PROTEINS: Structure, Function, and Bioinformatics 63:892–906 (2006)

On the Nature of Cavities on Protein Surfaces: Application to the Identification of Drug-Binding Sites

Murad Nayal and Barry Honig*

Howard Hughes Medical Institute, Center for Computational Biology and Bioinformatics and Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York

ABSTRACT **In this article we introduce a new method for the identification and the accurate characterization of protein surface cavities. The method is encoded in the program SCREEN (Surface Cavity REcognition and EvaluatioN). As a first test of the utility of our approach we used SCREEN to locate and analyze the surface cavities of a nonredundant set of 99 proteins cocrystallized with drugs. We find that this set of proteins has on average about 14 distinct cavities per protein. In all cases, a drug is bound at one (and sometimes more than one) of these cavities. Using cavity size alone as a criterion for predicting drug-binding sites yields a high balanced error rate of 15.7%, with only 71.7% coverage. Here we characterize each surface cavity by computing a comprehensive set of 408 physicochemical, structural, and geometric attributes. By applying modern machine learning techniques (Random Forests) we were able to develop a classifier that can identify drug-binding cavities with a balanced error rate of 7.2% and coverage of 88.9%. Only 18 of the 408 cavity attributes had a statistically significant role in the prediction. Of these 18 important attributes, almost all involved size and shape rather than physicochemical properties of the surface cavity. The implications of these results are discussed. A SCREEN Web server is available at http://interface.bioc.columbia.edu/screen. Proteins 2006;63:892–906.**

Key words: protein–drug-binding sites; protein–drug interactions; protein surface cavities; molecular recognition; molecular surface patches; protein function

INTRODUCTION

A detailed characterization of the molecular surface is a key element in understanding and predicting the binding preferences of different proteins. Proteins bind to a large array of different partners, including flat membrane surfaces, highly curved nucleic acids of different shapes, other proteins with variable surface properties, and small ligands. In each case, the interface must be designed in such a way that there is shape complementarity between the two surfaces and so that physicochemical interactions provide the requisite binding affinity. Individual protein complexes exhibit an intricate pattern of interactions that in some ways defies classification. For example, binding affinity can depend on networks of charged and polar groups whose precise placement is a crucial determinant of affinities.[1] Nevertheless, there has been great interest in determining whether there are general features on the protein surface that are characteristic of interfaces in general, as this would allow the prediction of interfacial regions in the absence of information about the structure of the complex.[2–16] The GRASP[17] program was an early attempt to focus on the molecular surface, and properties such as the presence of cavities, or patches, of significant electrostatic potential, have been used for some time to identify putative interfacial regions.[18] However, such applications have been largely anecdotal, and it is of interest to develop an objective and quantitative strategy for the segmentation and classification of molecular surface regions based on geometrical and physicochemical properties.

A number of studies focusing on the analysis of local regions of the molecular surface have been reported.[2,4–6,8–10,12,16,19–30] In these studies, either surface patches of interest were chosen based on a particular surface property such as electrostatic potential,[10] hydrophobicity,[8,19] hydrogen-bonding groups' density,[6] surface curvature,[21] surface cavities,[2,12,16,22,23,30] secondary structure type,[20] sequence conservation,[24–27] and a combination of properties,[29] or surface regions of a predefined size were simply chosen to cover the surface uniformly.[4,5,9,28] Often surface patches are defined not in terms of the molecular surface per se, but using approximate representations, for example, a rectangular grid associated with solvent-accessible surface atoms,[22,30] clusters of spheres[2] or Delaunay triangles[16] filling surface clefts, Gaussian-level surface,[21] surface residues,[4,5,19,23,28] or surface atoms.[9,12,20] Many surface segmentation procedures proposed so far have certain limitations, such as in the size of the detected

The Supplementary Materials referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat

Grant sponsor: National Science Foundation; Grant number: DBI 9904841.

*Correspondence to: Barry Honig, College of Physicians and Surgery, Columbia University/Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, 1130 St. Nicholas Avenue, ICRB Mail Box 200, New York, NY. E-mail: bh6@columbia.edu

Received 7 June 2005; Revised 10 November 2005; Accepted 15 November 2005

Published online 13 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20897

patches (either predefined rigidly[4,5,9,28] or effectively restricted to lie within a certain range[6,29]), or their shape (e.g., circular[4,5,9,28]). Here we describe a new surface segmentation method that is both accurate and free of the previous limitations on detectable patches. We develop a large number of descriptors to define and characterize cavities, and use them in conjunction with the Random Forests machine learning technique, to study the properties of known drug-binding sites. A specific goal is to identify such sites based on a protein's three-dimensional (3D) structure. More generally, we wish to establish an approach that can be used in various aspects of structure-based function prediction, such as the identification of possible interfaces with other macromolecules, and in the development of new docking techniques.

Many strategies used in the drug discovery process, such as virtual screening or de novo drug design, call for the identification of drug-binding sites based on structure. This is especially true if computationally intensive techniques, such as docking with receptor flexibility, are to be used. It is not uncommon that the target is a well-studied enzyme where drug-design efforts will be focused on the known active site. In other instances, the discovery of new sites and the assessment of their suitability for drug binding will be very helpful. One possible application of particular interest involves the targeting of a drug to a protein–protein interface. Identifying potentially "druggable" sites in such regions would be of great interest. Another application that is likely to be of increasing importance is the identification of the best target in a particular pathway that one wishes to perturb.

Small ligands are known to bind proteins at surface crevices, often the largest.[15] Hence, current ligand-binding site prediction methods are designed to locate the largest protein surface cavity. However, as we discuss below, this strategy identifies only 71.7% of the drug sites in our data set (reported results in the literature range from 74% to 84%[2,16]). The difference may be due to earlier focus on enzymes for which active sites are typically located at a large, often the largest, surface cavity. Hence, our estimated coverage of 71.7% is probably more representative of the general case. Another problem with choosing the largest surface cavity is that this criterion is not useful when one wishes to select the best target for drug design out of a number of possible targets. Rather, the goal in this case would be to identify the most promising drug-binding site on the surfaces of different proteins.

Ligands must satisfy a number of requirements, in addition to strong and specific binding to their biological target, to be useful pharmaceutical agents. Among other considerations, drugs must have the appropriate pharmacokinetics to allow them to be absorbed from the point of administration, must reach their intended target tissue, must not get secreted or degraded too quickly or too slowly, must not aggregate, and must be stable in plasma and other body fluids. Many of these requirements were found to be related to physicochemical properties of the ligand, such as solubility, permeability, and molecular weight.[31,32] As summarized by Lipinski's "rule of 5," drugs should have a molecular weight less than 500 dalton, have less than 5 H-bond donors and 10 H-bond acceptors, and a calculated Log P (octanol–water partition coefficient) greater than 5.[33] It follows that drug-binding sites will themselves have to be endowed with properties that are somehow compatible with the general physicochemical profile to which a pharmaceutical agent is required to adhere.

A question we address here is whether these properties are reflected in some generally identifiable features on the protein surface. Since most small ligands bind proteins in surface indentations or cavities, addressing this question requires an accurate surface cavity–finding method in terms of both cavity identification and delineation, and characterization of the cavity's physicochemical and geometric properties.

A wide range of surface cavity-finding algorithms have been proposed.[11,12,16,30,34–49] Geometry-based cavity detection approaches can be grouped into two major categories: volumetric and surface-based. Volumetric methods aim to identify spaces in the vicinity of the protein sequestered by protruding protein atoms, while surface-based approaches locate surface cavities by an analysis of the geometry of the molecular surface itself. Volumetric approaches can in turn be divided into "on-grid" and "off-grid" methods. In the former category, the protein is embedded in a 3D grid. Grid points in the interstitial spaces of the protein (points not overlapped by protein atoms) are detected, and certain criteria are applied to determine whether or not they are part of a surface pocket. Examples include the method of Voorintholt et al.,[34] Cavity Search,[35] POCKET,[36] LIGSITE,[37] and the closely related method of Stahl et al.,[30] VOIDOO,[38] and finally the method of Delaney,[39] based on cellular logic operations, later generalized using the different formulation of mathematical morphology by Masuya and Doi.[40] The second category of volumetric cavity identification programs do not use a regular 3D lattice. Instead, two types of geometric constructions are used to delineate the empty space comprising a surface cavity. In the first approach, the molecular surface is covered with spheres. Surface cavities are then identified as dense clusters of spheres squeezed in surface invaginations. Examples of programs that use this general strategy are DOCK,[41] SURFNET,[42] and PASS.[43] The second type of off-grid volumetric methods uses Voronoi tessellation,[50,51] such as the program FindSite[44] or the related α shapes theory,[52,53] as in CAST[16,54] and APROPOS.[12]

Surface-based methods rely on geometric analysis of the molecular surface itself to define surface cavities. These methods use solid angles,[45] the surface fractal dimension,[11,13,46] and surface contour maps, where surface cavities appear as valleys between mountain ranges. Essential to the construction of a contour map is the choice of a sea-level surface. Various choices have been explored, including the average sphere,[47] the ellipsoid of inertia,[48] or the convex hull.[49] Using these procedures, significant distortions result when the protein's overall shape deviates from the simple sphere, ellipsoid, or polytope used as a reference sea-level.

In addition to approaches based on geometry, several groups have used energy calculations to locate small ligand-binding sites on the protein surface.[55–57] The general strategy employed is to compute the interaction energy between the protein and a probe or chemical group positioned appropriately close to protein surface atoms or at grid points surrounding the protein. Contiguous clusters of probe positions characterized by favorable interaction energy with the proteins are then determined, and the best one is chosen (based on size or energy criteria) as a prediction of the ligand-binding site. An et al.[56] have added an interesting variation on this theme. Instead of clustering probe positions, they contour the interaction energy grid (after a smoothing step) at a chosen level, thus circumscribing a region in space where favorable interactions of the ligand with the protein could occur. The largest region is then used to predict the ligand-binding site. While no geometrical notions of a surface cavity are explicitly considered in these approaches, in practice, the identified binding sites seem often to occur in surface cavities.

Each of the cavity detection methods proposed so far has inherent limits. The results obtained using grid-based cavity-finding methods are sensitive to grid spacing, as well as the position and orientation of the protein in the grid. Grid-based methods also suffer from the ambiguity in defining a cavity ceiling that separates cavity space from the free space outside cavities. Cavity detection methods that depend on fitting spheres in the cavity space are not well suited for the detection of wide cavities. A wide cavity requires large spheres to span it, which are then likely to spill over ridges and merge with spheres belonging to other cavities. The estimated volume in this case will also be greatly exaggerated as large spheres protrude far beyond what could reasonably be considered the cavity-enclosed space. α-Shape/discrete flow methods, while attractive because they allow analytical computation of cavity volume and area, can only detect surface cavities where all cavity mouths are smaller in circumference than any cross section through the cavity.[16] Methods that use a simplified protein envelope, such as an ellipsoid or convex hull, are susceptible to the presence of protein extensions or arms, which will stretch the protein envelope away from the molecular surface, distorting the computed contour lines and atom depth. And finally, methods that define cavities in terms of surface curvature will not work for surface cavities that deviate somewhat in shape from a paraboloid, for example, cavities that feature a flat bottom.

Despite the large number of proposed methods for the detection of protein surface pockets, few have been applied to more than a small number of anecdotal cases. Of the geometry-based methods, three methods were tested in large studies: Peters et al.[12] used APROPOS to study 309 proteins with known small ligand-binding sites. Laskowski et al.[2] used SURFNET to examine 67 enzyme–small inhibitors, while Liang et al.[16] applied CAST on the enzyme–small inhibitor complexes set studied by Laskowski et al.,[2] as well as to a few additional examples from the Peters et al.[12] set. Two of the energy-based methods were tested on large data sets: Q-SiteFinder[57] was used to predict ligand-binding sites in 134 ligand–protein complexes from the GOLD docking data set,[58] while PocketFinder[56] was tested using the largest data set so far composed of 5616 bound and 11,510 unbound structures. The APROPOS algorithm does not provide a measurement of the surface cavity's volume per se. Hence, the Peters et al.[16] study mostly confirmed that small ligands bind at *a* surface cavity 95% of the time. Both Laskowski et al.[2] and Liang et al.[16] reported that, in the majority of cases, small ligands bind in the largest cavity on the protein surface (in 83.6% and 74% of the cases, respectively). Laurie and Jackson[57] achieved a 71% prediction coverage of ligand-binding sites by selecting the probe cluster with the most favorable interaction energy, while the prediction coverage reported by An et al.,[56] computed on their bound structures data set, was 78.3%.* No attempt has been made so far to fully characterize drug-binding cavities in terms of physicochemical properties, with the exception of the An et al. study, where the hydrophobicity and electrostatic charge of the predicted small-molecule envelope were computed.

Here we propose a new method for surface cavity finding, SCREEN: **S**urface **C**avity **RE**cognition and **E**valuatio**N**, that circumvents many of the limitations inherent in other approaches. SCREEN defines surface cavities geometrically in terms of the empty space between the protein's molecular surface and an envelope surface constructed by rolling an intermediate size spherical probe (of dimensions akin to that of a typical small ligand). Each cavity is represented by the patch of the molecular surface forming the cavity floor and the corresponding region of envelope surface forming a cavity ceiling. As such, cavity volume and surface area are well defined and can be computed precisely. The molecular surface construction process establishes an unambiguous correspondence between the molecular surface vertices and protein atoms. This allows SCREEN to characterize each surface cavity using measures of the surface shape and geometry, as well as by mapping various properties of the underlying atoms and residues on the cavity's floor. This data-rich description constitutes a cavity *property profile* or *feature vector* that is then used as a data substrate for this analysis.

While no extensive study of the physicochemical properties of surface cavities have been reported so far, the computational biology literature is rich with analyses of the properties of the protein molecular surfaces, especially at various types of interaction sites. Many physical, chemical, and structural properties have been considered, including amino acid composition,[3,5,9,10,14] hydrophobicity,[3,5,59,60] hydrogen-bonding density,[10,29] charge distribution, the electrostatic potential and field,[10,61,62] polarizability,[63,64] the shape of the titration curve,[65,66] curvature,[17,45] local shape descriptors,[60,67] planarity and protrusion,[3,5] surface

*Prediction coverage computed from the reported results by multiplying the probability of the ligand-binding site contacting an identified cavity (96.8%) by the probability that the cavity overlapping the ligand is the largest one (80.9%).

depth,[6] relative accessibility,[5] fractal dimension,[11,13] secondary structure,[3,9,10] side-chain energy,[68–70] sequence distance,[9] and temperature factors.[9] Typically only a few properties are examined in any one study. This makes it difficult to establish the role that individual properties play in binding, especially that many of these properties are correlated, often in ways that are hard to foresee. Moreover, no standard parameterization is universally agreed upon for many of the characteristics considered relevant to molecular recognition, such as hydrophobicity or binding-site shape. Definitive statements about the relative importance of the various properties can only be made when they are compared simultaneously on the same data set. Consequently, we decided to be as comprehensive as possible in our selection of properties to use in the analysis, sometimes using multiple parameterization for the same concept, so as to minimize the impact of the "feature selection bias" on the results. Instead, statistical methods are used to assess the importance of different properties and the appropriateness of various corresponding parameterization as predictors of drug binding.

In total 408 attributes were computed for each cavity. The properties considered include various measures of cavity size (6), electrostatics (94), hydrogen bonding (34), hydrophobicity and polarity (42), amino acid composition (21), rigidity (26), secondary structure (5), and cavity shape (180). The machine learning technique, Random Forests, was chosen to train a classifier to distinguish drug-binding from non-drug-binding cavities using the computed cavity property profile. Random Forests builds a predictor from an ensemble of decision trees using an modified version of the bootstrap aggregation method (bagging)[71] and has a prediction accuracy that compares favorably to other modern, supervised learning algorithms such as neural networks, support vector machines, and boosting.[72] We chose to use Random Forests for this study because it is robust to the presence of a large number of irrelevant variables, it does not require prior scaling of input variables, and it can detect and take advantage of high-order interactions.[72] All are important requirements for our purposes, since we expect that a number of the calculated properties may not be relevant to drug binding, while a number of others are useful only in combinations. In addition, the properties considered here are measured on different scales, and no good a priori reason exists for making a particular rescaling choice.

In this study we used SCREEN to analyze 99 nonredundant, comprehensive protein–ligand complexes from the Protein Data Bank (PDB) selected by Perola et al.[73] The Random Forests algorithm was used to train a classifier that predicted drug-binding cavities with a balanced error rate of 7.2%, coverage of 88.9%, and Matthews correlation coefficient of .77. To investigate the role that various cavity properties play in the prediction of druggability, we used variable importance measures implemented in the Random Forests package.[72] These measures compute the reduction in the classification accuracy when the variable in question is permuted. Only 18 cavity attributes out of 408 had a statistically significant variable importance (at the $\alpha = .1$ level). Most of these attributes were related to the size and shape of the surface cavity, with cavity rigidity and types of amino acids involved playing a role as well.

RESULTS

Protein Surface Cavities

SCREEN identified 1347 cavities on the surface of the 99 proteins in our data set, 13.6 ± 5 per protein, on the average. In all cases (100%), a drug binds at least one, but sometimes as many as four, surface cavities (160 surface cavities in total are filled either partially or completely by a bound drug, or 1.6 ± 0.8 cavities per drug). However, many of these cavities have only minor contact with the drug. We refer to the surface cavity where the percentage of its surface in contact with the bound drug is larger than other cavities in the protein as the *primary drug-binding cavity*; the rest of the surface cavities are called *secondary drug-binding cavities*. On average, 74 ± 24% of the surface of the primary drug-binding cavity is in contact with the drug. In 45% of the complexes the drug also binds at one or more secondary drug-binding surface cavities. On average, only 25 ± 28% of the surface of secondary cavities is in contact with the drug. As this renders the importance of the small number of drug interactions at the secondary surface cavities somewhat unclear, we decided not to include the secondary surface cavities, 61 in all, in the data set used to train the classifier. The reduced data set contains 99 drug-binding cavities and 1187 non-drug-binding cavities. The analysis has been repeated without excluding secondary drug-binding cavities, and the conclusions were essentially the same as these reported here.

Surface Cavity Volume Is Uncorrelated With the Volume of the Bound Drug

Figure 1 shows a plot of the relationship between drug volumes and the volume of the primary drug-binding cavity. It is clear from the plot that there is no correlation between the two volumes (correlation coefficient of 0, even if the two outliers were removed). This is probably due to the fact that in many cases, only a portion of the drug is immersed into the surface cavity. Conversely, our data set contains many examples of surface cavities that are only partially filled with the bound drug. Other studies have reported a weak relationship between ligand volume and cavity volume observed when the drug-binding cavity size is small.[16]

Drugs Bind at Large Surface Cavities

Drugs have been observed in previous studies to bind at large surface cavities.[2,12,16] This is also borne out in our data set. In Figure 1, the volume distribution of all surface cavities is plotted as a blue line, while the volumes of drug-binding cavities are represented by red points. The average volume of a surface cavity is 229 Å^3, while the average volume of a drug-binding cavity is 930 Å^3. Figure 2 shows the rank distribution of drug-binding and non-drug-binding cavities, where cavities are ranked by the area of their floor surface. In 71.7% of the complexes, the drug

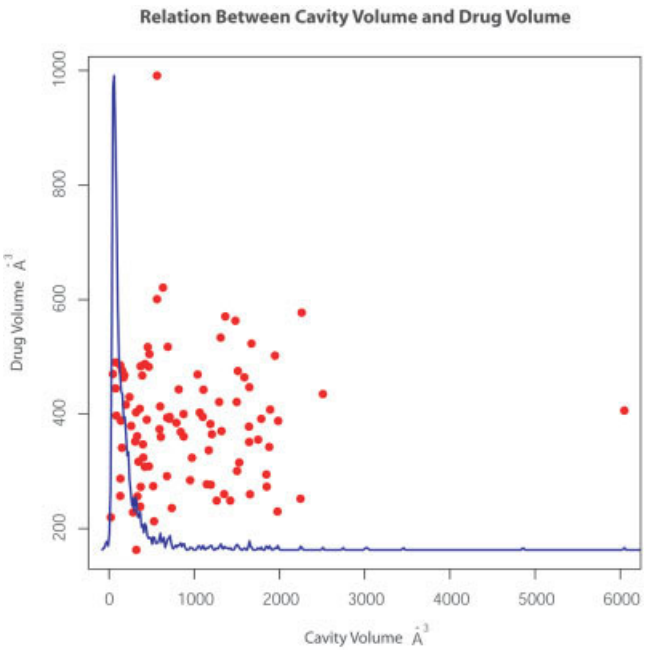

Fig. 1. The relation between cavity volume and the volume of the bound drug. In blue is the distribution of the volumes of all surface cavities. We note that volumes of most drug-binding cavities are larger than volumes of typical surface cavities.

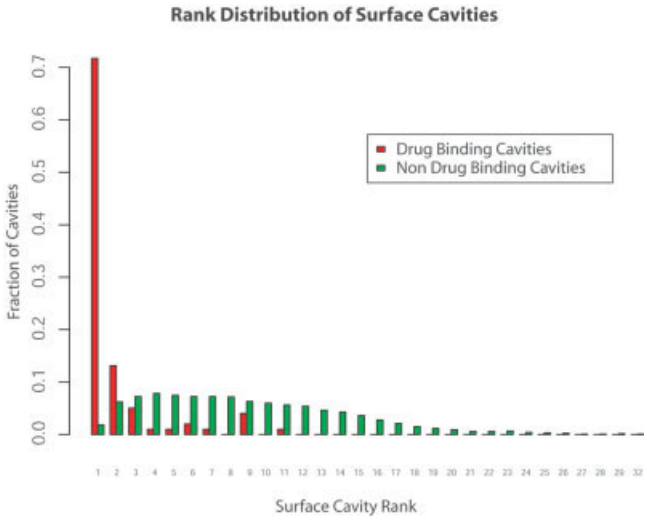

Fig. 2. The distribution of the ranks (by surface area) of drug-binding and non-drug-binding cavities.

binds at the largest protein surface cavity (i.e., rank 1). If volume is used instead as a measure of cavity size, we find that 64% of ligands bind at the surface cavity largest in volume.

Using Cavity Size to Predict Drug-Binding Cavities

Based on the tendency for drugs to bind at large surface cavities, many researchers have proposed to predict drug binding sites by locating the largest surface cavity.[2,12,16] This simple criterion yields a balanced error rate (BER) of 15.7%, coverage of 71.7%, and a correlation coefficient of

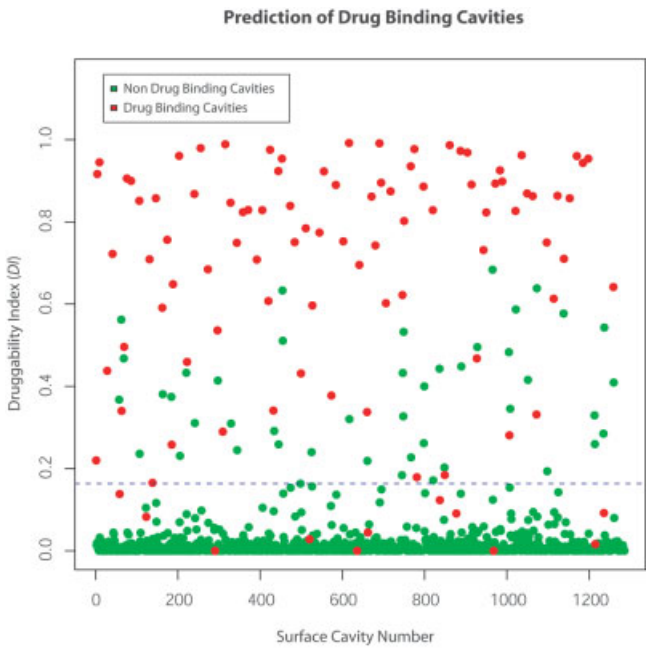

Fig. 3. The Druggability Index (*DI*, the fraction of the decision trees ensemble voting for a drug-binding classification) for each surface cavity in the data set. For each cavity, only trees trained with this cavity in the test set (OOB) are counted. Drug-binding cavities are plotted in red; non-drug-binding cavities are plotted in green. The dotted line represents the maximum mutual information *DI* cutoff (i.e., the classification threshold where the predicted classification conveys the most information about the true classification).

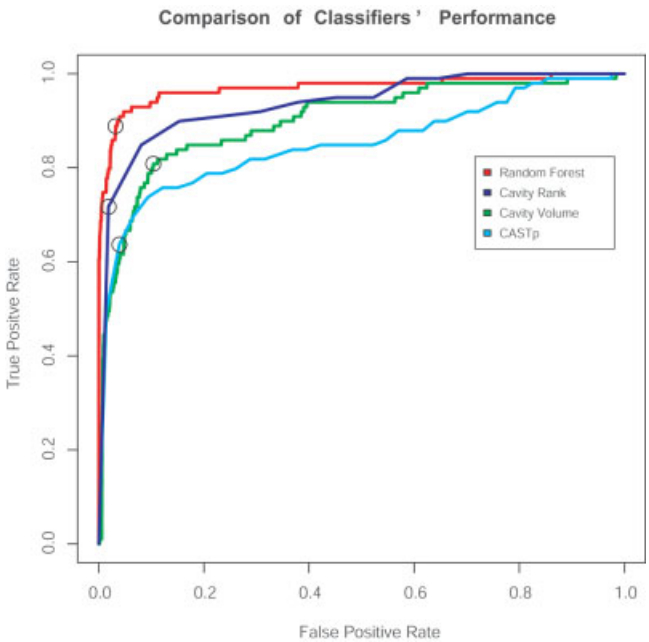

Fig. 4. ROC curves for the Random Forests classifier, as well as for classifiers using cavity rank or volume alone, as computed by SCREEN. Also compared is the performance of a classifier based on CASTp cavities and their ranks. The circles highlight the points corresponding to the maximum mutual information thresholds for each classification strategy. The AUC for the Random Forests, as well as classifiers using SCREEN cavity rank, SCREEN cavity volume, or CASTp cavities, are 0.97, 0.935, 0.90 and 0.85, respectively.

.72. Other indicators of surface cavity size, such as cavity volume, could potentially be used instead of cavity rank. To use cavity volume, one has to choose a threshold above which cavities are classified as drug-binding. One reasonable strategy for determining such a threshold is to choose a volume cutoff that maximizes the mutual information[74] between the predicted and the observed classification. In our data set, the cavity volume classification threshold, chosen using maximum mutual information criteria, is 328 Å^3. The resulting classification rule yields a BER of 14.7%, coverage of 80.8%, and a Matthews correlation coefficient of .52. These quick estimates of classification accuracy are rather optimistic, since they are not estimated using a proper training/testing design. Incidentally, the prediction performance of cavity size–based classifiers would deteriorate significantly if secondary drug-binding cavities were included in the analysis. In this case, prediction using cavity rank yields a BER of 26.9%, coverage of 48.1%, and correlation coefficient of .57, while cavity volume yields a BER of 26.2%, coverage of 56.9%, and a correlation coefficient of 0.43.

Using a Comprehensive Set of Cavity Properties and Random Forests to Predict Drug Binding

We used the Random Forests algorithm to train a classifier capable of distinguishing drug-binding from non-drug-binding cavities. Surface cavities were each represented by feature vectors containing 408 attributes. An ensemble of 10,000 unpruned decision trees was grown using the Random Forests algorithm and bootstrap samples of the data set (sampling with replacement). For each cavity, about one third of the trees were grown while this cavity was in the *out-of-bag* set, and this cavity was thus used to test the prediction of these trees. The percentage of trees that vote to classify a cavity as a drug-binding cavity represents a measure of the classifier's assessment of the *druggability* of this cavity and is referred to here, for brevity, as the Druggability Index (*DI*), that ranges from 0 (non-drug-binding cavity) to 1 (drug-binding cavity). Figure 3 shows a plot of the computed *DI* for each of the 1286 cavities in our data set. A classification *DI* threshold, DI_{cutoff}, has to be chosen if a crisp prediction of drug binding is needed. Usually this choice involves a trade-off between the sensitivity and the specificity of the classifier. Setting the DI_{cutoff} to the value that maximizes mutual information between the predicted and observed classification (in our case, 0.156), we obtain a BER for the Random Forests classifier of 7.2%, coverage of 88.9%, and a correlation coefficient of 0.77, a significant improvement over prediction performance measures reported so far, as well as predictions using cavity size alone on our data set.

The Imbalanced Class Size Problem

Training a classifier when the available data are significantly unbalanced presents particular difficulties. When the classifier is presented with many more examples of one class compared to the other, then the classifier is likely to optimize the overall error rate by biasing the predictions toward the abundant class, hence missing too many cases of the under-represented class. In our study, most of the surface cavities in our data set are non-drug-binding cavities (92.3%). As is often the case, the under-represented class, here the drug-binding cavities, is the class of real interest. Two strategies were attempted in this study to balance the class-specific error rates. One is to modify the behavior of the Random Forests algorithm by assigning a higher weight to the under-represented class. The second strategy is to balance the data set using resampling techniques, either by down-sampling the abundant class or up-sampling the under-represented class. Unfortunately, these strategies were either ineffective in balancing the class-specific error rates or resulted in a notable degradation in prediction accuracy. As a result no special measures were taken to address the data imbalance in this study. In the end, the class-specific error rates for our classifier were unequal. The error rate for the prediction of drug-binding cavities is 11.1%, while the error rate predicting non-drug-binding cavities is 3.2%.

Comparing Random Forests With Druggability Prediction Using Cavity Size Criteria Alone

Figure 4 shows Receiver Operating Characteristic (ROC) curves for the Random Forests classifier, as well as for classifiers using cavity rank (when cavities are ranked in each protein by surface area) or volume alone. ROC curves allow a complete depiction of the performance of a classifier at all sensitivities—specificity trade-off choices. The ROC curve of a good classifier will tend to start from the origin and climb up the true-positive axis (the *y* axis) much more rapidly than it advances in the direction of the false-positive rate axis (the *x* axis). This quality can be summarized succinctly by computing the area under the ROC curve (AUC). The AUC for the Random Forests, cavity rank, and cavity volume classifiers are 0.97, 0.94, and 0.90, respectively, again an indication of a significant improvement in classification performance when multiple cavity properties combined with Random Forests are used to predict drug-binding cavities. The maximum mutual information points for each classifier are highlighted on the ROC curves in Figure 4 with a circle.

Comparing SCREEN to CASTp

CASTp is a leading and widely used protein cavity detection program.[16] For the most part, cavities detected by SCREEN and CASTp coincided well in our data set, albeit the extent and size of the corresponding cavities can differ. Many times a cavity identified by SCREEN is broken up into two or more cavities by CASTp. As a result CASTp finds 36 cavities per protein on average compared to around 14 cavities for SCREEN. In all complexes the drug was bound at one or more CASTp-identified cavities (as was the case for SCREEN). CASTp output does not predict cavity druggability per se. However, if cavities are ranked by volume, we find that the drug is associated with the largest cavity only 72% of the time. Here the association of the drug with a cavity is defined as the presence of a drug atom less than 4 Å from a cavity atom. To compare the performance of a CASTp-based druggability predictor

to SCREEN, we repeat the statistical analysis used previously on the CASTp results: First we compute the cavity rank threshold that yields the maximum mutual information between CASTp cavity rank and the *primary* drug-binding cavity. In our data set, this threshold is $\text{rank}_{\text{cutoff}} = 2$ (i.e., the maximum mutual information prediction strategy classifies both the largest and the second largest CASTp-identified cavities as druggable cavities). Next we compute the coverage (63.6%) and the BER (20.1%) of this classification strategy when used to predict the primary drug-binding cavity in the protein. We note that both the coverage and the error rate are significantly worse than what is obtained using SCREEN. This point is made clearer in Figure 4, which plots the ROC curves of all four drug-binding predictors considered in this work: SCREEN/Random Forests, SCREEN-computed cavity rank and volume, and CASTp.

What Makes a Drug-Binding Cavity a Drug-Binding Cavity?

The success of the Random Forests classifier in predicting drug-binding cavities to a high accuracy suggests that it has learned something about cavity characteristics important for drug binding. This knowledge, however, is encoded rather opaquely in the thousands of decision trees that make up the classifier. To try to understand the inner workings of the classifier we resorted to variable importance measures computed on the test set. Importance measures for variable x_j are based on the increase in the classification error rate per tree when x_j is permuted, averaged over all trees. If x_j happens not to be useful for the prediction of drug binding, then the expected change in error rate would be zero. Figure 5 shows the computed variable importance measure for each of the 408 properties considered in this study. Only 18 variables have an importance measure that is statistically significant at $\alpha = 0.1$ level (highlighted by drawing circles around the data points) assuming a normal distribution (with zero mean) of the measure under the null hypothesis of lack of correlation between the variable in question and drug binding.[72]

Table I, which lists all 18 significant predictors of drug binding in decreasing order of importance, conveys an overall picture of drug-binding cavities: Drug binding cavities are large, deep, have an intricate curvature profile, are rigid, and have a relatively small number of prolines, as well as amino acids with small but negative octanol-to-water transfer free energies (Asn, Gln, Glu).

The classifier actually uses these properties in a manner that is more complex than what is suggested by this simple picture. To gain an appreciation of this complexity we present an analysis of the prediction role of four important variables in Figure 6. In the left panel we plot the probability density of particular surface property P in drug-binding and non-drug-binding cavities, while in the right panel we plot the expectation of the DI conditioned on that property $E(DI \mid P)$ (by integrating over all others). The curve is colored red when the $E(DI \mid P) \geq 0.59$, a DI threshold that yields a high precision (99%) prediction of drug-binding cavities,

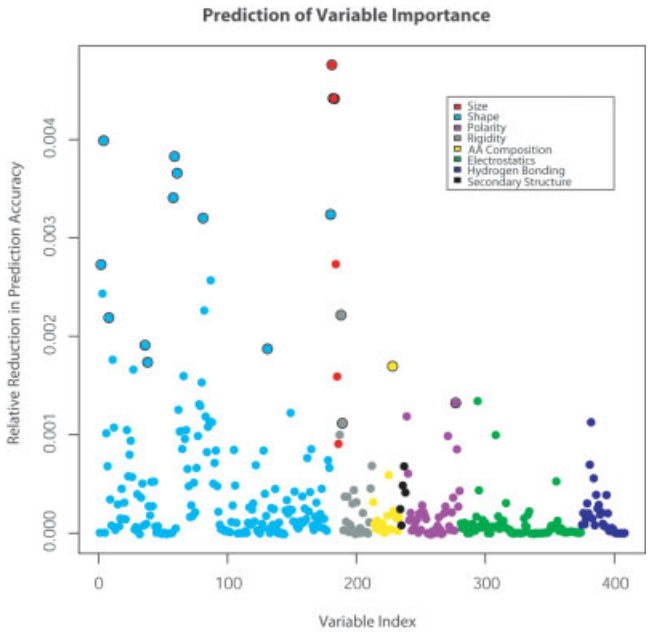

Fig. 5. The variable importance (relative reduction in classification accuracy upon permutation) for all 408 cavity attributes. The attributes are color coded by category: cavity size (red), shape (cyan), polarity (purple), rigidity (gray), amino acid frequencies (yellow), electrostatics (green), hydrogen bonding (blue), and secondary structure (black). Only 18 variables, mostly related to the size and shape of the surface cavity, have an importance measure that is statistically significant at $\alpha = 0.1$ level and are highlighted by circles around the data points.

green when the $E(DI \mid P) \leq 0.155$, a second DI threshold that yields, this time, a high precision (99%) prediction of non-drug-binding cavities. The plot is colored gray otherwise, indicating indecisive prediction.

We see from Figure 6 that cavity rank is decisive at cavity rank 1 (predicts the cavity as drug binding) and when the cavity has a rank greater than 3 (cavity is then predicted as non-drug-binding) but somewhat agnostic for cavities ranked 2. The role of the number of cavity residues is somewhat subtler. Judging by the conditional expected value of DI, we observe that cavities composed of small number of residues are not likely to be predicted as drug-binding cavities, but this is also the case for cavities composed of a very large number of residues (very large cavities). We can see from the plot that there is an ideal range of cavity sizes (as measured by the number of cavity residues) where the classifier is most confident about classifying the cavity as drug binding. A similar situation is illustrated for the average cavity depth. Neither shallow nor very deep cavities are ideal candidates for drug binding; rather, drug-binding cavities have an average depth that is roughly between 6.8 Å and 11.4 Å. A different scenario emerges when we examine the role of the proportion of the cavity surface related to proline (P_{proline}) in the prediction of drug binding. We note that the distribution of the proportion of proline in drug-binding and non-drug-binding cavities seem similar, and $E(DI \mid P_{\text{proline}})$ seems to mostly be independent of this attribute [$E(DI \mid P_{\text{proline}})$] is a horizontal line as a function of P_{proline}. This probably indicates that the proportion of proline, by itself, is not a

TABLE I. Important Predictive Drug-Binding Cavity Attributes

Surface cavity property	Category	Drug-binding cavities	Non drug-binding cavities
Cavity rank[a]	Size	1.89 ± 2.07	8.88 ± 5.4
Number of residues[b]	Size	22.8 ± 14.3	7.31 ± 5.4
Number of atoms[c]	Size	85.0 ± 62.4	18.7 ± 21.2
Smallest moment of inertia[d]	Size/shape	$1.7 \times 10^4 \pm 2.5 \times 10^4$	$1.2 \times 10^3 \pm 8.3 \times 10^3$
Depth standard deviation[e]	Size/shape	2.3 ± 1.1 (Å^3)	0.75 ± 0.45
Maximum depth[f]	Size/shape	10.5 ± 4.0 (Å)	4.75 ± 1.67
Average depth[g]	Size/shape	5.3 ± 1.9 (Å)	3.2 ± 0.7
Normalized smallest moment of inertia[h]	Shape	17.0 ± 11.7	3.9 ± 5.3
Proportion of cavity at depth between [6.5, 6.75)[i]	Shape	0.02 ± 0.013	0.003 ± 0.001
Largest moment of inertia[j]	Size/shape	$1.6 \times 10^4 \pm 8.4 \times 10^4$	$2.8 \times 10^3 \pm 1.6 \times 10^4$
Average side-chains residual entropy[k]	Rigidity	-0.41 ± 0.18 (kcal)	-0.55 ± 0.25
Average curvature[l]	Shape	-49.0 ± 8.3	-57.0 ± 13.1
Maximum curvedness[m]	Shape	6.4 ± 2.9	4.0 ± 4.9
Maximum mean curvature[n]	Shape	5.3 ± 2.6	3.5 ± 4.2
Curvedness < 0.5°[o]	Shape	0.35 ± 0.04	0.29 ± 0.08
Proportion of proline[p]	Amino acid composition	0.019 ± 0.028	0.04 ± 0.09
Proportion of cavity with logP between $[-1, 0)$[q]	Hydrophobicity	0.09 ± 0.07	0.15 ± 0.16
Side-chain residual entropy standard deviation[r]	Rigidity	0.43 ± 0.18 (kcal)	0.55 ± 0.17

[a]Cavity rank by the floor surface area.
[b]Number of residues forming the cavity.
[c]Number of atoms forming the cavity.
[d]The smallest moment of inertia of the cavity vertices.
[e]The standard deviation of the depth of cavity vertices.
[f]The maximum depth of cavity vertices.
[g]The average depth of cavity vertices.
[h]The smallest moment of inertia of cavity vertices divided by the number of vertices.
[i]The proportion of cavity vertices that lie at a depth between 6.5Å and 6.75 Å.
[j]The largest moment of inertia of cavity vertices.
[k]The average residual entropy of cavity side chains (see Methods section).
[l]The average curvature, computed over cavity vertices.
[m]The maximum curvedness, computed over cavity vertices.
[n]The maximum of the mean curvature, computed over cavity vertices.
[o]The proportion of cavity vertices with curvedness less than 0.5.
[p]The proportion of cavity vertices corresponding to a proline residue.
[q]The proportion of cavity vertices with mapped logP between -1 and 0.
[r]The standard deviation of the residual entropy of cavity side-chains.

good predictor of drug binding. The fact that this variable has a high importance measure suggests that it plays its predictive role in close association with other variables, or possibly is useful in predicting certain subtypes of drug-binding cavities.

Complete analysis of all 18 cavities' attributes is presented in Figures 10, 11, and 12 in the Supplementary Material. In summary, nine attributes yield a high-confidence prediction of drug binding: a cavity rank of 1, a number of contributing residues between 32 and 54, a number of contributing atoms between 120 and 261, an average depth between 6.8 Å and 11.4 Å, a maximum depth between 13 Å and 22.9 Å, a standard deviation of surface vertices depth distribution between 3.0 Å and 5.6 Å, a largest moment of inertia between 10^5 and 2.2×10^5, a minimum moment of inertia between 3.8×10^4 and 10^5, or a normalized minimum moment of inertia between 26.6 and 48.7. The remaining nine important attributes seem to operate in combination with other predictors. Note that all nine high-confidence attributes describe cavity shape, whereas none describe physicochemical properties.

Random Forests Predicts Drug-Binding Cavities Missed by Cavity Size Criteria

We present here three examples of instances where the Random Forests–based classifier was able to predict drug-binding cavities that would have been missed using the simple "largest surface cavity" criterion:

1. Protein–tyrosine phosphatase 1B, PTP1B[75] (PDB code: 1l8g). SCREEN identifies 18 cavities on the surface of PTP1B. Only one cavity is predicted correctly as a drug-binding cavity. This corresponds to the second largest cavity (area, 170 Å^2; volume, 259 Å^3). The largest surface cavity (area, 184 Å^2; volume, 400 Å^3) is about 20 Å from the ligand-binding site (Figure 7).
2. Human carbonic anhydrase II (CA II).[76] SCREEN identifies 12 surface cavities on CA II. Here again, the second largest cavity is correctly predicted as a drug-binding cavity (area, 194 Å^2; volume, 281 Å^3). The largest cavity (area, 281 Å^2; volume, 679 Å^3), shown in green in Figure 8, is predicted not to be suitable for drug binding with a *DI* of 0.08. Here, criteria related to cavity

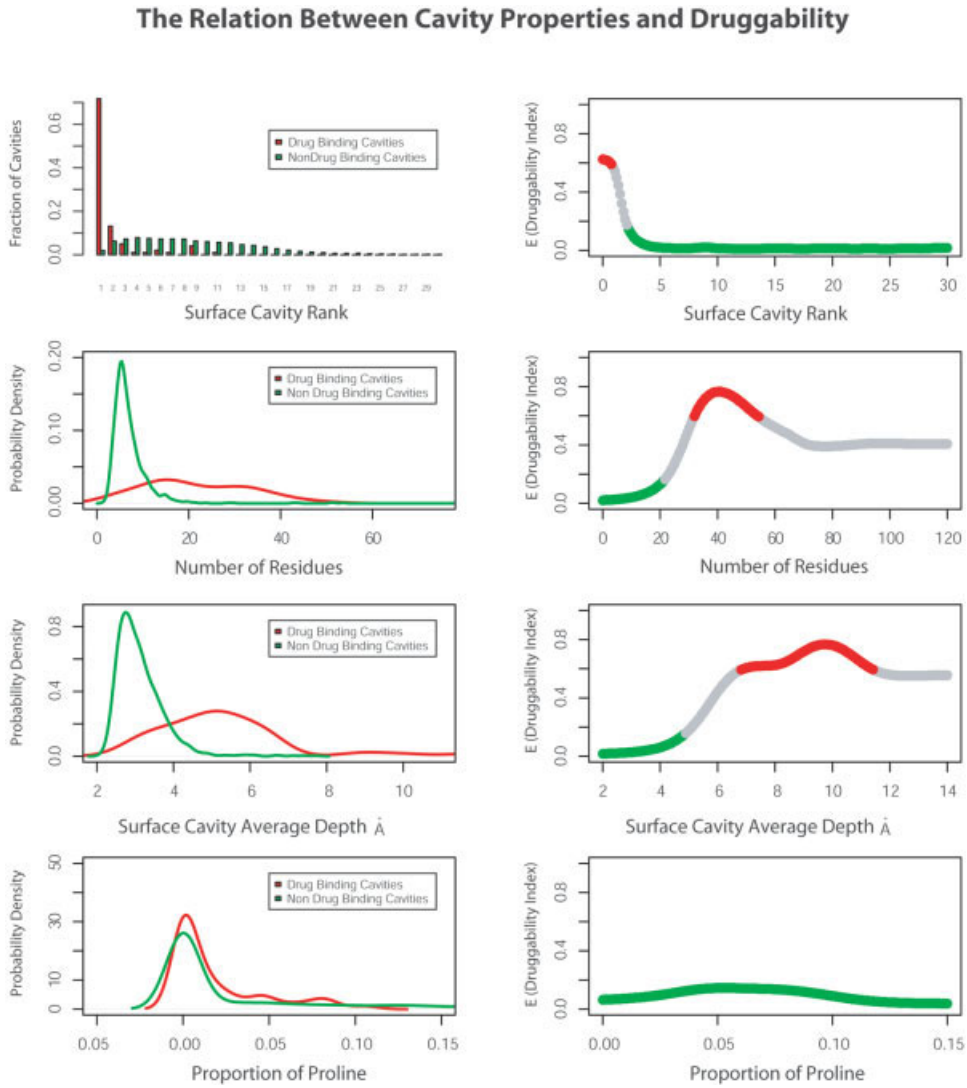

Fig. 6. Analysis of the role of cavity rank, number of residues, average depth, and the fraction of the cavity surface contributed by proline. The left-hand column contains the distributions of each surface property *P* in drug-binding and non-drug-binding cavities. In the right-hand column we plotted the expectation of the *DI* conditioned on that variable $E(DI \mid P)$. The curve is colored red when the expected value of *DI* implies a high-precision prediction of a drug-binding cavity, green when the expected value of *DI* implies a high-precision prediction of a non-drug-binding cavity.

shape were probably decisive in making this classification. The largest surface cavity is actually rather shallow with an average and maximum depth of 3.3 Å and 5.78 Å, respectively, versus 5.19 Å and 9.8 Å, respectively, for the drug-binding second largest cavity.

3. Human factor Xa complexed with inhibitor RPR128515[77] (PDB code: 1ezq). SCREEN identifies 14 surface cavities, four of which (cavities ranked 1, 2, 3, and 9) are predicted to be potential drug-binding cavities, shown in red in Figure 9. The ligand actually binds at two cavities, rank 3 (the S1 substrate-binding pocket[77]: area, 274 $Å^2$; volume, 384 $Å^3$) and rank 9 (the S4 substrate binding pocket[77]: area, 69 $Å^2$; volume, 155 $Å^3$). In this case, S1 pocket was considered a secondary drug-binding cavity (and hence was excluded from the Random Forests training set). Nonetheless, the trained classifier was able to predict it correctly as a drug-binding cavity.

DISCUSSION

The fact that most small ligands' binding sites are found in surface pockets or clefts has been noted since the early days of X-ray studies of protein–ligand complexes.[15] Binding at a cavity maximizes the number of possible interactions a given ligand can achieve with the protein. In a sense, the propensity of ligands to bind at surface cavities hints at the multiplicity of noncovalent interactions necessary to achieve a stable multimolecular complex in an aqueous milieu. Protein surface grooves and pockets vary greatly in shape and size, from minor indentations between surface atoms to large cavities between secondary structure elements and protein domains. Existing geometri-

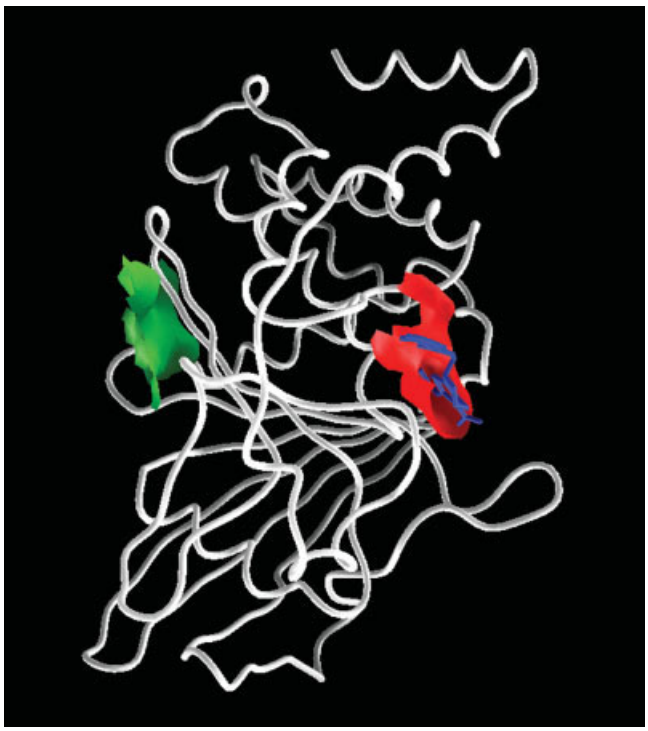

Fig. 7. Protein–tyrosine phosphatase 1B, PTP1B[75] (PDB code: 1l8g). The largest surface cavity (colored green: area, 184 Å^2; volume, 400 Å^3; residues Gln78, Arg79, Ser80, and Pro210) is about 20 Å from the ligand-binding site. The drug binds at the second largest cavity, colored red, as predicted (area, 170 Å^2; volume, 259 Å^3; residues Gln262, Ala217, Ile219, Val49, and Asp181).

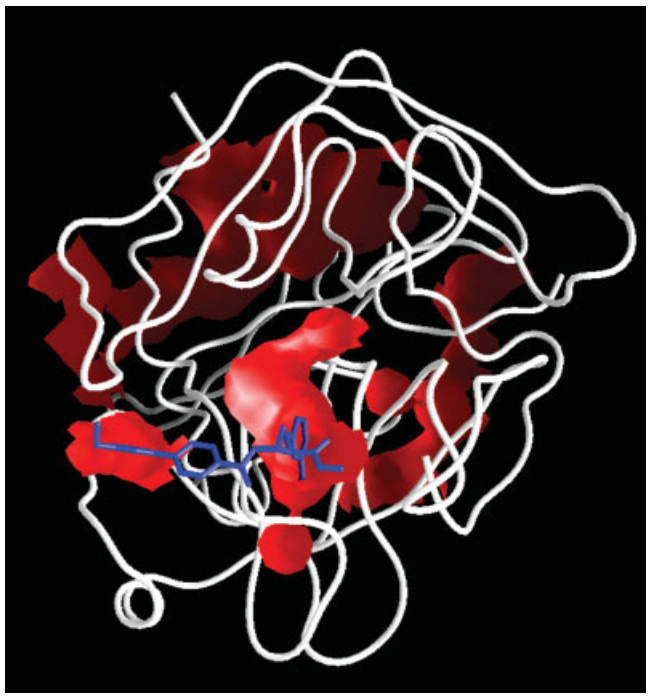

Fig. 9. Human factor Xa complexed with inhibitor RPR128515[77] (PDB code: 1ezq). Four cavities ranked 1, 2, 3, and 9, shown here in red, were predicted to be potential drug-binding cavities. The ligand actually binds at two cavities, 3 (the S1 pocket: area, 274 Å^2; volume, 384 Å^3; residues Gln192, Trp215, Ser195, Cys191, Gly216, and Asp) and 9 (the S4 pocket: area, 69 Å^2; volume, 155 Å^3; residues Trp215, Phe174, Thr98, Tyr99, and Ile175). In this case, S1 pocket was considered a secondary drug-binding cavity (and was excluded from the Random Forests training set).

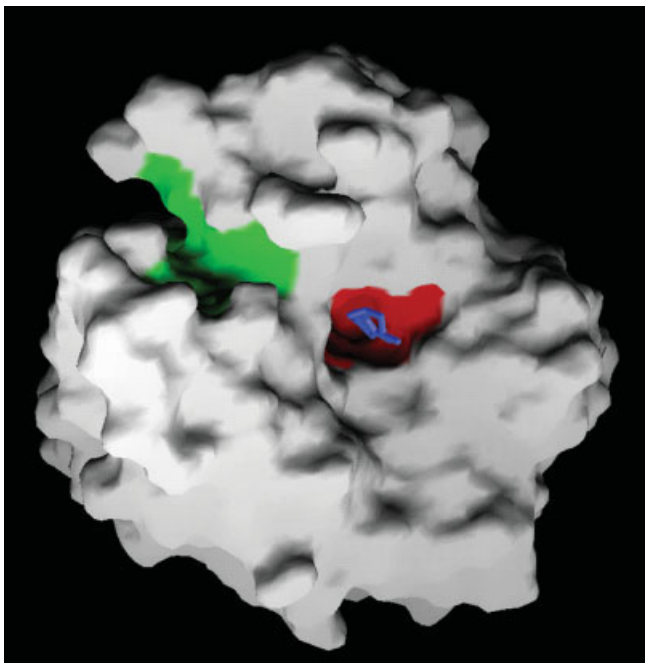

Fig. 8. Human carbonic anhydrase II (CA II).[76] The largest cavity (area, 281 Å^2; volume, 679 Å^3; residues Phe213, Tyr7, Gly8, Asp243, and Lys170), shown in green, is rather shallow and is predicted not to bind a drug. Instead, the second largest cavity (area, 194 Å^2; volume, 281 Å^3; residues Leu198, Thr200, His94, Val121, and His64) is the one predicted correctly to bind the drug.

cal strategies for surface cavity–finding often impose limits on the shape and size of surface cavities that can be detected or use a cavity representation that does not lend itself to accurate, chemically meaningful characterization. SCREEN addresses many of the shortcomings encountered in earlier methods. Using a large probe sphere of a size approximate to that of a medium-size ligand allows the definition of a "sea-level" that adapts to the local shape of each surface region and is hence less likely to be distorted by distant protein arms or extensions. The usage of single-link clustering enhanced with refinement steps ensures that SCREEN does not impose prior constraints on the shape or size of the cavities detected. Last, the representation of surface cavities in terms of molecular surface regions allows a rich, chemically meaningful characterization of the physicochemical, evolutionary, structural, and geometric properties of the cavity by evaluating these attributes on the cavity surface. Some quantities of interest, such as curvature, are only meaningful using a surface representation. For others, the molecular surface is a natural locus for the evaluation of properties likely to be relevant to molecular recognition. For example, the electrostatic potential or field will be infinite at the center of atoms that carry a nonzero charge, while it remains finite and has a useful interpretation on the molecular surface. In addition, the contribution of the underlying residues to cavity properties can be rationally apportioned

by using their share of the cavity surface as a weighting factor.

The energy-based approaches to surface cavity localization offer a promising alternative to geometry-based methods, as these methods directly attempt to assess the binding affinity at the evaluated sites.[55] The successful methods reported so far that use this strategy have relied entirely[56,57] or mostly[55] on the van der Waals interaction energy, a nonspecific, short-range energy term. As such, this energy term is closely related to the geometry of the surface cavity. In practice, we saw here that a classifier that used careful parameterization of the size and shape of surface cavities, and also considered other physicochemical properties, achieved better prediction performance than the energy-based methods that have been reported so far.

Many of the 408 cavity properties used in this study are correlated. This is partly due to the fact that some of the concepts we wanted to consider are either inherently vague (e.g., cavity shape) or do not have a single, fully satisfactory parameterization (e.g., hydrophobicity). Hence multiple, overlapping measures were used to assess these concepts. In addition, the protein structures used for the analysis can be fairly inaccurate, especially for surface residues. This is either because the structure is underdetermined by the experimental measurements or because of its inherent flexibility.[78] Using multiple cavity properties is expected to improve the robustness of the classifier to structural variability and inaccuracy than when a single property, such as cavity volume, is used. The sensitivity of our classification methodology to structure flexibility, as well as to the expected structural differences in the binding site between the bound and the apo forms, will be considered systematically in forthcoming reports.

Correlations among predictor variables present some difficulties for many statistical modeling techniques. Strategies that can be used to reduce variate correlations, for example, principal component analysis (PCA), exist. However, among other concerns, PCA results are notoriously difficult to interpret. Random Forests is possibly more robust to variate correlation than to other, simpler statistical methods such as linear regression.[72] Nonetheless, it is prudent to keep this fact in mind when interpreting the variable importance results. We clearly see from Table I that, in general, variables related to primarily cavity size and shape are important predictors of cavity druggability. However, the fact that the number of cavity residues and atoms in particular were found to be more important as predictors than, say, cavity volume or surface area, variables that are equally indicative of cavity size, may simply be an idiosyncrasy of the sample selected for this study.

Figure 6 suggests that many of the 408 cavity attributes included in this study make some contribution, albeit not necessarily a decisive one, toward the observed improvement in prediction accuracy of drug-binding cavities. This is not particularly surprising. Molecular recognition is a complex process likely to involve a multitude of observables, some more important than others. This multiplicity of causes is sometimes referred to as *effect size tapering* and is commonly encountered in the study of complex systems.[79] On the other hand, it is clear that the most important properties are related to the size and shape of the surface cavity. In fact, preliminary analysis indicates that a BER of 7.4% and coverage of 88.9% in the prediction of drug-binding cavities can be obtained using only size and shape cavity properties, a classification performance that is essentially identical to that obtained using the full complement of cavity properties computed in this study. Given that the question addressed here is about the prediction of drug-binding cavities irrespective of the nature of the drug, it is not surprising that generic attributes such as size and shape are shown to be the most useful overall. The same type of analysis could be repeated to identify and characterize surface cavities that recognize a particular ligand or are associated with a specific protein function. In this case, it is likely that physicochemical cavity properties, attuned to specific molecular interactions, would play a more prominent role.

As more genomic and structural information become available, the number of potential targets for pharmaceutical intervention multiplies.[80] A special need is emerging for a reliable utility that could aid in directing efforts toward the most promising (i.e., druggable) targets. This is especially important when the goal is to try to inhibit protein–protein interactions that so far have been notoriously difficult to disrupt.[81] In subsequent reports, we will discuss the application of SCREEN to study protein–protein interfaces available in the PDB and the assessment of their druggability potential. Surface cavities are likely important sites of specific and high-affinity interactions. The accurate detection, characterization, and classification of protein surface cavities afforded by SCREEN should help elucidate structural signatures that are diagnostic of certain protein function and thus aid in structure-based function prediction. A SCREEN web server that will allow the detection, evaluation, and visualization of protein surface cavities is available at http://interface.bioc.columbia.edu/screen/.

METHODS

The Data Set

Our data set is derived from a collection of 100 nonredundant protein–ligand complexes selected by Perola et al.[73] based on the following criteria: (1) diverse set of proteins and ligands, (2) ligands that are drugs or drug/lead-like (molecular weight between 200 and 600, 1–12 rotatable bonds), (3) high-quality structures (crystallographic resolution < 3.0 Å, no severe atomic clashes between the protein and the bound ligand), and (4) availability of experimental binding constant measurements. One of the complexes in the data set (PDB code: 1ly7) seems to no longer be available from the PDB, as the PDB code currently points to an unrelated structure. The remaining 99 structures were used in this study.

Identification of Surface Cavities

We developed a new method, SCREEN, for the accurate identification and characterization of molecular surface

cavities. SCREEN works by first constructing two molecular surfaces using GRASP[17]: a conventional molecular surface, MS, using a 1.4 Å probe radius and a second, low-resolution envelope surface using a large probe sphere. The low-resolution envelope surface, ME. serves as a "sea-level" surface from which elevation and depth can be measured and is used to define a ceiling for surface cavities. A probe sphere 5.0 Å in radius was chosen for the construction of the envelope surface. A depth value for each vertex of the MS surface is assigned by computing the shortest distance between this vertex and the ME surface. Surface cavities are then identified as contiguous MS surface regions where all vertices are below a surface depth threshold from the ME surface. In this work, a threshold of 2.0 Å was chosen to define cavity surface based on graphical analysis, as it yielded surface cavities compatible with intuition. The identification of surface cavities was not sensitive to the value of this parameter within the range 1.5–2.5 Å, although the measured cavity size differed somewhat.

The surface cavity identification algorithm proceeds by first clustering the MS surface vertices that are deeper than the threshold, using single-link clustering and the topology implied by the surface triangulation. Single-link clustering was used here, as it is the only clustering method that does not make prior assumptions, explicitly or implicitly, about the size of detectable clusters. Protein surface cavities tend to span a wide range of sizes, and we felt it was important not to bias our search in favor of cavities of a particular size. A caveat of single-link clustering is that it can generate noncompact, meaninglessly large clusters connected only by long, thin chains of vertices. This problem was circumvented effectively by using an image processing technique reminiscent of Delaney's cellular logic operations.[39] The clusters found by single-link clustering were postprocessed by applying one round of contraction followed by expansion operations. The contraction operation removed all boundary vertices from the clusters (vertices connected, by a surface triangle edge, to others that did not pass the surface depth criteria). This leads to the removal of all single-vertex chains connecting neighboring cavities, should they exist. Next, a one-step expansion operation is applied to the clusters, where by vertices that pass the depth criteria and have been removed in the previous step, are re-enrolled in the contracted cluster.

Our method allows accurate and physically meaningful definition of various properties of surface cavities. The surface area of the cavity is computed as the summation of the areas of the cavity surface triangles. The cavity volume is delineated accurately as the space between the cavity surface floor and its corresponding molecular envelope ceiling. Every surface vertex is unambiguously associated with a protein atom during the construction of the molecular surface. This allows us to map various physicochemical properties from the protein onto the cavity surface. Hence, the surface cavities can be characterized in multiple ways.

Cavity Properties

For each surface cavity, 408 attributes were computed. The complete list is provided in Table II in the Supplementary Material. Many of the attributes are statistics of the measured cavity properties, such as the average, standard deviation, entropy, minimum, maximum, and histogram. The computed attributes cover eight broad categories:

1. *Cavity size:* Cavity surface area, volume, diameter, number of residues and number of atoms. Cavity surface area is computed by summing the area of the surface triangles comprising the cavity floor. Cavity volume is computed using Gauss's theorem:

$$\int\int_S \mathbf{F} \cdot \mathbf{n} dS = \int\int\int_V \nabla \cdot \mathbf{F} dV, \quad (1)$$

where the left-hand integral is over a closed surface constructed by connecting the cavity floor and ceiling surfaces, **n** is the surface normal, and **F** is a vector field. The right hand integral is over the cavity volume. If **F** is chosen to have a constant divergence of 1 (in this work, $\mathbf{F} = (x/3, y/3, z/3)$, then the right-hand side would be equal to the cavity volume, which can then be obtained by means of the surface integral in the left-hand part of Gauss's law.

2. *Cavity shape:* Cavity moments of inertia, moments of inertia normalized by the number of surface vertices, depth, curvature, curvedness, mean curvature, Gaussian curvature, shape index. Surface curvature is defined as the solid angle on the accessible surface of a probe water molecule in contact with the molecular surface that is removed by the surface[17] and is computed using GRASP. Gaussian curvature (K), mean curvature (H), Curvedness (R), and the shape index (S) are computed as follows[82]:

$$K = \kappa_{max}\kappa_{min} \quad (2)$$

$$H = \frac{\kappa_{max} + \kappa_{min}}{2} \quad (3)$$

$$R = \sqrt{\frac{\kappa_{min}^2 + \kappa_{max}^2}{2}} \quad (4)$$

$$S = -\frac{2}{\pi} \arctan\left(\frac{\kappa_{max} + \kappa_{min}}{\kappa_{max} - \kappa_{min}}\right), \quad (5)$$

where κ_{max} and κ_{min} are the principal curvatures calculated by differentiating the Gaussian density approximation of the protein[83] at the molecular surface.

3. *Hydrophobicity:* proportions of hydrophobic and hydrophilic cavity surface, atom solvation parameters (ASPs),[84] residue-specific octanol to water transfer energy,[85] and overall cavity solvation energy, E_{solv}, calculated as follows:

$$E_{solv} = \sum_{i=1}^{n} ASA_i ASP_i, \quad (6)$$

where n is the number of atoms lining the cavity, ASA_i is the accessible surface area of atom i, and ASP_i is the atomic solvation parameter of atom i.

4. *Electrostatics:* the atomic partial charges assigned according to the CHARMM 22 charge set, and the electrostatic potential and field, calculated using the program DELPHI[86] at the locations of surface vertices.
5. *Hydrogen bonding:* The proportion of the cavity surface made up of hydrogen bond donor or acceptor groups, and the density of hydrogen-bonding groups, $HB_{density}$, calculated at surface vertex v, in a manner analogous to that proposed by Exner et al.,[29] as the fraction of the molecular surface within distance of 5.0 Å from v that is associated with a chemical group capable of making a hydrogen bond.
6. *Amino acid composition:* the proportion of the cavity surface due to each of the 20 standard amino acids.
7. *Secondary structure:* the proportion of the cavity surface overlying an α-helix, a β-sheet, a turn, or a coil secondary structure, as determined by DSSP.[87]
8. *Rigidity:* cavity rigidity evaluated by estimating the conformational entropy of cavity residues using an approach proposed by Pickett and Sternberg,[88] and extended by Vajda et al.[89] Briefly, residue i of type r is assumed to retain its full conformational entropy in the unfolded state S_r^{free}(obtained from the Pickett and Sternberg scale) if its relative accessibility, ASA_i in the protein structure (compared to the unfolded state) is greater than 60%. Otherwise, the conformational entropy of residue i, S_i, is proportional to its relative accessibility as follows[89]:

$$S_i = \frac{10\ ASA_i S_r^{free}}{6}. \quad (7)$$

We intentionally did not use B factors as a possible predictor in our analysis, as our sample consisted of complexes of drug-bound proteins. The presence of the bound drug is likely to lower the B factors of the protein atoms at the binding site, hence creating a signal that is not present in the unbound form. Since we intended to use our prediction strategy on proteins that were not cocrystallized with drugs, we opted to exclude B factors from our analysis.

Random Forests

Random Forests is a modern, ensemble-based classification strategy, where the role of the weak predictors is played by decision trees.[72] Key to prediction accuracy is the reduction of both the bias and the correlation of the weak predictors.[72] Random Forest reduces the correlation between the ensemble trees by (1) growing each tree using a random, bootstrap sample of the training data; and (2) using a small subset of the available predictors, also chosen randomly, to select the best split at each node. Each tree is grown to maximum depth without regularization or pruning. This helps reduce the bias of the individual weak predictors. Random Forests has certain strengths: It does not require variable rescaling, it detects and handles high order interactions, it seems to be more robust to noise and data mislabeling than some of the other classification techniques like boosting, and it features prediction accuracy that compares favorably with other, state-of-the-art classification techniques such as boosting, support vector machines, and neural networks.[72]

Training the Random Forests Classifier

The surface cavities in our data set are represented by a set of feature vectors $\{(x_i,y_i)\}_{i=1}^n$, where $\mathbf{x}_i$ is a vector of 408 attributes computed for cavity i, y_i is label indicating whether cavity i is a drug-binding cavity or not, and n is the number of observations, in this case, 1286. The Random Forests classifier is trained by repeatedly drawing a sample of n observations with replacement (bootstrap sampling) from the data set. These drawn observations will be used to construct a decision tree that learns the label y_i based on the values of the properties $\mathbf{x}_i$. Since the sampling is performed with replacement, some observations will be drawn more than once and others, not at all. One tree is constructed at each iteration. The observations not used to train this tree are called the *out-of-bag* (OOB) cases and are used as a test set for that tree. Typically about two thirds of the cases will be drawn and used as a training set, while the remaining one third of the cases will constitute the test set. In this sense this training/testing scheme is essentially equivalent to the more familiar three-fold cross-validation design. The OOB error rate is considered a good estimate of the generalization error rate.

Evaluating the Classification Performance

Given that the data set is class-imbalanced [i.e., many more non-drug-binding cavities (1187) than drug-binding cavities (99)], an overall error rate is not a useful indicator of the performance of the classifier. In an extreme case, a trivial classifier can achieve an accuracy of 92% simply by classifying every cavity as a non-drug-binding cavity, while being completely useless for all practical purposes. Instead we evaluate the classifier using the BER coverage (also referred to as sensitivity, or recall) and the correlation coefficient r:

$$\mathrm{BER} = \frac{1}{2}\left(\frac{FN}{FN+TP} + \frac{FP}{FP+TN}\right) \quad (8)$$

$$\mathrm{coverage} = \frac{TP}{TP+FN} \quad (9)$$

$$r = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}, \quad (10)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. Sometimes r is referred to as the Matthews correlation coefficient. But essentially it is the standard Pearson's correlation coefficient between the observed and predicted classification encoded as indicator variables (taking values of {0,1}).

CASTp Analysis

CASTp results were obtained by submitting the protein structures in our data set to the CASTp web server at http://cast.engr.uic.edu/cast using a perl script and parsing the HTML response of the server.

ACKNOWLEDGMENTS

Our thanks to Natasja Brooijman for suggesting the data set used in this work and for informative discussions on rational drug design, and to Phil Long for simulating conversations on machine learning.

REFERENCES

1. Sheinerman FB, Honig B. On the role of electrostatic interactions in the design of protein–protein interfaces. J Mol Biol 2002;318: 161–177.
2. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci 1996;5:2438–2452.
3. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
4. Jones S, Thornton JM. Prediction of protein–protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
5. Jones S, Thornton JM. Analysis of protein–protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
6. Keil M, Exner TE, Brickmann J. Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. J Comput Chem 2004;25:779–789.
7. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
8. Lijnzaad P, Berendsen HJ, Argos P. Hydrophobic patches on the surfaces of protein structures. Proteins 1996;25:389–397.
9. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. J Mol Biol 2004;338:181–199.
10. Stawiski EW, Gregoret LM, Mandel-Gutfreund Y. Annotating nucleic acid-binding function based on protein structure. J Mol Biol 2003;326:1065–1079.
11. Pettit FK, Bowie JU. Protein surface roughness and small molecular binding sites. J Mol Biol 1999;285:1377–1382.
12. Peters KP, Fauck J, Frommel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 1996;256:201–213.
13. Aqvist J, Tapia O. Surface fractality as a guide for studying protein–protein interactions. J Mol Graph 1987;5:30–34.
14. Ofran Y, Rost B. Analysing six types of protein–protein interfaces. J Mol Biol 2003;325:377–387.
15. Lewis RA. Clefts and binding sites in protein receptors. Methods Enzymol 1991;202:126–156.
16. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 1998;7:1884–1897.
17. Nicholls A, Sharp KA, Honig B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 1991;11:281–296.
18. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
19. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R. Optimal docking area: a new method for predicting protein–protein interaction sites. Proteins 2005;58:134–143.
20. Preissner R, Goede A, Frommel C. Dictionary of interfaces in proteins (DIP): data bank of complementary molecular surface patches. J Mol Biol 1998;280:535–550.
21. Cosgrove DA, Bayada DM, Johnson AP. A novel method of aligning molecules by local surface shape similarity. J Comput Aided Mol Des 2000;14:573–591.
22. 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.
23. Ferre F, Ausiello G, Zanzoni A, Helmer-Citterich M. SURFACE: a database of protein surface regions for functional annotation. Nucleic Acids Res 2004;32(Database issue):D240–D244.
24. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 2003;19:163–164.
25. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 2002;18(Suppl 1):S71–S77.
26. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
27. Nayal M, Hitz BC, Honig B. GRASS: a server for the graphical representation and analysis of structures. Protein Sci 1999;8:676–679.
28. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein–protein recognition. Protein Sci 1994;3:717–729.
29. Exner TE, Keil M, Brickmann J. Pattern recognition strategies for molecular surfaces: I. Pattern generation using fuzzy set theory. J Comput Chem 2002;23:1176–1187.
30. Stahl M, Taroni C, Schneider G. Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. Protein Eng 2000;13:83–88.
31. Di L, Kerns EH. Profiling drug-like properties in discovery research. Curr Opin Chem Biol 2003;7:402–408.
32. Oprea TI. Current trends in lead discovery: are we looking for the appropriate properties? J Comput Aided Mol Des 2002;16:325–334.
33. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 2001;46:3–26.
34. Voorintholt R, Kosters MT, Vegter G, Vriend G, Hol WG. A very fast program for visualizing protein surfaces, channels and cavities. J Mol Graph 1989;7:243–245.
35. Ho CM, Marshall GR. Cavity search: an algorithm for the isolation and display of cavity-like binding regions. J Comput Aided Mol Des 1990;4:337–354.
36. Levitt DG, Banaszak LJ. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 1992;10:229–234.
37. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 1997;15:359–363, 389.
38. Kleywegt GJ, Jones TA. Detection, delineation, measurement and display of cavities in macromolecular structures. Acta Crystallogr D Biol Crystallogr 1994;50:178–185.
39. Delaney JS. Finding and filling protein cavities using cellular logic operations. J Mol Graph 1992;10:174–177, 163.
40. Masuya M, Doi J. Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations. J Mol Graph 1995;13:331–336.
41. Kuntz I, Blaney J, Oatley S, Langridge R, Ferrin T. A geometric approach to macromolecule–ligand interactions. J Mol Biol 1982; 161:269–288.
42. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 1995;13:323–330, 307–328.
43. Brady GP Jr, Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 2000;14:383–401.
44. Lewis RA. Determination of clefts in receptor structures. J Comput Aided Mol Des 1989;3:133–147.
45. Connolly ML. Measurement of protein surface shape by solid angles. J Mol Graphics 1986;4:3–6.
46. Lewis M, Rees DC. Fractal surfaces of proteins. Science 1985;230: 1163–1165.
47. Wodak SJ, Janin J. Computer analysis of protein–protein interaction. J Mol Biol 1978;124:323–342.
48. Fanning DW, Smith JA, Rose GD. Molecular cartography of globular proteins with application to antigenic sites. Biopolymers 1986;25:863–883.
49. Badel-Chagnon A, Nessi J, Buffat L, Hazout S. "Iso-depth contour map" of a molecular surface. J Mol Graph 1994;12:162–168, 193.
50. Voronoi G. New applications of continuous parameters of the theory of quadratic forms. Journal für die Reine und Angewandte Mathematik 1907;133:97–178.

51. Rogers CA. Packing and covering. Cambridge UK: Cambridge University Press; 1964.
52. Edelsbrunner H, Mucke EP. Three dimensional alpha shapes. ACM Trans Graph 1994;13:43–72.
53. Edelsbrunner H. The union of balls and its dual shape. Discrete Comput Geom 1995;13:415–440.
54. Edelsbrunner H, Facello M, Liang J. On the definition and the construction of pockets in macromolecules. Discrete Appl Math 1998;88:83–102.
55. Ruppert J, Welch W, Jain AN. Automatic identification and representation of protein binding sites for molecular docking. Protein Sci 1997;6:524–533.
56. An J, Totrov M, Abagyan R. Pocketome via comprehensive identification and classification of ligand-binding envelopes. Mol Cell Proteomics 2005;4:752–761.
57. Laurie AT, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 2005;21:1908–1916.
58. Nissink JW, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R. A new test set for validating predictions of protein–ligand interaction. Proteins 2002;49:457–471.
59. Eisenhaber F, Argos P. Hydrophobic regions on protein surfaces: definition based on hydration shell structure and a quick method for their computation. Protein Eng 1996;9:1121–1133.
60. Heiden W, Brickmann J. Segmentation of protein surfaces using fuzzy logic. J Mol Graph 1994;12:106–115.
61. Klapper I, Hagstrom R, Fine R, Sharp K, Honig B. Focussing of electric fields in the active site of CuZn superoxide dismutase: effects of ionic strength and amino acid modification. Proteins 1986;1:47–59.
62. Bate P, Warwicker J. Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. J Mol Biol 2004;340:263–276.
63. Nayeem A, Krystek S, Jr., Stouch T. An assessment of protein–ligand binding site polarizability. Biopolymers 2003;70:201–211.
64. Alkorta I, Perez JJ, Villar HO. Molecular polarization maps as a tool for studies of intermolecular interactions and chemical reactivity. J Mol Graph 1994;12:3–13.
65. Ondrechen MJ, Clifton JG, Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure. Proc Natl Acad Sci USA 2001;98:12473–12478.
66. Shehadi IA, Yang H, Ondrechen MJ. Future directions in protein function prediction. Mol Biol Rep 2002;29:329–335.
67. Duncan B, Olson A. Approximation and characterization of molecular surfaces. Biopolymers 1993;33:219–229.
68. Liang S, Zhang J, Zhang S, Guo H. Prediction of the interaction site on the surface of an isolated protein structure by analysis of side chain energy scores. Proteins 2004;57:548–557.
69. Elcock AH. Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 2001;312:885–896.
70. Korolev S, Nayal M, Barnes WM, Di Cera E, Waksman G. Crystal structure of the large fragment of *Thermus aquaticus* DNA polymerase I at 2.5-Å resolution: structural basis for thermostability. Proc Natl Acad Sci USA 1995;92:9264–9268.
71. Breiman L. Bagging predictors. Machine Learning 1996;24:123–140.
72. Breiman L. Random Forests. Machine Learning 2001;45:5–32.
73. Perola E, Walters WP, Charifson PS. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 2004;56:235–249.
74. Cover TM, Thomas JA. Elements of information theory. New York: Wiley-Interscience; 1991.
75. Iversen LF, Andersen HS, Moller KB, Olsen OH, Peters GH, Branner S, Mortensen SB, Hansen TK, Lau J, Ge Y, Holsworth DD, Newman MJ, Hundahl Moller NP. Steric hindrance as a basis for structure-based design of selective inhibitors of protein–tyrosine phosphatases. Biochemistry 2001;40:14812–14820.
76. Smith GM, Alexander RS, Christianson DW, McKeever BM, Ponticello GS, Springer JP, Randall WC, Baldwin JJ, Habecker CN. Positions of His-64 and a bound water in human carbonic anhydrase II upon binding three structurally related inhibitors. Protein Sci 1994;3:118–125.
77. Maignan S, Guilloteau JP, Pouzieux S, Choi-Sledeski YM, Becker MR, Klein SI, Ewing WR, Pauls HW, Spada AP, Mikol V. Crystal structures of human factor Xa complexed with potent inhibitors. J Med Chem 2000;43:3226–3232.
78. DePristo MA, de Bakker PI, Blundell TL. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure (Camb) 2004;12:831–838.
79. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. 2nd ed. New York: Springer-Verlag; 2002.
80. Henry CM. Structure-based drug design. Chem Eng News 2001;79: 69–74.
81. Cochran AG. Antagonists of protein–protein interactions. Chem Biol 2000;7:R85–R94.
82. Koenderink JJ. Solid shape. Cambridge, MA: MIT Press; 1990.
83. Diamond R. Real-space refinement of the structure of hen egg-white lysozyme. J Mol Biol 1974;82:371–391.
84. Eisenberg D, Wesson L. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci 1992;1:227–235.
85. Radzicka A, Wolfenden R. Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry 1988;27:1644–1670.
86. Sharp KA, Honig B. Applications of the finite difference Poisson–Boltzmann method to proteins and nucleic acids. In: Sarma RH, Sarma MH, editors. DNA protein complexes and proteins: Vol. 2. Structure and methods. Adenine Press; 1990. p 211.
87. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
88. Pickett SD, Sternberg MJ. Empirical scale of side-chain conformational entropy in protein folding. J Mol Biol 1993;231:825–839.
89. Vajda S, Weng Z, Rosenfeld R, DeLisi C. Effect of conformational flexibility and solvation on receptor–ligand binding free energies. Biochemistry 1994;33:13977–13988.