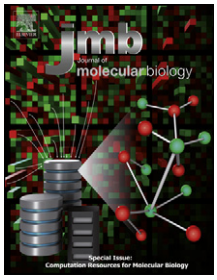

CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites

Peter Cimermancic[1,2], Patrick Weinkam[1], T. Justin Rettenmaier[3,4], Leon Bichmann[1], Daniel A. Keedy[1], Rahel A. Woldeyes[1,3], Dina Schneidman-Duhovny[1], Omar N. Demerdash[5], Julie C. Mitchell[6], James A. Wells[4,7], James S. Fraser[1] and Andrej Sali[1,4]

*1 - **Department of Bioengineering and Therapeutic Sciences,** University of California, San Francisco, San Francisco, CA 94158, USA*
*2 - **Graduate Group in Biological and Medical Informatics,** University of California, San Francisco, San Francisco, CA 94158, USA*
*3 - **Graduate Group in Chemistry and Chemical Biology,** University of California, San Francisco, San Francisco, CA 94158, USA*
*4 - **Pharmaceutical Chemistry and California Institute for Quantitative Biosciences,** University of California, San Francisco, San Francisco, CA 94158, USA*
*5 - **Department of Chemical and Biomolecular Engineering,** University of California, Berkeley, Berkeley, CA 94720, USA*
*6 - **Departments of Biochemistry and Mathematics,** University of Wisconsin–Madison, Madison, WI 53706, USA*
*7 - **Cellular and Molecular Pharmacology and California Institute for Quantitative Biosciences,** University of California, San Francisco, San Francisco, CA 94158, USA*

***Correspondence to Peter Cimermancic and Andrej Sali:** P. Cimermancic is to be contacted at: University of California, San Francisco, 1700 4th Street, Byers Hall 501A, San Francisco, CA 94158, USA; A. Sali, is to be contacted at: University of California, San Francisco, 1700 4th Street, Byers Hall 503B, San Francisco, CA 94158, USA. peter.cimermancic@ucsf.edu; http://salilab.org*
http://dx.doi.org/10.1016/j.jmb.2016.01.029
Edited by M. Sternberg

Abstract

Many proteins have small-molecule binding pockets that are not easily detectable in the ligand-free structures. These cryptic sites require a conformational change to become apparent; a cryptic site can therefore be defined as a site that forms a pocket in a *holo* structure, but not in the *apo* structure. Because many proteins appear to lack druggable pockets, understanding and accurately identifying cryptic sites could expand the set of drug targets. Previously, cryptic sites were identified experimentally by fragment-based ligand discovery and computationally by long molecular dynamics simulations and fragment docking. Here, we begin by constructing a set of structurally defined *apo–holo* pairs with cryptic sites. Next, we comprehensively characterize the cryptic sites in terms of their sequence, structure, and dynamics attributes. We find that cryptic sites tend to be as conserved in evolution as traditional binding pockets but are less hydrophobic and more flexible. Relying on this characterization, we use machine learning to predict cryptic sites with relatively high accuracy (for our benchmark, the true positive and false positive rates are 73% and 29%, respectively). We then predict cryptic sites in the entire structurally characterized human proteome (11,201 structures, covering 23% of all residues in the proteome). CryptoSite increases the size of the potentially "druggable" human proteome from ~40% to ~78% of disease-associated proteins. Finally, to demonstrate the utility of our approach in practice, we experimentally validate a cryptic site in protein tyrosine phosphatase 1B using a covalent ligand and NMR spectroscopy. The CryptoSite Web server is available at http://salilab.org/cryptosite.

Introduction

Biological function often involves binding of proteins to other molecules, including small ligands and macromolecules. Usually, these interactions occur at defined binding sites in the protein structure [1]. Knowledge of binding site location has a number of applications [2]. For example, in drug discovery,

binding site localization is often the starting point followed by virtual screening or *de novo* ligand design [3]; in cell biology, it facilitates prediction of protein substrates, especially when the target protein cannot be reliably related to homologs of known function [4].

Binding sites, particularly those for small molecules, are often located in exposed concave pockets, which provide an increased surface area that, in turn, maximizes intramolecular interactions [5]. A concave pocket can already exist in a ligand-free structure of a protein; such binding sites are called here binding pockets. Sometimes, however, a binding site is flat in the absence of a ligand and only forms in the presence of a ligand (i.e., induced fit) or only opens transiently for short periods of time (i.e., conformational selection); such binding sites are called cryptic sites (Fig. 1a) [6–11].

Many computational methods have been developed to localize binding pockets on proteins. These methods are based on a variety of principles [12]: (i) concavity of the protein surface, (ii) energy functions including van der Waals terms, (iii) geometrical and physicochemical similarity to known binding pockets, and (iv) composite approaches that use a combination of different features [13–15]. Unfortunately, only ~60% of protein structures were judged to

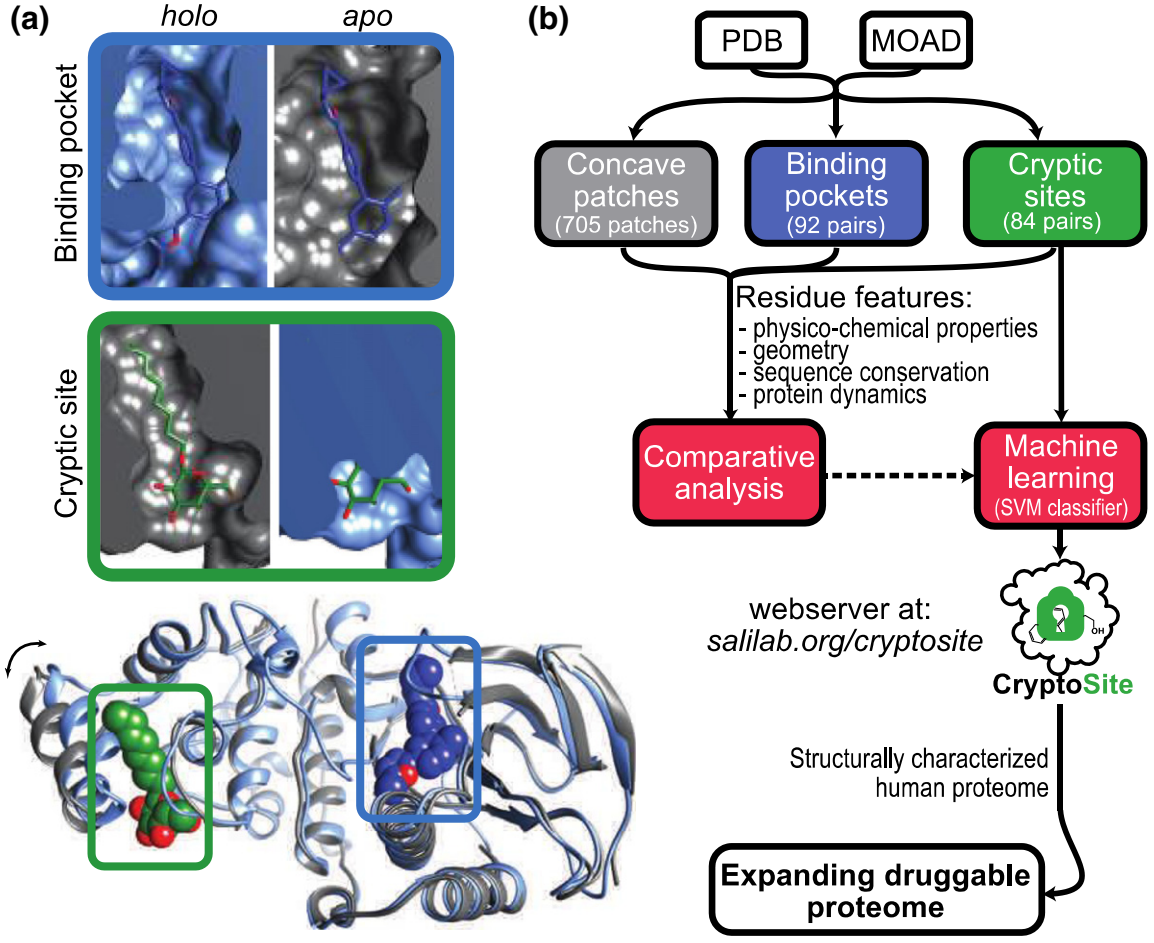

Fig. 1. (a) Examples of a pocket and cryptic site in p38 MAP kinase. The nucleotide binding site of the p38 MAP kinase is a pocket visible in both bound (*holo*; blue ribbon; PDB ID: 2ZB1) and unbound (*apo*; gray ribbon; PDB ID: 2NPQ) conformations. The ligand, biphenyl amide inhibitor, is depicted as blue spheres. On the other hand, the site in the C-lobe domain that binds octyglucoside lipid (green spheres) becomes a visible pocket only after the movement of the α-helix at the left of the structure (marked with the double-headed arrow). The small molecules are shown as they bind in the *holo* structures. UCSF Chimera software was used for the visualization [67]. (b) Flowchart summarizing the analyses in this study. We started by creating a representative dataset of 84 known examples of cryptic binding sites, 92 binding pockets, and 705 concave surface patches from the Protein Data Bank [31] and the MOAD database [32]. Next, we designed a set of 58 features that describe sequence, structure, and dynamics of individual residues and their neighbors. We then compared these attributes between the three types of a site to better understand the underlying characteristics of each site. Next, we used machine-learning algorithms to classify residues as belonging to a cryptic site or not. We then predicted cryptic sites in the entire structurally characterized human proteome (Materials and Methods and SI Text).

have pockets larger than 250 $Å^3$ (many of which may not be druggable) and could potentially be subjected to ligand discovery based on binding pocket knowledge [16,17].

In contrast to binding pockets, cryptic sites are not easily detectable in a ligand-free structure of a protein because they, by definition, require ligand-induced conformational changes to become apparent. For example, large and flat interfaces between interacting proteins were considered undruggable, although several examples of protein interfaces undergoing a conformational change coupled with binding a small molecule were recently described [18,19]. Similarly, allosterically regulated sites are sometimes not apparent in the absence of a small-molecule allosteric regulator (e.g., p38 MAP kinase [7] and TEM1 β-lactamase [9]).

Currently, the only approaches to cryptic site discovery are exhaustive site-directed small-molecule tethering by experiment [20–22], long-timescale molecular dynamics simulations by computation [6,8,23,24], flexible docking [25,26], and computational tools for identification of small-molecule binding hot spots [10,27–30]. All of these approaches are time consuming, expensive, and/or not always successful. Therefore, there is a need for an accurate, automated, and efficient method to predict the location of cryptic pockets in a given ligand-free protein structure. Such a method would offer several advantages. First, a cryptic site may be the only suitable binding site on the target protein; for example, when activation is required and thus the active site cannot be targeted, the active site is not druggable, or active site ligands need to be avoided due to adverse off-target effects. Second, binding sites may be discovered on structures determined or computed at only moderate resolution.

Here, we analyze known cryptic sites and develop a method for predicting cryptic site locations to address a number of questions: what are the sequence, structure, and dynamics attributes of a cryptic site, especially in comparison to binding pockets? Can we accurately, automatically, and efficiently predict cryptic sites? How common are cryptic sites? Are they common enough to significantly expand the druggable proteome? Can we predict cryptic sites in specific proteins of clinical significance?

Results and Discussion

Our method development and analysis proceeded according to Fig. 1b. In outline, we started by creating a representative dataset of 84 known examples of cryptic binding sites, 92 binding pockets, and 705 concave surface patches from the Protein Data Bank [31] and the MOAD database [32] (Materials and Methods, SI Text, and Table S1). We selected cryptic sites and binding pockets whose ligands are biologically relevant [32]. Next, we designed a set of 30 features that describe sequence, structure, and dynamics of individual residues and their neighbors (SI Text and Table S2), based on the crystal structures [15,33,34]. We then compared these attributes between the three types of a site to better understand the underlying characteristics of each site. Based on these comparisons, we expanded the set of features for proteins containing cryptic sites to 58 (Table S2), describing their crystal structures and their alternative conformations obtained by molecular dynamics simulations using AllosMod [35] (SI Text). Next, we put to test 11 supervised machine-learning algorithms [36,37] to classify residues as belonging to a cryptic site or not; the accuracy of the best predictive model was assessed using leave-one-out cross-validation on a training set and using an independent test set. We then predicted cryptic sites in the entire structurally characterized human proteome. Finally, we focused on a detailed characterization of protein tyrosine phosphatase 1B (PTP1B), a protein that is involved in the insulin signaling pathway and is considered a validated therapeutic target for treatment of type 2 diabetes [38].

Pocket formation at a cryptic site is driven by small changes in the structure, resulting in a conformationally conserved cryptic site regardless of the ligand type

First, we set out to analyze structural changes needed for a binding pocket formation at a cryptic site. The dataset of cryptic sites reveals mostly minor structural changes required for formation of a detectable pocket. The all-atom RMSD of cryptic binding sites between *apo* and *holo* conformations ranges between 0.45 Å and 22.45 Å (Fig. S1a) with 67% *apo–holo* pairs differing less than 3 Å in RMSD. The only two *apo–holo* pairs whose differences in RMSD exceed 10 Å are calcium ATPase and calmodulin (PDB IDs: 1SU4–3FGO and 1CLL–1CTR, respectively). Loop movement is the most prominent type of conformational changes (observed in 45% of the binding sites), followed by side-chain rotation (18%), domain motion (17%), displacement of secondary structure elements (16%), and N- or C-terminal flexibility (4%).

To determine whether or not a cryptic site assumes the same bound conformation irrespective of the ligand type, we computed similarities between cryptic site conformations in a protein bound to at least five different ligands (58 proteins). Interestingly, only 26% of such cases have an average RMSD exceeding 2 Å (Fig. S1b), even though the average Tanimoto distance (calculated by Open Babel [39]; SI Text) is low (0.8). This finding suggests that the conformation of a given cryptic site generally does not depend strongly on the ligand type (similar analysis of binding pockets yields 9% of cases with an average RMSD exceeding 2 Å, as well as an

average Tanimoto distance of 0.7). Moreover, the magnitude of the conformational difference within a group of *holo* structures is not significantly correlated with ligand similarity (the correlation coefficient between the all-atom binding site RMSD and Tanimoto distance is 0.01; Fig. S1c and S1d). Finally, the average RMSD of 1.7 Å between bound cryptic binding sites is significantly lower than the average RMSD of 3.0 Å between the unbound and bound conformations ($P = 1.4 \times 10^{-3}$, based on two-sample Kolmogorov–Smirnov statistics). Thus, the bound form of the cryptic site is surprisingly conformationally conserved with respect to the ligand type (the average RMSD values of bound conformations of cryptic sites and binding pockets are 1.7 and 2.0 Å, respectively). These observations are consistent with a limited number of protein conformational states, as well as with the variability in allosterically regulated proteins, where the binding of the effector alters the conformational distribution between two or more conformational states [40]. Indeed, 24 of the 58 cryptic sites are found in proteins that are known to be allosterically regulated, with 17 of the 24 annotated as effector binding sites [41]. Of the remaining 34 cryptic sites, 20 are found on proteins with two or more different binding sites that may or may not be allosteric. The remaining 14 cryptic sites occur on enzymes with flexible active sites and receptors for large hydrophobic ligands, where cryptic site residues modulate binding site accessibility (e.g., the "portal" hypothesis for glycolipid transfer protein, lactoglobulin, and adipocyte lipid binding protein) [42]. In other words, a cryptic site does not convert from flat to concave to accommodate a number of different ligands; rather, cryptic sites may have evolved the ability to convert from flat to concave to modulate ligand binding kinetics, specificity, affinity, and allostery.

Cryptic sites are as flexible as random concave surface patches but evolutionarily as conserved as binding pockets

Next, we analyzed the differences between the sequence, structure, and dynamics attributes of cryptic sites, binding pockets, and concave surface patches. While the differences between cryptic sites and binding pockets are generally small, four characteristics distinguish a cryptic site from a binding pocket and/or a concave surface patch: first, a cryptic site predominantly localizes at concave protein regions, even though the site itself is not as concave in the unbound form as a binding pocket. For example, while the average number of protruding atoms at a cryptic site and a binding pocket is 170 and 183 ($P = 8.0 \times 10^{-3}$) and the average convexity value is 2.4 and 1.9 ($P = 0.8$), the average pocket score is 0.07 and 0.42 ($P = 1.7 \times 10^{-31}$), respectively (Table S3). Second, a cryptic site tends to be less hydrophobic than a binding pocket due mostly to an increased frequency of charged residues (arginine, in particular; $P = 1.8 \times 10^{-5}$) (Fig. S2a and Table S3). Third, a cryptic site is more flexible than a binding pocket, as indicated by significantly higher normalized *B*-factors (Fig. S2b). Finally, cryptic site residues are evolutionarily as conserved as those of a binding pocket (Fig. S2c), suggesting a similar degree of evolutionary pressure and selection on the function of many of these two types of binding sites. Evolutionarily conserved residues have been previously associated with low *B*-factors [43–45]; low *B*-factors are an indicator of residue rigidity. Both evolutionarily conserved residues and residues with low *B*-factors are often found in functionally important regions of a protein, including binding pockets [13,46]. In contrast to binding pockets, cryptic sites conserve conformational flexibility to convert from flat into concave. We found no statistically significant differences between properties of ligands of cryptic sites and binding pockets (Fig. S3).

Molecular dynamics simulations based on a simplified energy landscape, sequence conservation, and fragment docking are sufficient to predict cryptic sites

To test if cryptic sites could be predicted accurately, automatically, and efficiently, we used the dataset of *apo* structures with cryptic sites to train 10 different machine-predictive models for the prediction of cryptic site residues, based on the extended set of 58 features (Table S2); the datasets with binding pockets and concave surface patches were not used as training sets for machine learning. The optimal predictive model and its parameter values were selected by maximizing the sensitivity (true positive rate) and the specificity (true negative rate) of cryptic site residue prediction, using leave-one-out cross-validation on the training set of proteins with 84 cryptic binding sites (SI Text and Fig. S4a). The optimal predictive model is a support vector machine (SVM) with a quadratic kernel function. By removing redundant and irrelevant features using greedy-forward selection that maximizes the area under the curve (AUC) and by testing the statistical significance of the improvement in the prediction accuracy, we selected three features, resulting in the AUC of 0.77 (Fig. S4b–d).

Although an SVM operates as a "black box", the relative importance of different features can be inferred from the order of selection and may be informative about the cryptic site characteristics [47]. We find the average pocket score from the molecular dynamics simulations is the most informative single feature according to greedy-forward selection (AUC = 0.73), as well as the two-sample Kolmogorov–Smirnov test ($P = 4.3 \times 10^{-138}$) (Fig. S2d and Table S2). This feature alone is almost as informative as a subset of 30 crystal structure features combined (AUC = 0.74)

(Table S2). Therefore, molecular dynamics simulations on a simplified energy landscape, which is significantly more computationally efficient than a traditional all-atom molecular dynamics simulation [6], often provide sufficient information for localizing cryptic sites. The second feature added to the subset of the three features by the greedy-forward approach was sequence conservation (AUC = 0.74). Cryptic site residues are significantly more conserved than the rest of a protein ($P = 3.4 \times 10^{-67}$). The third feature, likelihood of binding small-molecule fragments (SI Text), also significantly improves the accuracy of the model (AUC = 0.77). Despite the relatively small magnitude of the increase in the accuracy, the improvement of adding two additional features to the single most informative feature is statistically significant (Fig. S4c and S4d). In summary, a cryptic site can be predicted relatively accurately based primarily on pocket formation in molecular dynamics simulations, evolutionary conservation, and likelihood of binding small-molecule fragments. Independent predictions based on different molecular dynamics trajectories are highly similar, with the cross-correlation coefficient larger than 0.9 and the average residue score difference of the most variable decile smaller than 0.04 (Fig. S5a); the predicted scores vary the most for residues that reside on α-helices or β-sheets and are adjacent to flexible parts of a protein. Similarly, predictions for a subunit on its own or in the context of a biological assembly are also highly similar, except for the subunit–subunit interface residues (Fig. S5b).

CryptoSite accurately localizes over 96% of cryptic binding sites, outperforming other computational methods

To assess the performance of our predictive model, we applied it to the training set using leave-one-out cross-validation and to the test set of 14 *apo* structures with one or more known cryptic sites that were not used during the training or any of the analyses above. The prediction capability of the SVM model is satisfactory; we measure an overall AUC of 0.83, with respective true positive and false positive rates of 79% and 29% at the residue score threshold of 0.05 (Fig. 2a). At higher score thresholds of 0.1 and 0.15, the respective true positive and false positive rates are 15% and 55%, as well as 6% and 28% (in other words, in an experimental test of a prediction, on average, 7.6, 5.9, and 4.9 residues with the predicted residues score higher than 0.05, 0.1, and 0.15, respectively, would need to be tested to find at least one true cryptic site residue—a significant improvement over the need to test 19 randomly chosen residues for the same outcome). CryptoSite can also be applied to low-resolution atomic structures and comparative models, in addition to high-resolution X-ray structures, without a large loss of accuracy. For example, the average cross-correlation between cryptic site predictions for a high-resolution X-ray structure and its comparative model based on a template with at least 50% sequence identity is approximately 0.7 (SI Text, Fig. S6, and Table S4). To further dissect the performance of the learning algorithm, we evaluated predictions for individual proteins from our training and test sets (Fig. 2b). We define a prediction of a cryptic site to be accurate when at least one-third of its residues are identified (sensitivity >33%). Predictions above this threshold can arguably guide small-molecule tethering experiments and more detailed molecular dynamics simulations. Remarkably, all 14 proteins in the test set and 75 out of 79 proteins in the cross-validation/training set have all of their cryptic sites identified accurately, resulting in 96% recall (Tables S1 and S5); even for 50% sensitivity, the recall is still 88%. The predictions are particularly accurate when a large and hydrophobic ligand binds to a cryptic site. For example, we identified 98% of cryptic site residues in the acyl-CoA binding site of the fatty acid responsive transcription factor and 89% of cryptic site residues in the lipid binding site of β-lactoglobulin (Fig. S7). Our predictive model also accurately predicted cryptic sites in 18 out of 20 proteins (including the proteins from the cross-validation set) that undergo domain movements to expose small-molecule binding sites. For example, more than half of the cryptic site residues of GluR2 receptor (100%), exportin 1 (68%), and biotin carboxylase were predicted correctly (56%) (Fig. 2c and Fig. S7).

Our predictive model also accurately predicts known allosteric cryptic sites in TEM1 β-lactamase that are buried in the *apo* conformation (60%) (Fig. 2c and Fig. S7d) and were previously studied using extensive molecular dynamics simulations in explicit solvent and Markov state models [6,24]. Moreover, both molecular dynamics simulations [23,25,48] and CryptoSite also successfully predicted known binding sites at difficult-to-drug protein–protein interaction interfaces, including in interleukin-2 (specificity of 79%), Bcl-X_L (73%), FK506 binding protein (FKBP12; 73%), HPV regulatory protein E2 (50%), and cell division protein ZipA (60%) (Fig. 2c and Figs. S7e and S8c). Finally, we used our testing set to benchmark CryptoSite against FTFlex [28,49], a computational solvent mapping approach for prediction of small-molecule binding hot spots that takes into account side chain flexibility. CryptoSite is more accurate (Fig. 2a and Fig. S8a), especially when a cryptic site is buried (TEM1 β-lactamase and β-lactoglobulin) or resides in a large protein (exportin 1; 68%) (Fig. 2 and Fig. S8). In conclusion, CryptoSite is as accurate as approaches based on extensive molecular dynamics simulations but is significantly faster (a calculation on an average-sized protein takes 1–2 days on our Web server) and completely automated. In comparison to approaches

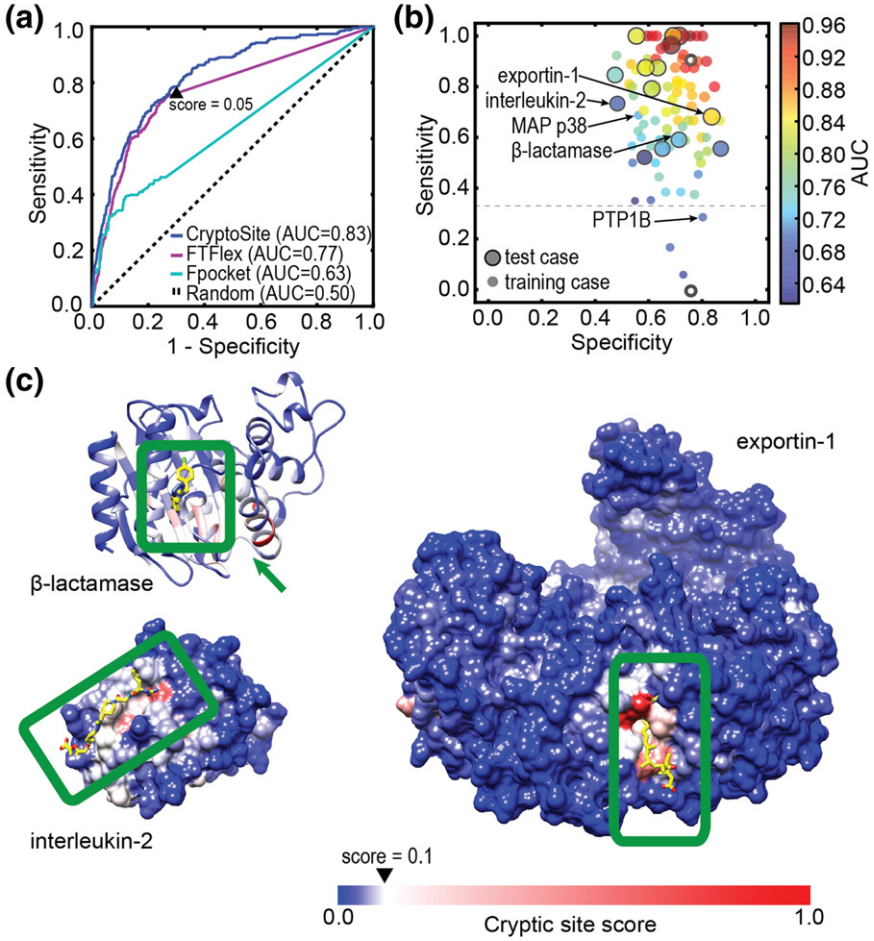

Fig. 2. The accuracies of our predictive model, FTFlex, and Fpocket are measured as the area under the receiver-operating characteristic (ROC) curve based on predictions on all proteins in the test set (a), as well as based on sensitivity (true positive rate) and specificity (true negative rate) values from predictions on individual proteins (b). (a) Only ~45% and ~80% of cryptic site residues were detected by Fpocket and FTFlex, respectively; the area under the ROC curve was calculated by connecting the end of the ROC curve and the upper-right corner as a straight line. The accuracy of CryptoSite is comparable to that of FTFlex when small pockets that could fit small-molecule fragments are already present in the *apo* state of a cryptic site (this is the case in 10 out of 14 testing examples). However, CryptoSite is more accurate than FTFlex when a cryptic site is buried or resides in a large protein (Fig. S8a). (b) Sensitivities and specificities were determined for each protein in our test set (larger data points with black circle) and training set (smaller data points) based on leave-one-out cross-validation. The classification of the residues is based on the score threshold of 0.05. The two empty circles denote two predictions (one failed) of cryptic sites in proteins with more than one cryptic site. (c) The cryptic sites from our dataset are marked by green rectangles, and the computed scores that a residue is in a cryptic site are shown on the blue-to-red color scale. The small molecules that bind into the known cryptic sites are superposed from the alignment to the bound conformations and represented as yellow sticks.

of similar efficiency [25,28], CryptoSite is generally more accurate, particularly when the location of a cryptic site is buried in the *apo* state.

False negatives result from large rearrangements

Next, we analyze false negatives and false positives (defined based on the cryptic sites annotated in MOAD). Our predictive model failed to predict most cryptic sites that undergo large conformational changes and whose pockets are difficult to sample with current molecular dynamics approaches, as well as partial sites that require binding to another protein chain to become functional (Fig. S9). In particular, we failed at predicting the cryptic site for stabilizing substrates (e.g., cyclopiazonic acid) in Ca-ATPase (sensitivity of 6%) that resides at the interface between three domains, two of which are ~50 Å apart in the *apo* conformation (Fig. S9a). Similarly, we also failed at predicting an allosteric site in the thumb site of HCV RNA polymerase (sensitivity of 0%), a site between two chains of kynurenine aminotransferase II (sensitivity of 17%), and an allosteric site in PTP1B (sensitivity of 29%) (Fig. S9). In the future, inadequate

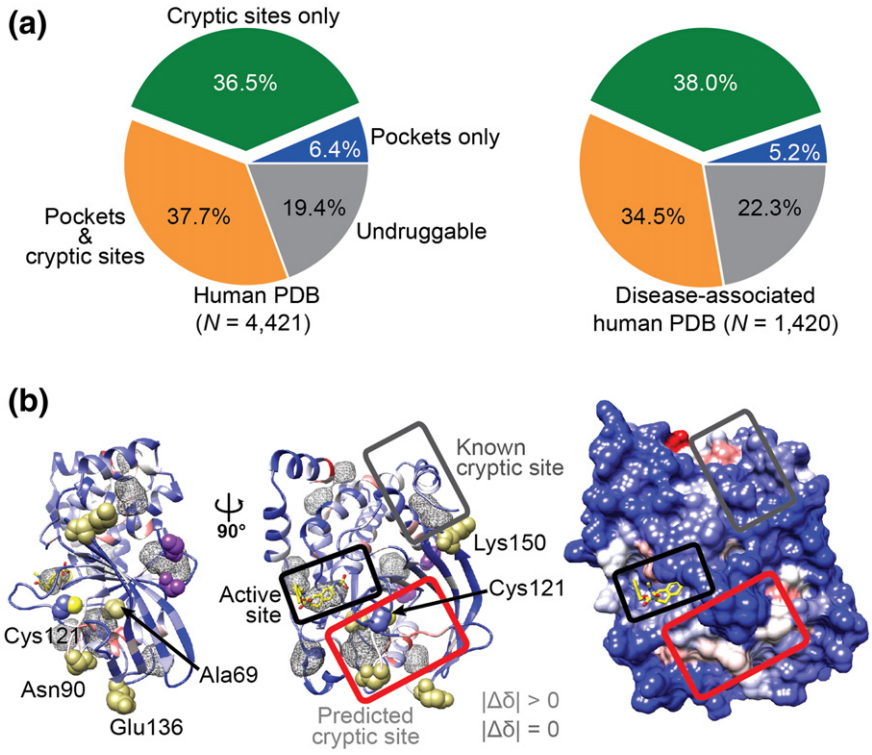

Fig. 3. Cryptic binding sites are predicted to expand the size of the druggable proteome. (a) The percentage of proteins for which no binding sites (gray), only cryptic sites (green), only binding pockets (blue), and both cryptic sites and binding pockets (orange) were predicted for all human proteins with known structure (left pie chart) and for a subset of disease-associated proteins (right pie chart). Shown are the results of the fast version of our predictive model that does not take into account features based on molecular dynamics simulations. (b) Cryptic binding sites in PTP1B. Ribbon (left and center) and surface (right) representations of the PTP1B structure (PDB ID: 2F6V) are colored based on the cryptic site score as in Fig. 2c. Residues with definitive chemical shift changes ($|\Delta\delta|$) upon ABDF labeling (khaki) cluster around the cryptic and ABDF binding sites, whereas residues whose chemical shifts definitively do not change (purple) are more distal. The panel also shows positions and average volumes of the pockets (gray mesh) that are at least partially open more than 50% of the time, as observed in the molecular dynamics simulation at 300 K.

sampling in AllosMod will be addressed by using multiple input structures and/or restraints from experimental data (e.g., small-angle X-ray scattering profiles [50], chemical cross-links [51], hydrogen/deuterium exchange with mass spectrometry, and electron microscopy density maps [52]).

A false positive prediction can be an unknown cryptic site

While it is difficult to be certain that a predicted cryptic site does not bind a ligand, potential false positives include high-scoring isolated residues or terminal regions of truncated proteins, which may not be as flexible in full-length proteins. However, our benchmark probably overestimates the false positive rate because some predicted cryptic sites are in fact true binding sites, even though they are not annotated as such in the MOAD database (e.g., proteins that bind peptides or other proteins). For example, our predictive model identifies the binding site for the light chain of coagulation factor VII in the heavy chain of coagulation factor VII, the binding site for guanine–nucleotide exchange factor DBS in CDC42 protein, the dimer interfaces in fructose-1,6-bisphosphate aldolase and estrogen-related receptor γ, the docking site for its N-terminal motif in Bcl-X_L, and the phosphate binding site in acid-β-glucosidase (Fig. 2c and Fig. S7). Excluding protein–protein interface residues from the prediction of cryptic sites may reduce the number of false positives; however, the improvement appears to be modest, case dependent, and comes at a cost of ignoring cryptic sites that are located at such interfaces (Fig. S5c). In summary, the analysis of successes and failures demonstrates the potential of our approach to guide the experimental identification of new sites in difficult small-molecule targets.

The druggable proteome is significantly larger than estimated previously

Given the overall accuracy of our approach (above), a large number of predicted cryptic sites that are not yet annotated as such in our benchmark

might also indicate that there are many cryptic sites yet to be discovered. If so, our predictive model could facilitate finding novel binding sites in "undruggable" proteins and, hence, expand the druggable proteome space. It has been suggested that the human proteome of approximately 20,000 proteins contains ~3000 proteins associated with disease and ~3000 druggable proteins, with the overlap between the two sets of only ~600–1500 [16,53,54]. To predict how much cryptic binding sites expand the druggable proteome space, we first applied a faster version of our predictive model (based on a subset of features that are not extracted from molecular dynamics simulations, resulting in the speedup factor of 1000 and AUC of 0.74) on 4421 human proteins with at least one domain of known structure (11,201 structures in total). Next, we counted the numbers of cryptic sites and pockets in each structure (SI Text). Pockets were predicted in ~1900 (43%) proteins, and cryptic sites were predicted in ~3300 (74%) proteins. Among the 1420 disease-associated proteins of known structure, 40% have pockets in their crystal structures (in agreement with the previous estimate that the fraction of proteins that are both disease-associated and druggable is 20–50% [55]). In contrast to pockets, cryptic sites were predicted in 72% of the disease-associated proteins, 38% of which have no apparent pockets (Fig. 3a). However, some of the predictions may be false positives (the sites may in fact not bind any ligands). Moreover, for some sites, it may be very difficult to find a ligand (even if it does exist), and even if the ligand is found, it may not be a drug because it does not target the disease-modifying function of a protein or because it does not meet clinical development criteria. Nevertheless, the prediction of cryptic sites on the disease-associated proteins of known structure indicates that small molecules might be used to target significantly more disease-associated proteins than were previously thought druggable.

If cryptic sites are more abundant than previously estimated, why does high-throughput screening not identify them more often than it does? It has been shown that small-molecule libraries are biased toward traditional drug targets, such as G-protein-coupled receptors, ion channels, and kinases, while they are not as suitable for antimicrobial targets and those identified from genomic studies [56]. It is conceivable that the existing libraries are also less suitable for cryptic sites. Moreover, cryptic sites may tend to bind ligands more weakly than binding pockets due to the need to compensate for the free energy of site formation [57] and may thus be ranked lower on the high-throughput screening lists. Therefore, different approaches based on larger and more diverse chemical libraries, including small fragments [20,58,59], peptides, peptidomimetics, and natural products, may be needed for more efficient discovery of cryptic site ligands. A case in point is the discovery of a number of ligands for cryptic allosteric sites and cryptic sites at protein–protein interfaces, such as interleukin-2, caspases, kinase PDK1, and PTP1B, by fragment-based tethering [20–22,59]. Our data suggest that cryptic sites are much more prevalent than previously expected. However, while such sites do provide additional opportunities for drug discovery, they may not ultimately lead to drugs.

Experimental characterization of a predicted cryptic site in PTP1B by NMR spectroscopy

Finally, to demonstrate the practical utility of our approach, we focused on the clinically significant protein PTP1B. Targeting PTP1B with small molecules has been challenging due to the lack of specificity and bioavailability of substrate mimetics, as well as the presence of only a single known allosteric pocket [38,59,60]. In addition to identify 4 of the 14 residues in the known allosteric cryptic site [59], our predictive model also suggested two additional putative cryptic sites, a site near the N-terminus and a site relatively close to the active site (Fig. 3b). The latter site is interesting for several reasons. First, the predicted cryptic site residues form an internal cavity (between residues Ile67 and Phe95) in crystal structures of PTP1B that is large enough to accommodate a small molecule (volume of ~150 $Å^3$). Our molecular dynamics simulations suggest that small conformational changes in the cavity-forming loops could make the cavity accessible to the solvent and expand its size (up to 430 $Å^3$). Second, the site is in proximity of two cysteine residues, Cys92 and Cys121, that could be targeted covalently in small-molecule fragment screening by tethering [20]. In fact, Cys 121 is an already known target of a covalent small-molecule modifier and an allosteric inhibitor of PTP1B, ABDF, but its mechanism of action remains unclear [61]. Third, this cryptic site in PTP1B differs from the corresponding region in the closely related tyrosine-protein phosphatase non-receptor type 2 (TCPTP) at, for example, position 97 (glutamate instead of leucine). This difference between PTP homologs could be exploited to develop selective inhibitors that avoid the serious adverse effects associated with TCPTP inhibition in mice [59]. Finally, the cryptic site may be allosterically coupled to the catalytic site; examining contacts between pairs of residues [35] suggests extensive coupling between the cryptic and catalytic sites (Fig. S10a).

We experimentally studied the binding of ABDF to PTP1B to determine whether or not it involves the putative cryptic site. Although PTP1B has three other surface-exposed cysteine residues, ABDF covalently attaches specifically to the side chain of Cys121 (Fig. 3b and Fig. S10b). The Cys121 side chain points toward the interior of the unlabeled protein; thus, binding of ABDF likely requires a conformational change in the protein. We were unable to obtain a crystal structure of ABDF-labeled

PTP1B, in agreement with other reports that ABDF-labeled PTP1B, unlike *apo* PTP1B, is recalcitrant to crystallization [61]. To determine whether or not the covalent label causes specific local conformational changes or globally perturbs the protein, we collected $^{1}H, ^{15}N$ transverse relaxation optimized spectroscopy heteronuclear single quantum coherence NMR spectra of both *apo* and ABDF-labeled protein (SI Text and Fig. S10c–f). Using previously published backbone resonance assignments [62], we observed no perturbation of chemical shifts for a number of residues distal to the predicted cryptic site, indicating that the effects are local and that the protein remains folded. In contrast, a cluster of residues nearby the predicted cryptic site were significantly perturbed (Fig. 3b and Fig. S10c–f). Many other residues near the predicted cryptic site that would need to move for ligand binding, including the adjacent β-sheet and Cys121 loop, were unassigned due to resonance broadening, which is indicative of conformational exchange. Collectively, these results point to structural flexibility in the vicinity of the predicted cryptic site and the specific perturbation of residues surrounding the predicted binding pocket, validating our prediction.

To conclude, we describe cryptic sites and a method that accurately, automatically, and efficiently predicts their locations in protein structures. Our results support the hypothesis of ubiquitous cryptic sites and suggest many new small-molecule protein targets, including those that are associated with diseases. Moreover, we illustrate how chemical tethering can be used to validate cryptic site predictions by discovering cryptic site ligands. Cryptic sites can also be characterized by experimental techniques that measure protein dynamics, such as NMR spectroscopy and room-temperature X-ray crystallography [63,64], as well as by discovery of ligands through virtual screening against conformations with pockets computed by AllosMod or molecular dynamics simulations. Our approach provides a convenient first step for such characterizations.

Materials and Methods

We started by finding cryptic sites in the Protein Data Bank (PDB) [65,66], as follows. First, we gathered structures of protein–ligand complexes and structures of proteins in ligand-free (unbound) conformations. We define binding residues as the residues with at least one atom within 5 Å from any atom of a ligand in the bound conformation (a binding site). Second, we removed the redundant protein occurrences in the dataset by applying sequence identity threshold of 40% (SI Text). Finally, we evaluated each binding site in the unbound conformation using pocket scores based on two pocket-detection algorithms, Fpocket and ConCavity [13,14]. Binding sites with bad pocket sores in the unbound conformation and good pocket scores in the bound conformation were defined as cryptic sites, whereas those with good pocket scores in both conformations were defined as binding pockets (Tables S1 and S5). More details and methods are available in SI Text†.

Acknowledgments

The authors thank Hao Fan, Marcus Fischer, Nir London, Avner Schlessinger, and other members of Sali laboratory for their comments and feedback. P.C. is supported by a Howard Hughes Predoctoral Fellowship; T.J.R. is supported by a predoctoral fellowship from the National Institutes of Health (F31 CA180378) and the Krevans Fellowship; L.B. is supported by Bayer Science and Education Foundation; D.A.K. is supported by an A. P. Giannini Foundation Postdoctoral Research Fellowship; R.A.W. is supported by a National Science Foundation Graduate Research Fellowship; J.S.F. is supported by the National Institutes of Health (DP5 OD009180, R21 GM110580, and P30 DK063720) and National Science Foundation (STC-1231306); A.S. is supported by the National Institutes of Health (R01 GM083960, U54 RR022220, U54 GM094662, P01 AI091575, and U01 GM098256).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2016.01.029.

Received 24 June 2015;
Received in revised form 29 January 2016;
Accepted 30 January 2016
Available online 5 February 2016

Keywords:
cryptic binding sites;
protein dynamics;
undruggable proteins;
machine learning

†The Web server for predicting cryptic binding sites is available at http://salilab.org/cryptosite

Abbreviations used:
PTP1B, protein tyrosine phosphatase 1B; SVM, support vector machine; AUC, area under the curve.

References

[1] B. Nisius, F. Sha, H. Gohlke, Structure-based computational analysis of protein binding sites for function and druggability prediction, J. Biotechnol. 159 (3) (2012) 123–134.

[2] S.J. Campbell, N.D. Gold, R.M. Jackson, D.R. Westhead, Ligand binding: Functional site location, similarity and docking, Curr. Opin. Struct. Biol. 13 (3) (2003) 389–395.
[3] A.T. Laurie, R.M. Jackson, Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites, Bioinformatics 21 (9) (2005) 1908–1916.
[4] J.C. Hermann, et al., Structure-based activity prediction for an enzyme of unknown function, Nature 448 (7155) (2007) 775–779.
[5] R.A. Laskowski, N.M. Luscombe, M.B. Swindells, J.M. Thornton, Protein clefts in molecular recognition and function, Protein Sci. Publ. Protein Soc. 5 (12) (1996) 2438–2452.
[6] G.R. Bowman, P.L. Geissler, Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites, Proc. Natl. Acad. Sci. U. S. A. 109 (29) (2012) 11681–11686.
[7] R. Diskin, D. Engelberg, O. Livnah, A novel lipid binding site formed by the MAP kinase insert in p38 alpha, J. Mol. Biol. 375 (1) (2008) 70–79.
[8] J.D. Durrant, J.A. McCammon, Molecular dynamics simulations and drug discovery, BMC Biol. 9 (2011) 71.
[9] J.R. Horn, B.K. Shoichet, Allosteric inhibition through core disruption, J. Mol. Biol. 336 (5) (2004) 1283–1291.
[10] K.W. Lexa, H.A. Carlson, Full protein flexibility is essential for proper hot-spot mapping, J. Am. Chem. Soc. 133 (2) (2011) 200–202.
[11] S. Mitternacht, I.N. Berezovsky, Binding leverage as a molecular basis for allosteric regulation, PLoS Comput. Biol. 7 (9) (2011), e1002148.
[12] S. Henrich, et al., Computational approaches to identifying and characterizing protein binding sites for ligand design, J. Mol. Recognit.: JMR 23 (2) (2010) 209–219.
[13] J.A. Capra, R.A. Laskowski, J.M. Thornton, M. Singh, T.A. Funkhouser, Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure, PLoS Comput. Biol. 5 (12) (2009), e1000585.
[14] V. Le Guilloux, P. Schmidtke, P. Tuffery, Fpocket: An open source platform for ligand pocket detection, BMC bioinformatics 10 (2009) 168.
[15] A. Rossi, M.A. Marti-Renom, A. Sali, Localization of binding sites in protein structures by optimization of a composite scoring function, Protein Sci. Publ. Protein Soc. 15 (10) (2006) 2366–2380.
[16] A.L. Hopkins, C.R. Groom, The druggable genome, Nat. Rev. Drug Discov. 1 (9) (2002) 727–730.
[17] R.P. Sheridan, V.N. Maiorov, M.K. Holloway, W.D. Cornell, Y.D. Gao, Drug-like density: A method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank, J. Chem. Inf. Model. 50 (11) (2010) 2029–2040.
[18] M.R. Arkin, J.A. Wells, Small-molecule inhibitors of protein–protein interactions: Progressing towards the dream, Nature reviews. Drug discovery 3 (4) (2004) 301–317.
[19] J.A. Wells, C.L. McClendon, Reaching for high-hanging fruit in drug discovery at protein–protein interfaces, Nature 450 (7172) (2007) 1001–1009.
[20] J.A. Hardy, J.A. Wells, Searching for new allosteric sites in enzymes, Curr. Opin. Struct. Biol. 14 (6) (2004) 706–715.
[21] J.M. Ostrem, U. Peters, M.L. Sos, J.A. Wells, K.M. Shokat, K-ras(G12C) inhibitors allosterically control GTP affinity and effector interactions, Nature 503 (7477) (2013) 548–551.
[22] J.D. Sadowsky, et al., Turning a protein kinase on or off from a single allosteric site via disulfide trapping, Proc. Natl. Acad. Sci. U. S. A. 108 (15) (2011) 6056–6061.
[23] D.K. Johnson, J. Karanicolas, Druggable protein interaction sites are more predisposed to surface pocket formation than the rest of the protein surface, PLoS Comput. Biol. 9 (3) (2013), e1002951.
[24] G.R. Bowman, E.R. Bolin, K.M. Hart, B.C. Maguire, S. Marqusee, Discovery of multiple hidden allosteric sites by combining Markov state models and experiments, Proc. Natl. Acad. Sci. U. S. A. 112 (9) (2015) 2734–2739.
[25] K.A. Loving, A. Lin, A.C. Cheng, Structure-based druggability assessment of the mammalian structural proteome with inclusion of light protein flexibility, PLoS Comput. Biol. 10 (7) (2014), e1003741.
[26] A. Bakan, N. Nevins, A.S. Lakdawala, I. Bahar, Druggability assessment of allosteric proteins by dynamics simulations in the presence of probe molecules, J. Chem. Theory Comput. 8 (7) (2012) 2435–2447.
[27] R. Brenke, et al., Fragment-based identification of druggable "hot spots" of proteins using Fourier domain correlation techniques, Bioinformatics 25 (5) (2009) 621–627.
[28] L.E. Grove, D.R. Hall, D. Beglov, S. Vajda, D. Kozakov, FTFlex: Accounting for binding site flexibility to improve fragment-based identification of druggable hot spots, Bioinformatics 29 (9) (2013) 1218–1219.
[29] D. Kozakov, et al., The FTMap family of Web servers for determining and characterizing ligand-binding hot spots of proteins, Nat. Protoc. 10 (5) (2015) 733–755.
[30] C.H. Ngan, et al., FTMAP: Extended protein mapping with user-selected probe molecules, Nucleic Acids Res. 40 (2012) W271–W275 (Web Server issue).
[31] F.C. Bernstein, et al., The Protein Data Bank: A computer-based archival file for macromolecular structures, J. Mol. Biol. 112 (3) (1977) 535–542.
[32] M.L. Benson, et al., Binding MOAD, a high-quality protein-ligand database, Nucleic Acids Res. 36 (Database issue) (2008) D674–D678.
[33] O.N. Demerdash, M.D. Daily, J.C. Mitchell, Structure-based predictive models for allosteric hot spots, PLoS Comput. Biol. 5 (10) (2009), e1000531.
[34] X. Zhu, J.C. Mitchell, KFC2: A knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features, Proteins 79 (9) (2011) 2671–2683.
[35] P. Weinkam, Y.C. Chen, J. Pons, A. Sali, Impact of mutations on the allosteric conformational equilibrium, J. Mol. Biol. 425 (3) (2013) 647–661.
[36] F. Pedregosa, et al., Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[37] T. Schaul, et al., PyBrain, J. Mach. Learn. Res. 11 (2010) 743–746.
[38] A.P. Combs, Recent advances in the discovery of competitive protein tyrosine phosphatase 1B inhibitors for the treatment of diabetes, obesity, and cancer, J. Med. Chem. 53 (6) (2010) 2333–2344.
[39] N.M. O'Boyle, et al., Open Babel: An open chemical toolbox, J. Cheminformatics 3 (2011) 33.
[40] K. Gunasekaran, B. Ma, R. Nussinov, Is allostery an intrinsic property of all dynamic proteins? Proteins 57 (3) (2004) 433–443.
[41] Z. Huang, et al., ASD: A comprehensive database of allosteric proteins and modulators, Nucleic Acids Res. 39 (Database issue) (2011) D663–D669.
[42] A.E. Jenkins, J.A. Hockenberry, T. Nguyen, D.A. Bernlohr, Testing of the portal hypothesis: Analysis of a V32G, F57G, K58G mutant of the fatty acid binding protein of the murine adipocyte, Biochemistry 41 (6) (2002) 2022–2027.

[43] A. Schlessinger, B. Rost, Protein flexibility and rigidity predicted from sequence, Proteins 61 (1) (2005) 115–126.
[44] C.H. Shih, C.M. Chang, Y.S. Lin, W.C. Lo, J.K. Hwang, Evolutionary information hidden in a single protein structure, Proteins 80 (6) (2012) 1647–1657.
[45] L.S. Swapna, R.M. Bhaskara, J. Sharma, N. Srinivasan, Roles of residues in the interface of transient protein–protein complexes before complexation, Sci. Rep. 2 (2012) 334.
[46] I. Bartova, J. Koca, M. Otyepka, Functional flexibility of human cyclin-dependent kinase-2 and its evolutionary conservation, Protein Sci. Publ. Protein Soc. 17 (1) (2008) 22–33.
[47] D. Martens, J. Huysmans, R. Setiono, J. Vanthienen, B. Baesens, Rule extraction from Support Vector Machines: An overview of issues and application in credit scoring, Stud. Comput. Intell. 80 (2008) 33–63.
[48] S.P. Brown, P.J. Hajduk, Effects of conformational dynamics on predicted protein druggability, ChemMedChem 1 (1) (2006) 70–72.
[49] D. Kozakov, et al., Structural conservation of druggable hot spots in protein–protein interfaces, Proc. Natl. Acad. Sci. U. S. A. 108 (33) (2011) 13528–13533.
[50] P. Weinkam, et al., (TBD) Mapping protein allosteric mechanisms with small angle X-ray scattering profiles, 2016 submitted for publication.
[51] K.S. Molnar, et al., Cys-scanning disulfide crosslinking and Bayesian modeling probe the transmembrane signaling mechanism of the histine kinase, PhoQ, 2014 submitted for publication.
[52] M. Liao, E. Cao, D. Julius, Y. Cheng, Structure of the TRPV1 ion channel determined by electron cryo-microscopy, Nature 504 (7478) (2013) 107–112.
[53] J.P. Overington, B. Al-Lazikani, A.L. Hopkins, How many drug targets are there? Nat. Rev. Drug Discov. 5 (12) (2006) 993–996.
[54] A.P. Russ, S. Lampel, The druggable genome: An update, Drug Discov. Today 10 (23–24) (2005) 1607–1610.
[55] P. Schmidtke, X. Barril, Understanding and predicting druggability. A high-throughput method for detection of drug binding sites, J. Med. Chem. 53 (15) (2010) 5858–5867.
[56] J. Hert, J.J. Irwin, C. Laggner, M.J. Keiser, B.K. Shoichet, Quantifying biogenic bias in screening libraries, Nat. Chem. Biol. 5 (7) (2009) 479–483.
[57] D.L. Mobley, K.A. Dill, Binding of small-molecule ligands to proteins: "What you see" is not always "what you get", Structure 17 (4) (2009) 489–498.
[58] L.N. Makley, J.E. Gestwicki, Expanding the number of "druggable" targets: Non-enzymes and protein–protein interactions, Chem. Biol. Drug Des. 81 (1) (2013) 22–32.
[59] C. Wiesmann, et al., Allosteric inhibition of protein tyrosine phosphatase 1B, Nat. Struct. Mol. Biol. 11 (8) (2004) 730–737.
[60] N. Krishnan, et al., Targeting the disordered C terminus of PTP1B with an allosteric inhibitor, Nat. Chem. Biol. 10 (7) (2014) 558–566.
[61] S.K. Hansen, et al., Allosteric inhibition of PTP1B activity by selective modification of a non-active site cysteine residue, Biochemistry 44 (21) (2005) 7704–7712.
[62] S. Meier, et al., Backbone resonance assignment of the 298 amino acid catalytic domain of protein tyrosine phosphatase 1B (PTP1B), J. Biomol. NMR 24 (2) (2002) 165–166.
[63] J.S. Fraser, et al., Accessing protein conformational ensembles using room-temperature X-ray crystallography, Proc. Natl. Acad. Sci. U. S. A. 108 (39) (2011) 16247–16252.
[64] M. Fischer, B.K. Shoichet, J.S. Fraser, One crystal, two temperatures: Cryocooling penalties alter ligand binding to transient protein sites, Chembiochem Euro. J. Chem. Biol. (2015).
[65] H.M. Berman, et al., The Protein Data Bank, Acta Crystallogr. D Biol. Crystallogr. 58 (Pt 6 No 1) (2002) 899–907.
[66] H.M. Berman, et al., The Protein Data Bank, Nucleic Acids Res. 28 (1) (2000) 235–242.
[67] E.F. Pettersen, et al., UCSF Chimera—A visualization system for exploratory research and analysis, J. Comput. Chem. 25 (13) (2004) 1605–1612.