Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Molecular Graphics and Modelling 24 (2006) 414–425

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Exploring protein–ligand recognition with Binding MOAD

Richard D. Smith [a,1], Liegi Hu [b,1], Jayson A. Falkner [c,1], Mark L. Benson [c], Jason P. Nerothin [b], Heather A. Carlson [a,b,c,*]

[a] *Biophysics Research Division, University of Michigan, Ann Arbor, MI 48109-1055, USA*
[b] *Department of Medicinal Chemistry, College of Pharmacy, University of Michigan, 428 Church Street, Ann Arbor, MI 48109-1065, USA*
[c] *Bioinformatics Program, University of Michigan, Ann Arbor, MI 48109-0621, USA*

Received 2 July 2005; received in revised form 5 August 2005; accepted 9 August 2005
Available online 15 September 2005

Abstract

We have recently announced the largest database of protein–ligand complexes, Binding MOAD (Mother of All Databases). After the August 2004 update, Binding MOAD contains 6816 complexes. There are 2220 protein families and 3316 unique ligands. After searching 6000+ crystallography papers, we have obtained binding data for 1793 (27%) of the complexes. We have also created a non-redundant set of complexes with only one complex from each protein family; in that set, 630 (28%) of the unique complexes have binding data. Here, we present information about the data provided at the Binding MOAD website. We also present the results of mining Binding MOAD to map the degree of solvent exposure for binding sites. We have determined that most cavities and ligands (70–85%) are well buried in the complexes. This fits with the common paradigm that a large degree of contact between the ligand and protein is significant in molecular recognition. GoCAV and the GoCAVviewer are the tools we created for this study. To share our data and make our online dataset more useful to other research groups, we have integrated the viewer into the Binding MOAD website (www.BindingMOAD.org).

Keywords: Molecular recognition; Binding affinity; Cavity; Surface area; Protein–ligand database

1. Introduction

A growing trend in computational biology is the development of large datasets to provide the scientific community with various information on protein–ligand structures. Of course, the definitive online resource for structural data of these complexes is the Protein Data Bank (PDB) [1]. It is constantly being improved through the addition of online tools and links to complementary datasets [2]. Most recently, Ligand Depot was created by the curators of the PDB to facilitate searching the HET groups via chemical substructures and text-based searches [3]. There are many other examples of databases and websites that analyze and augment protein–ligand complexes from the PDB. The following discussion is by no means an exhaustive listing of such derivatives of the PDB.

Our own contribution in this area is Binding MOAD (Mother of All Databases) [4]. Our goal for Binding MOAD is to create the largest resource of high-quality protein–ligand complexes and augment those structures with binding affinity data and online analytical tools. We took a top–down approach to create Binding MOAD. Starting with the entire PDB, we selected only crystal structures of high resolution (≤ 2.5 Å). We ensured that Binding MOAD contained only appropriate protein–ligand structures through extensive hand curation. Structures were required to contain at least one valid, non-covalently bound ligand. Chains of four nucleic acids or less and 10 amino acids or less were treated as ligands. In our original creation of the 2003 version of Binding MOAD, we eliminated any structure with a heme group because of the difficulty in distinguishing non-covalently bound ligands. With the August 2004 update, all heme-containing proteins

[*] Corresponding author. Tel.: +1 734 615 6841.
E-mail address: carlsonh@umich.edu (H.A. Carlson).
[1] These authors have contributed equally to this work.

1093-3263/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2005.08.002

have been examined by hand and appropriate structures are now part of Binding MOAD. The 2004 version of Binding MOAD contains 6816 complexes. We read over 6000 crystallography papers to confirm the validity of the protein–ligand complexes and to gather binding affinity data. As a result of this process, we have binding data for 1793 (27%) of the complexes.

We wanted to mine Binding MOAD to provide general patterns of molecular recognition to the scientific community. How exposed binding sites are across all protein–ligand complexes? To answer this, we needed a resource that could properly treat any binding site—regardless of size, shape, degree of solvent exposure, the inclusion of bridging water molecules, or the occurrence of side chains with multiple resolved orientations (partial atom occupancy). For this, we have developed GoCAV and the GoCAVviewer to calculate and display molecular surfaces for the ligands and for the protein cavities.

A number of online tools are already available to view atomic coordinates, secondary structure, and cavities. We are not presenting GoCAV as a breakthrough to supercede these programs. We simply feel that GoCAV and the GoCAVviewer are complementary alternatives to these other excellent resources, and by incorporating the viewer into our website (www.BindingMOAD.org), we have a means to share the data from this study with the scientific community. The discussion below highlights some of the most useful online resources created by other research groups for analyzing and viewing protein–ligand complexes. Generally, these databases describe a ligand as any molecule that is not one of the common 20 amino acids or eight common nucleic acids. They make no distinction between valid and invalid ligands like crystallographic additives or covalent modifications to the protein.

PDBsum is the most comprehensive resource. It provides data on the entire collection of structures from the PDB [5–8]. Chemical, enzymatic, and genomic information is available for all PDB structures, even if they are not proteins and even if they do not contain ligands. One of the most powerful features of PDBsum for understanding the molecular recognition of ligands is its analysis of macromolecule–ligand interactions. The information is provided via 2D pictures and several 3D viewers. Most relevant to this work is the fact that PDBsum provides analysis of potential cavities using an updated version of SURFNET [9].

CASTp is an online database that uses rigorous analytical techniques to analyze all proteins in the PDB for interior cavity voids and surface pockets [10]. Using the program CAST [11], it calculates the volumes and surface areas of the sites, and it also determines the size of the openings in solvent-exposed pockets. It is not limited to proteins with bound ligands, so it has the benefit of identifying previously unknown binding sites, but it also identifies many small surface pockets that do not bind ligands. The online viewer displays the residues that make up the cavities and pockets, but it does not show the bound ligands. This makes it difficult to understand the molecular recognition that controls binding in that site.

MSDsite [12] provides information on ligand interactions with any macromolecule, not just proteins. MSDsite provides various analyses of the macromolecular environment surrounding ligands. The dataset can be mined by matching patterns based on the ligand or on the binding-site environment. PDB-ligand [13] is a new resource that is very similar to MSDsite, but strictly focuses on analyses of protein residues and nucleic acids within 6.5 Å of a HET group. Relibase [14,15] is a resource that specifically focuses on the protein–ligand complexes in the PDB. It allows for text-based and sequence-based searching of the PDB. SMILES strings can be used to search ligand substructures. It also provides graphics tools to examine the structures. Relibase+ [16] is a newer version that allows for additional 2D and 3D similarity searches. NCBI's Entrez resource for 3D structures is the Molecular Modeling Database (MMDB) [17]. MMDB is based on pregenerated relationships, found by comparing each PDB entry with various structure and sequence databases. Their viewer can be used to compare any individual PDB entry to its structural homologs. This reveals their similar tertiary structure and can be used to examine common binding motifs of bound ligands. So though the focus of MMDB is the comparison of folds and domains, it can provide valuable information on protein–ligand recognition. Each of the four online databases mentioned above has very useful features, but as mentioned above, they make no distinction of which HET groups are proper ligands.

Two additional datasets, PDBbind and sc-PDB, are similar to Binding MOAD and also focus on valid ligands. These databases do not provide viewers to examine protein–ligand complementarity, but the atomic coordinates of the proteins and ligands are available for download and can be examined offline. PDBbind is a large set of protein–ligand complexes from the PDB, focusing on binary structures with a single ligand in a protein binding site [18]. PDBbind also provides binding affinity data obtained from reading the crystallography papers. As of its latest update in January 2004, it contains binding data on 1622 complexes (a subset of 900 complexes makes up the "refined" set) [19]. PDBbind provides graphical interfaces, similar to those used with Ligand Depot, to view the ligands and perform substructure searches to find related systems. The other database, sc-PDB [20], was created in a fashion similar to Binding MOAD and PDBbind, but it does not provide binding data. The set of structures is used for "inverse screening," a procedure where a ligand is docked to a series of binding sites to determine its appropriate target. sc-PDB is a set of 5634 protein binding sites and 7109 ligands at the time of writing this paper [personal communication, Esther Kellenberger, Université Louis Pasteur, Strasbourg]. The online interface to the dataset allows for text-based searches of much of the information within the PDB files (PDB ID, HET group name, authors, EC numbers, deposition date, resolution, etc.). The data can also be accessed by information based on other resources like Swiss-Prot [21] data and NCBI taxonomy notation [22].

2. Methods

Rather than simple PDB files, ‘‘corrected biounit files’’ were used for all protein–ligand complexes. Biounit files are available from the PDB, and they represent the appropriate multimer for biological activity. For instance, if only a monomer appears in the unit cell, but a trimer is the appropriate biounit, the other two monomers are generated through symmetry operations. We found that HET groups and water molecules frequently were not properly treated in the PDB’s biounit files. They were not propagated where necessary, and they were not removed in cases where their corresponding protein was deleted from the unit cell. We corrected all biounit files by propagating the water and ligands as necessary using the program PyMOL [23]. We also removed any molecules that were more than 10 Å away from the protein. Covalent links were checked to avoid truncating sugar chains and other post-transcriptional modifications that were longer than 10 Å. These corrected biounit files are the structures that are available for download on the Binding MOAD website.

It is straightforward to calculate the surface of an enclosed binding site [24,25], but many interesting ligands are bound in open clefts. Molecular surface area (MSA) is calculated by ‘‘rolling a solvent probe’’ on the van der Waals (vdw) surface of the atoms. With an exposed binding site, the probe escapes and maps out the entire protein surface. Ho and Marshall suggested that some cut-off distance to the ligand might be a reasonable way to determine the boundary of an open site [26]. We were not able to find code to do this, so we wrote GoCAV to accomplish the task and provide a consistent treatment for any type of binding site in Binding MOAD.

GoCAV, uses an ‘‘enlarged ligand surface’’ (ELS) to create a boundary for the binding site (Fig. 1). It calculates MSAs using a grid-based method. Voronoi tessellations are more accurate than grids for enclosed sites [24,27,28], but the method does not work as well on surfaces [29,30]. We use a very fine, 0.2-Å grid (0.008 $Å^3$ cubes) to minimize the errors as much as possible. Codes have been developed by other groups that calculate surfaces and cavities (for example, POCKET [31], SURFNET [9], CAST [11], PASS [32], and an unnamed grid-based technique by Schneider and co-workers [33]). Many of these have the benefit of finding pockets without needing bound ligands to guide them, which means they can identify new binding sites (a definite advantage over GoCAV). However, in the process of analyzing/identifying all possible cavities, some of these codes produce pockets that are not true binding sites. Some have poorly defined boundaries that do not encapsulate all of a bound ligand. Some do not identify all types of pockets, and others tend to create large networks of interconnected cavity spaces over the surface of the protein. For our purposes with Binding MOAD, we needed a code, which focuses on defining a cavity within the local vicinity of a bound ligand.

To create the ELS, we extended the ligand’s vdw radii by 2.8 Å (the equivalent of one layer of water as the exterior boundary for an open binding site). We wanted to define a boundary for open binding sites, but not hinder the calculation of enclosed binding sites that incorporate bridging water molecules. To verify our description of the ELS, we examined several binding sites with bridging water molecules (see Fig. 2). Appropriate boundaries of these binding sites were identified with the ELS radius of 2.8 Å. The MSA for ligands are straightforward to calculate and were also part of the GoCAV output.

The overwhelming majority of Binding MOAD’s structures do not contain hydrogen atoms, so we needed to use united-atom radii in our analyses. Our chosen radii were based on averaged OPLS united-atom vdw parameters: C = 1.925 Å, N = 1.655 Å, O = 1.52Å, S = 1.81 Å, P = 1.87 Å [34]. (We

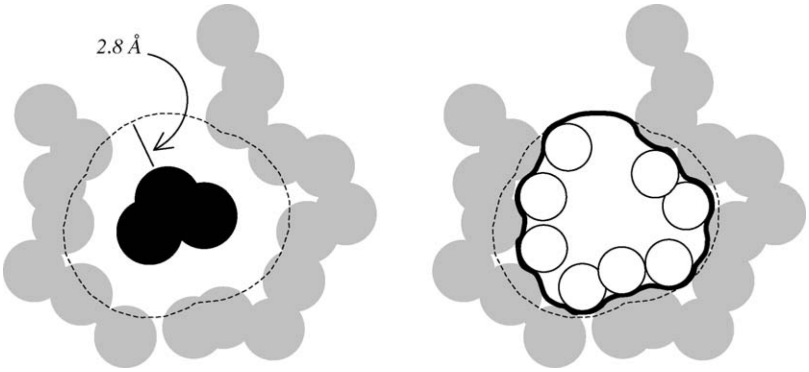


Fig. 1. Determining the boundary of an open cavity using ELS. (Left) A ligand molecule (black) is bound in an open protein cleft (gray). The dashed line is the ELS, determined by adding 2.8 Å to the radii. A probe rolls over the vdw surfaces of the protein atoms and the inward-facing surface of the ELS. The resulting surface of the cavity is shown as a bold, black line. The solvent-exposed portion of the cavity surface is defined as the section of bold, black line that is defined only by the ELS in the opening of the binding site.

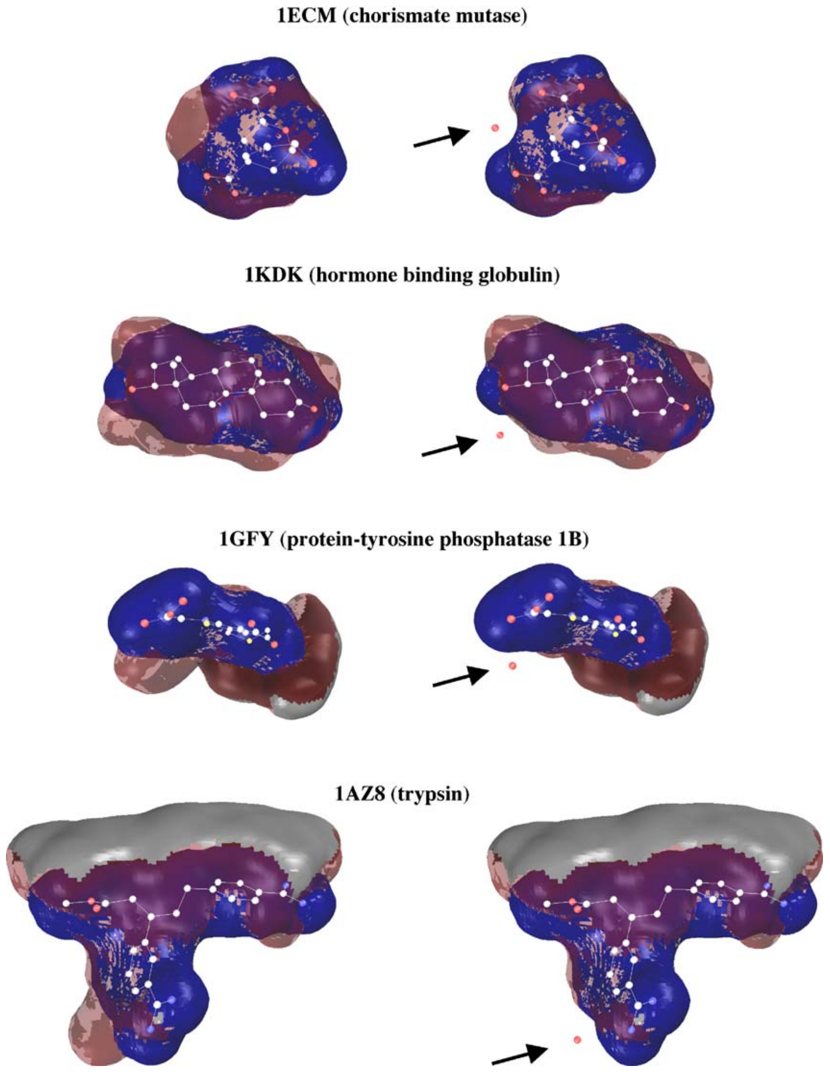

1ECM (chorismate mutase)
1KDK (hormone binding globulin)
1GFY (protein-tyrosine phosphatase 1B)
1AZ8 (trypsin)

Fig. 2. The use of an ELS does not create inappropriate boundaries for open or closed cavities that contain bridging water molecules. Examples are given for completely buried cavities (1ECM and 1KDK) and solvent-exposed pockets (1AZ8 and 1GFY). (Left) Binding site and ligand surfaces calculated with GoCAV, employing an ELS cut-off. (Right) The resulting surfaces when the noted bridging water molecules within the cavity are included in the calculation as additional protein atoms. The ligand surface is blue, and the binding site surface is red and gray. The red regions are buried, and the gray region denotes the solvent-exposed or ELS surface of the cavity. Protein atoms are not shown for clarity. This figure was created using the GoCAVviewer on the Binding MOAD website.

estimated radii for other less-typical atoms as 2.0 Å.) OPLS parameters were carefully developed to reproduce thermodynamic properties in condensed phases. We are confident in the choice of OPLS radii because of their good agreement with Li and Nussinov's radii set which was determined in an entirely different fashion [35]. Li and Nussinov derived radii through contact distance distributions in a set of 1405 protein crystal structures (C = 1.92 Å, N = 1.66 Å, O = 1.51 Å, S = 1.92 Å). Though we do not present the data here, a user can use a second set of radii in GoCAV. We made the Fleming and Richards' radii [29] (C = 1.9 Å, N = 1.5 Å, O = 1.4 Å, S = 1.85 Å) an available option because they are well established and many groups support smaller radii. Gerstein and co-workers determined similar, smaller radii using contact distance distributions from crystal structures of small organic molecules (C = 1.88 Å, N = 1.64 Å, O = 1.44 Å, S = 1.77 Å) [36].

Invalid ligands are not included in the calculations unless they are a covalent modification of the protein or a structural element like a catalytic/structural zinc ion or a heme (these are treated as additional protein atoms). When mining a large dataset, the code must properly treat unusual cases. GoCAV was created with several "filters" to analyze structures before performing the surface calculations. With these filters, GoCAV was able to properly process >98% of the structures in Binding MOAD. In the case of ligands with warnings (too many or too few atoms), those complexes were not included in this study (even when GoCAV was able to calculate their surfaces).

Unusual protein–ligand complexes include the following situations: (1) Side chains within a binding site can be solved in multiple orientations (as denoted by partial occupancy). In these cases, GoCAV automatically calculates the surfaces twice, with the side chain in either orientation. Appropriate combinations are generated if more than one side chain in the binding site requires this treatment. All solutions are presented in our analysis, providing averaged data points with error bars for those complexes. (2) Some sugar-binding proteins actually contain both enantiomers of the sugar in the binding site, superimposed with 50%–50% occupancy. Again, GoCAV recognizes the two solutions inherent in the structure and does two independent calculations, each with a single enantiomer. Both ligands are presented independently in our plots and histograms (no error bars because they are not the same ligand). (3) When two separate ligands are accommodated in a large binding site (such as a cofactor and an inhibitor bound in close proximity), GoCAV actually does three calculations: (a) both ligands are treated as one large molecule; (b) the first ligand is treated as part of the protein while the surfaces around the second ligand are calculated; (c) the second ligand is part of the protein while the first ligand is calculated independently. The later two calculations, where each ligand is treated independently, are the values included in the plots and histograms in this study.

To verify that the patterns calculated with GoCAV are appropriate and comparable to other standard techniques, we have also calculated the solvent accessible surface area (SASA) of the ligands with the program NACCESS [37] (SASA of the ligand should be roughly comparable to the MSA of the cavity), NACCESS is based on Lee and Richard's analytical method [38] as opposed to our grid-based approach. It uses radii based on Chothia's [39] but with more subtypes for carbon, nitrogen, and oxygen with slightly different radii (carbons range 1.76–2.0 Å, nitrogens range 1.5–1.65 Å, oxygens are 1.35 or 1.4 Å, S = 1.85 Å, P = 1.9 Å, Fe = 1.47 Å). NACCESS provides SASA on a "per residue" basis. If the ligand is completely buried, the SASA calculated with NACCESS is zero. The SASA of each ligand in each complex was calculated in the presence and absence of the protein (again, we included any other appropriate ligands as part of the protein environment). We calculated the buried surface area of the ligand as SASA (no protein) − SASA (with protein) and percent buried surface area as 100 × (1 − SASA (with protein)/SASA (no protein)). NACCESS is not able to treat the unusual cases that we describe above for GoCAV. Those systems are not included in the NACCESS plots and histograms.

3. Results and discussion

3.1. Binding MOAD

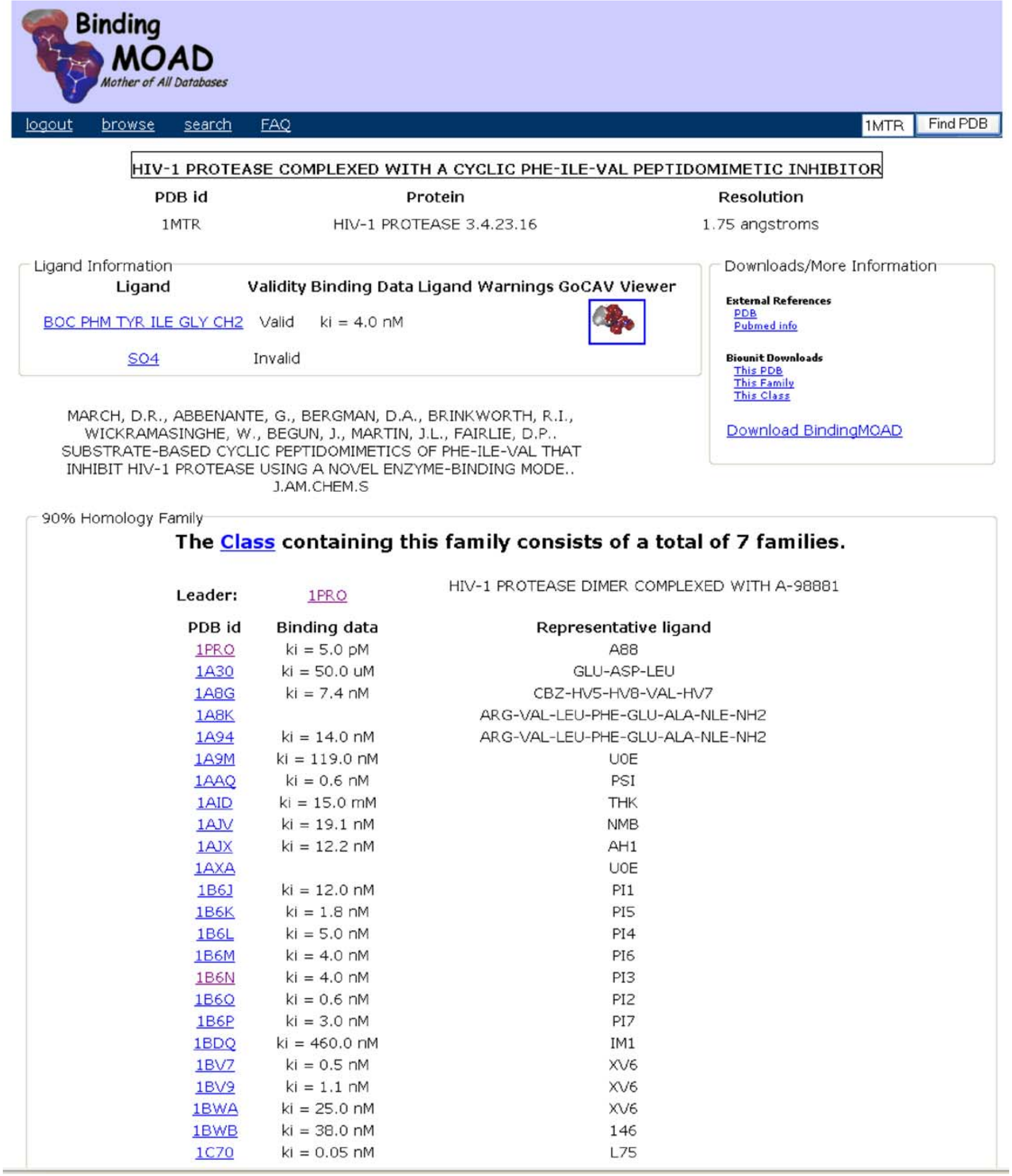
Several features of Binding MOAD make it particularly useful for examining the degree of solvent exposure of all protein–ligand binding sites. First, the dataset has been carefully curated to identify valid and invalid ligands in each structure. Only the valid ligands are included in our analysis. Without this analysis, any broad mining of the structures would reflect real binding patterns skewed by the less relevant patterns seen for crystallographic additives. (We have also excluded any ligands with warnings of too many or too few atoms from the analysis, though they are part of the MOAD dataset.)

Second, the dataset has been analyzed for redundancy. The proteins have been grouped into families by 90% sequence identity. The non-redundant set of structures from Binding MOAD contains only one complex from each protein family. The representative for the family is the tightest binder when binding data is known. In cases where there is no binding data for any of the complexes in the family, the representative is chosen based on best resolution and other structural considerations [4]. This allows us to present the data without some of the inherent bias of structures deposited within the PDB.

3.2. Sharing the data on the Binding MOAD website

Each entry's datapage on the Binding MOAD website is organized to help users identify related protein systems and compare binding data, see Fig. 3. Entries are cross-linked by function (classes for both enzymes and non-enzymes), sequence identity, and ligand content. All HET groups in the complex are identified as valid or invalid, and warnings are provided when too few or too many ligand atoms appear in the PDB entry (unresolved atoms or multiple resolved orientations for parts of the ligand, respectively). Binding data is provided when available. Text-based searches can be used to identify entries based on PDB i.d., EC number, protein name, three-letter HET codes, and authors. Wildcards are permitted. The results can be limited to a user-defined range of crystallographic resolution. The user can also limit the search to the 1793 structures in Binding MOAD with available binding data, the 2220 structures of the non-redundant Binding MOAD dataset (where each protein family is only represented once), or the 630 structures in the non-redundant set that have binding data. There is also a browse feature to allow users to page through functional classes of structures. When a user clicks the "class" link on a datapage (seen in Fig. 3), they are taken to the browse page for that functional class where all protein families within the class are shown and ligand/binding information is also provided (Fig. 4).

Clicking the blue and red thumbnail on a datapage, see Fig. 3, launches a version of the GoCAVviewer that interactively displays the atomic coordinates and the surfaces calculated with GoCAV. The binding-site surfaces and the ligand surfaces calculated with GoCAV are grid points, so the "raw" surfaces look like LEGO building blocks. A smooth surface is created by a graphics trick, applying a Gaussian filter to the image. The GoCAVviewer is written entirely in standards compliant Java, and the code will work on any operating system that provides an implementation of the standard Java Runtime Environment and Java3D API. GoCAVviewer is interactive, allowing the user to rotate, zoom, or translate the structures in real time.



HIV-1 PROTEASE COMPLEXED WITH A CYCLIC PHE-ILE-VAL PEPTIDOMIMETIC INHIBITOR

PDB id	Protein	Resolution
1MTR	HIV-1 PROTEASE 3.4.23.16	1.75 angstroms

Ligand Information

Ligand	Validity	Binding Data	Ligand Warnings	GoCAV Viewer
BOC PHM TYR ILE GLY CH2	Valid	ki = 4.0 nM		
SO4	Invalid			

Downloads/More Information

External References
PDB
Pubmed info

Biounit Downloads
This PDB
This Family
This Class

Download BindingMOAD

MARCH, D.R., ABBENANTE, G., BERGMAN, D.A., BRINKWORTH, R.I.,
WICKRAMASINGHE, W., BEGUN, J., MARTIN, J.L., FAIRLIE, D.P..
SUBSTRATE-BASED CYCLIC PEPTIDOMIMETICS OF PHE-ILE-VAL THAT
INHIBIT HIV-1 PROTEASE USING A NOVEL ENZYME-BINDING MODE..
J.AM.CHEM.S

90% Homology Family

The Class containing this family consists of a total of 7 families.

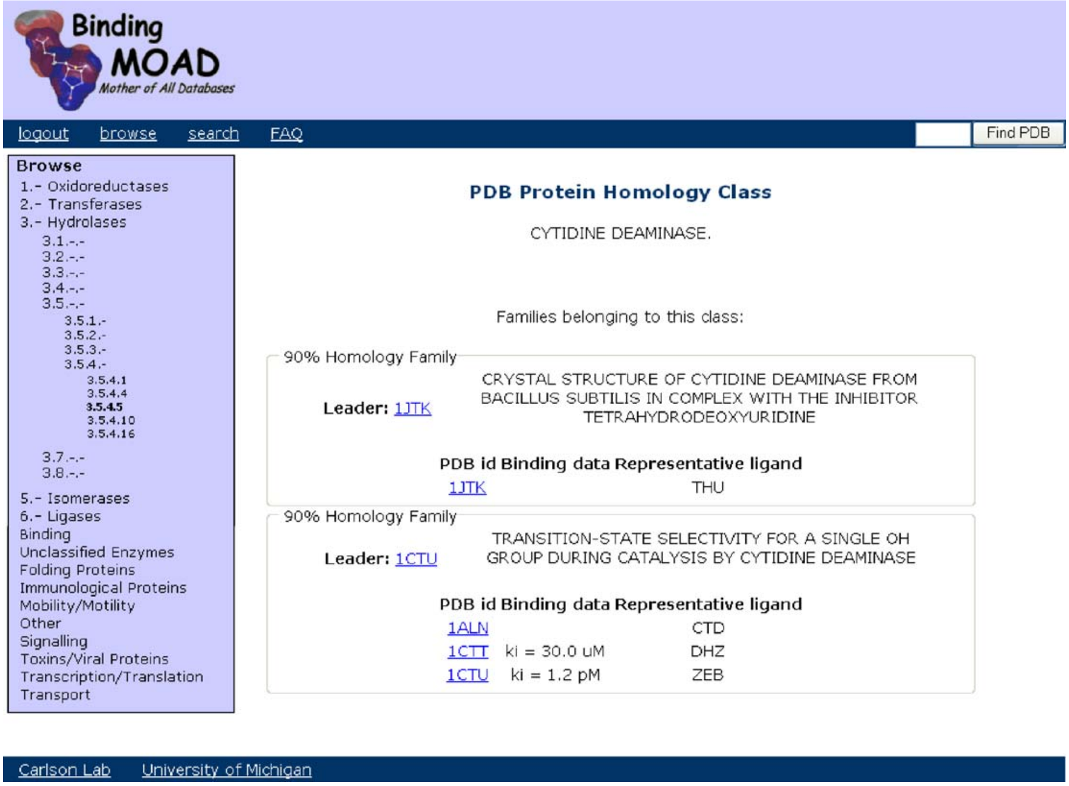
Leader: 1PRO HIV-1 PROTEASE DIMER COMPLEXED WITH A-98881

PDB id	Binding data	Representative ligand
1PRO	ki = 5.0 pM	A88
1A30	ki = 50.0 uM	GLU-ASP-LEU
1A8G	ki = 7.4 nM	CBZ-HV5-HV8-VAL-HV7
1A8K		ARG-VAL-LEU-PHE-GLU-ALA-NLE-NH2
1A94	ki = 14.0 nM	ARG-VAL-LEU-PHE-GLU-ALA-NLE-NH2
1A9M	ki = 119.0 nM	U0E
1AAQ	ki = 0.6 nM	PSI
1AID	ki = 15.0 mM	THK
1AJV	ki = 19.1 nM	NMB
1AJX	ki = 12.2 nM	AH1
1AXA		U0E
1B6J	ki = 12.0 nM	PI1
1B6K	ki = 1.8 nM	PI5
1B6L	ki = 5.0 nM	PI4
1B6M	ki = 4.0 nM	PI6
1B6N	ki = 4.0 nM	PI3
1B6O	ki = 0.6 nM	PI2
1B6P	ki = 3.0 nM	PI7
1BDQ	ki = 460.0 nM	IM1
1BV7	ki = 0.5 nM	XV6
1BV9	ki = 1.1 nM	XV6
1BWA	ki = 25.0 nM	XV6
1BWB	ki = 38.0 nM	146
1C70	ki = 0.05 nM	L75

Fig. 3. The datapage for the HIV-1 protease complex 1MTR. The page starts with the general information from the PDB file. The ligand HET codes are single-click searches that pull up all other structures with that ligand. All ligands are listed as valid or invalid, and binding affinity data is provided when available. Warnings are provided when the number of atoms in the structure do not match the formula section of the PDB file. Clicking the thumbnail launches the GoCAVviewer. Links to the right of the thumbnail take the user to the equivalent datapage at the PDB and to the crystallography paper on Pubmed. Various sets of structural and binding data are available for download. At the bottom of the page, the structure is linked to other entries with the same functional class, and all other members of its protein family are listed with ligand information (over 100 HIV-1 protease structures are included in Binding MOAD and the user needs to scroll down the page to see all the data).

The cavity surfaces are transparent and near-by protein atoms can be displayed, so the user can look at the complex in detail. At this time, the most critical issue is speeding up the viewer. We are committed to improving it, but we wanted to make the data available to the rest of the community as soon as possible.

Tight complementarity between the protein and ligand is highlighted by the ligand surface projecting through the cavity surface (see Fig. 2). We have found that these intersections only occur at positions with strong hydrogen bonding or very specific vdw interactions. We have also configured the viewer to display a second set of surface information calculated with bridging water molecules. It was easy to include water molecules as additional protein atoms in a GoCAV calculation and determine their influence on creating a surface to complement the ligand. Fig. 2 shows how the surfaces change when bridging waters are treated as part of the protein. The shape complementarity between the



PDB Protein Homology Class

CYTIDINE DEAMINASE.

Families belonging to this class:

90% Homology Family

Leader: 1JTK CRYSTAL STRUCTURE OF CYTIDINE DEAMINASE FROM BACILLUS SUBTILIS IN COMPLEX WITH THE INHIBITOR TETRAHYDRODEOXYURIDINE

PDB id	Binding data	Representative ligand
1JTK		THU

90% Homology Family

Leader: 1CTU TRANSITION-STATE SELECTIVITY FOR A SINGLE OH GROUP DURING CATALYSIS BY CYTIDINE DEAMINASE

PDB id	Binding data	Representative ligand
1ALN		CTD
1CTT	ki = 30.0 uM	DHZ
1CTU	ki = 1.2 pM	ZEB

Fig. 4. The user can find information by browsing through the complexes within Binding MOAD. The structures are organized by function: EC numbers for enzymes and our own classifications for entries without EC numbers. All protein families within a class are displayed for the user to compare related systems and their binding affinity data.

ligand and the pocket is often more evident when waters are included.

3.3. Mining Binding MOAD

Fig. 5 provides histograms of the size of the ligands in the redundant and non-redundant Binding MOAD sets. The distribution of ligand sizes is similar in the two sets. In Fig. 6, the plots of size vs. affinity show the wide range of data available in Binding MOAD. One issue that should be noted is that the ‘‘affinities’’ used in our plots are a simplistic translation of the K_d, K_i, and IC_{50} data using the formula RT ln(data). This is not strictly correct for K_i or IC_{50}, but it is a way to do a standardized treatment of a large dataset. The K_d data in the plots is highlighted in black because the affinities should be more reliable and better reflect true free energies of binding. Complexes with K_i and IC_{50} data are in gray. The data available for download from the website is the

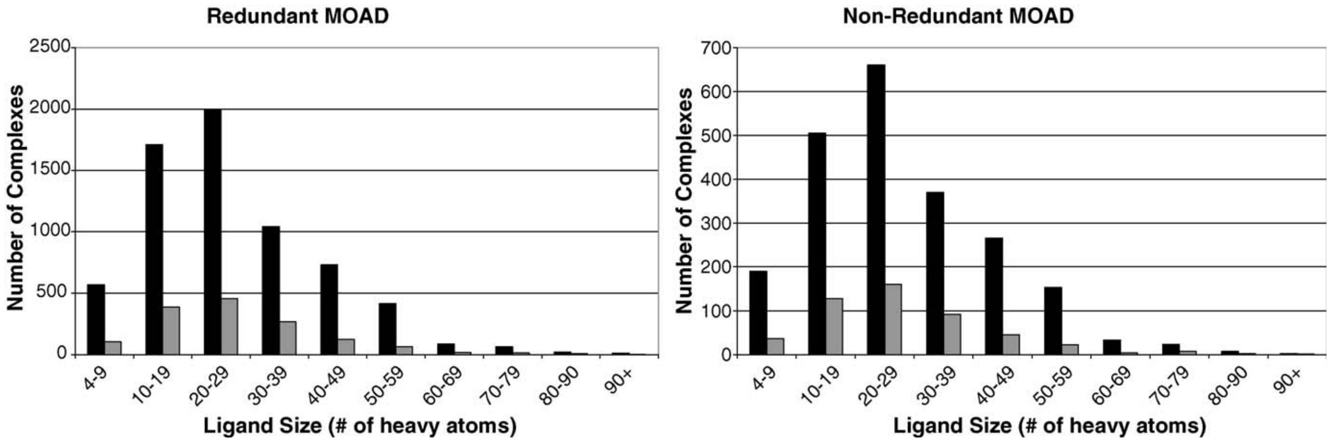


Fig. 5. Distribution of ligand size within the complexes in redundant and non-redundant Binding MOAD, note the larger scale for the redundant complexes. Black bars represent all complexes in Binding MOAD; gray bars represent only the complexes with affinity data.

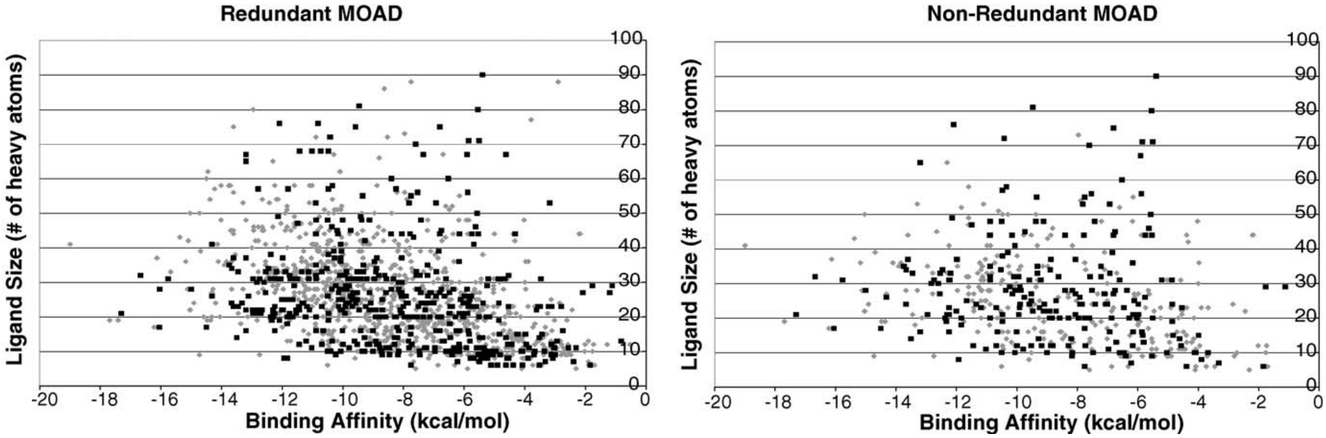


Fig. 6. Plots of ligand size vs. binding affinity for the complexes in redundant and non-redundant Binding MOAD. The data points in black squares are from complexes with K_d data, and gray diamonds are used for complexes with K_i or IC_{50} data.

original K_d, K_i, and IC_{50} data from the crystallography papers.

The ranges in Fig. 6 are approximately the same for the redundant and non-redundant sets, but the averages for both sets are slightly different. The average binding affinity for the redundant set is −8 kcal/mol, but the average for the non-redundant set is −9 kcal/mol. Both sets have a standard deviation of 3 kcal/mol. The average numbers of heavy

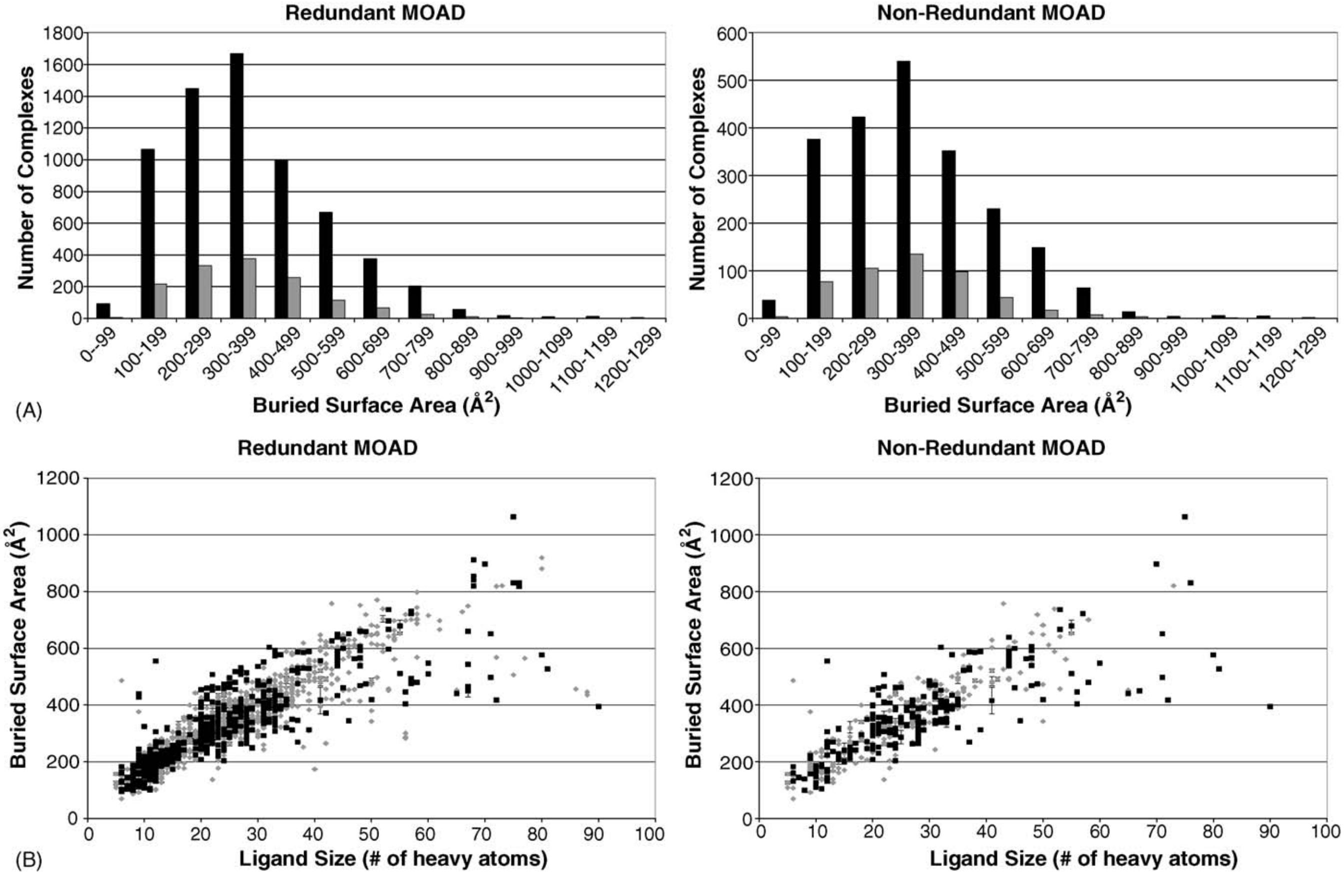


Fig. 7. (A) Distribution of the buried surface area (Å^2) for cavities within Binding MOAD as calculated with GoCAV, note the larger scale for the redundant data. Black bars represent all complexes in Binding MOAD; gray bars represent only the complexes with affinity data. (B) Plots of buried surface area of the cavity (Å^2) vs. ligand size. The data points in black squares are from complexes with K_d data, and gray diamonds are used for complexes with K_i or IC_{50} data. Error bars for data points were available in two cases. First, if a side chain in the active site was resolved in more than one orientation. Second, some multimer complexes are solved with slight differences in the independent binding sites (for instance, the atomic coordinates of the binding sites within a dimer will not be the exactly same if symmetry was not imposed while fitting the electron density).

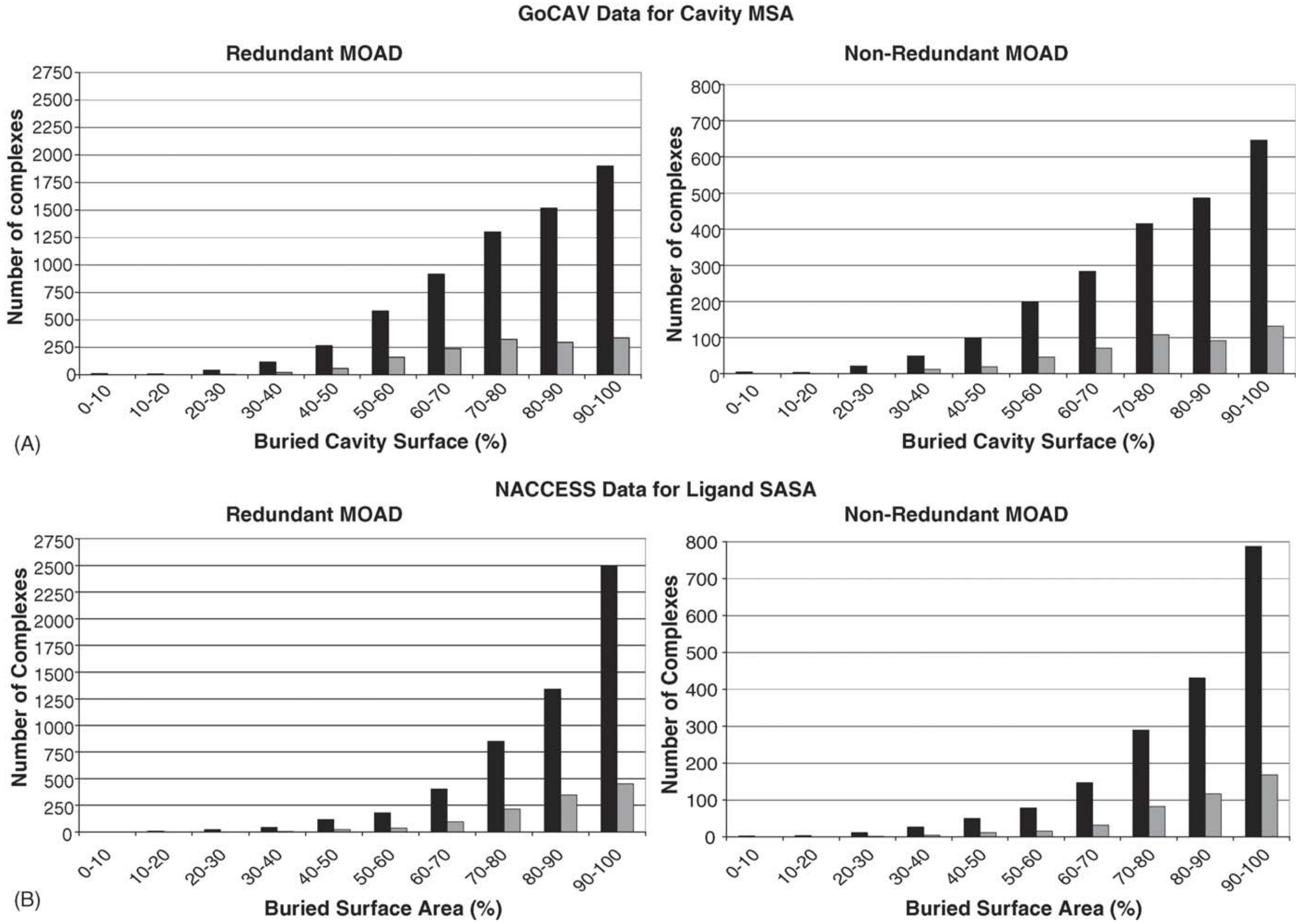


Fig. 8. Histograms of the percent of surface area that is buried. (A) Percentage of buried MSA of the cavity and (B) percentage of buried SASA of the ligand. Black bars represent all complexes in Binding MOAD; gray bars represent only the complexes with affinity data.

atoms for the ligands in these sets are 26 and 27, respectively, both with a standard deviation of 14 atoms. A size range of 12–41 heavy atoms corresponds to drug-like molecular weights of approximately 150–700. It should be noted that these are the averages for just the complexes with binding affinity data (all points in Fig. 6, but only the gray bars in Fig. 5). The average number of heavy atoms for ligands in all of the Binding MOAD complexes is 31 (black bars in Fig. 5).

Fig. 7A presents histograms of buried MSA for the binding-site cavities as calculated by GoCAV. The distribution of buried surface area parallels the size distribution of ligands, and in Fig. 7B, a plot of the cavity's buried MSA versus ligand size shows a good correlation, simply reflecting the relationship between increasing size of the ligand and increasing surface of the cavity it occupies. (As expected, the distributions for buried SASA of the ligands, as calculated with NACCESS, were very similar and also well correlated to ligand size, data not shown.)

Liang et al. [11] found that a linear correlation exists between ligand volume and binding site volume, provided that the pockets were small ($\leq$700 Å^3). Fig. 7B also shows that the correlation is not as tight for the larger ligands and pockets. Others have found that binding sites tend to be the largest pockets/cavities in a protein [11,40–42]. We have not examined other cavities within our proteins, but we plan to compare the patterns of valid and invalid ligands in the future. One would assume that the crystal additives on the surfaces of the protein are in shallow pockets with little buried surface area, but covalent cofactors and structural elements of proteins will occupy both surface and buried positions. Patterns of valid versus invalid ligands of both types should help current efforts in the field to identify binding sites in apo structures. At this time, groups are focusing on the analysis of occupied versus unoccupied pockets and having good success [11,32,33,41–49], but the methods could be further refined with data on the invalid ligands identified within Binding MOAD.

Liang et al. [11] also found that binding sites are either buried cavities or more often pockets with one, occasionally two, exposed openings [11]. In agreement with that study, our histograms in Fig. 8 show that most ligand-binding sites have limited exposure to solvent; GoCAV data shows that 70% of the cavities have $\geq$70% of their MSA buried, and NACCESS data shows that 85% of the ligands have $\geq$70% of their SASA buried. The high degree of burial also

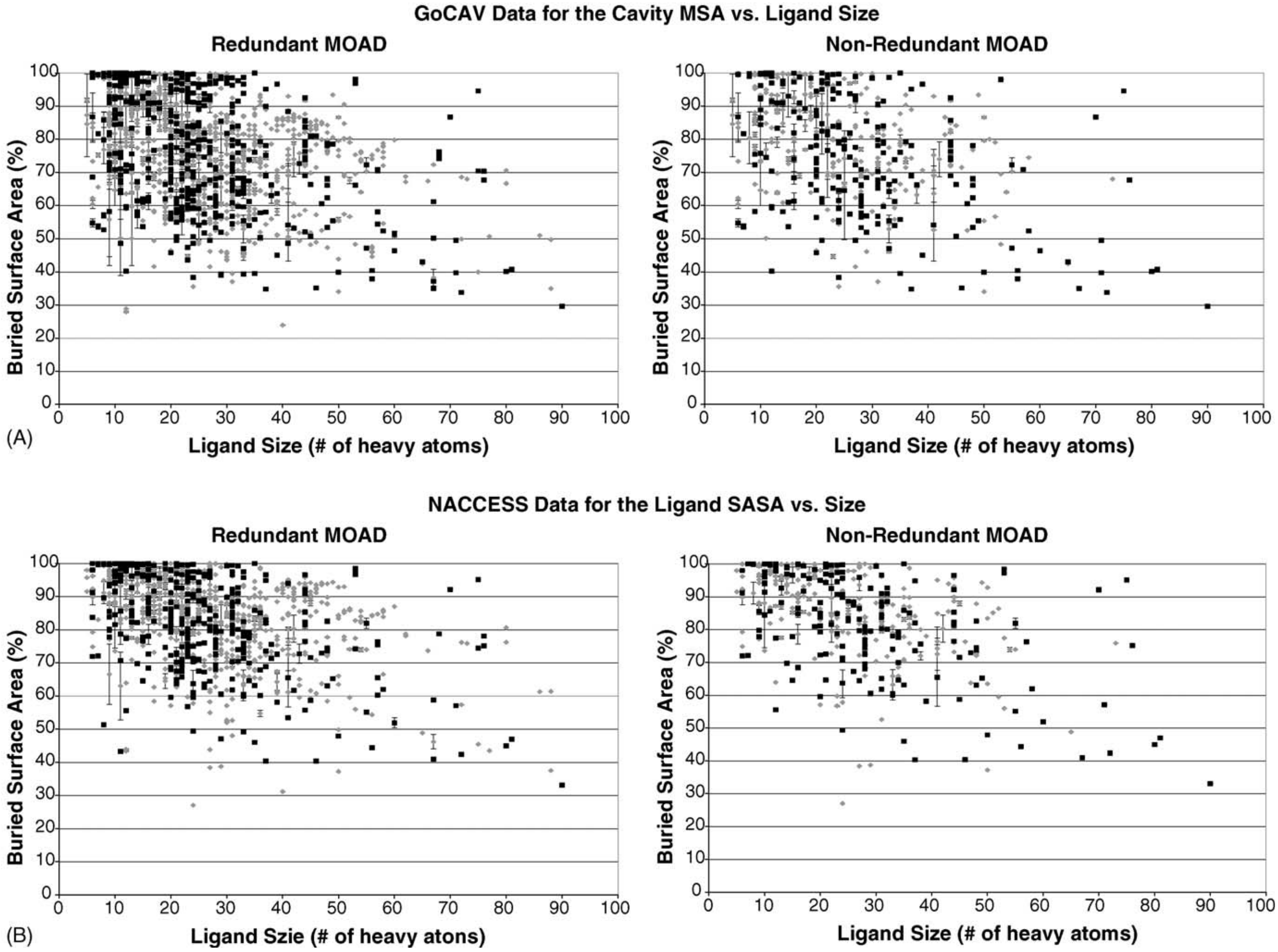


Fig. 9. The largest ligands tend to have much of their surface area exposed to solvent (low % buried). (A) Percentage of buried MSA of the cavity and (B) percentage of buried SASA of the ligand. The data points in black squares are from complexes with K_d data, and gray diamonds are used for complexes with K_i or IC_{50} data.

parallels findings by Keil et al. [44] where they show that binding sites for ligands are deeper and more concave than binding sites for protein–DNA or protein–protein associations. We found that the largest ligands are rarely well buried. They tend to have less percent buried MSA of the cavity and less percent buried SASA of the ligand (Fig. 9); many of them are short peptide or nucleic acid chains, again fitting with the findings that such binding sites are more shallow.

4. Conclusions

The histograms in Figs. 7A and 8 tell us that most ligands are well buried. This fits the common paradigm that many contacts between the ligand and the protein are a significant factor in molecular recognition. Fig. 9 shows that largest ligands tend to have more exposed surface area. These large ligands are typically peptide, nucleic acid, or sugar chains, and one would expect the patterns of binding such molecules to start to resemble the patterns of proteins binding macromolecules.

The general trends found here do not change with the choice of MSA of the pocket versus SASA of the ligand. Also, GoCAV and NACCESS use different methodologies and radii, so the patterns appear to be independent of how the calculation is performed. We do want to note that surfaces of the binding site calculated with GoCAV are not completely independent of the ligand because of the use of an ELS to ‘‘bound’’ the binding site. However, the probe never reaches the ELS boundary in a buried binding site. Most of our sites are highly buried, so the majority of the cavity surface is defined only by contacts to the protein. This typically makes the portion of the surface defined by the ELS only a small percentage.

In closing, future efforts with Binding MOAD will allow us to compare – broadly, for the first time – the binding affinity data to the patterns of molecular recognition mined from the PDB. Past studies have mined subsets of the PDB with various structural analyses of proteins and ligands

[11,15,25,32,33,41–55], but now, we will be able to add another layer of depth to such studies. There is more to binding affinity than just burying a ligand inside a protein, and all of the complex issues that go into creating an effective scoring function [56] will need to be considered in our analyses. Both shape and chemical complementarity are thought to be the basis of molecular recognition. Our future analyses will have to consider the chemical complementarity or what ‘‘types’’ of surfaces are solvent-exposed or interact with the protein. We will also need to address the very complex issue of entropic changes upon binding.

Acknowledgements

This work has been funded through the generosity of the Beckman Young Investigator Award Program (HAC). RDS would like to thank the University of Michigan's Molecular Biophysics Training Program for support (NIGMS grant GM008270). We thank Michael G. Lerner for his help with Fig. 2. We also thank Prof. Shaomeng Wang and Dr. Renxiao Wang (Medical School, University of Michigan, Ann Arbor) for many helpful discussions about PDBbind.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucl. Acids Res. 28 (2000) 235–242.

[2] N. Deshpande, K.J. Addess, W.F. Bluhm, J.C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, R. Kramer Green, J.L. Flippen-Anderson, J. Westbrook, H.M. Berman, P.E. Bourne, The RCSB protein data bank: a redesigned query system and relational database based on the mmCIF schema, Nucl. Acids Res. 33 (2005) D233–D237.

[3] Z. Feng, L. Chen, H. Maddula, O. Akcan, R. Oughtred, H.M. Berman, J. Westbrook, Ligand Depot: a data warehouse for ligands bound to macromolecules, Bioinformatics 20 (2004) 2153–2155.

[4] L. Hu, M.L. Benson, R.D. Smith, M.G. Lerner, H.A. Carlson, Binding MOAD (Mother of All Databases), Protein: Struct. Funct. Bioinfo. 60 (2005) 333–340.

[5] R.A. Laskowski, E.G. Hutchinson, A.D. Michie, A.C. Wallace, M.L. Jones, J.M. Thornton, PDBsum: a web-based database of summaries and analyses of all PDB structures, Trends Biochem. Sci. 22 (1997) 488–490.

[6] N.M. Luscombe, R.A. Laskowski, D.R. Westhead, D. Milburn, S. Jones, M. Karmirantzou, J.M. Thornton, New tools and resources for analysing protein structures and their interactions, Acta Crystallogr. D Biol. Crystallogr. 54 (1998) 1132–1138.

[7] R.A. Laskowski, PDBsum: summaries and analyses of PDB structures, Nucl. Acids Res. 29 (2001) 221–222.

[8] R.A. Laskowski, V.V. Chistyakov, J.M. Thornton, PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids, Nucl. Acids Res. 33 (2005) D266–D268.

[9] R.A. Laskowski, SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions, J. Mol. Graph. 13 (1995) 323–330.

[10] T.A. Binkowski, S. Naghibzadeh, J. Liang, CASTp: computed atlas of surface topography of proteins, Nucl. Acids Res. 31 (2003) 3352–3355.

[11] J. Liang, H. Edelsbrunner, C. Woodward, Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design, Protein Sci. 7 (1998) 1884–1897.

[12] A. Golovin, D. Dimitropoulos, T. Oldfield, A. Rachedi, K. Henrick, MSDsite: a database search and retrieval system for analysis and viewing of bound ligands and active sites, Protein: Struct. Funct. Bioinfo. 58 (2005) 190–199.

[13] J.-M. Shin, D.-H. Cho, PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures, Nucl. Acids Res. 33 (2005) D238–D241.

[14] M. Hendlich, Databases for protein–ligand complexes, Acta Crystallogr. D Biol Crystallogr. 54 (1998) 1178–1182.

[15] A. Bergner, J. Günther, M. Hendlich, G. Klebe, M. Verdonk, Use of Relibase for retrieving complex three-dimensional interactions patterns including crystallographic packing effects, Biopolymers 61 (2002) 99–110.

[16] M. Hendlich, A. Bergner, J. Günther, G. Klebe, Relibase: design and development of a database for comprehensive analysis of protein–ligand interactions, J. Mol. Biol. 326 (2003) 607–620.

[17] J. Chen, J.B. Anderson, C. DeWeese-Scott, N.D. Fedorova, L.Y. Geer, S. He, D.I. Hurwitz, J.D. Jackson, A.R. Jacobs, C.J. Lanczycki, C.A. Liebert, C. Liu, T. Madej, A. Marchler-Bauer, G.H. Marchler, R. Mazumder, A.N. Nikolskaya, B.S. Rao, A.R. Panchenko, B.A. Shoemaker, V. Simonyan, J.S. Song, P.A. Thiessen, S. Vasudevan, Y. Wang, R.A. Yamashita, J.J. Yin, S.H. Bryant, MMDB: Entrez's 3D-structure database, Nucl. Acids Res. 31 (2003) 474–477.

[18] R. Wang, X. Fang, Y. Lu, S. Wang, The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures, J. Med. Chem. 47 (2004) 2977–2980.

[19] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, S. Wang, The PDBbind database: methodologies and updates, J. Med. Chem. 48 (2005) 4111–4119.

[20] N. Paul, E. Kellenberger, G. Bret, P. Müller, D. Rognan, Recovering the true targets of specific ligands by virtual screening of the protein data bank, Protein: Struct. Funct. Bioinfo. 54 (2004) 671–680.

[21] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucl. Acids Res. 31 (2003) 365–370.

[22] D.L. Wheeler, C. Chappey, A.E. Lash, D.D. Leipe, T.L. Madden, G.D. Schuler, T.A. Tatusova, B.A. Rapp, Database resources of the National Center for Biotechnology Information, Nucl. Acids Res. 28 (2000) 10–14.

[23] W.L. DeLano, The PyMOL molecular graphics system, DeLano Scientific LLC, San Carlos, CA, USA, 2002. http://www.pymol.org.

[24] F.M. Richards, Calculation of molecular volumes and areas for structures of known geometry, Meth. Enzymol. 115 (1985) 440–464.

[25] J. Liang, H. Edelsbrunner, P. Fu, P.V. Sudhakar, S. Subramaniam, Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins, Protein: Struct. Funct. Genet. 33 (1998) 18–29.

[26] C.M.W. Ho, G.R. Marshall, Cavity search: an algorithm for the isolation and display of cavity-like binding region, J. Comput-Aided Mol. Design 161 (1990) 269–288.

[27] A. Goede, R. Preissner, C. Frömmel, Voronoi cell: new method for allocation of space among atoms: elimination of avoidable errors in calculation of atomic volume and density, J. Comput. Chem. 18 (1997) 1113–1123.

[28] B.J. McConkey, V. Sobolev, M. Edelman, Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure, Bioinformatics 18 (2002) 1365–1373.

[29] P.J. Fleming, F.M. Richards, Protein packing: dependence on protein size, secondary structure and amino acid composition, J. Mol. Biol. 299 (2000) 487–498.

[30] A. Poupon, Voronoi and Voronoi-related tessellations in studies of protein structure and interaction, Curr. Opin. Struct. Biol. 14 (2004) 233–241.

[31] D.G. Levitt, L.J. Banaszak, POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids, J. Mol. Graphics 10 (1992) 229–234.

[32] G.P. Brady Jr., P.F.W. Stouten, Fast prediction and visualization of protein binding pockets with PASS, J. Comput. Aided Mol. Des. 14 (2000) 383–401.

[33] M. Stahl, C. Taroni, G. Schneider, Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network, Protein. Eng. 13 (2000) 83–88.

[34] W.L. Jorgensen, J. Tirado-Rives, The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin, J. Am. Chem. Soc. 110 (1988) 1657–1666.

[35] A.J. Li, R. Nussinov, A set of van der Waals and Coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking, Protein: Struct. Funct. Genet. 32 (1998) 111–127.

[36] J. Tsai, R. Taylor, C. Chothia, M. Gerstein, The packing density in proteins: standard radii and volumes, J. Mol. Biol. 290 (1999) 253–266.

[37] S.J. Hubbard, J.M. Thornton, NACCESS version 2.1.1, University of Manchester, UK, 1996.

[38] B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility, J. Mol. Biol. 55 (1971) 379–400.

[39] C. Chothia, The nature of accessible and buried surfaces in proteins, J. Mol. Biol. 105 (1976) 1–14.

[40] R.L. DesJarlais, R.P. Sheridan, G.L. Seibel, J.S. Dixon, I.D. Kuntz, R. Venkataraghavan, Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure, J. Med. Chem. 31 (1988) 722–729.

[41] J. An, M. Totrov, R. Abagyan, Pocketome via comprehensive identification and classification of ligand binding envelopes, Mol. Cell. Proteomics 4 (2005) 752–761.

[42] R.A. Laskowski, N.M. Luscombe, M.B. Swindells, J.M. Thornton, Protein clefts in molecular recognition and function, Protein Sci. 5 (1996) 2438–2452.

[43] T.A. Binkowski, L. Adamian, J. Liang, Inferring functional relationships of proteins from local sequence and spatial surface patterns, J. Mol. Biol. 332 (2003) 505–526.

[44] M. Keil, T.E. Exner, J. Brickmann, Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network, J. Comput. Chem. 25 (2004) 779–789.

[45] F. Ferrè, G. Ausiello, A. Zanzoni, M. Helmer-Citterich, Surface: a database of protein surface regions for functional annotation, Nucl. Acids Res. 32 (2004) D240–D244.

[46] C. Hofbauer, H. Lohninger, A. Aszodi, Surfcomp: a novel graph-based approach to molecular surface comparison, J. Chem. Inf. Comput. Sci. 44 (2004) 837–847.

[47] A. Gutteridge, G.J. Bartlett, J.M. Thornton, Using a neural network and spatial clustering to predict the location of active sites in enzymes, J. Mol. Biol. 330 (2003) 719–734.

[48] V.A. Ivanisenko, S.S. Pintus, D.A. Grigorovich, N.A. Kolchanov, PDBSite: a database of the 3D structure of protein functional sites, Nucl. Acids Res. 33 (2005) D183–D187.

[49] B. Ma, T. Elkayam, H. Wolfson, R. Nussinov, Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 5772–5777.

[50] B. Halle, Flexibility and packing in proteins, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 1274–1279.

[51] M.J. Betts, M.J.E. Sternberg, An analysis of conformational changes upon protein–protein association: implications for predictive docking, Protein Eng. 12 (1999) 271–283.

[52] S. Zhao, D.S. Goodsell, A.J. Olson, Analysis of a data set of paired uncomplexed protein structures: new metrics of side-chain flexibility and model evaluation, Protein: Sturct. Funct. Genet. 43 (2001) 271–279.

[53] R. Najmanovich, J. Kuttner, V. Sobolev, M. Edelman, Side-chain flexibility in proteins upon ligand binding, Protein: Struct. Funct. Genet. 39 (2000) 261–268.

[54] X. Fradera, X. de la Cruz, C.H.T.P. Silva, J.L. Gelpí, F.J. Luque, M. Orozco, Ligand-induced changes in the binding sites of proteins, Bioinformatics 18 (2002) 939–948.

[55] L. Mao, Y. Wang, Y. Liu, X. Hu, Molecular determinants for ATP-binding in proteins: a data mining and quantum chemical analysis, J. Mol. Biol. 336 (2004) 787–807.

[56] H.J. Böhm, M. Stahl, The use of scoring functions in drug discovery applications, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, 18, Wiley/VCH Inc., New York, 2002, pp. 41–88.