify the similarity between the thermosome and GroEL in the similar the bound conformation, but they may not identify its similarity with the apo conformation of GroEL. A) wRMSD superposition of the bound conformation of GroEL (thick, colored lines) onto the homologous thermosome (thin, black lines). Light gray regions of GroEL indicate residues within gaps in the alignment. B) wRMSD fit of the apo conformation of GroEL (thick, colored lines) onto its homolog thermosome (thin, gray lines). The wSUM gives the average weights of all paired residues, showing that the two bound conformations in *A* have greater similarity than the two conformations in B.


A
B
d_n Distances (Å)
∞
3.40
2.85
1.60
0
0
0.10
0.20
0.60
1
w_n Weights
Thermosome Aligned to Bound GroEL wSUM = 0.45
Thermosome Aligned to Apo GroEL wSUM = 0.35

Once a sufficient overlay is established, the seed extension (SE) structural alignment algorithm[214] lets the wRMSD superposition dictate a new and improved sequence alignment. The SE method extends the alignment of residue pairs in very close proximity (seeds) along the protein chain. Many modern structural alignment methods, such as CE, combine sequence and structure data in their alignment procedure. CE uses blocks of aligned fragment pairs to perform a combinatorial extension, considering both sequence and structural similarity as part of an optimal alignment determination. In contrast, our method is modular and allows the structural information to dominate the superposition producing a consistent structural alignment solution. The SE algorithm allows us to then convert the information from the structural alignment into an equally consistent sequence alignment. Below, we demonstrate the robustness of the procedure with respect to initial sequence alignment, and the ability to correct misalignments in the initial alignment. We then compare alignment performance to that of several popular structure alignment programs.

6.2 Methods

Homologous protein pairs were obtained from the HOMSTRAD database [194]. The protein coordinates were downloaded from the PDB [2], and the specific protein chains used were dictated by the pairings listed on the HOMSTRAD website. For this study, we chose to focus on the more difficult cases of homologous proteins with lower sequence identities (ID) (39-16%). We examined homolog pairs with <16% ID, but the sequence alignment algorithms used to obtain an initial alignment failed, giving nonsensical alignments. The other structural alignment programs used in this study also failed with these cases, making <16% ID a relatively universal cutoff for current methods.

Our technique is performed using Cα coordinates, but it is easily extended to any atom subset. The HwRMSD procedure consists of 4 sequential steps:

1. Use a simple sequence alignment to determine an initial list of paired residues.

2. Calculate an initial sRMSD alignment (non-weighted) to overlay the centers of mass and provide a rough initial orientation for the proteins.

3. Conduct iterative wRMSD fitting until convergence is reached.

4. Obtain a corrected sequence alignment from the structural superposition using SE.

The program needle, an implementation of the Needleman-Wunsch (NW) global alignment from the EMBOSS [210] package was used to generate the pair-wise sequence alignment. This alignment determines the residue correspondence between the two proteins, which is then used to guide the initial structural superposition.

To obtain a sequence alignment from the structural superposition, SE is used with default parameters [214]. Briefly, SE finds "seed" pairs of structurally equivalent residues from overlaid structures based on their physical proximity and chemical similarity. Consecutive triplets of seeds are then extended along the alignment matrix in both directions, using distance and amino acid similarity to resolve conflicts which arise during the extension of more than one diagonal.

To test the robustness of the method with respect to initial sequence alignments, water, an implementation of the local Smith-Waterman (SW) sequence alignment from EMBOSS [210], was also used. The four different scoring matrices used were BLOSUM50, BLOSUM62, PAM120, and PAM250, Each employed its optimal gap-open and gap-extension penalty parameters: (-10,-2), (-7,-1), (-16,-4), and (-10,-2), respectively. The optimal parameters for each scoring matrix were recommended by the European Bioinformatics Institute [215].

For each protein pair, the pair-wise distances between the superpositions obtained with the same sequence alignment algorithm were calculated, and the median of these values was chosen to represent the similarity of the solutions. The distances between the superpositions were calculated using a simple all-atom RMSD. In cases such as the PHBD-like proteins (1FOH [216] and 1PBE [217]) and several others, the initial sequence alignment was altogether too small to be considered a reasonable solution when using certain sequence alignment parameter sets. To avoid such invalid outliers, any solution with an alignment length of less than 10 residue pairs was discarded, and the median distance was calculated between the 2 or 3 remaining solutions. Such cases were mostly in the range of low sequence identity (< 20% ID), and they illustrate the practical limits of the current approach. For comparing our results to other tools, we used the EMBOSS-wrapped implementations of CE [204], FATCAT (flex) [218], and the native SSM [205] and DaliLite [190] servers. Default parameters were used for all structural alignment methods. For SSM and DaliLite, only the best alignment solution was considered, based on highest Q-Score or Z-Score, respectively. To obtain consistent weighing for alignments from different software packages, a weighting constant of c=5 was used for any wRMSD or %wSUM calculations (in the default wRMSD method this constant is set to the value of the initial sRMSD superposition).

PyMOL [116] was used for various visualization purposes and the creation of figures in this paper.

6.3 Results and Discussion

6.3.1 Overcoming Variation in the Initial Sequence Alignment

Aligning proteins with high sequence identity is straightforward, so for this study, we chose to focus on the more difficult cases of homologous proteins in the low to intermediate range (39-16% ID), as listed in Table VI-1. We first examined the robustness of the HwRMSD protocol with respect to a variety of sequence alignments. We chose the canonical implementation of global Needleman-Wunsch (NW) and local Smith-Waterman (SW) algorithms to perform the initial sequence alignments and varied their parameters by using four different scoring matrices and their respective optimal gap penalty values. For each test case, each of the sequence alignments was used to generate a standard and weighted superposition. The similarity of the superpositions per test case was then evaluated by computing all-atom RMSD distances among the superpositions generated from each of the sequence alignment algorithms (note that this use of RMSD is simply a measure of the difference in two sets of coordinates, not a fitting procedure).

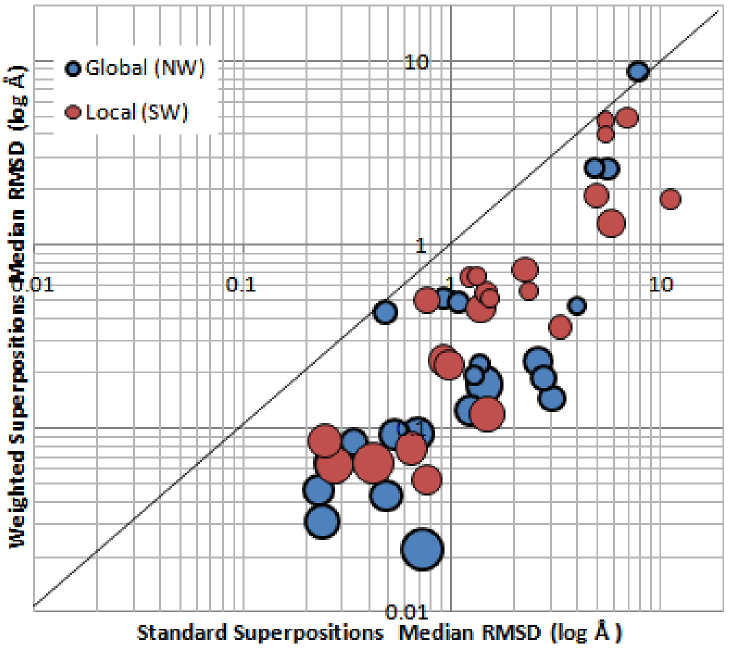
Table VI-1: Median RMSD differences (in Å)* between the structural superpositions generated utilizing sequence alignments with different parameters; using global (Needleman-Wunsch) or local (Smith-Waterman) sequence alignment algorithm, and standard or weighted superposition algorithm.

Homolog Proteins and PDB IDs	**%ID**	**Smith-Waterman**		**Needlman-Wunsch**	
		Standard	**Weighted**	**Standard**	**Weighted**
Serine/Threonine Phosphatase 1FJM & 1TCO	39%	0.42	0.07	0.72	0.02
Glutathione Synthase 1M0W & 2HGS	37%	0.27	0.06	0.27	0.07
Interferon 1AU1 & 1ITF	35%	1.48	0.12	1.43	0.17
Adenosylmethionine Decarboxylase 1I7B & 1MHM	33%	0.25	0.09	0.24	0.03
Clostridial Neurotoxin Zinc Protease 1EPW & 3BTA	31%	0.64	0.08	0.69	0.09
Sulfatase 1AUK & 1FSU	29%	0.91	0.24	0.54	0.09
Translation Initiation Factor 1AP8 & 1EJH	29%	0.97	0.22	0.23	0.05
Protocatechuate-3,4-Dioxygenase 3PCG (chain A) & 3PCG (chain M)	28%	0.76	0.05	0.49	0.04
Aminotransferase 1A3G & 5DAA	27%	1.38	0.45	1.22	0.13
SpoU rRNA Methylase 1IPA & 1GZ0	26%	5.83	1.30	2.59	0.23
FMN Oxidoreductase 1OYC & 2TMD	25%	2.26	0.73	0.30	0.08
Queuine tRNA-Ribosyltransferase 1IQ8 & 1K4G	25%	0.76	0.50	0.34	0.08
tRNA Synthestase 1GLN & 1QTQ	24%	4.96	1.86	3.01	0.15
DNA Methylase 1BOO & 2ENT	23%	3.31	0.36	2.73	0.19
DNA Topoisomerase 1AB4 & 1BJT	22%	1.46	0.55	0.48	0.43
Pyridoxal-Phosphate Enzymes 1TDJ & 2TYS	21%	6.89	4.91	5.54	2.59
Iron/Ascorbate Oxidoreductase 1BK0 & 1DCS	20%	11.08	1.77	0.92	0.51
Molybdopterin Dehydrogenase 1FFV & 1FO4	19%	1.21	0.66	1.08	0.49
Splicesomal Protein, Internalin B 1A9N & 1D0B	19%	2.33	0.56	**7.85**	**8.81**
Asp/Glu/Hydontoin Racemase 1B74 & 1JFL	18%	1.52	0.51	1.36	0.22
Polysaccharide Lyase 1CB8 & 1EGU	18%	1.32	0.68	1.27	0.20
PHBH-like Proteins 1FOH & 1PBE	17%	5.44	4.72	4.83	2.63
Adaptin, Clathrin Appendage Domain 1E42 & 1QTS	16%	5.51	3.98	3.94	0.47

*The sequence alignments were altered by varying the similarity matrix and gap penalties. The variation across the superpositions was measured by pair wise RMSD (Å) between all the solutions. Median differences are reported, but all calculated pair-wise RMSDs are included in the supplementary material. Smaller value denotes a greater agreement between the superpositions.

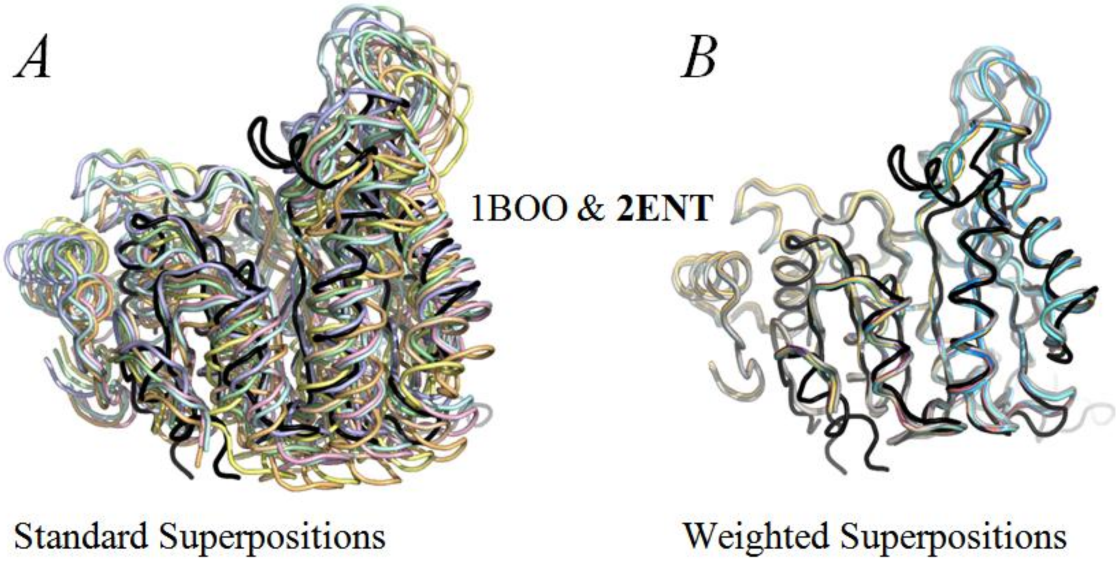
For each test case except one, the use of weighted superposition overcomes the variation in atom-pairing arising from the different sequence-alignment parameters to give a more consistent structural comparison. Figure VI-3 shows how the standard superpositions resulted in more variation across the solutions for a pair of proteins; the median RMSD ranged from 0.23 – 7.84 Å with roughly half of the test cases having variation of greater than 1 Å. In contrast, for the weighted fit, the median RMSD of the solutions ranged from 0.02 – 8.8 Å with only 9 out of 46 cases (3 for global and 6 for local) having greater than 1 Å variation. Both global NW and local SW sequence alignments resulted in similar variation in superpositions. Protein pairs with lower sequence identity showed greater variation (bubble size in Figure VI-3 is proportional to sequence identity, and the small bubbles are in the upper right quadrant).

Figure VI-3: Median RMSD distances (measures of agreement, not overlay fit) between superpositions obtained from varying sequence alignment parameters. Standard superposition values on the *x*-axis are higher than the weighted superposition values on the *y*-axis, indicating greater variation in solutions for the sRMSD superpositions and better agreement across the wRMSD superpositions. Blue bubbles represent medians of superpositions obtained from global Needleman-Wunsch alignments; red bubbles are from local Smith-Waterman alignments. The size of the bubbles is proportional to sequence identity of the aligned protein pairs (39% to 16%) with smaller bubbles indicating less similarity. In general, the smaller bubbles show higher variation due to difficulty of obtaining a consistent initial sequence alignment.



Standard superposition is sensitive to the initial alignment, and incorrectly paired residues skew the result even in cases where large structure similarity in a protein domain is visually obvious. The weighted superposition converges to a consistent result even when a wide variety of initial alignments are used, allowing regions of the structure in closest proximity – such as large similar domains – to drive the superposition. Figure VI-4 uses DNA methylase homologs [219] (23% ID) to visually show how the standard superpositions are noticeably different when varying the methods and parameters for sequence alignment (Figure VI-4A). For this example, the median difference was 3.69 Å among the SW alignments and 2.85 Å among the NW alignments. These variations in the superpositions result in different "corrected" sequence alignments from SE. Conversely, the weighted superpositions are indistinguishable by eye (Figure VI-4B); the median difference of the weighted superpositions is only 0.35 Å for the SW alignments and 0.16 Å for the NW. All the weighted superpositions generate the same corrected sequence alignment with SE. Most importantly, the weighted superpositions resulted in an improved fit over the standard superpositions, particularly in the core region which is structurally conserved between the homologous proteins. After all, a consistent superposition is only useful if it is also an improved superposition!

Figure VI-4: DNA methylase family (23% ID). Weighted structural superpositions are nearly independent of the sequence alignment method, but standard superpositions are greatly affected. A) Overlays of 2ENT (black ribbon) to 1BOO[219] (colored ribbons) from standard superpositions based on seven different sequence alignments. B) The seven weighted superpositions of 2ENT to 1BOO, based on the same varied sequence alignment routines collapse into a single converged solution.



There were several test cases where the weighted solutions of the SW and NW alignments were not a significant improvement over standard solutions. For example, in the case of spliceosomal protein [220] and internalin B [221] (1A9N and 1D0B) the NW PAM120 alignment produced a significantly different superposition, (median distance 15 Å from other solutions) skewing the median of both standard and weighted solutions. Similarly in the case of the adaptin [222] and clathrin [223] appendage domains (1E42 and 1QTS) the SW PAM120 alignment produced an outlier superposition (distances of 7 Å from other solutions). The PAM120 alignment parameter set has the most severe gap penalty of the four matrices tested (-16), and it produced alignments with very few gaps when used with the global NW method, and extremely short alignments with the local SW method. Both of these cases are in the < 20% sequence identity range and exemplify the current practical limits of the method, which requires at least a reasonable initial sequence alignment. A list of all the pairs used in the median RMSD calculations and the pair-wise distances of their respective superposition solutions is available in the Supplementary Materials.

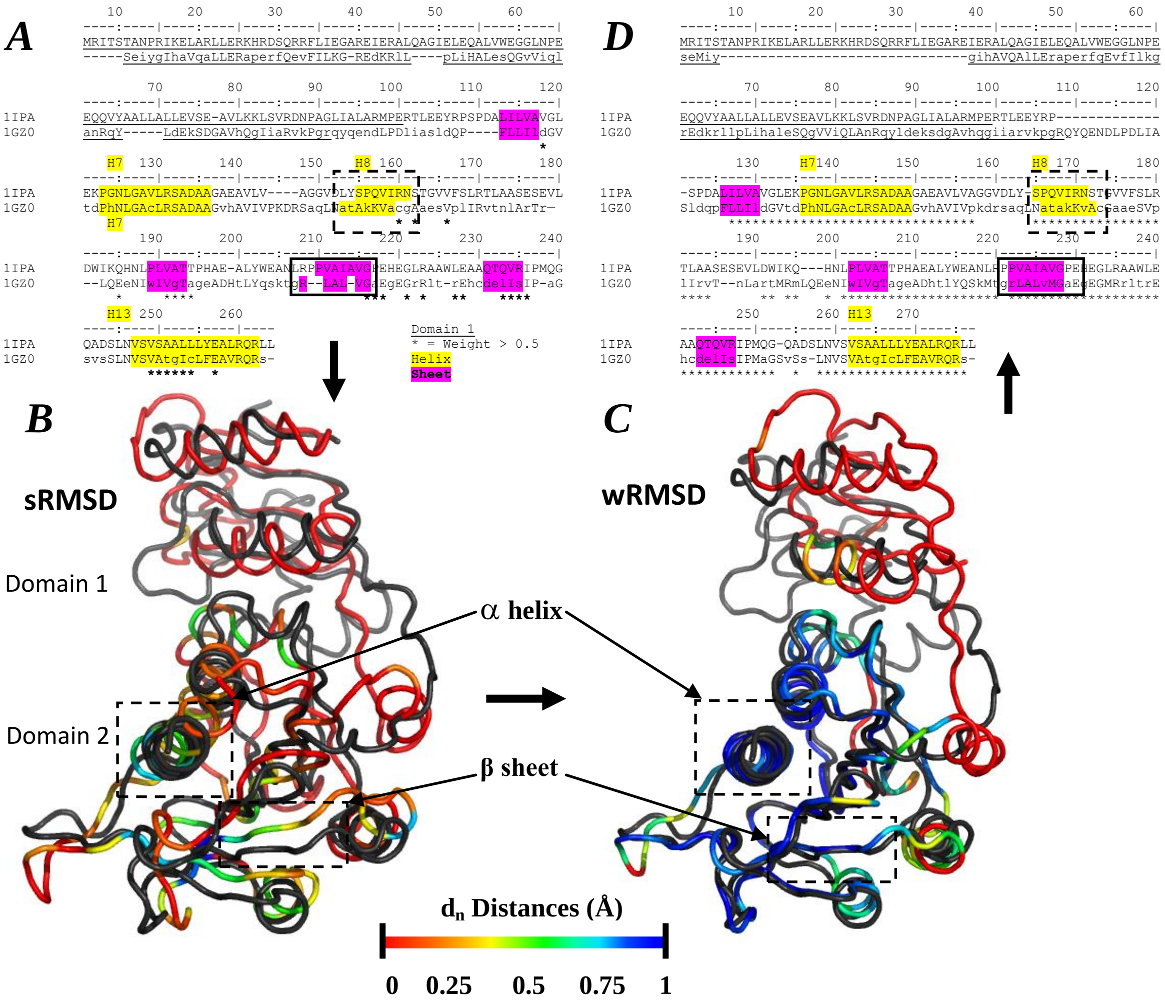
We found that wRMSD is more sensitive to the initial sequence alignment parameters – matrix, gap open penalty and gap extend penalty – rather than the choice of alignment algorithm or alignment length. As long as a reasonably long gapped alignment is provided, the weighted superposition technique can produce a consistent structural superposition. The NW alignment with a BLOSUM50 matrix produced the most appropriate alignments over all the test cases (the rest were too short or failed), and similar alignments were also obtained from BLOSUM62 and PAM250 with both NW and SW. Given its reliable performance, the NW BLOSUM50 alignment has been defined as the default initial alignment for our HwRMSD method and for the comparison to other methods below.

Of course, there may be situations where it is difficult to obtain an appropriate superposition with the weighted fitting, e.g., when a protein is large and has multiple domains. If two different initial sequence alignments obtain residue pairings each focused on a different domain, rather than spanning entire protein structure, then the weighted superpositions may not converge to the same solution. Another such case is when there is too little sequence or structural similarity, but this is when most comparison methods breakdown. For the test cases employed in this study, the sequence alignment tools broke down at ~16% ID, returning sporadic aligned segments that were too short and too infrequent. Homologs with so little sequence similarity are notoriously difficult to align [224], but it may be possible in some cases to compare them using methods based on structural information such as geometric comparisons of folds [225]. However, these techniques would be successful only when there is little structural variation or flexibility. Techniques such as wRMSD are absolutely required for large structural variation.

6.3.2 Correcting Sequence Misalignments

Residue pairings in regions of good structural agreement will be heavily weighted in the wRMSD calculation. Protein regions that have been brought into close spatial proximity, but have a low weighting with respect to the initial sequence alignment, indicate potentially incorrect pairings of residues. In Figure VI-5, we visually demonstrate how the spatial proximity of structures after a weighted superposition corrects a sequence alignment by using two homologs from the SpoU rRNA methylase family with 26% ID (1GZ0 and 1IPA) [226, 227].

Figure VI-5: SpoU rRNA methylase family (26% ID). A) NW sequence alignment of 1IPA and 1GZ0 using default parameters. Lower case represents sequence dis-similarity, and gaps are shown with dashes. The underlined region notes domain 1, yellow represents α-helices, purple represents β-sheets, and boxes represent misaligned residues corresponding to the labeled α-helix and β-sheet in B and C. Atom pairs with a weighting of 50% or greater in the wRMSD calculation are noted with asterisks. B) Standard superposition superpositions of 1GZ0 (colored ribbon) onto 1IPA (black ribbon) obtained using the initial alignment (from A), colored by weight. C) Weighted superposition obtained from the initial standard superposition. D) SE sequence alignment based on the wRMSD superposition, which now corrects the alignment of the secondary structure elements based on their spatial proximity in *C.*



```
                  10        20        30        40        50        60
          ----:----|----:----|----:----|----:----|----:----|----:----|
          MRITSTANPRIKELARLLERKHRDSQRRFLIEGAREIERALQAGIELEQALVWEGGLNPE
          -----SeiygIhaVqaLLERaperfQevFILKG-REdKRlL----pLiHALesQGvViql

                  70        80        90       100       110       120
          ----:----|----:----|----:----|----:----|----:----|----:----|
1IPA      EQQVYAALLALLEVSE-AVLKKLSVRDNPAGLIALARMPERTLEEYRPSPDALILVAVGL
1GZ0      anRqY-----LdEkSDGAVhQgIiaRvkPgrqyqendLPDliasldQP----FLLIldGV
                                                                     *
             H7    130       140       150   H8  160       170       180
          ----:----|----:----|----:----|----:----|----:----|----:----|
1IPA      EKPGNLGAVLRSADAAGAEAVLV----AGGVDLYSPQVIRNSTGVVFSLRTLAASESEVL
1GZ0      tdPhNLGAcLRSADAAGvhAVIVPKDRSaqLNatAkKVacgAaesVplIRvtnlArTr—
             H7                                    * *   *
                 190       200       210       220       230       240
          ----:----|----:----|----:----|----:----|----:----|----:----|
1IPA      DWIKQHNLPLVATTPHAE-ALYWEANLRPPVAIAVGPEHEGLRAAWLEAAQTQVRIPMQG
1GZ0      --LQEeNIwIVgTageADHtLYqsktgR--LAL-VGaEgEGrRlt-rEhcdelIsIP-aG
              *     ****                     ***  * *   **    ****
             H13   250       260
          ----:----|----:----|----
1IPA      QADSLNVSVSAALLLYEALRQRLL
1GZ0      svsSLNVSVAtgIcLFEAVRQRs-
                  ******  *
```

```
                  10        20        30        40        50        60
          ----:----|----:----|----:----|----:----|----:----|----:----|
          MRITSTANPRIKELARLLERKHRDSQRRFLIEGAREIERALQAGIELEQALVWEGGLNPE
          seMiy-------------------------------gihAVQAlLEraperfqEvfIlkg

                  70        80        90       100       110       120
          ----:----|----:----|----:----|----:----|----:----|----:----|
1IPA      EQQVYAALLALLEVSEAVLKKLSVRDNPAGLIALARMPERTLEEYRP-------------
1GZ0      rEdkrllpLihaleSQgVViQLAnRqyldeksdgAvhqgiiarvkpgRQYQENDLPDLIA

                 130     H7140       150       160   H8  170       180
          ----:----|----:----|----:----|----:----|----:----|----:----|
1IPA      -SPDALILVAVGLEKPGNLGAVLRSADAAGAEAVLVAGGVDLY-SPQVIRNSTGVVFSLR
1GZ0      SldqpFLLIldGVtdPhNLGAcLRSADAAGvhAVIVpkdrsaqLNatakKvAcGaaeSVp
                *******************************       ****************
                 190       200       210       220       230       240
          ----:----|----:----|----:----|----:----|----:----|----:----|
1IPA      TLAASESEVLDWIKQ---HNLPLVATTPHAEALYWEANLRPPVAIAVGPEHEGLRAAWLE
1GZ0      lIrvT--nLartMRmLQEeNIwIVgTageADhtlYQSkMtgrLALvMGaEgEGMRrltrE
          ****      ** **   * ******************** *******************
                 250       260 H13   270
          ----:----|----:----|----:----|----:--
1IPA      AAQTQVRIPMQG-QADSLNVSVSAALLLYEALRQRLL
1GZ0      hcdelIsIPMaGSvSs-LNVSVAtgIcLFEAVRQRs-
          ************  *  *******************
```

Figure VI-5A shows the initial global NW sequence alignment using default parameters (BLOSUM50), and the resulting standard and weighted superpositions are provided in Figure VI-5B and C. The final alignment generated by SE is shown in Figure VI-5D. Any residue pairs that received a weight of 0.5 or greater from the wRMSD calculation are noted with an asterisk. The underlined region of the sequence alignment in Figure VI-5A and D corresponds to a flexible domain between the proteins; as would be expected, none of these residues were significantly weighted to contribute to the superposition. The black boxes in Figure VI-5A indicate two regions of incorrect atom pairing. The first is due to an erroneous gap placement (in 1IPA) and corresponds to the residues of the denoted H8 α-helix in Figure VI-5B and C. The residues of the α-helix were not aligned properly, and hence, the appropriate Cα atoms were not paired together. However, after the weighted superposition, they are brought into close spatial proximity, and the final sequence alignment obtained by SE eliminates the gap to produce a correct pairing as evidenced by the high weights (Figure VI-5C). The β-sheet, noted in Figure VI-5B and C, is also a misalignment that is overcome by the wRMSD superposition. This initial error is caused by the default behavior of the Biopython parser [228], used to pull sequence information from the coordinates in the PDB files, which omits a modified methionine residue. While this is easily rectified programmatically, we allow the omission to serve as an example of parser error. Some parsers ignore non-standard amino acids (listed as HETATMs), and in the 1GZ0 structure, the methionines have been replaced with selenomethionine to aid in solving the structure. Once again, the structural superposition overrides the ambiguity, and the final alignment correctly pairs the beta sheet residues (with selenomethionine present this time due to the smarter parsing inherent to SE).

Correction of an alignment is made possible by the powerful combination of wRMSD-generated superposition and the “seed extension” algorithm used by SE to obtain a sequence alignment from a pair of protein structures. The SE method makes no inference about secondary structure elements of the aligned structures and considers residue similarity only in tie-breaking situations (using the BLOSUM62 scoring matrix to break ties). Additionally, the algorithm extends from a number of small “seed” pairings, so there are no gap penalties and no global cost optimization – two factors that are present in many structural alignment algorithms. The absence of these heuristics makes the SE algorithm a true What-You-See-Is-What-You-Get method for translating a structural superposition into a sequence alignment, and thus, it is a perfect fit for the HwRMSD protocol.

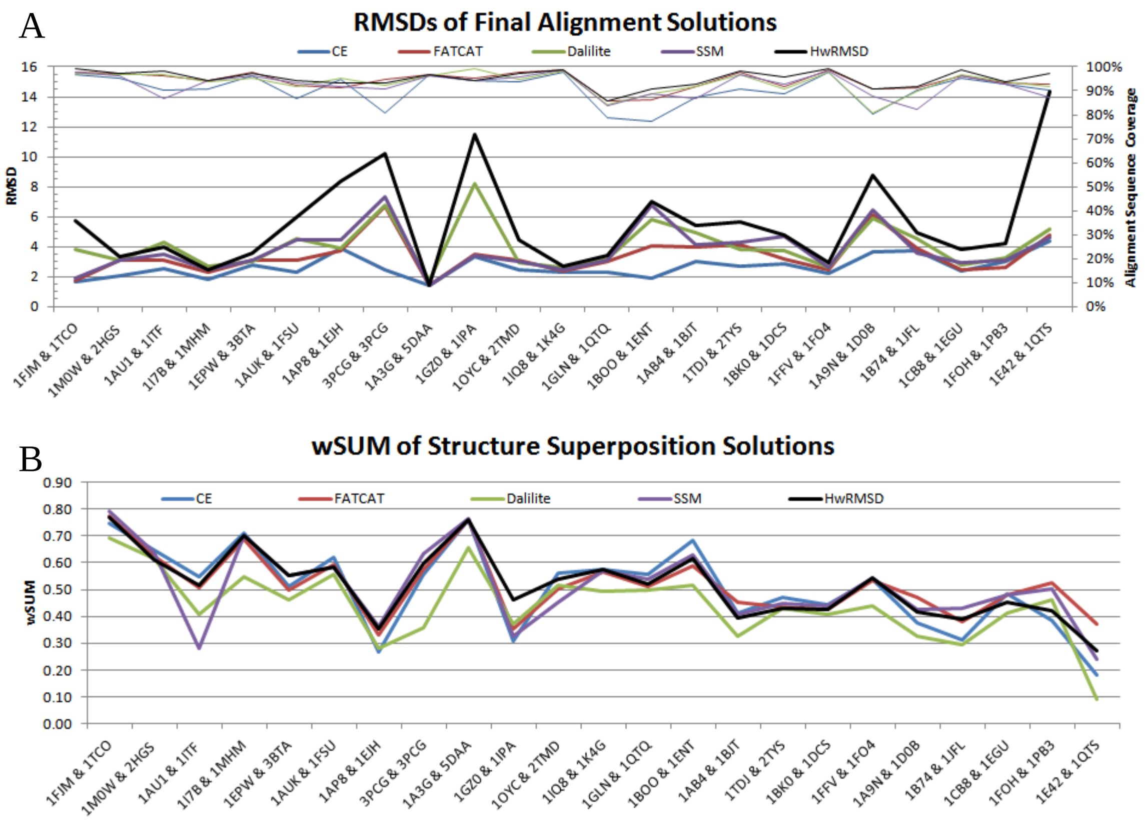
6.3.3 Comparison to Other Structural Alignment Methods

We compared the performance of our HwRMSD method against several leading structural alignment programs based on several benchmarks. First, we used the sequence alignments and structure solutions generated by structural alignment programs CE, FATCAT, DaliLite, and SSM to perform a robustness analysis similar to that performed with global and local sequence alignment methods described earlier. We then used the traditional raw RMSD metric and the previously described wSUM metric to directly compare the results of the programs to the final HwRMSD structural superpositions and structural alignments.

For the robustness analysis, we used the superposition and the final sequence alignment of a structural alignment program as the starting point for a weighted superposition (instead of the NW/SW sequence alignment and the standard superposition used earlier). As done previously, we computed the pair-wise RMSDs between the original solutions of CE, FATCAT, DaliLite, and SSM for each test pair of proteins before, and after, the weighted superposition. The median RMSDs for the original solutions ranged from 0.43 Å to 8.08 Å with 10 cases >1 Å, and the median RMSDs for the weighted alignments ranged from 0.021 Å to 8.46 Å with 6 cases >1 Å. Again, in most cases, the weighted fit had significantly lower median distance and was able to "collapse" the various solutions to one (or sometimes two very similar) consensus superpositions. Several of the test cases did not show a significant change in median distance between original and weighted solutions (Figure VI-3). In the cases of protocatechuate-3,4-dioxygenase[229] (3PCG), DNA topoiomerases [230, 231] (1AB4 and 1BJT), and the appendage domain pair [222, 223] (1E42 and 1QTS), the FATCAT program altered the solution structure by including a twist of a domain. This is an inherent functionality of the algorithm, which produced a structural superposition distinct from that of other algorithms. In the case of spliceosomal protein [220] and internalin B [221] (1A9N and 1D0B), the DaliLite algorithm produced an alternate, but quite reasonable solution, as the top result. This solution was equally far from the three other alignments, and thus, it increased the average distance. SSM also provided such an outlying solution in the case of the interferon homologs [232, 233] (1ITF and 1I7B). In each described case, removing the outlier solution greatly reduces the median distance between the remaining alignments, with the weighted solutions having a much smaller median distance. We chose to keep such solutions to demonstrate the variability of the results and the limits of any structural alignment approach. Again, this analysis does not indicate whether the final result of an HwRMSD alignment is better or worse than the other structural alignment programs; it is merely another illustration of the robustness of the weighted superposition method with respect to varied initial sequence alignments and superpositions.

A pair-wise RMSD score has long been used as a standard way to compare structural alignment solutions from different algorithms. While some structural alignment programs are tuned to minimize the final global RMSD, the HwRMSD method lets the structure dictate the alignment, so the distance between large similar domains is minimized. However, this is oftentimes accomplished at the expense of other, more flexible or more distantly related regions of the protein, thereby increasing the overall final RMSD measurement [7]. Figure VI-6A demonstrates the RMSD values of the final alignment solutions from CE, FATCAT, Dalilite, SSM, and wRMSD. Overall, the wRMSD solutions resulted in higher RMSD values as compared to the other tools in this study; however, aside from the half a dozen cases in which domain flexibility inflated the RMSD values, the wRMSD results were comparable to those of FATCAT and SSM. In our test cases, the CE solutions performed the best in terms of raw RMSD, but had slightly smaller alignment coverage than other methods in many of the cases (Figure VI-6).

Figure VI-6: Alignment results of HwRMSD (using BLOSUM50 global NW alignment) compared to other structural alignment programs using the RMSD A) and the wSUM B) metrics. Lower values of RMSD and higher values of wSUM represent a better alignment. Sequence similarity of the protein pairs decreases from left to right. Thin lines in A) indicate alignment coverage i.e. the number of residue pairs aligned with respect to the shortest protein chain.



Recent studies have pointed out the weakness of using the RMSD metric in cases where flexible or distantly related structures are considered [198, 199] and multiple alternate geometry-based scores have been proposed [198]. We chose to also compare HwRMSD to CE, FATCAT, DaliLite, and SSM using the previously described wSUM metric [7], which is the average weight of the paired residues. The weight is directly related to Cα-Cα distance via a scaling constant (set to 5 for all the comparative calculations); thus it represents an average similarity measure (see Appendix for more details). The wSUM metric is more informative for structural alignments than a pure

RMSD because it measures the extent of the ‘best aligned’ regions of a protein pair, rather than just the distance of all the paired atoms. If all the aligned residue pairs of two structures overlapped perfectly, the wSUM would be 1.0. For each of our test cases, the different programs generated very similar alignment lengths (Figure VI-6A), so even though the weights are normalized by the alignment length, the wSUM values are comparable among the methods.

CE, FATCAT, SSM, and HwRMSD all perform quite well (Figure VI-6B) with wSUM values ranging from 0.8 to 0.1. Values decreased as the %ID of the protein pairs decreased. DaliLite superpositions result in lowest wSUM values in general, with FATCAT, CE, and HwRMSD performing similarly well. Of course, the HwRMSD algorithm is built to minimize wRMSD which increases the weights, so we do not expect other structural alignment programs to get an optimal value in this metric. Rather, as seen here, we expect the HwRMSD result to be on par with other software in closely aligning a large portion of the protein in a structurally valid way. The wSUM performance provides confidence that the wRMSD algorithm and the HwRMSD protocol overall, not only overcomes the initial sequence alignment errors, but also produces structural superpositions that generate improved and valid alignments.

6.3.4 Local Alignments

Some alignment problems may have multiple solutions, especially if there are multiple similar domains that move relative to one another between the two structures being aligned (such as in the case of an apo vs. holo structure). The wRMSD algorithm can explore alternate multiple alignments by using a “local” alignment where sub-sets of the initial sequence alignment are used to produce multiple initial standard superpositions and, hence, multiple weighted superpositions. By using only small segments of the sequence alignment, the weighing is restricted to only a portion of the structure, allowing structural similarities that would have been washed out in the global alignment to drive the weighted superposition. The alternate solutions can then be ranked by wSUM to choose the best of these “local” alignments. The local alignment option is built into the current implementation of HwRMSD. The local alignment functionality is extensively described in the previous wRMSD publication [7].

6.4 Conclusion

We have now coupled our wRMSD method with a sequence alignment and seed extension (SE) algorithm. Our method is capable of preferentially selecting out the regions with the best structural agreement between homologous proteins and generating a superposition that can identify significant similarities and differences. The SE algorithm then generates a “corrected” sequence alignment based on the improved superposition. This algorithm combination, referred to as HwRMSD, provides a flexible and transparent structure alignment method. The HwRMSD technique can be used to superimpose homologs with low sequence identity and large conformational differences, an area where both sequence-based and structure-based methods may fail.

Employing homologs in the range of intermediate to low sequence identity, we have shown that applying a weighting term can overcome the dependence of a structural superposition on the initial sequence alignment used to determine the appropriate Cα pairs. The wRMSD superpositions are not significantly affected by the choice of the sequence alignment method or the employed parameters, but the sRMSD fits are highly dependent on both. The conserved regions of the structures are heavily weighted; thus, errors made in the initial sequence alignment are relatively discounted. The calculated weights can be used to determine potential mis-assignments in the initial sequence alignments. The wRMSD technique does not require prior knowledge of any protein system, and it removes the need to determine the best alignment method or parameters for each application. However, we must note that our tool, like any other, will breakdown when sequence or structural similarity is too low. Next, we aim to use this technique to align protein structures in our Binding MOAD database [108] to characterize ligand recognition across homologous families of protein structures.

The Appendix provides a short mathematical description of the technique and all calculated RMSD values used to determine the medians in Table VI-1 and Figure VI-3.

CHAPTER VII

Conclusion and Future Directions

The Binding MOAD database is the largest currently available source of protein-ligand binding sites with annotated ligand class and binding data. It is one of only several datasets cross-referenced by the PDB to provide binding data for individual protein-ligand complexes. However, the data in Binding MOAD is meant for more than a simple look-up of binding parameters. With proper mining and analysis, the database provides the scientific community with large, high-quality datasets for improving structure-based drug design methods. Regular updates to the data and the development of new functionalities can maintain Binding MOAD's relevance as a powerful scientific tool for exploring binding sites.

This thesis extends Binding MOAD to incorporate structural details of the binding sites currently annotated in the database. A robust relational-object model and efficient statistical routines were developed for mining this structural data in a dynamic and flexible way. Going forward, these extensions will be incorporated into the bindingmoad.org web server to make them available for public use. The binding-site data will be presented alongside the currently available information, and it will provide a user-friendly way to conduct binding-site composition analysis of the sort presented in Chapters III and IV of this thesis. These analyses can range from a simple display of residues contained in a single site to the calculation of residue propensities for a query-set of binding sites, with confidence intervals determined by sampling.

Specifically, the kinds of data views that will be available to the users are discussed below. Identities of residues contained in a specific binding site will be displayed, along with information about relative surface-accessibility of these residues. Visualization of a binding site will be provided either through an exportable PyMol script or through the Eolus Viewer that currently displays binding-site volume information. From an individual binding site, the user will be able to retrieve binding sites similar in composition, as determined by comparison of residue counts. This search can also be generalized to look for binding sites with custom patterns specifying the desired numbers of residues to be present.

A unique feature of Binding MOAD will be the ability to generate residue frequencies and propensities for an arbitrary set of binding sites. For example, a user may wish to see the propensities of residues for binding sites in a specific protein class or in proteins that bind the same ligand. Residue frequencies and propensities can be graphically and/or textually presented for any query returning a set of protein structures or for existing protein families and EC classes. The leave-10%-out sampling for proper interpretation of the propensities will be implicitly included in any such calculation. If desired, error bounds with respect to randomly-sampled propensities would be provided for any propensity calculation to let users to qualify the trends they obtained with their custom query. Of course, any query against the database will be subject to the various filtering criteria available based on Binding MOAD annotations. Current criteria include EC number, protein-source organism, ligand redundancy, resolution, and the presence of binding data. With respect to residue propensities the search criteria can be extended to include type of interaction (side chain or backbone), residue type, surface-accessibility, and/or binding-site size.

These features are only a few possible use cases for reporting binding-site data. Further work on the statistics of residue composition and propensities could allow Binding MOAD users to “drill down” into the data for more comparative analyses. The assessment of propensity variation with respect to ligand bias, and the significance of comparing binding-site composition are focus areas in need of further exploration.

Using the structural data obtained from Binding MOAD, the analysis of binding sites presented in this thesis reveals the broad trends in residue composition. In the spirit of a true top-down approach, a broad separation of binding sites is made using the biological relevance of the ligand as annotated in Binding MOAD. While further analyses of binding-site or ligand sub-classes are possible, we look to the broad trends of residue composition to give a backdrop for such classification analyses. We find that certain residues are over-represented in binding sites of biologically relevant “valid” ligands versus the sites of spurious “invalid” crystallographic additives. Trends like the over-representation of Arg in invalid ligand binding sites, or the relative scarcity of Lys in valid sites, stand up to tests that randomly shuffle the site labels. Examining the bias to propensity imparted by the types of ligands present in the PDB shows that propensities for valid sites retain overall trends even if the 20 most-frequent ligands are omitted from the calculation. The invalid ligand sets is much smaller and less diverse, and the omission of its most-frequent ligands from propensity calculations drastically alters the trends. We believe that valid binding-site propensities obtained by our analysis can help guide structure-based drug design in a way similar to previously-determined propensities of catalytic residues or protein-protein binding sites.

To examine the generality of the established propensity trends, their variation is explored systematically in random subsets of the data set. We observe that propensities for valid binding sites converge quicker than invalid due to the larger number of residues present in those sites. Moreover, we recommend that at least 1000 diverse protein complexes are needed for significant general conclusions for biologically relevant binding sites. While calculating propensities for sets below this size is appropriate for certain applications (e.g., those that focus on specific drug-binding sites or the analysis of a functional protein class), such calculations will not represent general trends of binding-site composition, given currently available structural data. Examining the propensity variation gives context to past and future studies attempting to calculate residue composition of protein surfaces. As the number of protein-ligand complexes in the PDB continues to grow, the available data for binding-site composition may strengthen or change the general trends. However, the propensity analysis presented in this thesis will still be applicable to evaluate the internal consistency of the trends in that data.

The scoring of binding-site predictions using residue-propensity data is one of the direct applications of general trends in binding-site composition. Geometry-based methods for binding-site prediction often make no assumption about the composition of the predicted pocket. The incorporation of a propensity-based score into the scoring schemes used to rank-order their predictions can help these methods identify potential pockets that contain residues frequently seen in biologically-relevant binding sites. This is a middle-of-the-road approach between *de novo* methods, based on pure geometry or energy criteria, and knowledge-based methods that explicitly compare predicted pockets to known binding sites. Chapter IV of this thesis demonstrates that a propensity-based score can perform as well as a native score of a binding-site prediction algorithm in ranking successful predictions. Moreover, for large proteins, where geometry-based algorithms may find several large binding-site-like pockets, a propensity-based or consensus score can significantly out-perform the native score in the success of top-ranked predictions.

Despite the success of propensity-based scoring for binding-site prediction, we find that the geometry-based algorithm we tested relied too heavily on the size of the binding site. The size of the predicted site was such an important factor for a successful prediction that the propensities only had impact in a limited number of cases. A better application of general residue propensities may lie in methods that rely on sub-structure matching for predicting protein regions relevant for ligand binding. For example, ProBiS [83] matches triplets of residue fragments from a query protein to a database of known sites, and it combines highly-conserved triplets to delineate potential binding sites. Since structurally-conserved residues may often appear outside of functional sites (such as surface patches repeatedly seen to bind invalid ligands with no known function), binding-site propensities can provide heuristics for weighing subs-structures containing residues known to preferentially participate in ligand binding. Using propensity-based scores or weights to improve the ProBiS algorithm is a potential future application of the general propensity trends. The aim will still be to discern binding-site-like regions from the rest of the protein surface, but the confounding effect of an additive score seen in the current study will be avoided.

Other knowledge-based algorithms for predicting binding sites may also benefit from binding-site composition information. Comparison of residue composition between a predicted site and propensities of classes of known sites is complementary to sequence-based motif-searching methods or geometric coordinate-matching methods employed by some current tools. However, more work needs to be done on the statistics involved in comparing propensities for small sets of binding sites, as well as the bias that ligand or protein similarity introduces into such comparisons. Proper comparisons to avoid system bias in ligands and/or proteins require careful statistical analysis and appropriate null scenarios for comparison. Also, a rigorous statistical framework will be required for establishing the significance of a residue-composition search against a database of known sites or sets of propensities. These are just some of the challenges that would be involved in shifting from general propensity trends to the comparison of propensities among small subsets of binding sites within the framework of the Binding MOAD database.

This thesis provides a framework and several guiding principles for the processing and analysis of binding sites in the Binding MOAD database. It takes a first step towards the understanding of trends of residue occurrence on protein surfaces and protein-ligand binding sites. Potential applications of these trends include, but are not limited to, the improvement of binding-site prediction methods and possibly binding-site comparison algorithms. The propensity trends and their applications are valuable contributions to the overarching goal of structure-based drug design; which hinges upon a thorough understanding of the general principals of protein-ligand binding.

Appendix

Supplementary Information for Chapter VI

A1. Additional Details of the HwRMSD procedure (see also: Damm & Carlson. 2000;49:457-466)

Given two proteins, X and Y, the two PDB files are parsed to compile a list of resolved residues in each crystal structure. Needle (EMBOSS) is used to align the sequences of X and Y. From the alignment, pairs of residues are matched and used in the overlay process. The superposition is based on paired Cα, but the code can easily be modified to incorporate more atoms if the user wishes.

First, the centers of mass of both proteins are placed at the origin, and a standard RMSD fit is used to give a rough, initial orientation for the overlay. Without this first step, the proteins would be too far apart and all weights would be zero. X and Y have n residues paired, and we calculate a Gaussian-weighting factor (w_n) for each pair, based on the distance between them.

$$w_n = e^{-(d_n)^2/c} \qquad (1)$$

where c is an arbitrary scaling factor and dn is determined as

$$d_n = \left((y_{nx}-x'_{nx})^2 + \left(y_{ny}-x'_{ny}\right)^2 + (y_{nz}-x'_{nz})^2\right)^{\frac{1}{2}} \qquad (2)$$

The scaling factor, c, was set to the RMSD of the initial sRMSD the weights need to have sufficient power to overcome the initial differences in the superposition. When the weights were calculated for alignments of other programs a scaling factor of c = 5 was used for consistency.

A wRMSD fit is an iterative process: after a rotation is applied to a protein, the distances between the residues change, which in turn changes the weights, which requires a recalculation. Convergence is straightforward. Each iteration starts by placing the Gaussian-weighted center of mass (wCM) of each protein at the origin.

$$wCM_x = \frac{\sum_n w_n m_n x_n}{\sum_n m_n} \quad \text{and} \qquad wCM_y = \frac{\sum_n w_n m_n y_n}{\sum_n m_n} \tag{3}$$

Weighting terms are used in the RMSD fit by simply incorporating them to the 3x3 covariance matrix (r_{ij}).

$$r_{ij} = \sum_n w_n y_{ni} x_{nj} \tag{4}$$

At this point, the rotation of the protein X onto Y is determined via the eigenvalues and eigenvectors of the square of the covariance matrix, as is standard practice. Rather than minimizing the sum of d_n^2, as is done in a standard RMSD fit, a wRMSD fit minimizes the sum of $w_n d_n^2$.

The goodness of fit can be measured in a sum of all weights. The maximum value occurs when all weighs are 1.0 and the sum is n (all atom pairs are perfectly overlaid). We write the sum of all weights (wSUM), normalized for the number of paired residues, as

$$wSUM = \frac{1}{n} \sum_n w_n \tag{5}$$

A2. Raw distances used to calculate medians for Table VI-1.

Test Cases	BLOSUM50 x BLOSUM62	BLOSUM50 x PAM120	BLOSUM50 x PAM250	BLOSUM62 x PAM120	BLOSUM62 x PAM250	PAM120 x PAM250	Median	BLOSUM50 x BLOSUM62	BLOSUM50 x PAM120	BLOSUM50 x PAM250	BLOSUM62 x PAM120	BLOSUM62 x PAM250	PAM120 x PAM250	Median
	Needleman-Wunsch (NW) Global Alignment Standard Fit							**Needleman-Wunsch (NW) Global Alignment Standard Fit**						
1FJM & 1TCO	1.16	0.11	0.29	1.18	1.24	0.18	**0.72**	0.01	0.00	0.03	0.01	0.03	0.04	**0.02**
1M0W & 2HGS	0.20	0.27	0.27	0.35	0.34	0.07	**0.27**	0.06	0.07	0.06	0.09	0.10	0.03	**0.07**
1AU1 & 1ITF	1.46	1.74	1.41	1.28	1.46	0.55	**1.43**	0.12	0.33	0.15	0.21	0.08	0.20	**0.17**
1I7B & 1MHM	0.33	0.10	0.15	0.38	0.37	0.13	**0.24**	0.02	0.02	0.04	0.04	0.06	0.03	**0.03**
1EPW & 3BTA	0.43	0.71	0.76	0.66	0.77	0.18	**0.69**	0.14	0.03	0.04	0.16	0.13	0.06	**0.09**
1AUK & 1FSU	0.23	0.71	0.34	0.72	0.39	0.69	**0.54**	0.09	0.11	0.11	0.03	0.08	0.10	**0.09**
1AP8 & 1EJH	0.00	0.23	0.23	0.23	0.23	0.00	**0.23**	0.00	0.05	0.05	0.05	0.05	0.00	**0.05**
3PCG & 3PCG	0.68	0.34	0.18	0.72	0.63	0.22	**0.48**	0.04	0.04	0.06	0.03	0.06	0.05	**0.04**
1A3G & 5DAA	0.58	0.61	1.90	1.01	2.18	1.43	**1.22**	0.12	0.07	0.17	0.13	0.07	0.15	**0.12**
1GZ0 & 1IPA	3.10		0.72		2.59		**2.59**	0.10		0.23		0.27		**0.23**
1OYC & 2TMD	0.30						**0.30**	0.08						**0.08**
1IQ8 & 1K4G	0.28	0.41	0.30	0.30	0.46	0.37	**0.34**	0.06	0.05	0.11	0.05	0.16	0.14	**0.08**

1GLN & 1QTQ	1.20		3.99		3.01		**3.01**	0.08		0.18		0.15		**0.15**
1BOO & 1EG2	2.38		2.73		3.44		**2.73**	0.10		0.19		0.21		**0.19**
1AB4 & 1BJT	0.48						**0.48**	0.43						**0.43**
1TDJ & 2TYS	1.51	7.32	4.08	7.00	3.68	8.50	**5.54**	0.17	4.94	0.74	4.91	0.77	4.41	**2.59**
1BK0 & 1DCS	0.92						**0.92**	0.51						**0.51**
1FFV & 1FO4	1.08		0.29		1.14		**1.08**	0.49		0.18		0.65		**0.49**
1A9N & 1D0B	2.85	13.60	0.83	12.81	2.88	13.25	**7.85**	3.56	15.00	1.19	13.22	4.39	15.93	**8.81**
1B74 & 1JFL	0.71		1.36		1.53		**1.36**	0.13		0.22		0.28		**0.22**
1CB8 & 1EGU	1.27		0.55		1.38		**1.27**	0.20		0.17		0.30		**0.20**
1FOH & 1PB3	4.83		2.21		5.38		**4.83**	1.83		2.63		3.59		**2.63**
1E42 & 1QTS	2.24		3.94		5.85		**3.94**	0.28		0.48		0.47		**0.47**
	Smith-Waterman (SW) Local Alignment Standard Fit							**Smith-Waterman (SW) Local Alignment Weighted Fit**						
1FJM & 1TCO	0.32	0.04	0.52	0.33	0.61	0.51	**0.42**	0.07	0.02	0.07	0.08	0.01	0.08	**0.07**
1M0W & 2HGS	0.19	0.27	0.28	0.34	0.34	0.06	**0.27**	0.06	0.06	0.06	0.10	0.10	0.02	**0.06**
1AU1 & 1ITF	0.00	1.57	1.48	1.57	1.48	0.28	**1.48**	0.00	0.11	0.18	0.11	0.18	0.13	**0.12**
1I7B & 1MHM	0.33	0.12	0.15	0.29	0.37	0.21	**0.25**	0.02	0.13	0.04	0.14	0.06	0.11	**0.08**
1EPW & 3BTA	0.43	0.62	0.64	0.64	0.65	0.91	**0.64**	0.14	0.04	0.04	0.11	0.13	0.05	**0.08**
1AUK & 1FSU	0.39	1.64	0.34	1.43	0.31	1.57	**0.91**	0.03	0.37	0.10	0.39	0.09	0.45	**0.24**
1AP8 & 1EJH		0.58	0.97	0.58	0.97	1.57	**0.97**	0.00	0.08	0.36	0.08	0.36	0.44	**0.22**
3PCG & 3PCG	0.19	1.37	0.18	1.33	0.17	1.37	**0.76**	0.03	0.06	0.06	0.03	0.05	0.05	**0.05**
1A3G & 5DAA	0.58	0.86	1.90	0.67	2.08	2.23	**1.38**	0.11	0.86	0.12	0.78	0.04	0.78	**0.45**
1GZ0 & 1IPA	3.30	7.54	7.54	5.83	5.83	5.83	**5.83**	0.10	1.31	1.31	1.30	1.30	1.30	**1.30**
1OYC & 2TMD	0.32	2.89	2.25	2.97	2.28	1.73	**2.26**	0.08	1.39	0.10	1.34	0.13	1.36	**0.73**
1IQ8 & 1K4G	0.29	1.16	0.30	1.22	0.48	1.05	**0.76**	0.07	0.85	0.11	0.83	0.17	0.92	**0.50**
1GLN & 1QTQ	0.46	4.96		5.30			**4.96**	0.14	1.95		1.86			**1.86**
1BOO & 1EG2	2.77	4.57	2.68	2.46	3.85	5.84	**3.31**	0.07	0.56	0.15	0.59	0.14	0.60	**0.36**
1AB4 & 1BJT	0.54	1.84	0.67	1.94	1.08	2.32	**1.46**	0.45	0.71	0.24	0.88	0.44	0.64	**0.55**
1TDJ & 2TYS	0.87	7.32		6.89			**6.89**	0.26	4.91		4.92			**4.91**
1BK0 & 1DCS	9.84	12.37	20.07	4.08	12.32	9.25	**11.08**	2.07	1.73	0.11	1.43	2.08	1.80	**1.77**
1FFV & 1FO4	0.65	1.59	0.77	1.81	0.84	1.76	**1.21**	0.60	0.73	0.28	1.18	0.82	0.48	**0.66**
1A9N & 1D0B	2.23	2.91	1.52	2.49	2.07	2.43	**2.33**	0.32	0.96	0.11	0.80	0.32	0.97	**0.56**
1B74 & 1JFL	1.11	1.52	0.85	2.14	1.51	1.73	**1.52**	0.07	0.76	0.33	0.76	0.30	0.70	**0.51**
1CB8 & 1EGU	1.30	2.02	0.73	1.14	1.34	1.91	**1.32**	0.24	1.01	0.25	0.92	0.43	1.07	**0.68**
1FOH & 1PB3	4.18	8.93	3.34	5.59	5.29	9.69	**5.44**	7.28	9.16	7.33	2.16	0.51	2.01	**4.72**
1E42 & 1QTS	2.03	6.90	2.80	7.73	4.12	6.91	**5.51**	0.45	7.51	0.15	7.76	0.39	7.53	**3.98**

A3. Raw RMSD and wSUM values used in Figure VI-6 for comparison of HwRMSD to other structural alignment programs.

Pair-wise RMSD (in Å) and wSUM* of HwRMSD and other Structural Alignment Program solutions

		wRMSD		CE		FATCATflex		DaliLite		SSM	
Protein Pair	**%ID**	**RMSD**	**wSUM**	**RMSD**	**wSUM**	**RMSD**	**wSUM**	**RMSD**	**wSUM**	**RMSD**	**wSUM**
1FJM & 1TCO	39%	5.74	0.77	1.69	0.75	1.75	0.78	3.79	0.69	1.90	0.79
1M0W & 2HGS	37%	3.36	0.61	2.07	0.65	3.09	0.62	3.06	0.61	3.08	0.64
1AU1 & 1ITF	35%	3.99	0.52	2.52	0.55	3.06	0.51	4.32	0.41	3.50	0.28
1I7B & 1MHM	33%	2.44	0.70	1.83	0.71	2.32	0.69	2.65	0.55	2.42	0.70
1EPW & 3BTA	31%	3.60	0.55	2.81	0.51	3.12	0.50	3.02	0.46	3.06	0.55
1AUK & 1FSU	29%	5.98	0.59	2.30	0.62	3.06	0.59	4.51	0.56	4.44	0.58
1AP8 & 1EJH	29%	8.39	0.35	3.88	0.27	3.70	0.33	3.92	0.28	4.41	0.36
3PCG (A) & 3PCG (M)	28%	10.22	0.60	2.44	0.56	6.68	0.57	6.76	0.36	7.30	0.63
1A3G & 5DAA	27%	1.44	0.76	1.44	0.76	1.43	0.76	1.71	0.66	1.39	0.77
1IPA & 1GZ0	26%	11.50	0.46	3.32	0.31	3.52	0.35	8.19	0.37	3.42	0.33
1OYC & 2TMD	25%	4.48	0.54	2.42	0.56	3.10	0.50	2.89	0.52	2.98	0.45
1IQ8 & 1K4G	25%	2.69	0.57	2.27	0.57	2.39	0.57	2.57	0.49	2.44	0.57
1GLN & 1QTQ	24%	3.38	0.52	2.27	0.56	3.00	0.51	3.18	0.50	3.12	0.54
1BOO & 2ENT	23%	7.03	0.62	1.88	0.69	4.02	0.59	5.80	0.52	6.78	0.63
1AB4 & 1BJT	22%	5.40	0.39	3.03	0.41	3.95	0.45	4.90	0.32	4.12	0.41
1TDJ & 2TYS	21%	5.66	0.43	2.70	0.47	4.12	0.43	3.85	0.43	4.32	0.45
1BK0 & 1DCS	20%	4.77	0.42	2.81	0.44	3.20	0.43	3.73	0.41	4.68	0.44
1FFV & 1FO4	19%	2.97	0.54	2.23	0.54	2.48	0.53	2.59	0.44	2.59	0.54
1A9N & 1D0B	19%	8.77	0.42	3.69	0.37	6.16	0.47	5.87	0.32	6.43	0.43
1B74 & 1JFL	18%	4.96	0.39	3.75	0.31	3.90	0.38	4.57	0.29	3.54	0.43
1CB8 & 1EGU	18%	3.78	0.45	2.37	0.48	2.46	0.48	2.73	0.41	2.96	0.48
1FOH & 1PBE	17%	4.18	0.42	3.02	0.39	2.63	0.52	3.23	0.46	3.11	0.50
1E42 & 1QTS	16%	14.38	0.27	4.35	0.18	4.74	0.37	5.15	0.09	4.62	0.24

*wSUM is the sum of distance-dependent weights for all the corresponding atom pairs in a structural alignment solution. Higher values indicate a higher fraction of pairs aligned, with the value of 1 signifying all atom pairs perfectly overlaid.

BIBLIOGRAPHY

1. Weigelt, J., et al., *Structural genomics and drug discovery: all in the family.* Current Opinion in Chemical Biology, 2008. **12**(1): p. 32-39.
2. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.
3. Hajduk, P., J. Huth, and S. Fesik, *Druggability Indices for Protein Targets Derived from NMR-Based Screening Data.* Journal of Medicinal Chemistry, 2005. **48**(7): p. 2518-2525.
4. Milletti, F. and A. Vulpetti, *Predicting Polypharmacology by Binding Site Similarity: From Kinases to the Protein Universe.* Journal of Chemical Information and Modeling, 2010. **50**(8): p. 1418-1431.
5. Haupt, J. and M. Schroeder, *Old friends in new guise: repositioning of known drugs with structural bioinformatics.* Briefings in Bioinformatics, 2011. **12**(4): p. 312-326.
6. Morin, A., J. Meiler, and L.S. Mizoue, *Computational design of protein–ligand interfaces: potential in therapeutic development.* Trends in Biotechnology, 2011. **29**(4): p. 159-166.
7. Damm, K.L. and H.A. Carlson, *Gaussian-Weighted RMSD Superposition of Proteins: A Structural Comparison for Flexible Proteins and Predicted Protein Structures.* Biophysical Journal, 2006. **90**(12): p. 4558-4573.
8. Feldman-Salit, A., R.C. Wade, and T.P. Begley, *Molecular Recognition: Computational Analysis and Modelling*, in *Wiley Encyclopedia of Chemical Biology*. 2007, John Wiley & Sons, Inc.
9. Cavasotto, C.N. and S.S. Phatak, *Homology modeling in drug discovery: current trends and applications.* Drug Discovery Today, 2009. **14**(13-14): p. 676-683.
10. Kahraman, A., et al., *Shape Variation in Protein Binding Pockets and their Ligands.* Journal of Molecular Biology, 2007. **368**(1): p. 283-301.
11. Henrich, S., et al., *Computational approaches to identifying and characterizing protein binding sites for ligand design.* J. Mol. Recognit., 2010. **23**(2): p. 209-219.
12. Smith, R.D., et al., *CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions.* Journal of Chemical Information and Modeling, 2011. **51**(9): p. 2115-2131.
13. An, J., M. Totrov, and R. Abagyan, *Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes.* Molecular & Cellular Proteomics, 2005. **4**(6): p. 752-761.
14. Hert, J., et al., *Quantifying the Relationships among Drug Classes.* Journal of Chemical Information and Modeling, 2008. **48**(4): p. 755-765.
15. Keiser, M., et al., *Predicting new molecular targets for known drugs.* Nature, 2009. **462**(7270): p. 175-181.

16. Petrey, D., M. Fischer, and B. Honig, *Structural relationships among proteins with different global topologies and their implications for function annotation strategies.* Proceedings of the National Academy of Sciences, 2009. **106**(41): p. 17377-17382.
17. Coleman, R. and K. Sharp, *Protein Pockets: Inventory, Shape, and Comparison.* Journal of Chemical Information and Modeling, 2010. **50**(4): p. 589-603.
18. Nayal, M. and B. Honig, *On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites.* Proteins: Structure, Function, and Bioinformatics, 2006. **63**(4): p. 892-906.
19. Smith, R.D., et al., *Exploring protein–ligand recognition with Binding MOAD.* Journal of Molecular Graphics and Modelling, 2006. **24**(6): p. 414-425.
20. Kawabata, T. and N. Go, *Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites.* Proteins, 2007. **68**(2): p. 516-529.
21. Pérot, S., et al., *Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery.* Drug Discovery Today, 2010. **15**(15-16): p. 656-667.
22. Carlson, H., et al., *Differences between High- and Low-Affinity Complexes of Enzymes and Nonenzymes.* Journal of Medicinal Chemistry, 2008. **51**: p. 6432-6441.
23. Gohlke, H., M. Hendlich, and G. Klebe, *Knowledge-based scoring function to predict protein-ligand interactions.* Journal of Molecular Biology, 2000. **295**(2): p. 337-356.
24. Huang, S.-Y. and X. Zou, *An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials.* Journal of Computational Chemistry, 2006. **27**(15): p. 1866-1875.
25. Muegge, I., *PMF Scoring Revisited.* Journal of Medicinal Chemistry, 2006. **49**(20): p. 5895-5902.
26. DeWitte, R.S. and E.I. Shakhnovich, *SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence.* Journal of the American Chemical Society, 1996. **118**(47): p. 11733-11744.
27. Schapira, M., M. Totrov, and R. Abagyan, *Prediction of the binding energy for small molecules, peptides and proteins.* J Mol Recognit, 1999. **12**(3): p. 177-90.
28. Stahl, M. and M. Rarey, *Detailed Analysis of Scoring Functions for Virtual Screening.* Journal of Medicinal Chemistry, 2001. **44**(7): p. 1035-1042.
29. Lins, L., A. Thomas, and R. Brasseur, *Analysis of accessible surface of residues in proteins.* Protein Science, 2003. **12**(7): p. 1406-1417.
30. Moelbert, S., E. Emberly, and C. Tang, *Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins.* Protein Sci, 2004. **13**(3): p. 752-762.
31. Bartlett, G.J., et al., *Analysis of Catalytic Residues in Enzyme Active Sites.* Journal of Molecular Biology, 2002. **324**(1): p. 105-121.
32. Davis, F. and A. Sali, *The Overlap of Small Molecule and Protein Binding Sites within Families of Protein Structures.* PLoS Comput Biol, 2010. **6**(2): p. e1000668.

33. Tseng, Y. and J. Liang, *Predicting Enzyme Functional Surfaces and Locating Key Residues Automatically from Structures.* Annals of Biomedical Engineering, 2007. **35**(6): p. 1037-1042.

34. Lopez, G., A. Valencia, and M. Tress, *FireDB—a database of functionally important residues from proteins of known structure.* Nucleic Acids Research, 2007. **35**(suppl 1): p. D219-D223.

35. Liang, J., H. Edelsbrunner, and C. Woodward, *Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.* Protein Sci, 1998. **7**: p. 1884 - 1897.

36. Laurie, A. and R. Jackson, *Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening.* Current protein & peptide science, 2006. **7**(5): p. 395-406.

37. Ansari, H. and G. Raghava, *Identification of NAD interacting residues in proteins.* BMC Bioinformatics, 2010. **11**(1): p. 160.

38. Kellenberger, E., C. Schalon, and D. Rognan, *How to Measure the Similarity Between Protein Ligand-Binding Sites?* Current Computer - Aided Drug Design, 2008. **4**(3): p. 209-220.

39. Levitt, D.G. and L.J. Banaszak, *POCKET: A computer graphies method for identifying and displaying protein cavities and their surrounding amino acids.* Journal of Molecular Graphics, 1992. **10**(4): p. 229-234.

40. Hendlich, M., F. Rippmann, and G. Barnickel, *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.* Journal of Molecular Graphics and Modelling, 1997. **15**(6): p. 359-363.

41. Connolly, M., *Analytical molecular surface calculation.* Journal of Applied Crystallography, 1983. **16**(5): p. 548-558.

42. Huang, B. and M. Schroeder, *LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.* BMC Structural Biology, 2006. **6**(1): p. 19.

43. Laurie, A. and R. Jackson, *Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.* Bioinformatics, 2005. **21**: p. 1908 - 1916.

44. Weisel, M., E. Proschak, and G. Schneider, *PocketPicker: analysis of ligand binding-sites with shape descriptors.* Chemistry Central Journal, 2007. **1**(1): p. 7.

45. Kleywegt, G.J. and T.A. Jones, *Detection, delineation, measurement and display of cavities in macromolecular structures.* Acta Crystallographica Section D, 1994. **50**(2): p. 178-185.

46. Kalidas, Y. and N. Chandra, *PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins.* Journal of Structural Biology, 2008. **161**(1): p. 31-42.

47. Wang, R., et al., *The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures.* Journal of Medicinal Chemistry, 2004. **47**(12): p. 2977-2980.

48. Laskowski, R., *SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions.* J Mol Graph, 1995. **13**: p. 323 - 330.

49. Laskowski, R., et al., *Protein clefts in molecular recognition and function.* Protein Science, 1996. **5**(12): p. 2438 - 2452.

50. Brady, G. and P. Stouten, *Fast prediction and visualization of protein binding pockets with PASS.* J Comput Aided Mol Des, 2000. **14**: p. 383 - 401.

51. Aurenhammer, F., *Voronoi diagrams\&mdash;a survey of a fundamental geometric data structure.* ACM Comput. Surv., 1991. **23**(3): p. 345-405.

52. Lee, D.T. and B.J. Schachter, *Two algorithms for constructing a Delaunay triangulation.* International Journal of Parallel Programming, 1980. **9**(3): p. 219-242.

53. Peters, K.P., J. Fauck, and C. Frömmel, *The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria.* J Mol Biol, 1996. **256**(1): p. 201-213.

54. Liang, J., C. Woodward, and H. Edelsbrunner, *Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design.* Protein Science, 1998. **7**(9): p. 1884-1897.

55. Le Guilloux, V., P. Schmidtke, and P. Tuffery, *Fpocket: An open source platform for ligand pocket detection.* BMC Bioinformatics, 2009. **10**(1): p. 168.

56. Group, T.C.C., *Molecular Operating Environment*, The Chemical Computing Group.

57. Schmidtke, P., et al., *Large-Scale Comparison of Four Binding Site Detection Algorithms.* Journal of Chemical Information and Modeling, 2010. **50**(12): p. 2191-2200.

58. Goodford, P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules.* Journal of Medicinal Chemistry, 1985. **28**(7): p. 849-857.

59. http://www.moldiscovery.com/. *Molecular Discovery*. 2011; Available from: http://www.moldiscovery.com/.

60. Nissink, J., et al., *A new test set for validating predictions of protein-ligand interaction.* Proteins, 2002. **49**: p. 457 - 471.

61. Halgren, T., *Identifying and Characterizing Binding Sites and Assessing Druggability.* Journal of Chemical Information and Modeling, 2009. **49**(2): p. 377-389.

62. Vajda, S. and F. Guarnieri, *Characterization of protein-ligand interaction sites using experimental and computational methods.* Current Opinion in Drug Discovery and Development, 2006. **9**(3): p. 354 - 362.

63. Clark, M., et al., *Grand Canonical Monte Carlo Simulation of Ligand−Protein Binding.* Journal of Chemical Information and Modeling, 2005. **46**(1): p. 231-242.

64. Chang, D.T.-H., Y.-J. Oyang, and J.-H. Lin, *MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm.* Nucleic Acids Research. **33**(suppl 2): p. W233-W238.

65. Hetényi, C. and D. van der Spoel, *Blind docking of drug-sized compounds to proteins with up to a thousand residues.* FEBS Letters, 2006. **580**(5): p. 1447-1450.

66. Kuhn, D., et al., *Functional Classification of Protein Kinase Binding Sites Using Cavbase.* ChemMedChem, 2007. **2**(10): p. 1432-1447.

67. Kuhn, D., et al., *From the similarity analysis of protein cavities to the functional classification of protein families using cavbase.* Journal of Molecular Biology, 2006. **359**(4): p. 1023-1044.

68. Hendlich, M., F. Rippmann, and G. Barnickel, *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.* J Mol Graph Model, 1997. **15**(6): p. 359 - 363.

69. Laskowski, R.A., J.D. Watson, and J.M. Thornton, *ProFunc: a server for predicting protein function from 3D structure.* Nucleic Acids Research, 2005. **33**(suppl 2): p. W89-W93.

70. Pupko, T., et al., *Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues.* Bioinformatics, 2002. **18**: p. s71 - s77.

71. Mayrose, I., et al., *Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Empirical Bayesian Methods Are Superior.* Molecular Biology and Evolution, 2004. **21**(9): p. 1781-1791.

72. Lichtarge, O., H.R. Bourne, and F.E. Cohen, *An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families.* Journal of Molecular Biology, 1996. **257**(2): p. 342-358.

73. Eric, M., *Protein Explorer: easy yet powerful macromolecular visualization.* Trends in Biochemical Sciences, 2002. **27**(2): p. 107-109.

74. Thornton, J.M., et al., *From structure to function: Approaches and limitations.* Nat Struct Mol Biol.

75. Sigrist, C.J.A., et al., *PROSITE: A documented database using patterns and profiles as motif descriptors.* Briefings in Bioinformatics, 2002. **3**(3): p. 265-274.

76. Sigrist, C.J.A., et al., *PROSITE, a protein domain database for functional characterization and annotation.* Nucleic Acids Research, 2010. **38**(suppl 1): p. D161-D166.

77. Brylinski, M. and J. Skolnick, *A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation.* Proceedings of the National Academy of Sciences, 2008. **105**(1): p. 129-134.

78. Laskowski, R.A., J.D. Watson, and J.M. Thornton, *Protein Function Prediction Using Local 3D Templates.* Journal of Molecular Biology, 2005. **351**(3): p. 614-626.

79. Kinoshita, K., J.i. Furui, and H. Nakamura, *Identification of protein functions from a molecular surface database, e{F}-site.* J Struct Funct Genomics, 2002. **2**(1): p. 9-22.

80. Kinoshita, K. and H. Nakamura, *Identification of the ligand binding sites on the molecular surface of proteins.* Protein Science, 2005. **14**(3): p. 711-718.

81. Shulman-Peleg, A., R. Nussinov, and H.J. Wolfson, *Recognition of Functional Sites in Protein Structures.* Journal of Molecular Biology, 2004. **339**(3): p. 607-633.

82. Shulman-Peleg, A., R. Nussinov, and H.J. Wolfson, *SiteEngines: recognition and comparison of binding sites and protein–protein interfaces.* Nucleic Acids Research, 2005. **33**(suppl 2): p. W337-W341.

83. Konc, J. and D. Janežič, *ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment.* Bioinformatics, 2010. **26**(9): p. 1160-1168.

84. Konc, J. and D. Janežič, *ProBiS: a web server for detection of structurally similar protein binding sites.* Nucleic Acids Research, 2010. **38**(suppl 2): p. W436-W440.

85. Stark, A. and R.B. Russell, *Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures.* Nucleic Acids Research, 2003. **31**(13): p. 3341-3344.
86. Stark, A., S. Sunyaev, and R.B. Russell, *A Model for Statistical Significance of Local Similarities in Structure.* Journal of Molecular Biology, 2003. **326**(5): p. 1307-1316.
87. Jones, G., et al., *Development and validation of a genetic algorithm for flexible docking.* Journal of Molecular Biology, 1997. **267**(3): p. 727-748.
88. Puvanendrampillai, D. and J.B.O. Mitchell, *Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein–ligand complexes.* Bioinformatics, 2003. **19**(14): p. 1856-1857.
89. Lee, B. and F.M. Richards, *The interpretation of protein structures: Estimation of static accessibility.* Journal of Molecular Biology, 1971. **55**(3): p. 379-IN4.
90. Shrake, A. and J.A. Rupley, *Environment and exposure to solvent of protein atoms. Lysozyme and insulin.* Journal of Molecular Biology, 1973. **79**(2): p. 351-371.
91. Hubbard, S., *NACCESS.* 1996.
92. Sanner, M.F., A.J. Olson, and J.-C. Spehner, *Reduced surface: An efficient way to compute molecular surfaces.* Biopolymers, 1996. **38**(3): p. 305-320.
93. Tsodikov, O.V., M.T. Record, and Y.V. Sergeev, *Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature.* Journal of Computational Chemistry, 2002. **23**(6): p. 600-609.
94. Verdonk, M.L., et al., *Improved protein–ligand docking using GOLD.* Proteins: Structure, Function, and Bioinformatics, 2003. **52**(4): p. 609-623.
95. Ewing, T., et al., *DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases.* Journal of Computer-Aided Molecular Design, 2001. **15**(5): p. 411-428.
96. Morris, G., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* J. Comput. Chem., 1998. **19**(14): p. 1639-1662.
97. Cornell, W.D., et al., *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.* Journal of the American Chemical Society, 1995. **117**(19): p. 5179-5197.
98. Brooks, B., et al., *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.* J. Comput. Chem., 1983. **4**(2): p. 187-217.
99. Rarey, M., et al., *A Fast Flexible Docking Method using an Incremental Construction Algorithm.* Journal of Molecular Biology, 1996. **261**(3): p. 470-489.
100. Wang, R., L. Lai, and S. Wang, *Further development and validation of empirical scoring functions for structure-based binding affinity prediction.* Journal of Computer-Aided Molecular Design, 2002. **16**(1): p. 11-26.
101. Feher, M., *Consensus scoring for protein–ligand interactions.* Drug Discovery Today, 2006. **11**(9-10): p. 421-428.
102. Wang, R., Y. Lu, and S. Wang, *Comparative Evaluation of 11 Scoring Functions for Molecular Docking.* Journal of Medicinal Chemistry, 2003. **46**(12): p. 2287-2303.

103. Dunbar, J.B., et al., *CSAR Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes.* Journal of Chemical Information and Modeling, 2011. **51**(9): p. 2036-2046.

104. Warren, G.L., et al., *A Critical Assessment of Docking Programs and Scoring Functions.* Journal of Medicinal Chemistry, 2005. **49**(20): p. 5912-5931.

105. Leach, A.R., B.K. Shoichet, and C.E. Peishoff, *Prediction of Protein−Ligand Interactions. Docking and Scoring: Successes and Gaps.* Journal of Medicinal Chemistry, 2006. **49**(20): p. 5851-5855.

106. Hu, L., et al., *Binding MOAD (Mother Of All Databases).* Proteins: Structure, Function, and Bioinformatics, 2005. **60**(3): p. 333-340.

107. Benson, M., et al., *Binding MOAD, a high-quality protein ligand database.* Nucl. Acids Res., 2007. **36**(Database issue): p. gkm911.

108. Benson, M.L., et al., *Binding MOAD, a high-quality protein–ligand database.* Nucleic Acids Research, 2008. **36**(suppl 1): p. D674-D678.

109. Liu, T., et al., *BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities.* Nucleic Acids Research, 2007. **35**(suppl 1): p. D198-D201.

110. Wang, R., et al., *The PDBbind Database: Methodologies and Updates.* Journal of Medicinal Chemistry, 2005. **48**(12): p. 4111-4119.

111. Porter, C.T., G.J. Bartlett, and J.M. Thornton, *The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.* Nucleic Acids Research, 2004. **32**(suppl 1): p. D129-D133.

112. Golovin, A. and K. Henrick, *MSDmotif: exploring protein sites and motifs.* BMC Bioinformatics, 2008. **9**(1): p. 312.

113. Raush, E., et al., *A New Method for Publishing Three-Dimensional Content.* PLoS One, 2009. **4**(10): p. e7394.

114. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Research, 1997. **25**(17): p. 3389-3402.

115. PDB. *Chemical Component Dictionary*. 2010; Available from: http://www.wwpdb.org/ccd.html.

116. Delano, W., *The PyMOL Molecular Graphics System.* 2002.

117. Holliday, G., J. Mitchell, and J. Thornton, *Understanding the Functional Roles of Amino Acid Residues in Enzyme Catalysis.* Journal of Molecular Biology, 2009. **390**(3): p. 560-577.

118. Yahalom, R., et al., *Structure-based identification of catalytic residues.* Proteins, 2011. **79**(6): p. 1952-1963.

119. Imai, Y., Y. Inoue, and Y. Yamamoto, *Propensities of polar and aromatic amino acids in noncanonical interactions: nonbonded contacts analysis of protein-ligand complexes in crystal structures.* Journal of Medicinal Chemistry, 2007. **50**(6): p. 1189-1196.

120. Ghersi, D. and R. Sanchez, *Beyond structural genomics: computational approaches for the identification of ligand binding sites in protein structures.* Journal of Structural and Functional Genomics, 2011. **12**(2): p. 109-117.

121. Soga, S., et al., *Use of Amino Acid Composition to Predict Ligand-Binding Sites.* Journal of Chemical Information and Modeling, 2007. **47**(2): p. 400-406.

122. Mehio, W., et al., *Identification of protein binding surfaces using surface triplet propensities.* Bioinformatics, 2010. **26**(20): p. 2549-2555.

123. Miller, S., et al., *Interior and surface of monomeric proteins.* Journal of Molecular Biology, 1987. **196**(3): p. 641-656.

124. Jones, S. and J.M. Thornton, *Analysis of protein-protein interaction sites using surface patches.* Journal of Molecular Biology, 1997. **272**(1): p. 121-132.

125. Cheung, M.S., A.E. García, and J.N. Onuchic, *Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse.* Proceedings of the National Academy of Sciences, 2002. **99**(2): p. 685-690.

126. Rank, J.A. and D. Baker, *A desolvation barrier to hydrophobic cluster formation may contribute to the rate-limiting step in protein folding.* Protein Science, 1997. **6**(2): p. 347-354.

127. Johan, W., *Structural genomics—Impact on biomedicine and drug discovery.* Experimental Cell Research, 2010. **316**(8): p. 1332-1338.

128. Bostrom, J., A. Hogner, and S. Schmitt, *Do Structurally Similar Ligands Bind in a Similar Fashion?* Journal of Medicinal Chemistry, 2006. **49**(23): p. 6716-6725.

129. Kupas, K., A. Ultsch, and G. Klebe, *Large scale analysis of protein-binding cavities using self-organizing maps and wavelet-based surface patches to describe functional properties, selectivity discrimination, and putative cross-reactivity.* Proteins, 2008. **71**(3): p. 1288-1306.

130. Keiser, M., J. Irwin, and B. Shoichet, *The chemical basis of pharmacology.* Biochemistry, 2010. **49**(48): p. 10267-10276.

131. Halgren, T., *New Method for Fast and Accurate Binding-site Identification and Analysis.* Chemical Biology & Drug Design, 2007. **69**(2): p. 146-148.

132. Glaser, F., et al., *A method for localizing ligand binding pockets in protein structures.* Proteins, 2006. **62**: p. 479 - 488.

133. Kalinina, O., M. Gelfand, and R. Russell, *Combining specificity determining and conserved residues improves functional site prediction.* BMC Bioinformatics, 2009. **10**(1): p. 174.

134. Nagao, C., N. Nagano, and K. Mizuguchi, *Relationships between functional subclasses and information contained in active-site and ligand-binding residues in diverse superfamilies.* Proteins: Structure, Function, and Bioinformatics, 2010. **78**(10): p. 2369-2384.

135. Soga, S., et al., *Identification of the Druggable Concavity in Homology Models Using the PLB Index.* Journal of Chemical Information and Modeling, 2007. **47**(6): p. 2287-2292.

136. Schmidtke, P., et al., *Large-Scale Comparison of Four Binding Site Detection Algorithms.* Journal of chemical information and modeling, 2010. **0**(0).

137. Soga, S., et al., *Chemocavity: Specific Concavity in Protein Reserved for the Binding of Biologically Functional Small Molecules.* Journal of Chemical Information and Modeling, 2008. **48**(8): p. 1679-1685.

138. Shoichet, B., *Lead discovery using molecular docking.* Current Opinion in Chemical Biology, 2002. **6**(4): p. 439-446.

139. Brooijmans, N. and I.D. Kuntz, *Molecular recognition and docking algorithms.* Annual Review Biophysics Biomolecular Structure, 2003. **32**: p. 335-373.

140. Halperin, I., et al., *Principles of docking: an overview of search algorithms and a guide to scoring functions.* Proteins, 2002. **47**: p. 409 - 443.

141. Joseph-McCarthy, D., et al., *Lead optimization via high-throughput molecular docking.* Current opinion in drug discovery & development, 2007. **10**(3): p. 264-274.

142. Rajamani, R. and A. Good, *Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development.* Current opinion in drug discovery & development, 2007. **10**(3): p. 308-315.

143. Seifert, M., J. Kraus, and B. Kramer, *Virtual high-throughput screening of molecular databases.* Current opinion in drug discovery & development, 2007. **10**(3): p. 298-307.

144. Case, D., et al., *The Amber biomolecular simulation programs.* Journal of Computational Chemistry, 2005. **26**(16): p. 1668-1688.

145. Wang, W., et al., *BIOMOLECULAR SIMULATIONS: Recent Developments in Force Fields, Simulations of Enzyme Catalysis, Protein-Ligand, Protein-Protein, and Protein-Nucleic Acid Noncovalent Interactions.* Annual Review of Biophysics and Biomolecular Structure, 2001. **30**(1): p. 211-243.

146. Rocchia, W., et al., *Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects.* Journal of Computational Chemistry, 2002. **23**(1): p. 128-137.

147. Grant, A., B. Pickup, and A. Nicholls, *A smooth permittivity function for Poisson–Boltzmann solvation methods.* J. Comput. Chem., 2001. **22**(6): p. 608-640.

148. Baker, N.A., et al., *Electrostatics of nanosystems: application to microtubules and the ribosome.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(18): p. 10037-10041.

149. Wei, B., et al., *A Model Binding Site for Testing Scoring Functions in Molecular Docking.* Journal of Molecular Biology, 2002. **322**(2): p. 339-355.

150. Chen, J., C. Brooks, and J. Khandogin, *Recent advances in implicit solvent-based methods for biomolecular simulations.* Current Opinion in Structural Biology, 2008. **18**(2): p. 140-148.

151. Zou, X., Yaxiong, and I. Kuntz, *Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model.* Journal of the American Chemical Society, 1999. **121**(35): p. 8033-8043.

152. Liu, H.-Y. and X. Zou, *Electrostatics of Ligand Binding: Parametrization of the Generalized Born Model and Comparison with the Poisson−Boltzmann Approach.* The Journal of Physical Chemistry B, 2006. **110**(18): p. 9304-9313.

153. Liu, H.-Y., I. Kuntz, and X. Zou, *Pairwise GB/SA Scoring Function for Structure-based Drug Design.* The Journal of Physical Chemistry B, 2004. **108**(17): p. 5453-5462.

154. Meng, E., B. Shoichet, and I. Kuntz, *Automated docking with grid-based energy evaluation.* J. Comput. Chem., 1992. **13**(4): p. 505-524.

155. Jain, A., *Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities.* Journal of Computer-Aided Molecular Design, 1996. **10**(5): p. 427-440.

156. Head, R., et al., *VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands.* Journal of the American Chemical Society, 1996. **118**(16): p. 3959-3969.

157. Eldridge, M., et al., *Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes.* Journal of Computer-Aided Molecular Design, 1997. **11**(5): p. 425-445.

158. Böhm, H.-J., *The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure.* Journal of Computer-Aided Molecular Design, 1994. **8**(3): p. 243-256.

159. Böhm, H.-J., *Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs.* Journal of Computer-Aided Molecular Design, 1998. **12**(4): p. 309-309.

160. Sippl, M., et al., *An attempt to analyse progress in fold recognition from CASP1 to CASP3.* Proteins: Structure, Function, and Genetics, 1999. **37**(S3): p. 226-230.

161. Vajda, S., *Empirical potentials and functions for protein folding and binding.* Current Opinion in Structural Biology, 1997. **7**(2): p. 222-228.

162. Verkhivker, G., et al., *Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity.* Protein Engineering, Design and Selection, 1995. **8**(7): p. 677-691.

163. Mitchell, J., et al., *BLEEP - potential of mean force describing protein-ligand interactions: I. Generating potential.* Journal of Computational Chemistry, 1999. **20**(11): p. 1165-1176.

164. Brown, M., et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(1): p. 262-267.

165. Zhang, C., et al., *A Knowledge-Based Energy Function for Protein−Ligand, Protein−Protein, and Protein−DNA Complexes.* Journal of Medicinal Chemistry, 2005. **48**(7): p. 2325-2335.

166. Muegge, I. and Y.C. Martin, *A general and fast scoring function for protein-ligand interactions: a simplified potential approach.* Journal of Medicinal Chemistry, 1999. **42**(5): p. 791-804.

167. Huang, S.-Y. and X. Zou, *An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function.* Journal of Computational Chemistry, 2006. **27**(15): p. 1876-1882.

168. DeWitte, R. and E. Shakhnovich, *SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence.* Journal of the American Chemical Society, 1996. **118**(47): p. 11733-11744.

169. Huey, R., et al., *A semiempirical free energy force field with charge-based desolvation.* J. Comput. Chem., 2007. **28**(6): p. 1145-1152.

170. Huang, S.-Y. and X. Zou, *Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking.* Proteins: Structure, Function, and Bioinformatics, 2007. **66**(2): p. 399-421.

171. Pettersen, E., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* Journal of Computational Chemistry, 2004. **25**(13): p. 1605-1612.

172. Gasteiger, J. and M. Marsili, *Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges.* Tetrahedron, 1980. **36**(22): p. 3219-3228.

173. Christianson, D.W., *Structural biology of zinc.* Advances in protein chemistry, 1991. **42**: p. 281-355.

174. Gutteridge, A. and J. Thornton, *Conformational Changes Observed in Enzyme Crystal Structures upon Substrate Binding.* Journal of Molecular Biology, 2005. **346**(1): p. 21-28.

175. Gloster, T., et al., *Structural, Kinetic, and Thermodynamic Analysis of Glucoimidazole-Derived Glycosidase Inhibitors†,‡.* Biochemistry, 2006. **45**(39): p. 11879-11884.

176. Gomez, G., et al., *Structure of Human Epoxide Hydrolase Reveals Mechanistic Inferences on Bifunctional Catalysis in Epoxide and Phosphate Ester Hydrolysis†,‡.* Biochemistry, 2004. **43**(16): p. 4716-4723.

177. Ishijima, J., et al., *Free energy requirement for domain movement of an enzyme.* The Journal of biological chemistry, 2000. **275**(25): p. 18939-18945.

178. Anderson, M., *Imidazo61,2-a9pyridines: A potent and selective class of cyclin-Dependent kinase inhibitors identified through structure-Based hybridisation.* Bioorganic & Medicinal Chemistry Letters, 2003. **13**(18): p. 3021-3026.

179. Niedzwiecka, A., et al., *Biophysical Studies of eIF4E Cap-binding Protein: Recognition of mRNA 5′ Cap Structure and Synthetic Fragments of eIF4G and 4E-BP1 Proteins.* Journal of Molecular Biology, 2002. **319**(3): p. 615-635.

180. Lee, J.-O., et al., *Crystal Structure of the PTEN Tumor SuppressorImplications for Its Phosphoinositide Phosphatase Activity and Membrane Association.* Cell, 1999. **99**(3): p. 323-334.

181. Evrard, C., et al., *Crystal structure of the C47S mutant of human peroxiredoxin 5.* Journal of Chemical Crystallography, 2004. **34**(8): p. 553-558.

182. Johnston, J., V. Arcus, and E. Baker, *Structure of naphthoate synthase (MenB) from Mycobacterium tuberculosis in both native and product-bound forms.* Acta crystallographica. Section D, Biological crystallography, 2005. **61**(Pt 9): p. 1199-1206.

183. Kim, S.W., et al., *High-Resolution Crystal Structures of Δ5-3-Ketosteroid Isomerase with and without a Reaction Intermediate Analogue†.* Biochemistry, 1997. **36**(46): p. 14030-14036.

184. Holland, D.R., et al., *The crystal structure of a lysine 49 phospholipase A2 from the venom of the cottonmouth snake at 2.0-A resolution.* Journal of Biological Chemistry, 1990. **265**(29): p. 17649-56.

185. Watson, J.D., R.A. Laskowski, and J.M. Thornton, *Predicting protein function from sequence and structural data.* Curr Opin Struct Biol, 2005. **15**(3): p. 275-84.

186. Holm, L. and C. Sander, *Mapping the Protein Universe.* Science, 1996. **273**(5275): p. 595-602.

187. Marsden, R.L., et al., *Exploiting protein structure data to explore the evolution of protein function and biological complexity.* Philos Trans R Soc Lond B Biol Sci, 2006. **361**(1467): p. 425-40.

188. Andreeva, A., et al., *Data growth and its impact on the SCOP database: new developments.* Nucleic Acids Res, 2008. **36**(Database issue): p. D419-25.

189. Greene, L.H., et al., *The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution.* Nucleic Acids Res, 2007. **35**(Database issue): p. D291-7.

190. Holm, L. and P. Rosenstrom, *Dali server: conservation mapping in 3D.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W545-9.

191. Bhaduri, A., G. Pugalenthi, and R. Sowdhamini, *PASS2: an automated database of protein alignments organised as structural superfamilies.* BMC Bioinformatics, 2004. **5**: p. 35.

192. Wang, Y., et al., *MMDB: annotating protein sequences with Entrez's 3D-structure database.* Nucleic Acids Research, 2007. **35**(suppl 1): p. D298-D300.

193. Chandonia, J.M., et al., *The ASTRAL Compendium in 2004.* Nucleic Acids Research, 2004. **32**(suppl 1): p. D189-D192.

194. Mizuguchi, K., et al., *HOMSTRAD: A database of protein structure alignments for homologous families.* Protein Science, 1998. **7**(11): p. 2469-2471.

195. Schmidt, R., R.B. Altman, and M. Gerstein, *LPFC: An internet library of protein family core structures.* Protein Science, 1997. **6**(1): p. 246-248.

196. Orengo, C.A. and J.M. Thornton, *Protein families and their evolution - a structural perspective.* Annual Review of Biochemistry, 2005. **74**(1): p. 867-900.

197. Valas, R., S. Yang, and P. Bourne, *Nothing about protein structure classification makes sense except in the light of evolution.* Current Opinion in Structural Biology, 2009. **19**(3): p. 329-334.

198. Kolodny, R., P. Koehl, and M. Levitt, *Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures.* Journal of Molecular Biology, 2005. **346**(4): p. 1173-1188.

199. Taylor, W.R. and C.A. Orengo, *Protein structure alignment.* Journal of Molecular Biology, 1989. **208**(1): p. 1-22.

200. Gerstein, M. and M. Levitt, *Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins.* Protein Science, 1998. **7**(2): p. 445-456.

201. Subbiah, S., D.V. Laurents, and M. Levitt, *Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core.* Current Biology, 1993. **3**(3): p. 141-148.

202. Holm, L. and C. Sander, *Protein Structure Comparison by Alignment of Distance Matrices.* Journal of Molecular Biology, 1993. **233**(1): p. 123-138.

203. Kleywegt, G., *Use of Non-crystallographic Symmetry in Protein Structure Refinement.* Acta Crystallographica Section D, 1996. **52**(4): p. 842-857.

204. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Engineering, 1998. **11**(9): p. 739-747.

205. Krissinel, E. and K. Henrick, *Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions.* Acta Crystallographica Section D, 2004. **60**(12 Part 1): p. 2256-2268.

206. Mayr, G., F. Domingues, and P. Lackner, *Comparative Analysis of Protein Structure Alignments.* BMC Structural Biology, 2007. **7**(1): p. 50.

207. Roland L, D., Jr., *Sequence comparison and protein structure prediction.* Current Opinion in Structural Biology, 2006. **16**(3): p. 374-384.

208. Sam, V., et al., *Towards an automatic classification of protein structural domains based on structural similarity.* BMC Bioinformatics, 2008. **9**(1): p. 74.

209. Kabsch, W., *A solution for the best rotation to relate two sets of vectors.* Acta Crystallographica Section A, 1976. **32**(5): p. 922-923.

210. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: The European Molecular Biology Open Software Suite.* Trends in Genetics, 2000. **16**(6): p. 276-277.

211. Xu, Z., A.L. Horwich, and P.B. Sigler, *The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex.* Nature, 1997. **388**(6644): p. 741-750.

212. Braig, K., P.D. Adams, and A.T. Brünger, *Conformational variability in the refined structure of the chaperonin GroEL at 2.8 A resolution.* Nat Struct Mol Biol, 1995. **2**(12): p. 1083-1094.

213. Ditzel, L., et al., *Crystal Structure of the Thermosome, the Archaeal Chaperonin and Homolog of CCT.* Cell, 1998. **93**(1): p. 125-138.

214. Tai, C.-H., et al., *SE: an algorithm for deriving sequence alignment from a pair of superimposed structures.* BMC Bioinformatics, 2009. **10 Suppl 1**.

215. EMBL-EBI. *Optimal scoring matrix parameters.* 2010 [cited 2010; Available from: http://www.ebi.ac.uk/help/matrix.html.

216. Enroth, C., et al., *The crystal structure of phenol hydroxylase in complex with FAD and phenol provides evidence for a concerted conformational change in the enzyme and its cofactor during catalysis.* Structure, 1998. **6**(5): p. 605-17.

217. Mesecar, A.D. and D.E. Koshland, *Sites of Binding and Orientation in a Four-Location Model for Protein Stereospecificity.* IUBMB Life, 2000. **49**(5): p. 457-466.

218. Ye, Y. and A. Godzik, *Flexible structure alignment by chaining aligned fragment pairs allowing twists.* Bioinformatics, 2003. **19**(suppl 2): p. ii246-ii255.

219. Gong, W., et al., *Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment.* Nucleic Acids Res, 1997. **25**(14): p. 2702-15.

220. Price, S.R., P.R. Evans, and K. Nagai, *Crystal structure of the spliceosomal U2B"-U2A' protein complex bound to a fragment of U2 small nuclear RNA.* Nature, 1998. **394**(6694): p. 645-50.

221. Marino, M., et al., *Structure of the lnlB leucine-rich repeats, a domain that triggers host cell invasion by the bacterial pathogen L. monocytogenes.* Mol Cell, 1999. **4**(6): p. 1063-72.

222. Owen, D.J., et al., *The structure and function of the beta 2-adaptin appendage domain.* EMBO J, 2000. **19**(16): p. 4216-27.

223. Traub, L.M., et al., *Crystal structure of the alpha appendage of AP-2 reveals a recruitment platform for clathrin-coat assembly.* Proc Natl Acad Sci U S A, 1999. **96**(16): p. 8907-12.

224. Rost, B., *Twilight zone of protein sequence alignments.* Protein Engineering, 1999. **12**(2): p. 85-94.
225. Elofsson, A., *A study on protein sequence alignment quality*. Proteins: Structure, Function, and Bioinformatics, 2002. **46**(3): p. 330-339.
226. Michel, G., et al., *The structure of the RlmB 23S rRNA methyltransferase reveals a new methyltransferase fold with a unique knot.* Structure, 2002. **10**(10): p. 1303-15.
227. Nureki, O., et al., *An enzyme with a deep trefoil knot for the active-site architecture.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 7): p. 1129-37.
228. Project, B., *Biopython PDB parser*. 2006.
229. Orville, A.M., et al., *Structures of competitive inhibitor complexes of protocatechuate 3,4-dioxygenase: multiple exogenous ligand binding orientations within the active site.* Biochemistry, 1997. **36**(33): p. 10039-51.
230. Fass, D., C.E. Bogden, and J.M. Berger, *Quaternary changes in topoisomerase II may direct orthogonal movement of two DNA strands.* Nat Struct Biol, 1999. **6**(4): p. 322-6.
231. Morais Cabral, J.H., et al., *Crystal structure of the breakage-reunion domain of DNA gyrase*. Nature, 1997. **388**(6645): p. 903-6.
232. Klaus, W., et al., *The three-dimensional high resolution structure of human interferon alpha-2a determined by heteronuclear NMR spectroscopy in solution.* J Mol Biol, 1997. **274**(4): p. 661-75.
233. Tolbert, W.D., et al., *The structural basis for substrate specificity and inhibition of human S-adenosylmethionine decarboxylase.* Biochemistry, 2001. **40**(32): p. 9484-94.