JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

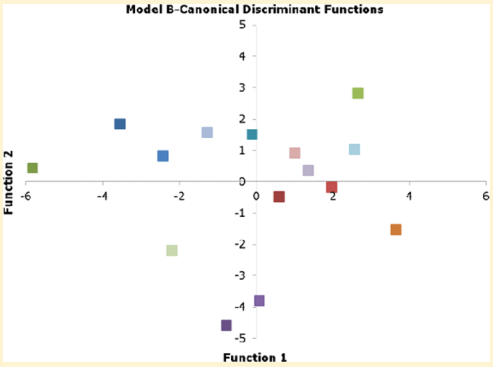
Functional Prediction of Binding Pockets

Maria Kontoyianni*,† and Christopher B. Rosnick§

†Department of Pharmaceutical Sciences and §Department of Psychology, Southern Illinois University Edwardsville, Edwardsville, Illinois 62026, United States

Supporting Information

ABSTRACT: Putative function for targets with no known ligands has typically been determined from liganded homologous proteins using sequence and structure comparisons. However, it is debatable what percentage of sequence identity implies similar function, whereas structural similarity is focused on global folds and could miss divergent structures and novel global folds. The present study describes an approach to classify a diverse set of proteins and predict their function. Descriptors corresponding to structural, physicochemical, and geometric properties of the ligand-binding cavities of a collection of 434 complexes (17 protein families) were calculated and analyzed by statistical methods. The best model using discriminant function analysis (DFA) consisted of 371 proteins (15 families) and had correct classification rates of 90% and cross-validation 86%. DFA with one protein and a random sample of the remaining proteins led to 100% correct prediction of putative protein function for 10 of the 15 protein families.



INTRODUCTION

The increasing number of X-ray, NMR, and model-built structures of receptors and enzymes has contributed greatly to the understanding of recognition events. This in turn has led to the recent successes of structure-based drug design (SBDD) and virtual screening. However, the rapid growth of SBDD-derived protein−ligand complexes has also created large amounts of data with information that can be leveraged by effectively mining the structural information and deriving knowledge applicable to drug discovery. Drug discovery has thus far focused on optimization of ligand physicochemical properties toward a single macromolecular target. With the advent of genomics, however, a noticeable growth in the number of protein sequences that lack experimental functional annotation has occurred. It is estimated that out of the 20,000 to 25,000 human genes coding for 600−1,500 druggable targets, a small subset of proteins (approximately 800) has been explored to date with medium affinity binders.[1−4] Of the above subset, only 140 human proteins are targeted by approved and orally bioavailable drugs.[1] Similarly, the chemical space has greatly expanded due to advances in parallel synthesis and combinatorial chemistry, but still only a fraction of these molecules are marketed drugs. It is thus apparent that pharmaceutical research has covered a mere portion of the available target and chemical spaces.

The breadth of genomic data has led to a disproportionate growth of uncharacterized genes and reinforced the importance of protein function prediction. Protein function is typically associated with the recognition and modification of endogenous ligands. However, defining function is not straightforward. For example, the biochemical function of a kinase is connected with phosphorylation of a substrate and, in turn, differs from its physiological function, which is associated with a signaling pathway, or its clinical relevance due to a mutation that can lead to a disease.[5] Regardless, functional prediction is fundamental to future advancements in drug discovery. However, even if function is adequately defined, its computational description can be a daunting task. Putative protein function is inferred from sequence or structural comparisons. Sequence homology has long been used to assign biological function, with the understanding that it only covers 50% of proteins, and it is accurate when the homology is more than 30%.[6−9] The underlying principle is that similarity in sequence suggests similarity in structure and function. However, it is not clear what percentage of identity is needed to imply that two proteins have similar function.[6,10] It has been claimed that a 40% sequence identity could be reflective of similar catalytic mechanisms, but information loss due to a high false-negative rate can still occur at such high identity percentages.[5,11] Consequently, pattern-based methods have been developed to recognize local regions of sequence similarity in the absence of full-length alignments. This is because only a small part of a structure is used to perform a certain function, and therefore identifying a sequence-based signature suffices to infer function. Whether based on a database of patterns, such as PROSITE,[12,13] or profiles constructed from whole protein families, as in SUPERFAMILY,[14−16] these methods provide greater sensitivity and often additional information about the possible function of a new protein.[17] Nevertheless, the resulting annotation is at the level of a probable superfamily or family at

Received: September 27, 2011
Published: February 21, 2012

dx.doi.org/10.1021/ci2005912 | J. Chem. Inf. Model. 2012, 52, 824−833

best, leading to a prediction of function with all the uncertainty that comes from erroneous database annotations.

Because proteins that show no sequence homology can still have similar overall folds, functional assignment has been augmented beyond sequence homology by using fold similarity based on structural comparisons. The reasoning is that tertiary structures are more evolutionarily stable than sequences. Most of the methodologies aiming at structural comparisons use the protein domain as the basic unit of classification. Domains are usually grouped into species and hierarchically classified into families, superfamilies, folds, and classes. Among the various programs developed within this framework, SCOP,[18−20] CATH,[21,22] and FSSP[23] are the most widely used. It should be mentioned that although these programs employ different methods and degrees of automation and are also based on distinct rules of protein structure and taxonomy, they seem to agree on the majority of their classifications.[24] Secondary Structure Matching,[25] Combinatorial Extension,[26] and DALI[23,27,28] are servers that compare 3D protein structures. One can submit the coordinates of a query protein structure, and the structural neighbors are identified with one of these services. However, problems arise for functionally diverse superfold families and for novel global folds.[11] It has actually been estimated that about 60% of the structures correspond to novel fold(s), and therefore assignment of function based on pre-existing data is rather challenging. In these cases, possible biochemical activities may be inferred which require a closer assessment of other characteristics, including surface, residue conservation, and active sites, thus making the whole process time-consuming and uncertain. Furthermore, none of the above methods takes into consideration how the residues surrounding the active site are oriented, which in turn is a geometric component affecting specificity.[29,30] Consequently, and although alternatives may be in place, there is information inherent in the active sites pertaining to charge, shape, size, etc., along with positioning of functionally important residues, which cannot be detected with convergent evolution. Structural and sequence comparisons "view" the proteins in a global fashion, while biological response is elicited from the interaction of a molecule with the protein's binding pocket.[31,32] Therefore, much information can be gained from studying and comparing active sites. Besides, it has been noted that divergent sequences and structures could have conserved binding sites that will be amiss when global comparisons of proteins are being made.[5,17,33]

As already mentioned, the human genome project has made available many potential new targets for drug intervention. Integration of new approaches in high-throughput genomics and chemistry disciplines led to chemogenomics,[34] which attempts to find all molecules capable of interacting with 'a' macromolecular target. It basically defines and matches target space with ligand space and vice versa. Chemogenomics suggests that data on targets with no known binders should be derived from the closest 'liganded' targets, and data on ligands with no respective targets should be derived from the closest 'targeted' ligands. Inherent to both directions is the reasoning that compounds sharing some similarity should also share targets, and targets sharing similar ligands should have similar binding sites. Therefore, it becomes obvious that given the number of available and unexplored targets, the contribution of the field can be substantial.[35] The inference that targets with similar ligands can potentially have similar binding sites has led to annotation of binding pockets in order to derive function. Toward that goal, several methods have been described in the literature for binding site definition. Rarey and co-workers did a thorough overview and grouped currently available software for binding pocket detection into categories, namely geometry-based, energy-based, and those that include combined approaches.[36] We will not discuss in detail any of these methods, as the interested reader is referred to relevant publications and reviews.[37−56] We will however present one representative of each group in order to provide a basic understanding of the different approaches employed. LigSite[47,57] (geometry-based method) uses a grid to detect pockets on the surface of proteins that could potentially be binding sites for small molecules. In GRID[58,59] (energy-based algorithm), the protein is embedded in a three-dimensional (3D) grid with the physicochemical properties of the atoms being mapped onto the grid points. SiteMap[42] (combined approach) employs a grid first and then uses geometric and energetic criteria to filter out irrelevant site-points, followed by calculations of hydrophobicity and hydrophilicity at key grid points.

An alternative approach for the description of binding pockets is the use of descriptors. Druggability mostly refers to the potential of a region to be therapeutically relevant upon ligand binding due to the presence of folds favoring interactions with drug-like molecules. Druggability and protein function have been used interchangeably. However, modulating a protein's function does not always lead to a therapeutic benefit, even if its structure dictates it is druggable.[3] Computational approaches for binding site identification using descriptors are relatively common;[60−66] however, descriptor-based comparison methods for protein function prediction are scarce. The most relevant work was reported by Honig and colleagues, who used 408 physicochemical attributes to select the drug-binding sites among all putative cavities in each of the 99 proteins in their data set.[41] These investigators defined a classifier that could detect regions on the protein surface with the potential for high binding affinity as opposed to nondrug binding cavities. Similarly, SiteMap utilizes properties, such as hydrophilicity, hydrophobicity, surface, donor, acceptor, and metal-binding to identify *druggable* sites.[42] Finally, Linusson and colleagues calculated 264 physicochemical properties for 239 protein cavities and mapped the ligand-binding pockets using Principal Component Analysis in order to identify important properties.[67]

The objective of our work is to use physicochemical properties to predict putative function of binding pockets. In navigating through a wide spectrum of 434 proteins (data set at the outset of this work), our work includes extraction of the protein features that are essential for eliciting biological response, followed by a representation of these features in computational terms and discriminant function analysis (DFA). Inherent to our approach is the fundamental tenet that function is linked to specific recognition elements of substrates or endogenous ligands with their respective binding pockets. Even if proteins fall into the same functional space, this does not necessarily imply they are similar geometrically. Besides, function is not confined to a particular fold. Our results indicate that we can indeed predict putative protein function with appropriate descriptors. The advantage of this work is that once a target is classified accurately, it makes hit identification easier by initially using ligands from targets of similar functions. The complete data set is shown in Table 1 and corresponds to a total of 434 proteins or 17 protein families. It should be noted

Table 1. Complete List of the Complexes (n = 434)

protein	PDB
carbonic anhydrases (n = 25)	1a42, 1azm, 1bcd, 1bnn, 1bnw, 1cim, 1czm, 1g4o, 1i8z, 1if7, 1jd0, 1kwq, 1okn, 1xpz, 2f14, 2foy, 2hd6, 2hnc, 2nmx, 2nn1, 2nn7, 2pou, 3caj, 3czv, 3da2
carboxypeptidases (n = 22)	1bav, 1cps, 1hdq, 1hdu, 1hee, 1iy7, 1zg7, 2c6c, 2ctc, 2jbj, 2jbk, 2pcu, 2piy, 2pj6, 2rfh, 3bi1, 3d7d, 3d7g, 3d7h, 6cpa, 7cpa, 8cpa
cysteine proteases (n = 22)	1gmy, 1ito, 1mem, 1qdq, 1sp4, 1tlo, 1tu6, 1vsn, 1zcm, 2aux, 2bdl, 2dc8, 2dcb, 2f7d, 2ftd, 2g8e, 2g8j, 2nqg, 2nqi, 2r6n, 2r9c, 3c9e
DHFRs (n = 25)	1aoe, 1dg5, 1dg7, 1dis, 1dlr, 1dr1, 1dre, 1drf, 1dyh, 1hfp, 1jol, 1klk, 1kms, 1kmv, 1ly3, 1mvt, 1ohk, 1pd8, 1rf7, 1rx4, 1s3u, 2ano, 2bl9, 4dfr, 7dfr
HIV (n = 25)	1aaq, 1ajv, 1bve, 1ec0, 1hps, 1hpv, 1hsg, 1hsh, 1htf, 1hvr, 1hwr, 1ida, 1kzk, 1ohr, 1pro, 1qbt, 1qbu, 1sbg, 1w5x, 2bpw, 2bpx, 2bpz, 3ebz, 3ecg, 7upj
immunoglobulins (n = 20)	1a4k, 1a6v, 1a8j, 1aj7, 1c5c, 1clz, 1d6v, 1fig, 1flr, 1i7z, 1ibg, 1igj, 1n7m, 1um4, 1uwg, 2cgr, 2dbl, 2mcp, 2pcp, 4fab
kinases (n = 40)	1ckp, 1m2q, 1ydr, 1yds, 1ydt, 2uzo, 2w1e, 2w7x, 2wmq, 2wmw, 2wqo, 2wtc, 3d2k, 3h0z, 3idp, 3ii5, 3iq7, 3jxw, 3jy0, 1byg, 1fgi, 1qpe, 2fgi, 2g1t, 2rgp, 2wd1, 2wqb, 3dpk, 3efj, 3ekn, 3et7, 3eta, 3ewh, 3f3v, 3f82, 3fzs, 3g0e, 3g5d, 3gql, 3kex
nuclear receptors (n = 27)	1dkf, 1ere, 1exa, 1fcx, 1fcz, 1fd0, 1fm6, 1fm9, 1gwq, 1gwx, 1h9u, 1k74, 1mvc, 1r5k, 1sj0, 1sr7, 1uhl, 1zuc, 2ama, 2ax9, 2axa, 2iog, 2iok, 3b5r, 3b65, 3erd, 3ert
serine proteases (n = 26)	1d3p, 1d4p, 1dfp, 1etr, 1ezq, 1f0r, 1f0t, 1fjs, 1k1j, 1k22, 1kye, 1lqd, 1mq5, 1mq6, 1pph, 1ql9, 1rgq, 1uvs, 1uvt, 1vgc, 2gv6, 2gv7, 2hvx, 4gch, 6gch, 7gch
isomerases (n = 33)	1a7x, 1f40, 1fkd, 1fkf, 1fkg, 1fkh, 1fki, 1j4h, 1j4i, 1j4r, 1nsg, 1qpf, 2dg3, 2fke, 2vn1, 3fap, 1gyx, 1gyy, 1nfs, 1nfz, 1oh0, 1ppv, 1ppw, 1x83, 2ick, 2ooh, 2oow, 2ooz, 3dji, 3jsf, 3jsg, 3jtu, 8cho
synthases (n = 36)	1f4e, 1f4f, 1f4g, 1jg0, 1jtq, 1lca, 1nce, 1nje, 1syn, 2bbq, 2vf0, 3bnz, 1rri, 1rrw, 1rry, 1rs4, 1rsd, 1rsi, 1u68, 1zaj, 2nm3, 3c52, 1a50, 1c29, 1c8v, 1c9d, 1cw2, 1cx9, 1k7x, 1kfj, 2clh, 2cli, 2clm, 2j9x, 2j9y, 2qjg
metalloproteases (n = 24)	1bqo, 1ciz, 1usn, 2d1o, 3usn, 1mmq, 1mmr, 1mmp, 2ddy, 1bzs, 1mnc, 1zs0, 1jh1, 1mmb, 1zp5, 1zvx, 3dng, 3dpe, 1thl, 1z9g, 5tln, 7tln, 1c3i, 1d5j
channels (n = 19)	1lb9, 1lbc, 1m5c, 1n0t, 1nnk, 1nnp, 1s9t, 1syh, 1vso, 2cmo, 2f35, 2ojt, 2pbw, 2qs1, 2qs4, 2wky, 2znu, 3en3, 3gba
hydrolases (n = 15)	1fdk, 1fx9, 1l8g, 1mkv, 2azz, 1li4, 2zj0, 3ce6, 3dhy, 3glq, 1a8t, 1hlk, 1l2s, 1q2p, 2a49
cytochrome P450s (n = 25)	1n6b, 1nr6, 1og5, 1r9o, 1w0f, 1w0g, 1z10, 1z11, 2bdm, 2fdu, 2fdv, 2fdw, 2fdy, 2hi4, 2j0d, 2nni, 2nnj, 2q6n, 2v0m, 2vn0, 3e4e, 3e6i, 3ebs, 3g5n, 3g93
gamma carboxylases (n = 16)	2v59, 2v5a, 2w6m, 2w6n, 2w6o, 2w6p, 2w6q, 2w6z, 2w70, 2w71, 3ff6, 3g8c, 3jrx, 3jzf, 3jzi, 3k8x
transferases (n = 34)	2aou, 2aox, 2g71, 2obf, 3fpd, 3hcc, 3k5k, 1mzs, 1wum, 2gfx, 2jev, 2qia, 2vba, 2vkz, 2wge, 3biy, 3cxq, 2f3m, 1twz, 2a57, 2c92, 2c97, 2c9d, 2dzb, 2o6h, 2r2l, 2vi5, 2zir, 3csj, 3erg, 3f63, 3h21, 3h2o, 3ksl

that the terms "families" and "classes" are used interchangeably throughout this work.

COMPUTATIONAL METHODS

Complex and Site Preparation. For all computations, Discovery Studio 2.5 was employed within the Accelrys suite of programs (Accelrys Inc., San Diego, CA 92121). The data set was compiled using the protein databank, the Structural Classification of Proteins (SCOP),[68] the Enzyme Classification (EC) system,[69] and the Washington University Basic Local Alignment Search Tool Version 2.0 (WU-Blast2).[70,71] Processing of complexes was as follows: proteins were imported and by default Gasteiger charges and formal charges were automatically calculated. Subsequently, ligands, subunits not involved in ligand binding and located far from the binding pocket, counterions, and other small molecules were removed. Metal ions were preserved whether in the binding pocket or not. Solvent molecules were preserved initially, in case they might be needed for subsequent descriptor calculations, but were removed in the end accordingly. The Clean Protein command in the Protein Reports and Utilities tool panel was first used, and then hydrogens were added and adjusted using the CHARMm force field from the Simulate Structures tool panel.[72,73] Assignment of the force field can change the formal charges calculated upon initial reading of the complex, according to the protonation state of the residues, while partial charges were assigned based on atom and residue templates. This final "clean" protein was used for all calculations. Binding pockets were created as groups and defined based on the ligand and a radius of 10 Å, using the "select" option in Discovery Studio.

Descriptors. Ligands were extracted from the active sites prior to calculating the descriptors. A total of 18 descriptors were calculated at first. A Connolly solvent surface was created with a probe atom of 1.4 Å radius. Type was set to "closed" so that the surface would completely enclose all atoms. Surface area and volume were calculated as properties of this Connolly surface encompassing and occupied by the binding site residues, respectively. Specifically, the surface area property was calculated by summing up the areas of all triangles in the triangle mesh representation of the surface. The volume property is estimated as 0.95 × (A - B), where A is the sum of the volumes of atomic spheres having van der Waals (vdW) radii, and B is the volume of all intersections of these atomic spheres for bonded pairs of atoms. For hydrophobicity a customized script was used (Jodi Shaulsky, personal communication) employing the standard Kyte and Doolittle scale.[74] The script creates a binding site object at the cavity of the protein designated as the binding pocket. It then sums the hydrophobicity of the residues within three angstroms of the site, and if the sum is >1.0 the cavity is considered hydrophobic and all waters within 3 Å of the binding site object are deleted. When a cavity is defined as hydrophobic, residues lining the cavity are added to a new group with the name Hydrophobic_Cavity_, followed by an incremented number.

dx.doi.org/10.1021/ci2005912 | J. Chem. Inf. Model. 2012, 52, 824−833

Partial charges were assigned using the CHARMm force field. For the molecular weight, radius of gyration, center of mass, density, number of rotatable bonds, number of hydrogen bond acceptors and donors, a script was used. For radius of gyration, the same definition as for small molecules was used that is, the mass weighted root-mean-square average distance of the atoms in the pocket from the center of mass of the pocket. The center of mass returns three numbers corresponding to the x, y, z coordinates of the center of mass of the site. Density is a 3D spatial descriptor that was defined as the ratio of molecular weight to molecular volume. It reflects the types of atoms and how tightly they are packed in the pockets. Of the acceptors, nitrogens must have at least one lone pair of electrons, if they are sp^2-hybridized they should not be connected by a single bond to another sp^2, while O, S, P must have at least one lone pair of electrons. For the scoring functions used as descriptors, the Score Ligand Poses protocol was used. First, cavities were defined as the grid points not occupied by the receptor. An "eraser" algorithm then removed all grid points "outside" the protein, with the boundary to the inside defined as points not reachable by the virtual erasers. The Score Ligand Poses protocol was then employed; it calculates scoring functions in order to evaluate ligand binding in the cavities. Similarly, the contact points were calculated using the Analyze Ligand Poses protocol. Another binding site object was based on the volume of the ligand and the volume occupied by the cavity. The space occupied by the ligand was extended slightly and built with potassium atoms representing a generic molecule occupying the cavity. The solvent accessible surface was the difference between the area when potassium atoms are or not present. The contact points were determined based on a threshold value which corresponds to vdW fraction. Descriptors employed are listed in Table 2.

Table 2. List of Calculated Descriptors for the Binding Pockets

spatial and electronic	structural	geometric
surface area[38,39]	molecular weight	contact points[40]
radius of gyration	hydrogen bonding	center of mass x, y, z
density	number of rotatable bonds	Ligscore1 components[80]
volume		PMF04[81]
hydrophobicity		PLP1[82]
electrostatics/partial charges		number of amino acids

Data Analyses. We conducted DFA to determine if the 18 descriptors presented in detail above could reliably discriminate between the 17 proteins classes (see Table 1). The discriminant function for the kth group (k = 1, 2,..., 17) is calculated as follows

$$\text{Protein class}_k = \text{constant}_{k0} + a_{k1}x_1 + a_{k2}x_2 + a_{k3}x_3 + \ldots + a_{kn}x_n$$

where x_1, x_2, x_3, ..., x_n represent the descriptor variables, and a_1, a_2, a_3, ..., a_n represent the coefficients associated with each descriptor. Leave-one-out cross validation was used to validate the models. For all analyses, prior probabilities were not assumed to be equal and were calculated based on group sizes.

RESULTS AND DISCUSSION

The primary objective of this work was to explore whether we could discriminate protein families by 2D and 3D descriptors reflecting respective binding pockets. Structural genomics is producing a growing number of protein structures, many of which have limited functional annotation. As discussed in the Introduction, function has historically been based upon and/or inferred by sequence or fold similarity using bioinformatics tools that rely on evolutionary methods. These approaches aim at recognizing similarity of the not-yet annotated protein with a protein whose function is known. However, known limitations of the evolutionary methods are as follows: (1) dependence on the quality of the sequence alignment tools and the number of sequences used, (2) lack of concrete evidence regarding the percentage of homology required for a viable annotation, and (3) inadequacy in addressing the high number of novel global folds and functionally diverse superfold families. An alternative approach to detect function is extracting information from sets of structurally related proteins, which can be subdivided into two steps: 1) identification of regions on the protein surface that could potentially be binding sites and 2) recognition of similarity in the binding pattern of these sites with a well-known protein which could lead to a better understanding of its function. The underlying assumption of this approach is proteins with structurally similar active sites will have similar functions.[36,75] Limitations pertaining to active-site identification techniques and drawbacks stemming from the simplified representations of pocket residues, as employed by search engines, have proven challenging. Quite often assessing the biological significance of the recognized similarity can add to the complexity of predicting putative function.

We took a different angle, in that our approach is not an active or binding site identification method, but it instead capitalizes on the knowledge base derived from existing binding sites in the form of 2D and 3D descriptors for each site in an effort to predict likely function. To our knowledge, the most similar work reported to date explored the protein surface cavities of a 99-member data set, characterized each cavity (approximately 14 cavities per protein) with a set of 408 descriptors, and subsequently developed a classifier to predict which of these were most likely involved in drug binding.[41] In contrast, our protein data set is much larger as it consists of 434 protein complexes, which in turn represent 17 protein families (classes), and the number of descriptors is substantially smaller. It should also be pointed out that we used the binding pocket location as our starting point rather than our objective. In essence, the work presented herein is based on the use of functional site descriptors for discrete protein biological functions. The descriptors are structural, geometric, and thermodynamic representations of the protein binding pockets in 3D space.

Complexes included in this study are presented in Table 1. The criteria used for target selection were rather straightforward: sufficient data points per protein family, preference for human species, exclusion of isozymes and mutants, noncovalent binding between the ligand and respective protein, resolution of the crystallographic complex should be less than 3.0 Å, and the bound ligand was preferably an inhibitor. It can be seen that not only representatives from each of the main divisions (classes) of the enzyme classification system are included, but we have also covered the majority of subclasses within each of these divisions. Hydrolases and gamma-carboxylases have the lowest

number of representatives, with 15 and 16 complexes, respectively. The reason for the reduced representation of these two families in our data set stems from their relatively limited presence in the protein databank. If we use SCOP's nomenclature, it also becomes obvious that our classification is not consistent hierarchically that is, certain categories are at the 'family' tree of SCOP, while others are at the 'protein' level. For example, the 'isomerase' class consists of a variety of proteins whose function is similar; however, they are geometrically and structurally very distinct. In contrast, families such as dihydrofolate reductase (DHFR) or HIV have representatives which are homologous proteins both functionally and structurally. We were also conflicted about the transferases (methyl-acyl and aryl-alkyl), which are grouped separately in our data set (Table 1) from the thymidylate synthases and kinases, although the latter two are both under the umbrella of transferases, based on the EC convention. We chose this division because there is enough structural identity in these individual classes to be independent categories, while the remaining are grouped as the respective EC general class. Another point of contention was the number of representative members in each protein family. In an ideal situation, all protein families would have 29 representative members (434/17) and the probability of being in any one protein family would be equal (approximately 6.7%). Even though we made every effort to have relatively equal sample sizes, availability of crystal protein complexes proved a deterrent. As can be seen in Table 3, the number of representatives per protein family ranged from 15 to 40, and this translated to a 3.5%–9.2% probability of being in any one protein family. Due to this large range of representatives per protein family (and probabilities), we used the protein family size to generate our classification rates. However, we will revisit the above as we present our statistical findings in the following sections.

Table 3. Sample Size Comparisons

protein class	original *n*	outliers	revised *n*	Model A	Model B
carbonic anhydrases	25		25	25	25
carboxypeptidases	22	2jbj	21	21	21
cysteine proteases	22	1tu6, 2ftd	20	20	20
DHFRs	25	2bl9	24	24	24
HIV	25	1bve	24	24	24
immunoglobulins	20	1a6v	19	19	19
kinases	40		40	40	40
nuclear receptors	27	1r5k	26	26	26
serine proteases	26	4gch	25	25	25
isomerases	33	1qpf, 1f40, 2dg3	30	30	30
synthases	36		36	36	36
metalloproteases	24		24	24	24
channels	19		19	19	19
hydrolases	15		15	---	---
cytochrome P450s	25	2j0d, 1n6b, 1nr6	22	22	22
gamma carboxylases	16		16	16	16
transferases	34	3k5k, 2qia, 2vkz, 3ksl, 2r2l	29	29	---
total	**434**		**415**	**400**	**371**

Regarding binding pockets, they were located based on the shape of the ligands and were subsequently adjusted for their size so that they would encompass the bound molecules. It should be pointed out that descriptors, such as number of hydrogen bonds, hydrophobicity, contact points, and the scoring functions (Ligscore 1, PLP1, and PMF), are implicitly describing the protein–ligand complexes, thus adding an element of molecular recognition. Furthermore, each of the aforementioned scoring functions places a different weight on receptor–ligand interactions, and therefore we expected that their contributions would be distinct.

Assumptions for DFA. It has been reported that some of the typical multivariate assumptions (e.g., multicollinearity) are more relaxed for DFA analyses when 1) the main goal is classification and 2) analyzing data sets with large sample sizes,[76] as is the case with the present data set. Therefore, we did not consider multicollinearity to be an area of concern for our analyses. However, we examined our data for multivariate outliers, as presented in the following section, and used Box's M test to examine differences in the variance-covariance matrices between protein classes in each analysis.

Model Refinement. The model refinement process is depicted in Figure 1. Description of each stage in the diagram is presented below.

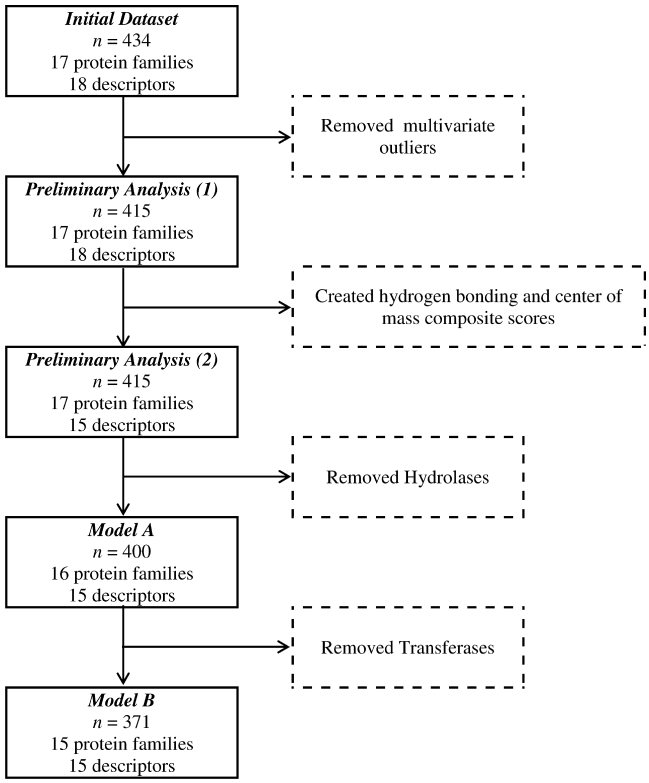


Figure 1. A schematic overview of the model refinement process.

Multivariate Outliers. Multivariate analyses such as DFA are sensitive to outliers; therefore, we tested to see if there were any multivariate outliers in our data. In order to do this, we conducted a regression using all 18 descriptors to calculate Mahalanobis distances.[77] The Chi-square critical value with 18 degrees of freedom and an alpha level of 0.001 is 42.31. The results suggested that there were 19 multivariate outliers, which were removed from all analyses to ensure reliable estimates (see Table 3). After removing these 19 complexes, our sample consisted of 415 protein complexes or 17 protein families.

dx.doi.org/10.1021/ci2005912 | J. Chem. Inf. Model. 2012, 52, 824–833

Preliminary Analyses. First, we ran the DFA with all 18 descriptors and the 17 proteins as the dependent variable to determine whether or not we had a sufficient number of complexes in each protein group to adequately fit the model (results not shown). As anticipated and alluded to earlier, the initial analysis revealed that reliable estimates could not be generated for the hydrolases (n = 15) and gamma carboxylases (n = 16). Because the sample size within each family should be greater than the number of descriptors, and in order to be more parsimonious we opted to decrease the number of descriptors by creating two new descriptor variables: 1) "hydrogen bonding", which is the sum of the hydrogen bond acceptors and donors, and 2) "center of mass composite", by z-scoring the three center of mass descriptors (x, y, and z coordinate space) and calculating the mean. We reasoned that the creation of the two composite scores were warranted because hydrogen bond acceptors and donors were highly correlated (r = 0.92), while using the mean of the center of mass coordinates made sense conceptually. We subsequently ran the DFA with the 15 new descriptors and found that a reliable estimate could now be generated for gamma carboxylases; however, we were still unable to generate a reliable estimate for hydrolases. Hydrolases are quite diverse at all levels of structural hierarchy. This in turn could prove detrimental to the overall quality of the model, in particular due to lack of adequate crystal protein complexes, should we decide to divide the hydrolases into subfamilies. We thus decided to remove hydrolases from all subsequent analyses,[78] which reduced the total data set size to 400 proteins.

Model A. As discussed above, after creating a more 'generalized' version of some of the descriptors, we now had 15 descriptors and 16 protein families or 400 proteins (Model A in Tables 3 and 4). In this model, 15 functions were generated in order to discriminate among the 16 protein families. Individually, all descriptors were able to distinguish between the 16 protein families. The Box's M test was significant (p < 0.001), leading to the conclusion that the variance-covariance matrices between protein families were not homogeneous. However, given our large sample size (n = 400) and the concern that using separate covariance matrices would overfit the data,[76,78,79] we did not consider this to be a violation of concern and proceeded to interpret the findings. It can be seen in Table 4 that this model correctly classified 84.5% of proteins, while cross-validation classified a large portion (77.0%) of the data successfully. Ligscore1 had the largest correlation with function one and density had the largest correlation with function two. All individual protein families had classification rates greater than 80% with the exception of channels (73.7%) and the transferases (31.0%). Furthermore, ten of the protein classes were correctly classified at greater than 80% in the cross-validation analysis. To summarize, Model A shows that the selected descriptors can indeed predict putative protein function at the 84% level. Figure 2 illustrates the discriminatory power of the first two functions generated in this analysis, in that there is no overlap of protein families. Specifically, the first two functions account for about 54% of the variance across protein families (data not shown). Based on these findings, and since transferases were poorly predicted, we felt we should explore the extent to which the model might improve if transferases were eliminated.

Table 4. Comparing the Correct Classification Rates Between Models A and B for the Overall DFA Models

protein class	Model A	validation	Model B	validation
carbonic anhydrases	88.0	84.0	88.0	84.0
carboxypeptidases	95.2	85.7	95.2	90.5
cysteine proteases	100.0	95.0	95.0	95.0
DHFRs	87.5	87.5	87.5	87.5
HIVs	95.8	95.8	100.0	95.8
immunoglobulins	94.7	78.9	94.7	89.5
kinases	85.0	75.0	87.5	85.0
nuclear receptors	92.3	92.3	92.3	92.3
serine proteases	84.0	80.0	84.0	80.0
isomerases	86.7	73.3	83.3	80.0
synthases	83.3	75.0	88.9	72.2
metalloproteases	95.8	91.7	95.8	95.8
channels	73.7	52.6	94.7	89.5
cytochrome P450s	86.4	81.8	86.4	81.8
gamma carboxylases	87.5	81.3	87.5	81.3
transferases	31.0	17.2	---	---
overall	**84.5%**	**77.0%**	**90.3%**	**86.0%**

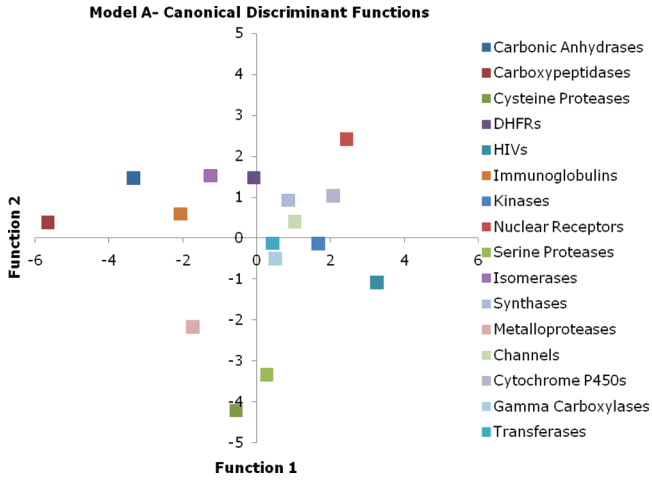


Figure 2. Plot of group centroids for Model A demonstrating the discriminatory properties of the first two discriminant functions.

Model B. When we eliminated the transferases, our sample size was reduced to 371. The DFA for Model B included 15 descriptors and 15 protein families (Table 3). Similar to Model A, all descriptors were able to discriminate between the 15 protein classes and the Box's M test was significant (p < 0.001). In this model, 14 (number of protein families-1) functions were generated in order to discriminate between the 15 protein families. In contrast to Model A, density had the largest correlation with both functions one and two. Overall, this model correctly classified 90.3% of the protein classes and the cross-validation analysis classified 86.0% of the data successfully (Table 4). This is a 5.8% increase in correctly classifying protein functions in the original data, and a 9% increase in correct classification in the cross-validation analysis. The individual proteins were correctly classified greater than 83% of the time in the original data and synthases were the only protein family to be correctly classified at less than 80% in the cross-validation analysis. This model does very well in discriminating between the 15 protein families used in the analysis (see Figure 3).

dx.doi.org/10.1021/ci2005912 | J. Chem. Inf. Model. 2012, 52, 824–833

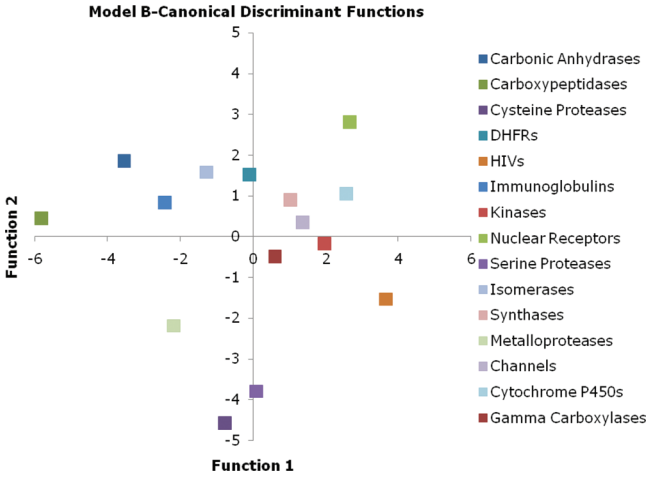


Figure 3. Plot of group centroids for Model B demonstrating the discriminatory properties of the first two discriminant functions.

Alternative Approach - ONE Protein versus All Others. *DFA Discriminating between ONE Protein and All Other Proteins Collapsed.* Based on the high discriminatory power of our findings, we next wanted to know if we would be able to predict the putative biological function of a novel protein. Our training data set obviously captures a wide spectrum of protein families, while the chosen descriptors are apparently capable of discriminating and predicting respective protein functions. Consequently, we next performed a DFA in order to determine if the 15 descriptors could discriminate between one protein versus all others. When discriminating among all proteins, the classification model is informative, in that it proves the concept of protein function annotation using descriptors for the binding pockets. However, if we were to run the DFA using just one protein family, carbonic anhydrase for example, versus all other proteins, the resultant analysis would offer a more qualitative impact. Ultimately, we conducted 15 DFAs with each protein pitted against all other proteins. There is only one function generated for each analysis, and all of these functions significantly discriminated between the protein of interest and all other proteins ($p < 0.001$). Results for the canonical correlations, correct classification rates, and cross-validation rates are depicted in Figures 4, 5, and 6, respectively.

Briefly, the results revealed that 10 of the 15 protein classes were correctly classified with greater than 80.0% accuracy.

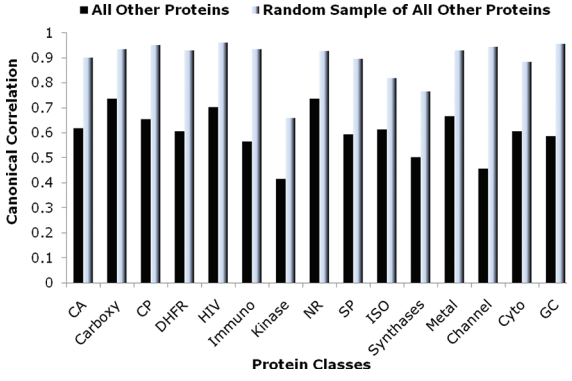


Figure 4. Canonical correlations for each of the functions when 1) comparing one protein and all other proteins collapsed (dark bars) and 2) comparing one protein and a random sample of all other proteins (light bars).

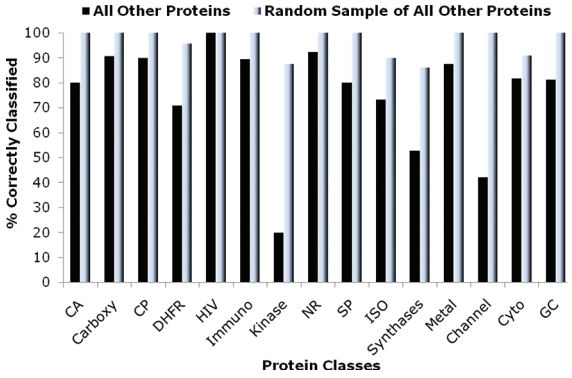


Figure 5. Percent correctly classified when 1) comparing one protein and all other proteins collapsed (dark bars) and 2) comparing one protein and a random sample of all other proteins (light bars).

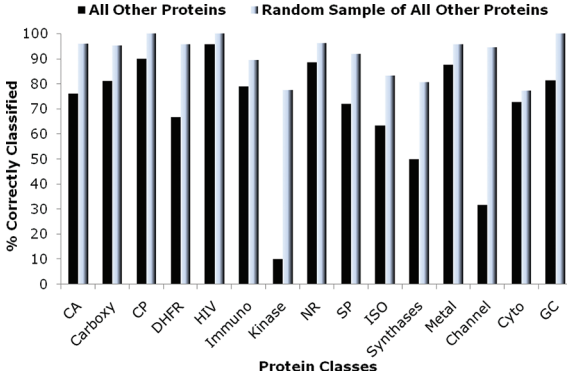


Figure 6. Percent correctly classified in the cross-validation when 1) comparing one protein and all other proteins collapsed (dark bars) and 2) comparing one protein and a random sample of all other proteins (light bars).

On the other hand, the cross-validation analyses were only able to classify 6 of the 15 protein families with greater than 80.0% accuracy. The Supporting Information provides the equations and group centroids for the 15 individual DFAs.

In summary, this model did not discriminate well between an individual protein and all other proteins collapsed. We believe the extremely discrepant sample sizes between the protein of interest and the "other protein" group resulted in biased estimates.[77] For example, the function generated by the DFA for carbonic anhydrases versus all other proteins collapsed would discriminate between 25 carbonic anhydrases and 346 "other proteins". Consequently, we elected to proceed by taking a random sample of complexes from the "other" protein group to try and equalize the sample sizes.

DFA Discriminating between ONE Protein and a Random Sample of Proteins. For the current model, we took a random sample of the 346 "other proteins" to obtain a sample size close to the sample size of the protein class under investigation (in the example above, as close to 25 as possible), and used this random sample as the "other" group of protein families (the "other" could consist of any combination of complexes).

Similar to the previous analyses, the model generated one function and all functions were significant at $p < 0.001$. Again, Figures 4, 5, and 6 depict the canonical correlations, correct classification, and cross-validation rates for the analysis. As can be seen in the figures, the current model does significantly better in discriminating between the individual proteins and a random sample of the other proteins compared to the previous model. There were 10 of the 15 proteins correctly classified

dx.doi.org/10.1021/ci2005912 | J. Chem. Inf. Model. 2012, 52, 824–833

with 100% accuracy, while the lowest cross-validation classification rate for these 10 proteins was 89.5%. Of the other five proteins, only two of the classification rates fell below 90%: kinases at 87.5% and synthases at 86.1%. Furthermore, only two of the 15 cross-validation classification rates fell below 80%: kinases and cytochrome P450s. The Supporting Information provides the equations generated by the current model with the corresponding group centroids.

In summary, we have described an approach that employs DFA in order to classify and predict putative protein function of a diverse set of protein complexes. Function of a novel protein can be predicted by using the equations and centroids in the Supporting Information. Once the descriptors of that novel protein are calculated, they can be entered into the equation and compared to respective centroids listed for that equation (we were able to generate 15 equations in the final model). However, despite its high predictive power, the final model has several limitations. First, as is the case with empirical methods, the ability to predict a novel protein's likely function will depend on whether its associated protein class is described in the data set. For example, we had to eliminate hydrolases and transferases because of their insufficient number of exemplars and the low classification rates, respectively. Therefore, our model would not be able to predict a novel protein falling into these families. Second, as discussed in the preceding section, our classification has some hierarchical inconsistencies, namely certain categories are at the 'family' tree of SCOP, while others are at the 'protein' level. Third, we did not test the possible nonlinear relationships in our data. Although we did look at the overall relationships (i.e., we did not break it down by protein), there were no extreme deviations from linearity. Lastly, as previously mentioned, the results discriminating one protein versus all others should be interpreted with caution due to unequal sample sizes and increased likelihood of unreliable estimates.

CONCLUSION

A lot of work has recently focused on chemogenomics, the discovery of all possible drugs for all possible protein targets. Because targets belonging to the same family are related but not identical, one way for the pharmaceutical sector to be more efficient is to have at its disposal distinct classes of compound libraries that can then be used for screening against multiple targets. Given the advances of structural genomics, if the function of a novel protein can be predicted by being grouped into a protein class of known function, then potentially by inference binders can be immediately used for drug discovery efforts. Our work was undertaken with this objective in mind.

We present here an approach in which binding pockets were described with structural, thermodynamic, and geometric attributes in order to predict putative protein function. By no means is this another method for binding site identification; instead we use the knowledge of where the pocket is located as a starting point. In the final model (B), the correct classification rates ranged from 83.3%−100% with slightly lower cross-validation correct classification rates. These findings indicate that the descriptors used in the current analysis were able to categorize the proteins in the data set according to their putative biological function. However, for a target of an unknown function, what would be more significant is the ability to predict its likely function independently of 'a' given training set. We were able to show that DFA with a random protein sample attains classification rates higher than 90% for almost all of the protein families. This, in turn, is reflective of the robustness of our method and emulates a more realistic scenario given the unexplored biological targets in the druggable genome.

ASSOCIATED CONTENT

Supporting Information

Equation methods, equations generated when conducting a DFA discriminating between ONE Protein and All Other Proteins Collapsed, equations generated when conducting a DFA discriminating between ONE Protein and a Random Sample of Proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: 618-650-5166. Fax: 618-650-5145. E-mail: mkontoy@siue.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank Nate Zack, Department of Mathematics and Statistics, SIUE, for some of the descriptor calculations. We are also thankful to Christina Hayes for reading the manuscript.

REFERENCES

(1) Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805−815.

(2) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727−730.

(3) Russ, A. P.; Lampel, S. The druggable genome: an update. *Drug Discovery Today* **2005**, *10*, 1607−1610.

(4) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5*, 993−996.

(5) Friedberg, I. Automated protein function prediction--the genomic challenge. *Briefings Bioinf.* **2006**, *7*, 225−242.

(6) Powers, R.; Copeland, J. C.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P. Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **2006**, *65*, 124−135.

(7) Finn, R. D.; Mistry, J.; Schuster-Bockler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R.; Eddy, S. R.; Sonnhammer, E. L. L.; Bateman, A. Pfam: clans, web tools and services. *Nucleic Acids Res.* **2006**, *34*, D247−D251.

(8) Sammut, S. J.; Finn, R. D.; Bateman, A. Pfam 10 years on: 10,000 families and still growing. *Briefings Bioinf.* **2008**, *9*, 210−219.

(9) Arakaki, A. K.; Tian, W.; Skolnick, J. High precision multi-genome scale reannotation of enzyme function by EFICAz. *BMC Genomics* **2006**, *7*, 315.

(10) Ofran, Y.; Punta, M.; Schneider, R.; Rost, B. Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discovery Today* **2005**, *10*, 1475−1482.

(11) Tian, W.; Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* **2003**, *333*, 863−882.

(12) Sigrist, C. J.; Cerutti, L.; de Castro, E.; Langendijk-Genevaux, P. S.; Bulliard, V.; Bairoch, A.; Hulo, N. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.* **2010**, *38*, D161−D166.

(13) Sigrist, C. J.; Cerutti, L.; Hulo, N.; Gattiker, A.; Falquet, L.; Paqni, M.; Bairoch, A.; Bucher, P. PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings Bioinf.* **2002**, *3*, 265−274.

(14) Gough, J. The SUPERFAMILY database in structural genomics. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 1897−1900.
(15) Gough, J.; Chothia, C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* **2002**, *30*, 268−272.
(16) Wilson, D.; Pethica, R.; Zhou, Y.; Talbot, C.; Vogel, C.; Madera, M.; Chothia, C.; Gough, J. SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **2009**, *37*, D380−D386.
(17) Adams, M. A.; Suits, M. D.; Zheng, J.; Jia, Z. Piecing together the structure-function puzzle: experiences in structure-based functional annotation of hypothetical proteins. *Proteomics* **2007**, *7*, 2920−2932.
(18) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536−540.
(19) Lo Conte, L.; Brenner, S. E.; Hubbard, T. J.; Chothia, C.; Murzin, A. G. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **2002**, *30*, 264−267.
(20) Brenner, S. E.; Chothia, C.; Hubbard, T. J.; Murzin, A. G. Understanding protein structure: using scop for fold interpretation. *Methods Enzymol.* **1996**, *266*, 635−643.
(21) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH--a hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093−1108.
(22) Bray, J. E.; Todd, A. E.; Pearl, F. M.; Thornton, J. M.; Orengo, C. A. The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Eng.* **2000**, *13*, 153−165.
(23) Holm, L.; Sander, C. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.* **1996**, *24*, 206−209.
(24) Hadley, C.; Jones, D. T. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* **1999**, *7*, 1099−1112.
(25) Krissinel, E.; Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 2256−2268.
(26) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739−747.
(27) Holm, L.; Sander, C. Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **1995**, *20*, 478−480.
(28) Holm, L.; Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **1997**, *25*, 231−234.
(29) Glasner, M. E.; Gerlt, J. A.; Babbitt, P. C. Evolution of enzyme superfamilies. *Curr. Opin. Chem. Biol.* **2006**, *10*, 492−497.
(30) Chiang, R. A.; Sali, A.; Babbitt, P. C. Evolutionarily conserved substrate substructures for automated annotation of enzyme superfamilies. *PLoS Comput. Biol.* **2008**, *4*, e1000142.
(31) Hegyi, H.; Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **1999**, *288*, 147−164.
(32) Kihara, D.; Skolnick, J. Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. *Proteins* **2004**, *55*, 464−473.
(33) Russell, R. B.; Sasieni, P. D.; Sternberg, M. J. Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **1998**, *282*, 903−918.
(34) Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S.; Murcko, M. A. Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* **2001**, *5*, 464−470.
(35) Bredel, M.; Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, *5*, 262−275.
(36) Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.* **2010**, *50*, 2041−2052.
(37) Ho, C. M.; Marshall, G. R. Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 337−354.
(38) Kleywegt, G. J.; Jones, T. A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1994**, *50*, 178−185.
(39) Brady, G. P. Jr.; Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383−401.
(40) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289−307.
(41) Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* **2006**, *63*, 892−906.
(42) Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377−389.
(43) Laurie, A. T.; Jackson, R. M. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **2005**, *21*, 1908−1916.
(44) An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell Proteomics* **2005**, *4*, 752−761.
(45) Glaser, F.; Morris, R. J.; Najmanovich, R. J.; Laskowski, R. A.; Thornton, J. M. A method for localizing ligand binding pockets in protein structures. *Proteins* **2006**, *62*, 479−488.
(46) Huang, B. MetaPocket: a meta approach to improve protein ligand binding site prediction. *OMICS* **2009**, *13*, 325−330.
(47) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics Modell* **1997**, *15*, 359−363.
(48) Brylinski, M.; Skolnick, J. FINDSITE: a threading-based approach to ligand homology modeling. *PLoS Comput. Biol.* **2009**, *5*, e1000405.
(49) Skolnick, J.; Brylinski, M. FINDSITE: a combined evolution/structure-based approach to protein function prediction. *Briefings Bioinf.* **2009**, *10*, 378−391.
(50) Henrich, S.; Salo-Ahen, O. M.; Huang, B.; Rippmann, F. F.; Cruciani, G.; Wade, R. C. Computational approaches to identifying and characterizing protein binding sites for ligand design. *J. Mol. Recognit.* **2010**, *23*, 209−219.
(51) Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M.; Funkhouser, T. A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.* **2009**, *5*, e1000585.
(52) Bray, T.; Chan, P.; Bougouffa, S.; Greaves, R.; Doig, A. J.; Warwicker, J. SitesIdentify: a protein functional site prediction tool. *BMC Bioinf.* **2009**, *10*, 379.
(53) Lee, D.; Redfern, O.; Orengo, C. Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 995−1005.
(54) Perot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A. C.; Villoutreix, B. O. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today* **2010**, *15*, 656−667.
(55) Konc, J.; Janezic, D. ProBiS: a web server for detection of structurally similar protein binding sites. *Nucleic Acids Res.* **2010**, *38*, W436−W440.
(56) Konc, J.; Janezic, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26*, 1160−1168.
(57) Huang, B.; Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6*, 19.
(58) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.
(59) von Itzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Van Phan, T.; Smythe, M. L.; White, H. F.; Oliver, S. W.; Colman, P. M.; Varghese, J. N.; Ryan, D. M.; Woods, J. M.; Bethell, R. C.; Hotham, V. J.; Cameron, J. M.; Penn, C. R. Rational design of

potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993**, *363*, 418−423.
(60) Minai, R.; Matsuo, Y.; Onuki, H.; Hirota, H. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. *Proteins* **2008**, *72*, 367−381.
(61) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387−406.
(62) Kuhn, D.; Weskamp, N.; Schmitt, S.; Hullermeier, E.; Klebe, G. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* **2006**, *359*, 1023−1044.
(63) Kupas, K.; Ultsch, A.; Klebe, G. 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*, 1288−1306.
(64) Schalon, C.; Surgand, J. S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71*, 1755−1778.
(65) Sael, L.; La, D.; Li, B.; Rustamov, R.; Kihara, D. Rapid comparison of properties on protein surface. *Proteins* **2008**, *73*, 1−10.
(66) Weisel, M.; Proschak, E.; Kriegl, J. M.; Schneider, G. Form follows function: shape analysis of protein cavities for receptor-based drug design. *Proteomics* **2009**, *9*, 451−459.
(67) Andersson, C. D.; Chen, B. Y.; Linusson, A. Mapping of ligand-binding cavities in proteins. *Proteins* **2010**, *78*, 1408−1422.
(68) Andreeva, A.; Howorth, D.; Chandonia, J. M.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. G. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **2008**, *36*, D419−D425.
(69) Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **2000**, *28*, 304−305.
(70) Gish, W.; States, D. J. Identification of protein coding regions by database similarity search. *Nat. Genet.* **1993**, *3*, 266−272.
(71) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(72) MacKerell, A. D. Jr.; Banavali, N.; Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2000**, *56*, 257−265.
(73) MacKerell, A. D. Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karlpus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(74) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105−132.
(75) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607−633.
(76) Tabachnick, B. G; Fidell, L. S. *Using Multivariate Statistics*, ed. 5; Pearson Education, Inc.: Boston, MA, 2007; pp 375−436.
(77) Pallant, J. Multivariate Analysis of Variance. *SPSS Survival Manual: A Step-by-Step Guide to Data Analysis Using SPSS*, ed. 3; Open University Press: New York, NY, 2007; pp 275−289.
(78) Burns, R. B.; Burns, R. A. Discriminant Analysis. *Business Research Methods and Statistics Using SPSS*; Sage Publication, Inc.: Thousand Oaks, CA, 2008; pp 589−608.
(79) Stevens, J. Assumptions in MANOVA. *Applied Multivariate Statistics for the Social Sciences*, ed. 5; Taylor and Francis, LLC: New York, NY, 2009; pp 217−244.
(80) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395−407.
(81) Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895−5902.
(82) Gehlhaar, D. K.; Moerder, K. E.; Zichi, D.; Sherman, C. J.; Ogden, R. C.; Freer, S. T. De novo design of enzyme inhibitors by Monte Carlo ligand generation. *J. Med. Chem.* **1995**, *38*, 466−472.