A critical comparative assessment of predictions of protein binding sites for biologically relevant organic compounds

Ke Chen[1], Marcin J. Mizianty[1], Jianzhao Gao[2] and Lukasz Kurgan[1*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, CANADA T6G 2V4

[2]College of Mathematics, Nankai University, Tianjin, PR CHINA 300071

*Correspondence should be addressed to Lukasz Kurgan

Department of Electrical and Computer Engineering, ECERF, 9107 116 Street, University of Alberta, Edmonton, AB, Canada T6G 2V4

Email lkurgan@ece.ualberta.ca

Phone (780) 492-5488

Fax (780) 492-1811

Running title: critical assessment of binding site predictors

Summary

Protein function annotation and rational drug discovery rely on the knowledge of binding sites for small organic compounds, and yet the quality of existing binding site predictors was never systematically evaluated. We assess predictions of ten representative geometry-, energy-, threading-, and consensus-based methods on a new benchmark dataset that considers apo and holo protein structures with multiple binding sites for biologically relevant ligands. Statistical tests show that threading-based Findsite outperforms other predictors when its templates have high similarity with the input protein. However, Findsite is equivalent or inferior to some geometry-, energy- and consensus-based methods when the similarity is lower. We demonstrate that geometry-, energy- and consensus-based predictors benefit from the usage of the holo structures and that the top four methods, Findsite, Q-SiteFinder, ConCavity, and MetaPocket, perform better for larger binding sites. Predictions from these four methods are complementary and our simple meta-predictor improves over the best single predictor.

Highlights

- Best-performing predictors accurately find important putative binding sites
- Geometrical and energy-based predictors outperform threading methods for novel folds
- Combining different types of predictors leads to significant improvements
- Quality of predictions is strongly positively correlated with size of binding sites

Introduction

Although the genomes of more than 1000 organisms have been sequenced (Adams et al., 2000) and the ongoing structural genomics efforts (Chandonia and Brenner, 2006) accelerate the determination of protein structures, the biological functions of many identified gene products are largely unknown. The rapid accumulation of protein sequences and structures motivates the development of computational tools for identification of protein's functions. Proteins carry out their functions through interactions with other molecules, including nucleic acids, other proteins, nucleotides, peptides, etc. These interactions are highly ubiquitous which is reflected in the Gene Ontology (GO) database that annotates nearly 1400 types of interactions (Ashburner et al., 2000). In the past two decades the rules that govern protein-protein (Jones and Thornton, 1996; Zhu et al., 2008), protein-DNA (Luscombe et al., 2001), protein-RNA (Ellis et al., 2007; Zhang et al., 2010) and protein-small ligand interactions (Chen and Kurgan, 2009) have been systematically investigated. Dozens of computational methods have been developed for the prediction of DNA and RNA binding sites, protein-protein interaction sites and binding sites for small-ligand (Jones and Thornton, 2004; López et al., 2009). Interactions with small organic compounds (organic molecules with less than 100 non-hydrogen atoms) are of particular interest because they find applications in elucidation of mechanism of numerous cellular activities, such as cellular signaling, growth of neurons and regulation of cell cycles (Whittard et al., 2006; Popova et al., 2010; Mukherjee et al., 2010). The small organic compounds constitute nearly 90% of the drugs approved by

the U.S. Food and Drug Administration (Wishart et al., 2008). Consequently, the knowledge of binding sites of these molecules plays a crucial role for molecular docking-based rational drug discovery (Brooijmans and Kuntz, 2003). The two most recent published CASP experiments, CASP7 and CASP8 (López et al., 2009), included evaluation of sequence-based predictors of residues that bind small ligands. Here we concentrate on evaluation of methods that use protein structure as their input. For convenience, we use ‘ligands’ and ‘binding sites’ to refer to the small organic compounds and the sites on the protein structure where they bind, respectively. Although some studies that introduced new binding site predictors have compared them to a few existing solutions, to date no systematic comparison between a comprehensive/representative set of methods was performed. Another drawback of prior comparative studies is that they consider datasets that are characterized by largely incomplete annotation of binding sites. Every protein in these datasets is usually annotated with a single binding site while in reality many of these proteins have multiple binding sites. Additionally, prior benchmark datasets include annotations of biologically “irrelevant” ligands, such as the glycol molecule that is introduced by the purification and crystallization procedures. We perform a comparative evaluation of the predictive quality of ten representative binding site predictors on a set of proteins that are annotated with multiple binding sites, which are confirmed to be biologically relevant. We selected prediction methods that offer either a web-server or a standalone program to generate the predictions. Overall, the structure-based predictors utilize three types of approaches including geometrical

analysis, calculation of binding energy, and threading using structural templates; one solution is based on a consensus of geometry- and energy-based approaches. The considered methods include the geometry-based SURFNET (Laskowski, 1995), PocketFinder (Hendlich et al., 1997), PASS (Brady and Stouten, 2000), LIGSITE^{csc} (Huang and Schroeder, 2006), PocketPicker (Weisel et al., 2007), ConCavity (Capra et al., 2009), and Fpocket (Le et al., 2009), the energy-based Q-SiteFinder (Laurie and Jackson, 2005), the threading-based Findsite (Skolnick and Brylinski, 2008), and the consensus-based MetaPocket (Huang, 2009). We evaluate their predictive quality using two criteria that were introduced in prior works and a new quality index that gives additional insights. Besides the overall predictive quality, we also assess the impact of ligand size and ligand type and we compare predictions from apo (ligand-unbound) and holo (ligand-bound) structures.

Results

We evaluate the performance of the ten prediction methods on a non-redundant benchmark dataset of 251 proteins; details are given in the Experimental Procedures section. These methods are also compared against a baseline predictor which randomly selects a surface patch on the target protein; the center of the patch is used as the prediction. Prior studies usually take top three or top five predictions and verify whether any of them are within a certain distance (which is used as a cutoff for calculation of prediction accuracy) to the actual binding site. If at least one of the top predictions is below the cutoff, then the binding site is assumed to be correctly

predicted. Since the previously used benchmark datasets contain proteins that are annotated with one binding site, the number of correctly predicted sites equals to the number of proteins and predictions were assessed "per protein". In our case majority of the proteins are annotated with multiple binding sites, and thus our assessment is "per binding site". For a protein with *n* binding sites we take the top *n* predictions for every considered method. A given binding site is correctly predicted if the minimal distance between this site and any of the *n* predictions from a given method is below a threshold *D*. The success rate is defined as the number of correctly predicted binding sites divided by the total number of sites.

Comparison of the overall prediction quality

The success rates of the ten methods and the random baseline predictor quantified using D_{CC}, which measures the distance from the center of the predicted site to the center of the ligand (details are given in the Experimental Procedures section), are shown in Figure 1A. Findsite outperforms all other considered predictors for thresholds *D* up to 10Å. The ConCavity achieves the "second-best" success rates and is chosen to represent the geometry-based approaches in the subsequent analysis. Several predictors, including Q-SiteFinder, MetaPocket and PocketPicker have comparable, "third-best" success rates. The SURFNET, which is the oldest method that was designed over a decade ago, has the lowest success rates but it still improves over the baseline. Supplementary Figure 1A online summarizes the success rates measured using D_{CA}, which is based on the distance from the center of the predicted

site to any atom of the ligand. The results are similar to the results obtained using D_{CC}, except for D = 1Å where the Q-SiteFinder is the top-performing predictor. For the cutoff D = 4Å, which was suggested by Skolnick and colleagues (Skolnick and Brylinski, 2008), the threading-based Findsite successfully predicts around 57% and 68% of the binding sites for the D_{CC} and D_{CA} distance definition, respectively, the geometry-based ConCavity identifies 28% and 51% of the binding sites, the energy-based Q-SiteFinder finds 26% and 44% of the binding sites, and the remaining methods cover 11-25% and 31-45% of the binding sites, respectively. To compare, the baseline random predictor correctly finds 5% and 9% of the binding sites when considering D_{CC} and D_{CA} distances, respectively.

The overlap index O_{PL}, which is defined as the ratio between the volume of the intersection of the predicted binding site and the ligand, and the union of the two volumes, expresses normalized spatial overlap between the predicted and the actual location of the ligand. This index is arguably more precise than the distance based indices since it considers spatial orientation. The O_{PL} could be calculated only for the ConCavity, Q-SiteFinder, PocketPicker and PocketFinder which generate a full set of grid points of the predicted pocket instead of just the center of the pocket that is outputted by the remaining predictors. We observe that about 60% of the binding sites predicted by ConCavity overlap with the predicted pocket while the coverage is only around 40% for Q-SiteFinder and PocketPicker; see Supplementary Figure 1B. However, in most cases, the overlapping volume measured using O_{PL} is rather small;

for instance, O_{PL} is above 20% only for about 16% of the binding sites.

We investigate significance of differences in the prediction quality measured with D_{CC} for all pairs of the considered prediction methods (details are given in the Experimental Procedures section); see Table 1. The Findsite is significantly better than all other methods. The ConCavity, Q-SiteFinder, MetaPocket and PocketPicker are second-best and not significantly different between each other (except for the ConCavity which significantly improves over the Q-SiteFinder), and this group is statistically better than LIGSITE^{csc}, SURFNET, PASS, PocketFinder and Fpocket.

The impact of structural similarity between query protein and template library

Findsite is a threading-based method that utilizes a library of template structures. Its predictions are generated by clustering binding sites of the template structures and they rely on the availability of templates that are similar to the predicted protein. To study the impact of the availability of similar templates, we use a threshold to limit the structural similarity between the query proteins and the templates used for the prediction. Only the template structures with a similarity score below the threshold are utilized. The structural similarity is measured with TM-score, which varies between 0 and 1 (Zhang and Skolnick, 2005); larger values indicate higher similarity. We vary the threshold between 0.5 and 1 with step of 0.1. Findsite also utilizes a default cut-off TM-score = 0.4 below which a given template is rejected. In case if Findsite does not find a suitable template above the 0.4 cut-off, we lower it by 0.1

until a template is found. This is done to assure that Findsite provides predictions for all targets in our benchmark dataset, even if in some cases only low quality templates are available; the other considered methods also provide predictions for all targets. We use the baseline predictions when Findsite cannot find a template for the cut-off TM-score = 0.1.

We compare Findsite with the Q-SiteFinder, which is the only energy-based method, the ConCavity, a representative (best-performing) geometry-based method, and with the MetaPocket that represents the consensus-based approaches. The success rates of these four methods are quantified using D_{CC}. For Findsite, we generate six sets of predictions that correspond to the consecutive values, between 0.5 and 1, of the maximal similarity threshold. The MetaPocket, ConCavity, and Q-SiteFinder do not utilize templates thereby they have one set of predictions. Figure 1B reveals that the predictive quality of Findsite improves with the increase of the similarity threshold. At a D = 4Å, its success rates equal 16%, 25%, 34%, 37%, 43% and 57% for the maximal TM-score threshold of 0.5, 0.6, 0.7, 0.8, 0.9, and 1, respectively. This indicates that the predictive quality of Findsite is largely dependent on the availability of structurally similar templates. We investigate significance of differences in the prediction quality measured with D_{CC} between the predictions generated by the four methods. Findsite is significantly better than the other three methods when the similarity threshold is 0.7 or higher, comparable to the other three methods when the threshold is set to 0.6, and significantly inferior for the threshold equal 0.5. These

results suggest that if Findsite identifies a template that shares a TM-score ≥ 0.7 with the query protein, then its predictions are expected to be better than the predictions of the ConCavity, MetaPocket and Q-SiteFinder. On the other hand, if the maximal TM-score between the Findsite's templates and the query protein ≤ 0.5, then the predictions generated by the ConCavity, MetaPocket or Q-SiteFinder are likely to be better.

We repeat the above evaluation when excluding the proteins for which Findsite cannot find a template with TM-score > 0.1; these proteins are removed from the benchmark set instead of being predicted using the baseline approach. We generated six subset of the original benchmark set for different maximal similarity thresholds, which are used to evaluate the four predictors. The results, which are summarized in Supplementary Figure 1C, are consistent with the above analysis. The predictive quality of Findsite depends on the availability of similar templates and the other three methods provide predictions with quality that, as expected, does not depend on the similarity threshold.

Comparison of prediction quality between Apo and Holo structures

The benchmark dataset consists of holo structures, i.e., structures that are bound to ligands. Since the protein-ligand interactions may lead to conformational changes at the vicinity of the binding site, we also investigate the binding site predictions performed on the apo structures, i.e., unbound-state proteins. We selected a subset of proteins, for which both apo structures and holo structures are known, from the

benchmark dataset. This results in two datasets, D_{Apo} that includes 104 of these apo structures and D_{Holo} that includes the corresponding set of the 104 holo structures (a subset of the benchmark dataset).

We assess predictions generated for the four representative methods, the threading-based Findsite, the energy-based Q-SiteFinder, the consensus-based MetaPocket, and the best performing geometry-based ConCavity on both datasets; see Figure 1C. Using the D_{CC} distance, the success rates of Findsite on the D_{Holo} dataset is on average (over different thresholds) about 1.6% higher than on the D_{Apo} dataset. For the MetaPocket, Q-SiteFinder, and ConCavity the success rates on the D_{Holo} dataset are on average 6.7%, 6.2%, and 7.3% higher than on the ligand-unbound dataset. Similar trends are observed when using the D_{CA}, see Supplementary Figure 1D. Specifically, Findsite, MetaPocket, Q-SiteFinder, and ConCavity achieve 1.1%, 6.7%, 7.5%, and 6.9% better success rates on the D_{Holo} dataset, respectively. The significance of the differences in the predictive quality between the D_{Holo} and D_{Apo} datasets was calculated using the Wilcoxon signed-rank test. The test reveals that MetaPocket, Q-SiteFinder, and ConCavity achieve significantly better predictions with $p < 0.01$, p<0.01, and $p < 0.05$, respectively, on the D_{Holo} dataset when compared with the D_{Apo} dataset, while Findsite achieves comparable results on both datasets. These results suggest that the geometry-, energy-, and consensus-based methods benefit from the usage of the holo structures, likely because the geometrical descriptors and the energy function can be calculated more accurately using these

structures.

Impact of the size of the binding sites

We assessed the impact of the size of the binding sites on the predictive quality. The size is approximated by the number of interacting atoms in the binding site. A non-hydrogen atom of a residue is considered as an interacting atom if it is within 3.9Å to a non-hydrogen atom of the ligand (Luscombe et al., 2001). The binding sites that are sorted by their sizes are divided into five subsets with equal number of sites. The success rates of the four representative predictors are calculated using these five subsets. Using D_{CC}, we observe a consistent trend that higher success rates are achieved for the larger binding sites, see Supplementary Figure 2. For instance, the average success rates for Findsite are 23%, 35%, 45%, 57% and 69% for the consecutive five subsets, respectively, when considering cutoff distances D between 1Å and 5Å. Similarly, the average success rates for Q-SiteFinder, MetaPocket, and ConCavity equal 3%, 4%, and 5%; 14%, 12%, and 11%; 22%, 18%, and 22%; 26%, 26%, and 24%; and 33%, 28%, and 34% on the five subsets, respectively. The Pearson correlations between the average success rates, over cutoff distances D between 1Å and 5Å, and the average size of the binding sites in each of the five subsets, see Figure 2, equal 0.98 for Q-SiteFinder and MetaPocket and 0.99 for Findsite and ConCavity. This shows that the predictive quality of these four methods is linearly correlated with the size of the binding sites. We measure the ratio between the solvent accessible area of the binding residues, computed with the DSSP program

(Kabsch and Sander, 1983), and the protein surface, i.e., the sum of the solvent accessible area of all residues, for each protein. The average ratios in each the five subsets are similar and they vary between 0.085 and 0.105. This shows that the improved success rates are not due to the larger binding areas, but rather due to inherent characteristics of these predictive models.

Predictive quality for different ligand groups

The benchmark dataset includes 475 biologically relevant (as defined in the Experimental Procedures section) ligands that are categorized into 253 types. We manually inspected the ligands that occur in the dataset at least 3 times and we grouped them into four categories, including acids, carbohydrates, mononucleotides and cofactors (excluding mononucleotides). The breakdown of the ligand types in each category is given in Supplementary Table 1. These ligands occur 219 times in the benchmark dataset and they cover 46% of all binding sites; see Supplementary Figure 3A. The remaining ligands are more unique and could not be clustered into sets that would allow for a statistically sound evaluation of the predictive quality. We compare the success rates of the four representative prediction methods on the four ligand categories. Using the D_{CC} measure, the Findsite and ConCavity achieve the highest success rates for the cofactors, followed by the mononucleotides and acids, and the lower accuracies for the carbohydrates; see Supplementary Figure 3B and 3D. These differences are quite substantial, e.g. at D = 4 Å the success rates for cofactors and carbohydrates differ by 50%. In contrast, the differences between the success rates for

different ligand groups for the Q-SiteFinder and MetaPocket are relatively minor; see Supplementary Figures 3F and 3H. Similar trends are observed when using D_{CA}; see Supplementary Figures 3C, 3E, 3G, and 3I. The above suggests that the predictions generated by Q-SiteFinder and MetaPocket are not sensitive to the ligand types, while the predictive quality of Findsite and ConCavity varies relatively widely between different ligand groups.

Complementarity of predictors

The four representative methods are based on different approaches, i.e., Findsite uses threading, Q-SiteFinder is based on the energy calculations, ConCavity utilizes geometrical descriptors, and MetaPocket combines geometrical descriptors and energy calculations. We investigate whether these differences result in complementarity in their predictions. A given binding site is regarded as covered by a combination of several methods if it is correctly predicted by any of these methods. Figure 1D demonstrates that combining predictions of the best performing Findsite with the other three methods results in a larger coverage. For the thresholds D between 1Å and 10Å, the coverage when using the four methods together increases between 4% and 10% when compared with the predictions of the Findsite. For the cutoff distance D = 4Å, 7% of binding sites that are not captured by the Findsite are successfully predicted by the Q-SiteFinder and 10% of the sites that are missed by the Findsite are correctly predicted by one of the three other methods. This shows that the four methods are complementary, which implies that they could be combined to build

a consensus-based method.

We developed a simple consensus predictor by re-ranking the predictions generated by Findsite using the predictions from Q-SiteFinder, ConCavity, and MetaPocket. This solution, in contrast to a straightforward merging of the predictions from the three methods, is motivated by overall high predictive quality of Findsite, when compared to the runner-up approaches. Moreover, we observe that for a protein with *n* binding sites, Findsite sometimes generates more than *n* predictions and some of the correct predictions are not ranked among the top *n* outputs. Predictions generated by Findsite are scored by comparing them to the predictions generated by the other three methods to improve the ranking. A Findsite's prediction receives score of 3 if it is within 4Å to the predictions from Q-SiteFinder, MetaPocket, and ConCavity. The score equals 2 if the Findsite's prediction is within 4Å to the predictions of the two other methods. The score of 1 corresponds to the case when the Findsite's prediction is within 4Å to a prediction from one of the other three methods, and the score equals 0 if the other three methods did not generate predictions within the 4Å radius. The predictions are sorted in the descending order by their scores, and ties are resolved by using the original order of the predictions from Findsite. The dashed line in Figure 1D reveals that the re-ranking improves the success rates of the original Findsite. When considering the cutoff distances *D* between 1Å and 5Å, the re-ranked predictions improve over the original Findsite on average by 2%. Although the magnitude of these improvements is relatively small, the Wilcoxon signed-rank test at the 0.05

significance level shows that they are statistically significant. This means that the distances between the native and the predicted positions of the ligand are consistently smaller when using our consensus approach. These preliminary results suggest that these four methods generate complementary predictions, and they motivate further research on the ensemble-based predictors.

Case studies

We use the chain A of Bcr-Abl protein (PDB code: 3K5V) (Zhang *et al*, 2010) and M2 proton channel of influenza A virus (PDB code: 2RLF) (Schnell and Chou, 2008) to demonstrate the utility of the four representative binding site predictors. These proteins were not included in our benchmark dataset and were subject to recent studies to reveal the atomic-level insights into their binding interactions (Schnell and Chou, 2008; Zhang *et al*, 2010). We superimpose the above two structures with other Bcr-Abl and M2 proton channels structures in the Protein Data Bank (PDB), respectively, using Fr-TM-align (Pandit and Skolnick, 2008). This is performed to assure a complete (to date) annotation of the native binding sites. As a result, both proteins are annotated with two binding sites. We used the web servers of Findsite, MetaPocket, ConCavity, and Q-SiteFinder to generate the predictions. The two structures with the ligands shown in black and the predictions from Findsite, ConCavity, MetaPocket and Q-SiteFinder that are colored green, pink, red and blue, respectively, are given in Figure 3.

For the Bcr-Abl protein, we evaluated the top 2 predictions from each predictor since this protein has two binding sites. Both of these sites are predicted correctly by Findsite and Q-SiteFinder, see Figure 3A. The distances between the predicted site and the center of the ligand are 1Å and 2Å for the Findsite and 1Å and 3Å for the Q-SiteFinder. The Q-SiteFinder predicts the grid points of the binding sites, which have more than 40% overlap, measured using O_{PL}, with the ligands. The predictions by the MetaPocket are less accurate; its D_{CC} for the two binding sites equals 6Å and 2Å. ConCavity generates one prediction for this structure with the D_{CC} equal 5Å. The pocket identified by ConCavity is not shown in Figure 3A because it would obstruct predictions from the other methods; this pocket is visualized in the Supplementary Figure 4A. We note that these two sites are biologically relevant; a recent study has shown that inhibitors that bind to these two sites lead to the inhibition of Bcr–Abl activity (Zhang *et al*, 2010).

The binding sites on the M2 proton channel of influenza A virus have recently attracted significant attention since a class of antiviral drugs, such as adamantane M2 inhibitors, interacts with this channel. The structure of the M2 proton channel in complex with inhibitors was solved in 2008 by two groups which proposed two distinct binding sites (Stouffer, et al., 2008; Schnell and Chou, 2008). A recent study confirmed that Adamantane and its derivatives are capable of interacting with both binding sites (Rosenberg and Casarotto, 2010). The sites on the M2 proton channel are difficult to predict due to two facts: 1) the channel is formed by 4 protein chains

while some predictors, including Findsite, are designed to predict using a single chain; and 2) the binding sites are located in the transmembrane regions (Rosenberg and Casarotto, 2010) while most of the complexes used to develop the binding site predictors concern globular proteins. Each of the four chains has two sites. The site located at the center of the channel is common to all four chains and the other sites are symmetrically distributed at the lipid-facing side of the four chains. As a result, this protein complex has total of five binding sites and thus we evaluated the top five predictions generated by each of the four prediction methods. The predicted binding sites and the ligands are shown in Figures 3B (side view) and 3C (top view). The binding site at the pore of the channel is predicted only by the MetaPocket. Although the distance between the predicted site and center of the ligand is around 6Å, the predicted site is at the center of four key binding residues (Ser31 on the four chains), which are depicted in yellow in Figures 3B and 3C. The other sites, which are targeted by Rimantadine, are located at the base of the transmembrane helix on each of the chains. Only one of these sites is correctly predicted by the MetaPocket, and none of the top five predictions by Q-SiteFinder is close to the ligand (D_{CC} > 8Å). The ConCavity predicts one pocket, which is shown in the Supplementary Figure 4B, and this prediction is relatively far from the actual site (D_{CC} > 8Å). We note that Findsite did not generate predictions for the M2 proton channel due to the unavailability of suitable templates. Overall, we conclude that majority of the considered binding sites were found by at least one of the top four methods, which suggests that they provide useful inputs for the atomic-level discovery of protein-ligand interactions.

Discussion

The knowledge of the location of the binding sites is crucial for protein function annotation, elucidation of the mechanism of cellular activities, molecular docking and rational drug discovery. We empirically compare ten structure-based binding site predictors which were developed in the past decade and which offer either a web-server or a standalone program to generate predictions. The more recent methods including Findsite, Q-SiteFinder, ConCavity, and MetaPocket are shown to provide improvements over the older solutions. This indicates that progress was made over the last several years. However, a considerable fraction of the binding sites is not identified by any of the considered methods. For instance, at a cutoff of 4Å and using the D_{CC} measure, about 33% of the binding sites are missed by the four best-performing methods. We demonstrate that the quality of the predictions is strongly positively correlated with the size of the binding sites. We also show that although Findsite is significantly more accurate than the other considered predictors and is more robust when performing predictions using the apo structures, this method is largely dependent on the completeness of its template library. When the maximal TM-score between a query protein and the best template identified by Findsite is below 0.5, then certain energy-, structure- and consensus-based predictors are shown to provide more accurate predictions. We developed a simple consensus-based approach that uses four complementary predictors, the threading-based Findsite, the energy-based Q-SiteFinder, the geometry-based ConCavity, and consensus-based

MetaPocket. This method is shown to provide success rates improved by 2% when compared with the best performing Findsite.

Since the threading-based method works by identifying a known similar fold for a given query protein, the templates that are used in the prediction are restricted to one structural fold. However, a recent study shows that conserved sugar-binding and aromatic-group binding fragments are found across multiple protein folds (Petrey et al., 2009). The phosphate-binding fragment that occurs in dozens of protein families was discovered already two decades ago (Saraste et al., 1990). This means that the approach taken by the Findsite may not work in these and related cases and it motivates further research in this area. One of the potential solutions would be to develop a measure of similarity between surface patches on the query protein and the surfaces of the known binding sites, which would be added into the threading library; this idea extends a recently proposed surface scanning method (Chen and Kurgan, 2009). By comparing relevant "sub-structures" (fragments of the fold concerning the binding sites), the above approach could overcome the constraint on the similarity of the overall fold between the query and the template structures.

Experimental Procedures

Benchmark dataset

The benchmark dataset is designed to cover a wide range of non-homologous protein structures and to include structures with the largest number of annotated binding sites.

We selected a representative chain for each SCOP family and we mapped the binding sites of other similar structures into this chain. Prior work shows that two chains from different SCOP families have less than 1% chance to share more than 25% sequence similarity (Levitt, 2007). Since every chain in our dataset comes from a different protein family, the included proteins should be dissimilar in both their tertiary structure and sequence. We downloaded all available protein-ligand complexes from the PDB as of August 18th, 2009 and we annotated these proteins with their corresponding SCOP families. One chain for each SCOP family is selected using the following procedure. First, sequence similarity and structural similarity expressed with TM-score (Zhang and Skolnick, 2005) are calculated for every pair of chains within a given SCOP family. Next, the two similarity scores are used to perform clustering. Two chains are assigned to the same cluster if their sequence similarity is above 80% and their TM-score is above 0.5, as suggested in (Zhang and Skolnick, 2005). We assume that the chains of the same cluster are homologous and that they share the same binding sites. Finally, we count the number of types of ligands that interact with the chains of each cluster. The cluster with the largest number of the ligand types is selected and this cluster is represented by the protein with the largest number of bound ligands. The latter choice is made to maximize the number and accuracy of the annotations of the binding sites. The ligands in the other chains in the selected cluster are superimposed into the representative structure using Fr-TM-align (Pandit and Skolnick, 2008). If the superimposed ligand structure clashes with the representative protein structure, then this ligand is removed. This step results in a

protein structure that includes a (large) number of bound ligands, where some of these ligands could be redundant. A single-linkage clustering was performed to remove the redundancy. The distance between two ligands is defined as the minimum distance between any atom of one ligand and any atom of the other ligand. The clustering is terminated using 5Å threshold to ensure that ligands from one cluster do not overlap with ligands from another cluster. The median structures are chosen for each cluster of ligands. These median structures form a set of non-redundant ligands that bind to the protein structure that represents a given SCOP family.

The resulting dataset contains 314 protein structures. These structures are manually inspected to filter out biological irrelevant ligands, such as the glycol molecule that is introduced by the purification and crystallization procedures. For the structures with a published reference, a ligand is considered as biologically relevant if it is mentioned in the title or the abstract of the reference or the interaction between the ligand and the target protein is discussed in the results section. The ligands that only appear in the materials section or are never mentioned in the reference are removed. In case of the structures with no published reference, we use rules that were recently suggested by Wodak and colleagues (Dessailly et al., 2008). A given ligand is considered as biologically relevant if 1) it includes at least 10 non-hydrogen atoms; 2) it establishes at least 70 inter-atomic contacts with the protein atoms; and 3) the interaction does not concern lipid and membrane proteins. Our benchmark dataset includes 251 proteins after removing the “irrelevant” ligands. These are ligand-bound (i.e., holo) structures.

Since the protein-ligand interactions could lead to conformation changes, we also generated a dataset that consists of the matching apo structures. For each protein in the benchmark dataset we searched for its corresponding apo structure in the PDB. An apo structure is assumed to correspond to a given holo structure if they belongs to the same SCOP family, they share more than 80% sequence similarity and their TM-score is above 0.5. We found 104 apo structures and we created two additional datasets, D_{Apo} dataset that includes the 104 apo structures and D_{Holo} dataset which is a subset of the corresponding 104 holo structures from the benchmark dataset.

The proteins in the benchmark dataset have diverse overall structural topology. Based on the annotation from the SCOP database, they cover 6 structural classes, 148 protein folds, 184 superfamilies, and 251 protein families. The maximal pairwise sequence similarity is between 11% and 24%; see Supplementary Figure 3J. The 251 proteins are annotated with 475 binding sites which interact with 253 types of ligands. All datasets including the benchmark dataset and the D_{Holo} and D_{Apo} datasets are available for download from http://biomine.ece.ualberta.ca/BindingSitesPredictors/main.htm. This web page also provides URLs for the considered ten binding site predictors.

Quality indices

We use three indices to evaluate predictions of the considered binding site predictors:

- D_{CA}, which is defined as the minimal *distance* between the *center* of the predicted binding site (pocket) and any *atom* of the ligand, was widely used to assess the

prediction quality in several prior studies. For instance, authors of LIGSITE^{csc}, PASS, PocketPicker and Fpocket assume that a predicted site is correct if its center is no farther that 4Å to any atom of the ligand. Instead of using one arbitrary threshold, we compute the success rates using D_{CA} values for integer thresholds between 1Å and 20Å.

- D_{CC}, which is defined as the minimal *distance* from the *center* of the predicted binding site to the *center* of the ligand, was proposed by Skolnick and colleagues (Skolnick and Brylinski, 2008). When compared with D_{CA}, this measure compensates for the size of the ligand, i.e., D_{CA} gives higher success rates for larger ligands. D_{CC} was recently used to compare Findsite and LIGSITE^{csc} (Skolnick and Brylinski, 2008). The success rates are computed using integer thresholds between 1Å and 20Å.

- O_{PL}, which quantifies *overlap* between the predicted *binding site* and the *ligand*, is proposed in this study. This measure is defined as the ratio between the volume of the intersection of the predicted site and ligand, and the volume of their union. In addition to being sensitive to the size of the ligand, this quality index improves over both D_{CA} and D_{CC} by compensating for the relative spatial orientation of the ligand and the binding site. It can be computed for the four methods that output the full set of grid points of the predicted site (Q-SiteFinder, PocketPicker, ConCavity, and PocketFinder) instead of just the center of the pocket that is predicted by the other considered predictors. To calculate this value, both the binding site (pocket) and ligand are represented using a set grid points in the

same grid scale. A grid point is assigned to the ligand/site if the distance between this point and ligand/site is smaller than half of the length of diagonal in the grid cube. The O_{PL} value is computed as the ratio between the number of grid points that are shared by the ligand and the binding site, and the number of grid points that belong to either the ligand or the site.

Statistical analysis

For a protein with *n* binding sites we take the top *n* predictions for every considered prediction method. We generate a set of minimal distances between each of the *m* binding sites (in the entire dataset) and the corresponding *n* predictions for each of the prediction methods. We assume that the predictions from different methods that are farther than 10Å away from the site are equally wrong, i.e., they are too far away to be meaningful, and thus we round them down to 10Å. The significance of the differences between a given pair of predictors was measured by evaluating the corresponding, for the same *m*, minimal distance values. Since the distances for the considered predictors are not normally distributed, per the Shapiro-Wilk test of normality at $p = 0.05$, we used the non-parametric Wilcoxon signed-rank test. We assume that the differences are significant if $p < 0.05$.

Acknowledgements

This work was sponsored in part by the Discovery grant from NSERC Canada to Lukasz Kurgan, the Alberta Ingenuity and iCORE scholarship in ICT to Ke Chen, and the Killam Memorial Scholarship to Marcin Mizianty. The authors declare no

competing interests.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F. et al. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185–2195.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. *25*, 25-9.

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

Brooijmans, N., and Kuntz, I.D. (2003). Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct. *32*, 335-73.

Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M., and Funkhouser, T.A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLos Comput Biol. *5*, 12.

Chandonia, J.M., and Brenner, S.E. (2006). The impact of structural genomics: expectations and outcomes. Science *311*, 347-51.

Chen, K., and Kurgan, L. (2009). Investigation of atomic level patterns in protein-small ligand interactions. PLoS ONE *4*, e4473.

Dessailly BH, Lensink MF, Orengo CA, Wodak SJ. (2008). LigASite--a database of biologically relevant binding sites in proteins with known apo-structures. Nucleic Acids Res. *36*, D667-73.

Ellis, J.J., Broom, M., and Jones S (2007). Protein-RNA interactions: structural analysis and functional classes. Proteins. *66*, 903–11.

Hendlich, M., Rippmann, F., and Barnickel G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model.*15*(*6*), 359-63.

Huang, B. (2009). MetaPocket: a meta approach to improve protein ligand binding site prediction. OMICS *13*, 325-30.

Huang, B., and Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol. *6*, 19.

Ivankov, D.N., Bogatyreva, N.S., Lobanov, M.Y., and Galzitskaya, O.V. (2009). Coupling between properties of the protein shape and the rate of protein folding. PLoS One *4*, e6476.

Jones, S., and Thornton, J.M. (1996). Principles of protein-protein interactions. Proc Natl Acad Sci U S A *93*, 13-20.

Jones, S., and Thornton, J.M. (2004). Searching for functional sites in protein structures. Curr Opin Chem Biol. *8*, 3-7.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. *22*, 2577-637.

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

Laurie, A.T., and Jackson, R.M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics *21*, 1908-16.

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

Levitt, M. (2007). Growth of novel protein structural data. Proc Natl Acad Sci U S A. 104, 3183-8.

López, G., Ezkurdia, I., and Tress, M.L. (2009). Assessment of ligand binding residue predictions in CASP8. Proteins *77*(*Suppl 9*), 138-46.

Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. *29*, 2860-74.

Mukherjee, S., Acharya, B.R., Bhattacharyya, B., and Chakrabarti, G. (2010). Genistein arrests cell cycle progression of A549 cells at the G(2)/M phase and depolymerizes interphase microtubules through binding to a unique site of tubulin. Biochemistry *49*, 1702-12.

Pandit, S.B., and Skolnick, J. (2008). Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. BMC Bioinformatics *9*, 531.

Petrey, D., Fischer, M., and Honig, B. (2009). Structural relationships among proteins with different global topologies and their implications for function annotation strategies. Proc Natl Acad Sci U S A. *106*, 17377-82.

Popova, Y., Thayumanavan, P., Lonati, E., Agrochão, M., and Thevelein, J.M. (2010). Transport and signaling through the phosphate-binding site of the yeast Pho84 phosphate transceptor. Proc Natl Acad Sci U S A. *107*, 2890-5.

Rosenberg, M.R., and Casarotto, M.G. (2010). Coexistence of two adamantane binding sites in the influenza A M2 ion channel. Proc Natl Acad Sci U S A. *107*, 13866-71.

Saraste, M., Sibbald, P.R., and Wittinghofer, A. (1990). The P-loop–A common motif in ATP-binding and GTP-binding proteins. Trends Biochem Sci. *15*, 430-434.

Schnell, J.R., and Chou, J.J. (2008). Structure and mechanism of the M2 proton channel of influenza A virus. Nature. *451*, 591-5.

Skolnick, J., and Brylinski, M. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci U S A. *105*, 129-34.

Stouffer, A.L., Acharya, R., Salom, D., Levine, A.S., Di, Costanzo L., Soto, C.S., Tereshko, V., Nanda, V., Stayrook, S., and DeGrado, W.F. (2008). Structural basis for the function and inhibition of an influenza virus proton channel. Nature.*451*, 596-9.

Weisel, M., Proschak, E., and Schneider, G. (2007). PocketPicker: analysis of ligand binding-sites with shape descriptors. Chem Cent J. *1*, 7.

Whittard, J.D., Sakurai, T., Cassella, M.R., Gazdoiu, M., and Felsenfeld, D.P. (2006). MAP kinase pathway-dependent phosphorylation of the L1-CAM ankyrin binding site regulates neuronal growth. Mol Biol Cell. *17*, 2696-706.

Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. *36*, D901-6.

Zhang, J., Adrián, F.J., Jahnke, W., Cowan-Jacob, S.W., Li, A.G., Iacob, R.E., Sim, T., Powers, J., Dierks, C., Sun, F. et al. (2010). Targeting Bcr-Abl by combining allosteric with ATP-binding-site inhibitors. Nature *463*, 501-6.

Zhang, T., Zhang, H., Chen, K., Ruan, J., Shen, S., Kurgan, L. (2010). Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. Curr. Prot. Peptide Sc. *11*(*7*), 609-628

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302-9.

Zhu, H., Sommer, I., Lengauer, T., and Domingues, F.S. (2008). Alignment of non-covalent interactions at protein-protein interfaces. PLoS ONE *3*, e1926.

Figure Legends

Figure 1. The success rates (*y*-axis) of the considered binding site predictors measured using D_{CC} (the minimal distance from the center of the predicted site to the center of the ligand) on the benchmark dataset.

(A) Results for the ten considered predictors. A given binding site is regarded as correctly predicted if the minimal distance between this site and the top *n* predictions is below the cutoff distance *D* (*x*-axis), where *n* is the number of binding sites of the protein that includes the evaluated binding site.

(B) Comparison of the success rates of Findsite using its entire template library for different cutoff distances *D* (*x*-axis) with the predictions where the maximal structural similarly between a query protein and the templates is limited to TM-score ≤ 0.9, ≤ 0.8, ≤ 0.7, ≤ 0.6, and ≤ 0.5. This panel also includes the success rates for the Meta-pocket, ConCavity, and Q-SiteFinder predictors.

(C) Comparison of success rates on the D_{Holo} and D_{Apo} datasets measured using D_{CC} for the Findsite, MetaPocket, ConCavity, and Q-SiteFinder. The two datasets contains structures of the same set of proteins where D_{Holo} includes ligand-bound structures and D_{Apo} includes structures at the ligand-unbound state. The *x*-axis shows the cutoff distance *D* that is used to calculate the success rates.

(D) The success rates of the Findsite, Q-SiteFinder, ConCavity, MetaPocket and a consensus-based method compared to the coverage of the binding sites predicted by combination of the four methods. The *x*-axis shows the cutoff distance *D* that is used to calculate the success rates and the dashed line shows the success rates of the

consensus-based re-ranking of the Findsite predictions.

Figure 2. Relation between the average, over cutoff distances *D* between 1Å and 5Å, success rates (*y*-axis) measured using D_{CC} and the size of the binding sites for Findsite, Q-SiteFinder ConCavity, and MetaPocket. The binding sites in the benchmark dataset are sorted by their sizes, which are approximated by the number of interacting atoms, in the ascending order and they are binned into five equally sized subsets. The *x*-axis shows the average size of the binding sites for the five consecutive subsets.

Figure 3. The binding sites predicted by the Findsite, MetaPocket, ConCavity, and Q-SiteFinder for chain A of the Bcr-Abl protein (panel A) and the M2 proton channel (panels B and C show the side and the top views, respectively). The predictions by Findsite, MetaPocket, ConCavity, and Q-SiteFinder are denoted with green, red, and pink spheres and blue mesh, respectively. The Q-SiteFinder predicted grid points of the pocket are shown using the mesh. The ligands are in the stick format and are colored in black. The M2 proton channel consists of 4 chains and has 5 binding sites. Each of the 4 chains is annotated with 2 sites, where the site at the center of the channel is common to all of them. The other 4 sites are symmetrically distributed at the lipid-facing side of the four chains. The key interacting residues for the central binding site, Ser31, on these four chains are colored in yellow in panels B and C.

Figure 1

Click here to download high resolution image

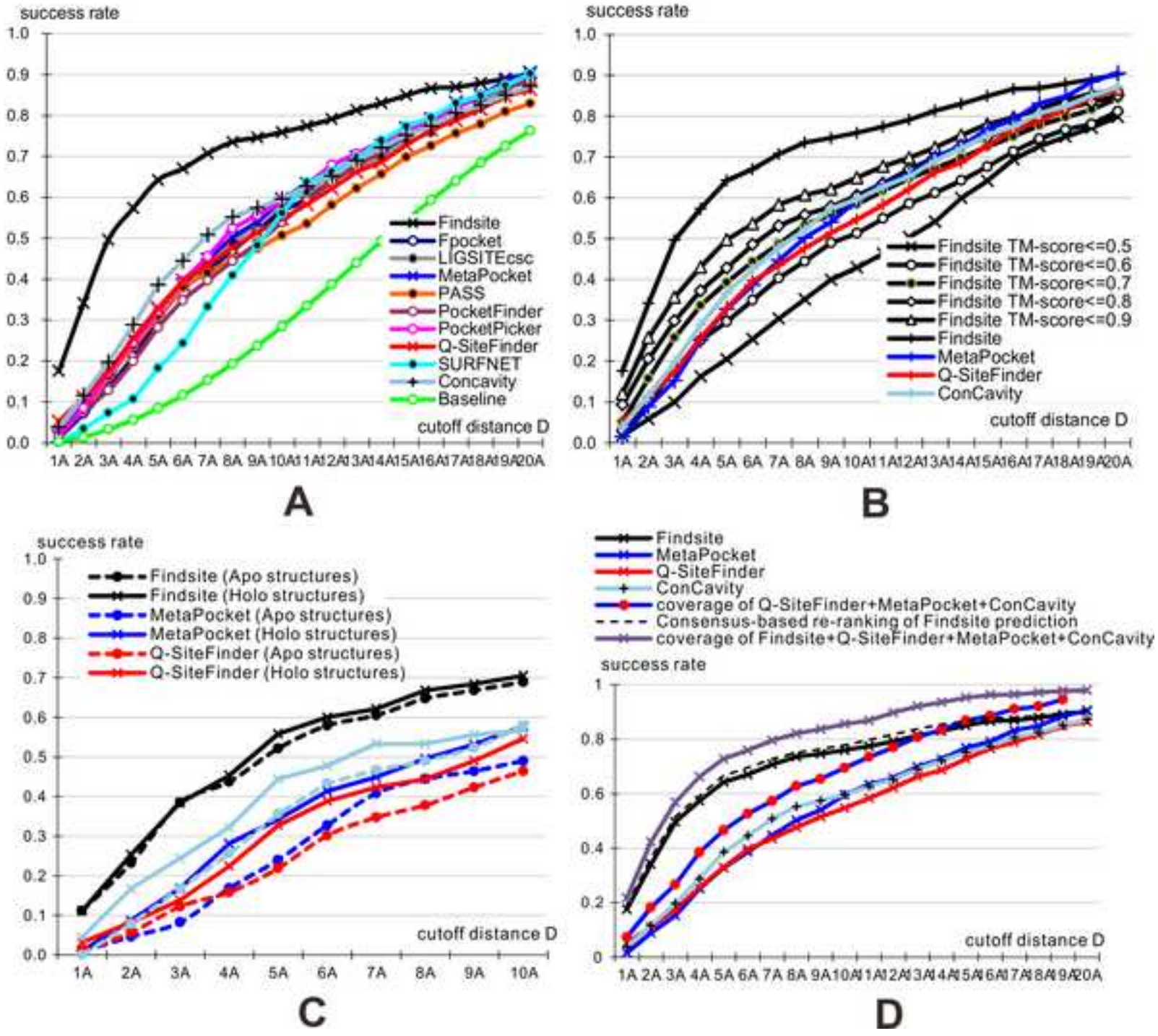


Figure 2

Click here to download high resolution image

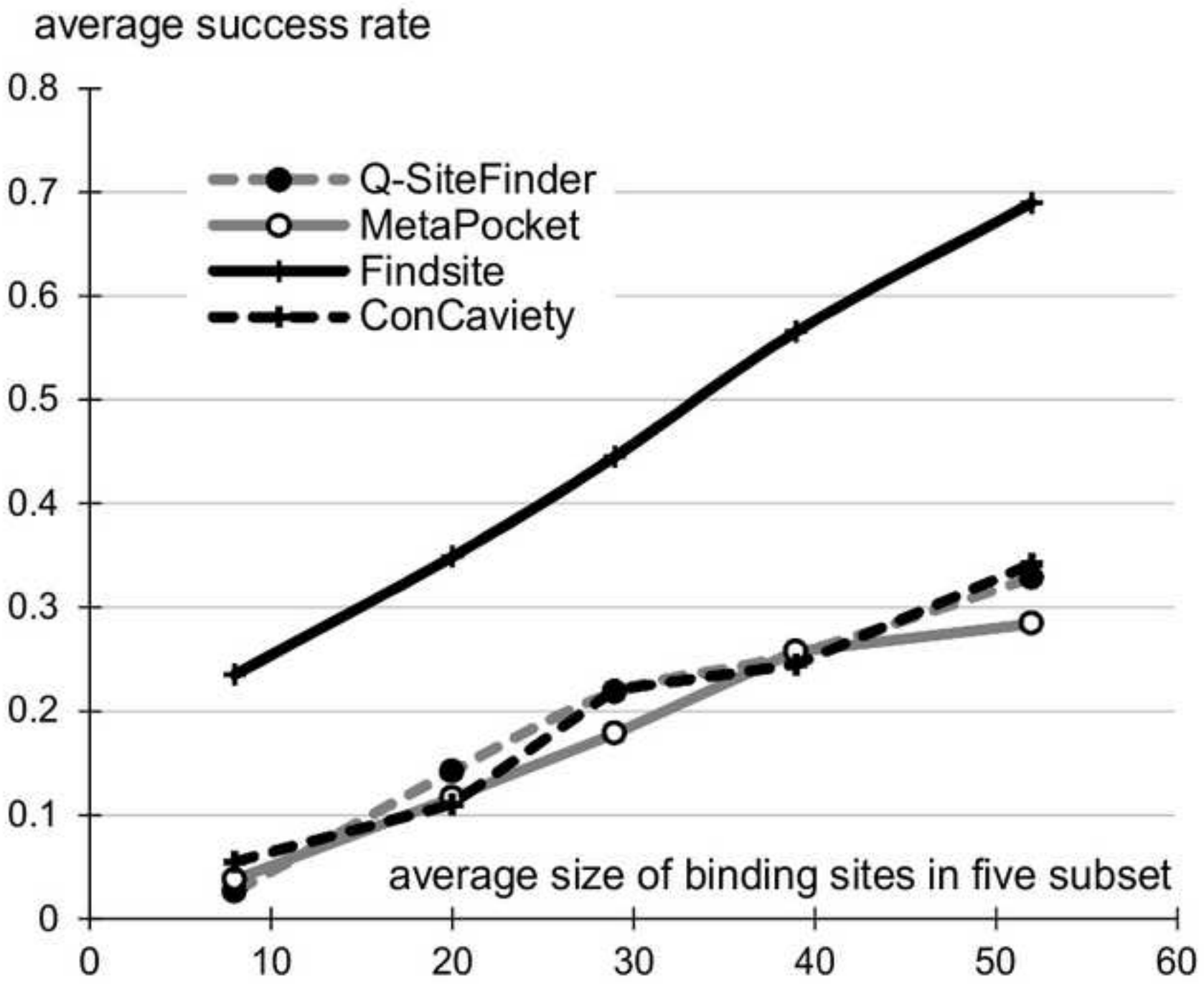


Figure 3

Click here to download high resolution image

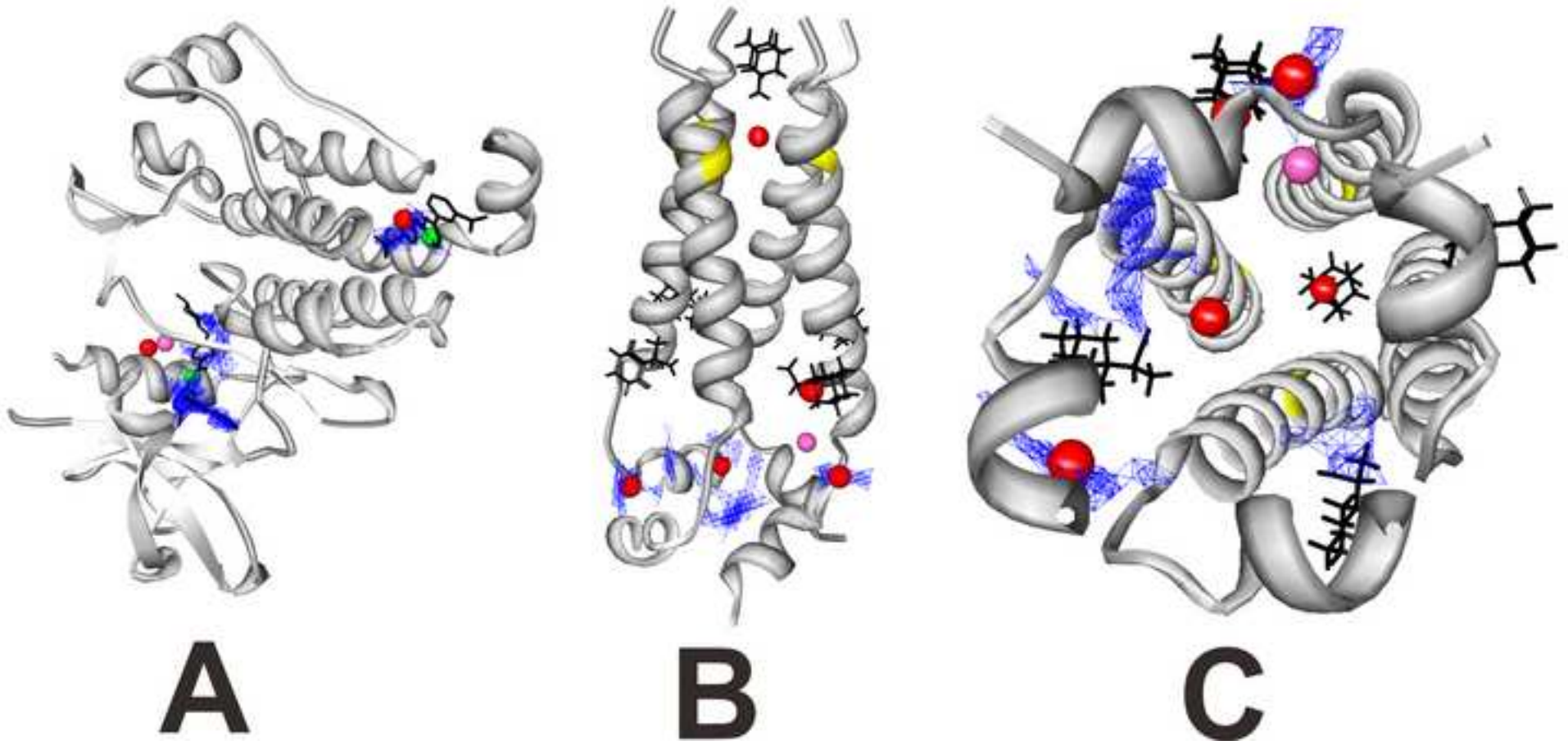


Table 1

Table 1. Statistical significance of the differences in distances measured using D_{CC} between the predicted and the actual location of the binding site for all pair of the considered ten prediction methods measured using Wilcoxon signed-rank test. The results are calculated on the entire benchmark dataset. The "+"/"–" indicates that a method in a given column is significantly better/worse than a methods in a given row with $p < 0.05$ and "=" denotes that a given pair of methods is not significantly different.

	PASS	SURFNET	Pocket-Finder	Fpocket	LIGSITE^{csc}	Q-SiteFinder	Pocket-Picker	Meta-Pocket	Con-Cavity	Findsite
PASS		=	+	+	+	+	+	+	+	+
SURFNET	=		+	+	+	+	+	+	+	+
PocketFinder	–	–		=	+	+	+	+	+	+
Fpocket	–	–	=		=	+	+	+	+	+
LIGSITE^{csc}	–	–	–	=		+	+	+	+	+
Q+SiteFinder	–	–	–	–	–		=	=	+	+
PocketPicker	–	–	–	–	–	=		=	=	+
MetaPocket	–	–	–	–	–	=	=		=	+
ConCavity	–	–	–	–	–	–	=	=		+
Findsite	–	–	–	–	–	–	–	–	–	

Inventory for the supplement for "A critical comparative assessment of predictions of protein binding sites for biologically relevant organic compounds"

Ke Chen[1], Marcin Mizianty[1], Jianzhao Gao[2] and Lukasz Kurgan[1*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, CANADA T6G 2V4
[2]College of Mathematics, Nankai University, Tianjin, PR CHINA 300071

Table 1 is related to Supplementary Table 1 and Supplementary Figure 3.

Figure 1 is related to Supplementary Figure 1.
Figure 2 is related to Supplementary Figure 2.
Figure 3 is related to Supplementary Figure 4.

Supplement for "A critical comparative assessment of predictions of protein binding sites for biologically relevant organic compounds"

Ke Chen[1], Marcin Mizianty[1], Jianzhao Gao[2] and Lukasz Kurgan[1*]

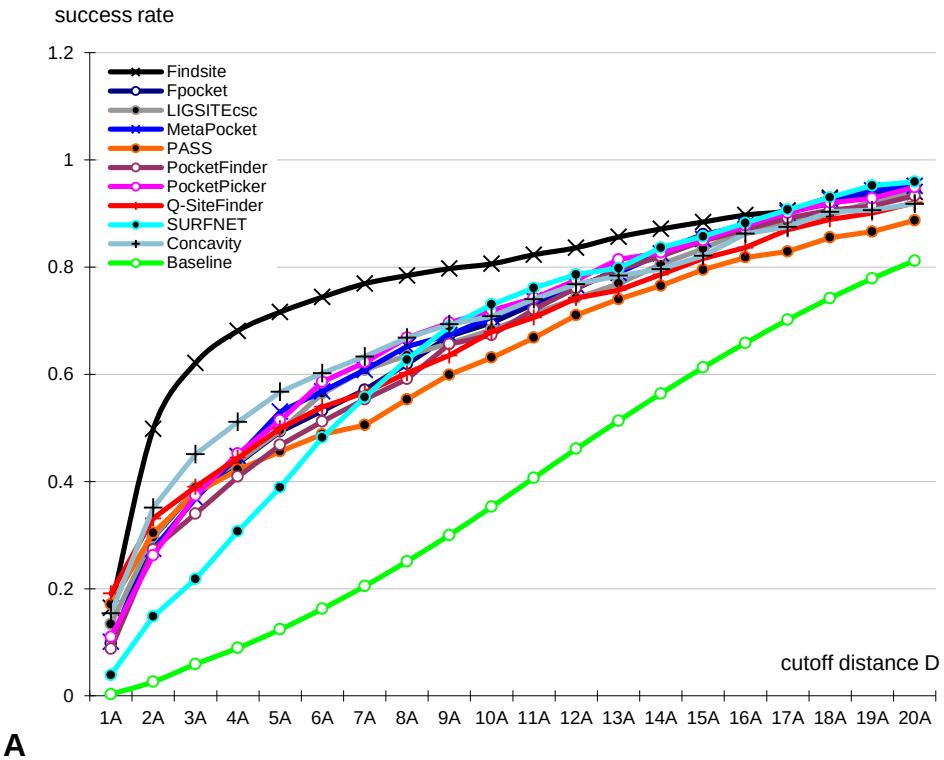
[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, CANADA T6G 2V4

[2]College of Mathematics, Nankai University, Tianjin, PR CHINA 300071

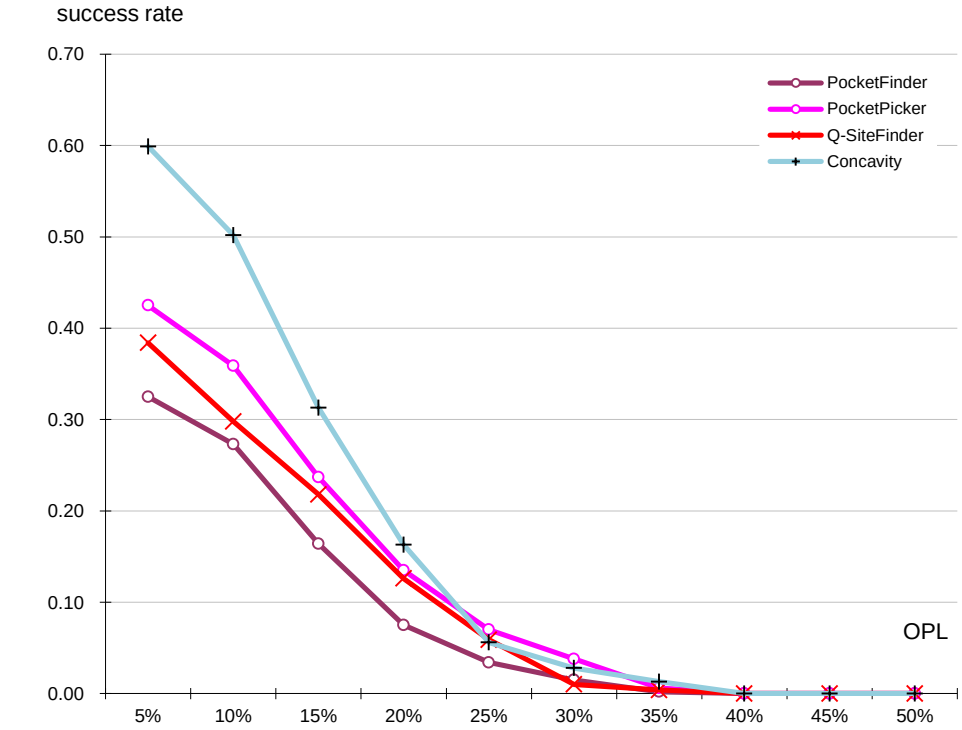
Supplemental Data

Supplementary Table 1. List of ligand types in the four considered major ligand categories.

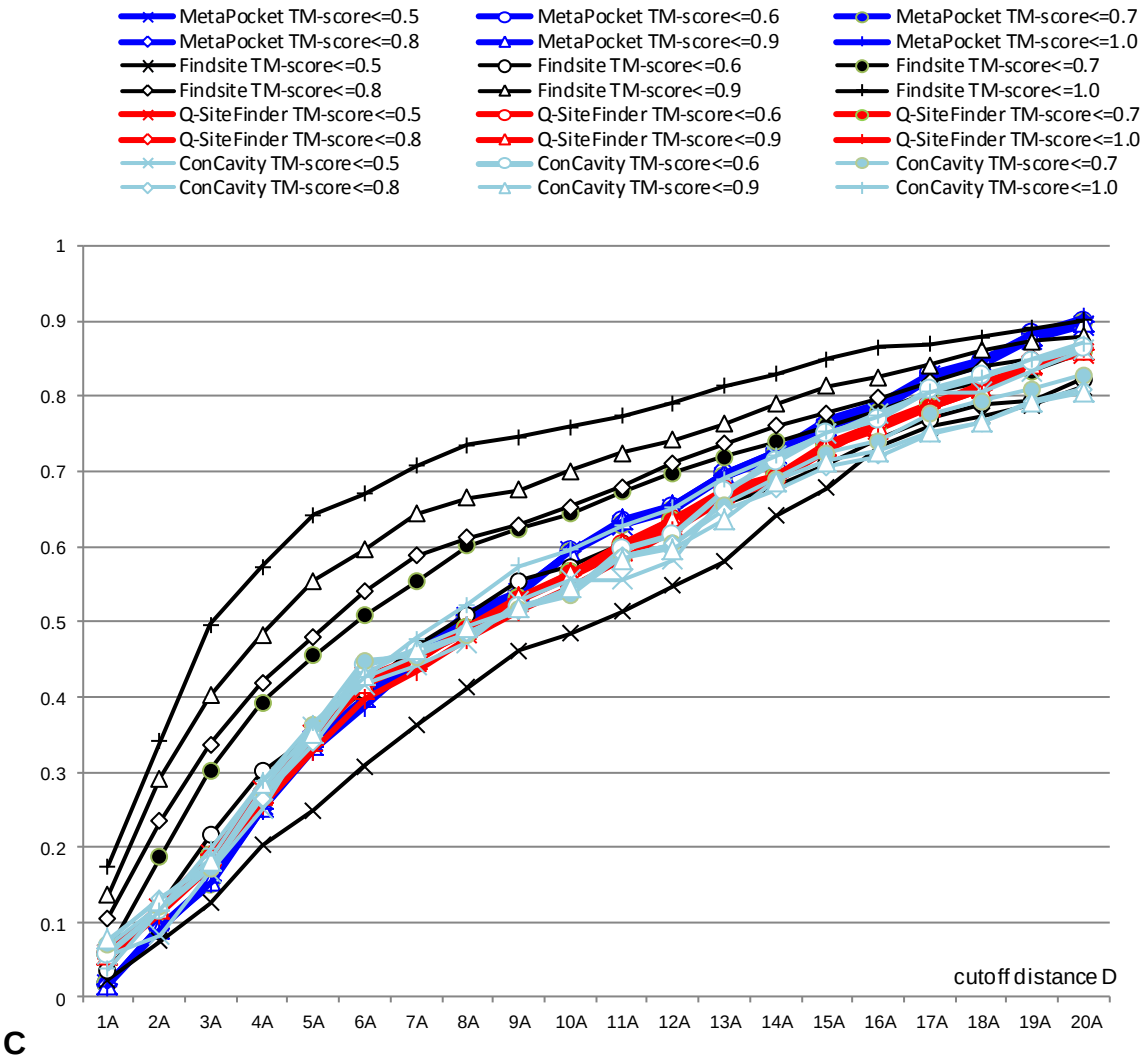
Ligand category	3-letter abbreviation of ligand name	Formula
Acids	BEZ	$C_7 H_6 O_2$
	SAH	$C_{14} H_{20} N_6 O_5 S$
	ASP	$C_4 H_7 N O_4$
	EPE	$C_8 H_{18} N_2 O_4 S$
	GLU	$C_5 H_9 N O_4$
	MYR	$C_{14} H_{28} O_2$
	PEB	$C_{33} H_{40} N_4 O_6$
	TRP	$C_{11} H_{12} N_2 O_2$
Carbohydrates	BGC	$C_6 H_{12} O_6$
	GLC	$C_6 H_{12} O_6$
	FUL	$C_6 H_{12} O_5$
	GAL	$C_6 H_{12} O_6$
	GLA	$C_6 H_{12} O_6$
	XYP	$C_6 H_{10} O_5$
	MAN	$C_6 H_{12} O_6$
	FUC	$C_6 H_{12} O_5$
	NAG	$C_8 H_{15} N O_6$
Nucleotides	2GP	$C_{10} H_{14} N_5 O_8 P$
	5GP	$C_{10} H_{14} N_5 O_8 P$
	A2P	$C_{10} H_{15} N_5 O_{10} P_2$
	A6P	$C_6 H_{13} O_9 P$
	ADP	$C_{10} H_{15} N_5 O_{10} P_2$
	AGP	$C_6 H_{16} N O_8 P$
	AMP	$C_{10} H_{14} N_5 O_7 P$
	AMZ	$C_9 H_{15} N_4 O_8 P$
	ANP	$C_{10} H_{17} N_6 O_{12} P_3$
	ATP	$C_{10} H_{16} N_5 O_{13} P_3$
	C5P	$C_9 H_{14} N_3 O_8 P$
	CMP	$C_{10} H_{12} N_5 O_6 P$
	G1P	$C_6 H_{13} O_9 P$
	GDP	$C_{10} H_{15} N_5 O_{11} P_2$
	GTP	$C_{10} H_{16} N_5 O_{14} P_3$
	NOS	$C_{10} H_{12} N_4 O_5$
	PAP	$C_{10} H_{16} N_5 O_{13} P_3$
	TMP	$C_{10} H_{15} N_2 O_8 P$
	TPP	$C_{12} H_{19} N_4 O_7 P_2 S$
	U5P	$C_9 H_{13} N_2 O_9 P$
	UDP	$C_9 H_{14} N_2 O_{12} P_2$
	UMP	$C_9 H_{13} N_2 O_8 P$
	UTP	$C_9 H_{15} N_2 O_{15} P_3$
Cofactors	ACO	$C_{23} H_{38} N_7 O_{17} P_3 S$
	COA	$C_{21} H_{36} N_7 O_{16} P_3 S$
	FAD	$C_{27} H_{33} N_9 O_{15} P_2$
	FMN	$C_{17} H_{21} N_4 O_9 P$
	HEM	$C_{34} H_{32} Fe N_4 O_4$
	NAD	$C_{21} H_{27} N_7 O_{14} P_2$
	NAP	$C_{21} H_{28} N_7 O_{17} P_3$
	PLP	$C_8 H_{10} N O_6 P$
	PQQ	$C_{14} H_6 N_2 O_8$
	SAM	$C_{15} H_{22} N_6 O_5 S$
	U2G	$C_{19} H_{24} N_7 O_{13} P$


success rate
1.2
1
0.8
0.6
0.4
0.2
0
Findsite
Fpocket
LIGSITEcsc
MetaPocket
PASS
PocketFinder
PocketPicker
Q-SiteFinder
SURFNET
Concavity
Baseline
cutoff distance D
1A 2A 3A 4A 5A 6A 7A 8A 9A 10A 11A 12A 13A 14A 15A 16A 17A 18A 19A 20A

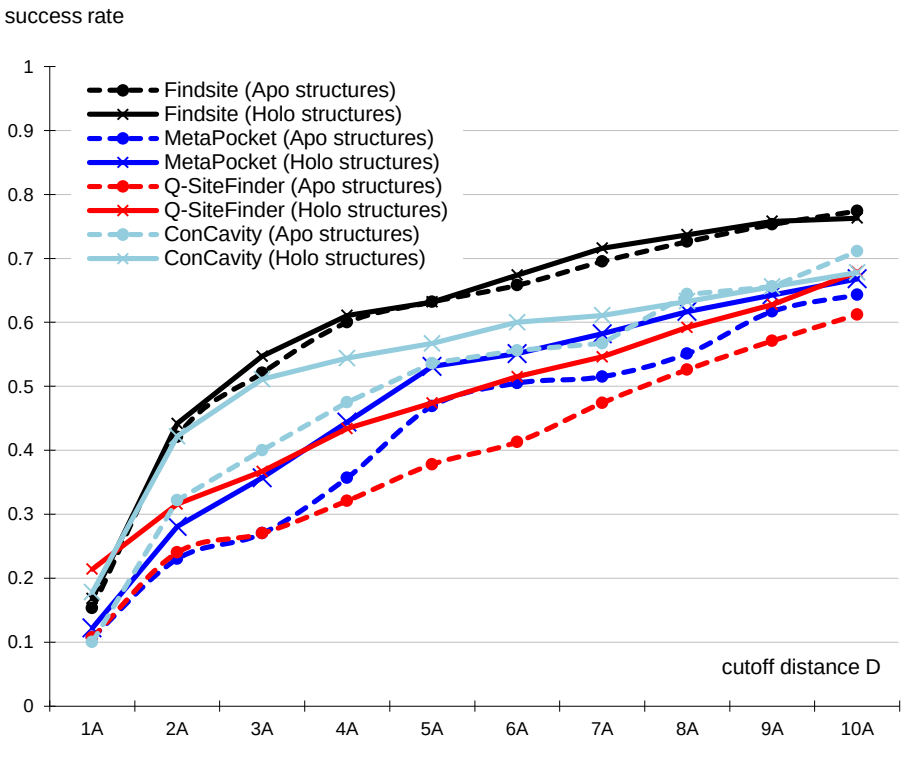
A


success rate
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
PocketFinder
PocketPicker
Q-SiteFinder
Concavity
OPL
5% 10% 15% 20% 25% 30% 35% 40% 45% 50%

B



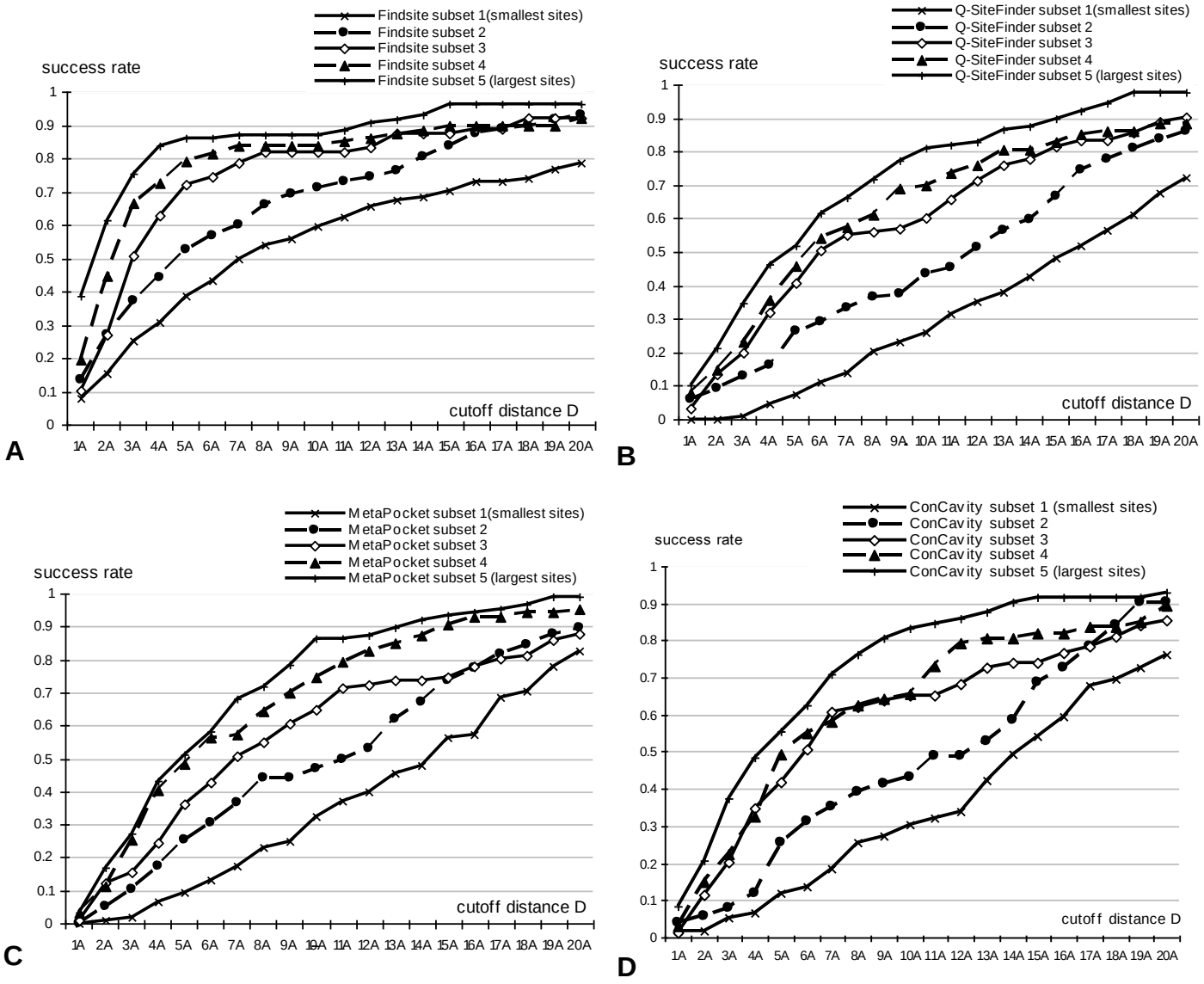
C



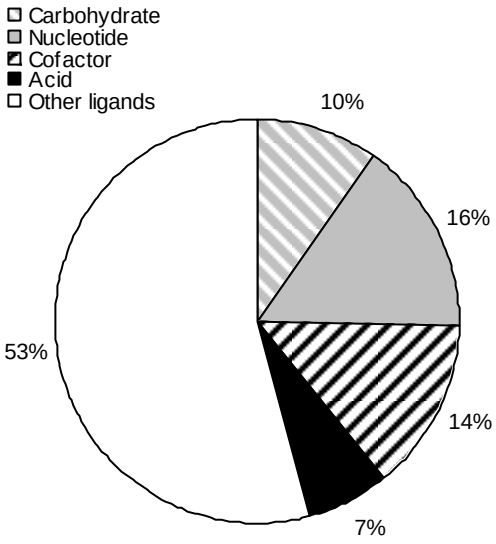
D

Supplementary Figure 1. The success rates (*y*-axis) of the considered binding site predictors on the benchmark dataset.

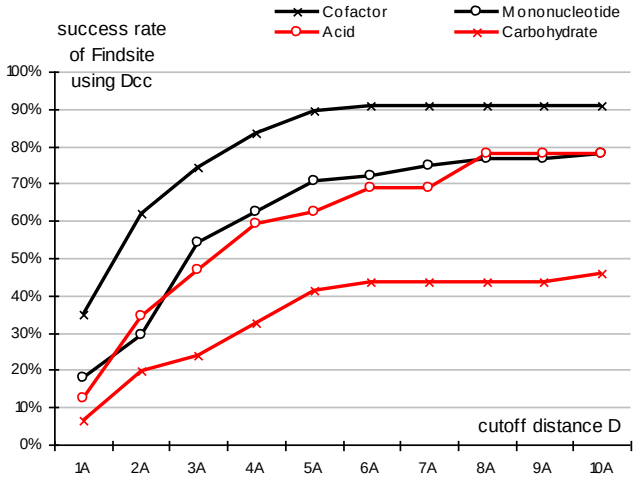
(A) The success rates of the ten representative methods measured using D_{CA} (the minimal distance from the center of the predicted site to any atom of the ligand). A given binding site is regarded as correctly predicted if the minimal distance between this site and the top *n* predictions is below the cutoff distance *D* (*x*-axis), where *n* is the number of binding sites of the protein that includes the evaluated binding site.
(B) The success rates of the PocketPicker, Q-SiteFinder, ConCavity and PocketFinder measured using O_{PL} (*overlap* between the predicted *pocket* and the *ligand*). The *y*-axis shows the percentage of binding sites that have their O_{PL} values equal or greater than value on the *x*-axis.
(C) Comparison of the success rates of Findsite using its entire template library measured using D_{CC} for different cutoff distances *D* (*x*-axis) on the benchmark dataset with the predictions where the maximal structural similarly between a query protein and the templates limited to TM-score ≤ 0.9, ≤ 0.8, ≤ 0.7, ≤ 0.6, and ≤ 0.5. The proteins for which Findsite could not find a template with TM-score > 1 were removed from each of the above five configurations. The figure also includes the success rates for Meta-pocket, ConCavity and Q-SiteFinder for the entire benchmark dataset and the five subsets.
(D) Comparison of the success rates (*y*-axis) on the D_{Holo} and D_{Apo} datasets measured using D_{CA} (the minimal distance from the center of the predicted site to any atom of the ligand) for the four representative methods, the threading-based Findsite, the energy-based Q-SiteFinder, the best performing geometry-based ConCavity, and the consensus-based MetaPocket. The two datasets include structures from the same set of proteins where D_{Holo} is composed of structures in the ligand-bound state and D_{Apo} in the ligand-unbound state. A given binding site is regarded as correctly predicted if the minimal distance between this site and the top *n* predictions is below the cutoff distance *D* (*x*-axis), where *n* is the number of binding sites of the protein that includes the evaluated binding site.



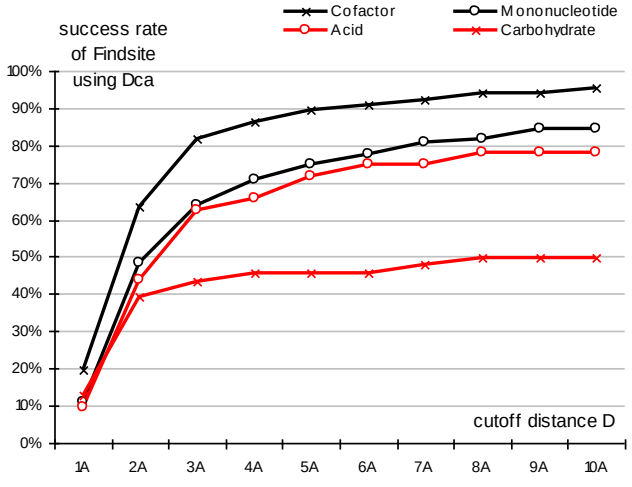
Supplementary Figure 2. Success rates (*y*-axis) measured using D_{CC} for different cutoff distances *D* (*x*-axis) for (A) Findsite, (B) Q-SiteFinder, (C) MetaPocket, and (D) ConCavity as a function of the size of the binding site, which is approximated by the number of interacting atoms. The binding sites in the benchmark dataset are sorted by their sizes in the ascending order and they are binned into five equally sized subsets. Each line corresponds to the results on one of these subsets, where subset 1 includes the smallest sites and subset 5 the largest sites.


Carbohydrate
Nucleotide
Cofactor
Acid
Other ligands
10%
16%
53%
14%
7%

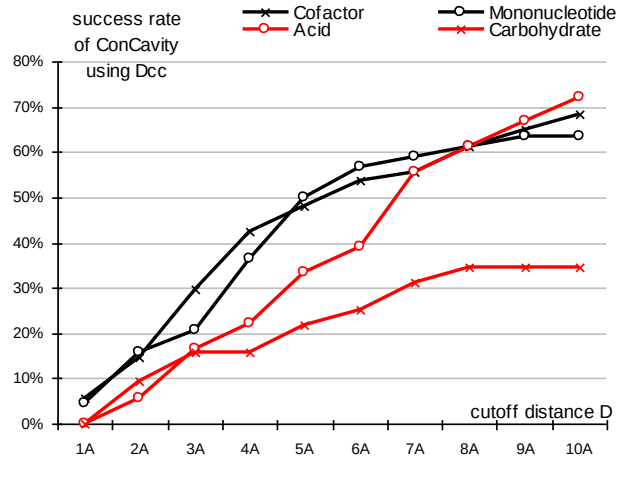
A


Cofactor
Mononucleotide
Acid
Carbohydrate
success rate
of Findsite
using Dcc
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
cutoff distance D
1A
2A
3A
4A
5A
6A
7A
8A
9A
10A

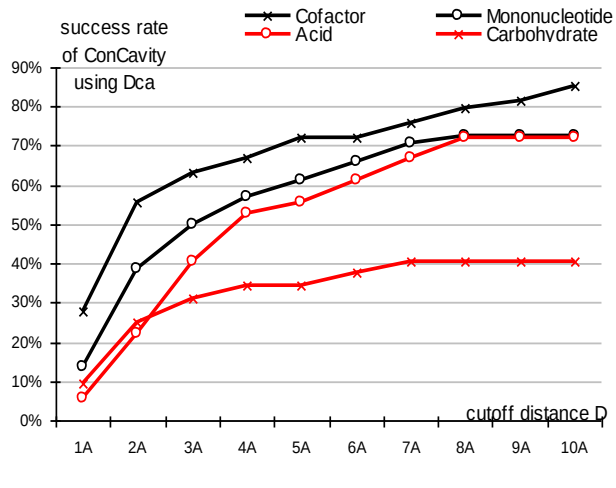
B


Cofactor
Mononucleotide
Acid
Carbohydrate
success rate
of Findsite
using Dca
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
cutoff distance D
1A
2A
3A
4A
5A
6A
7A
8A
9A
10A

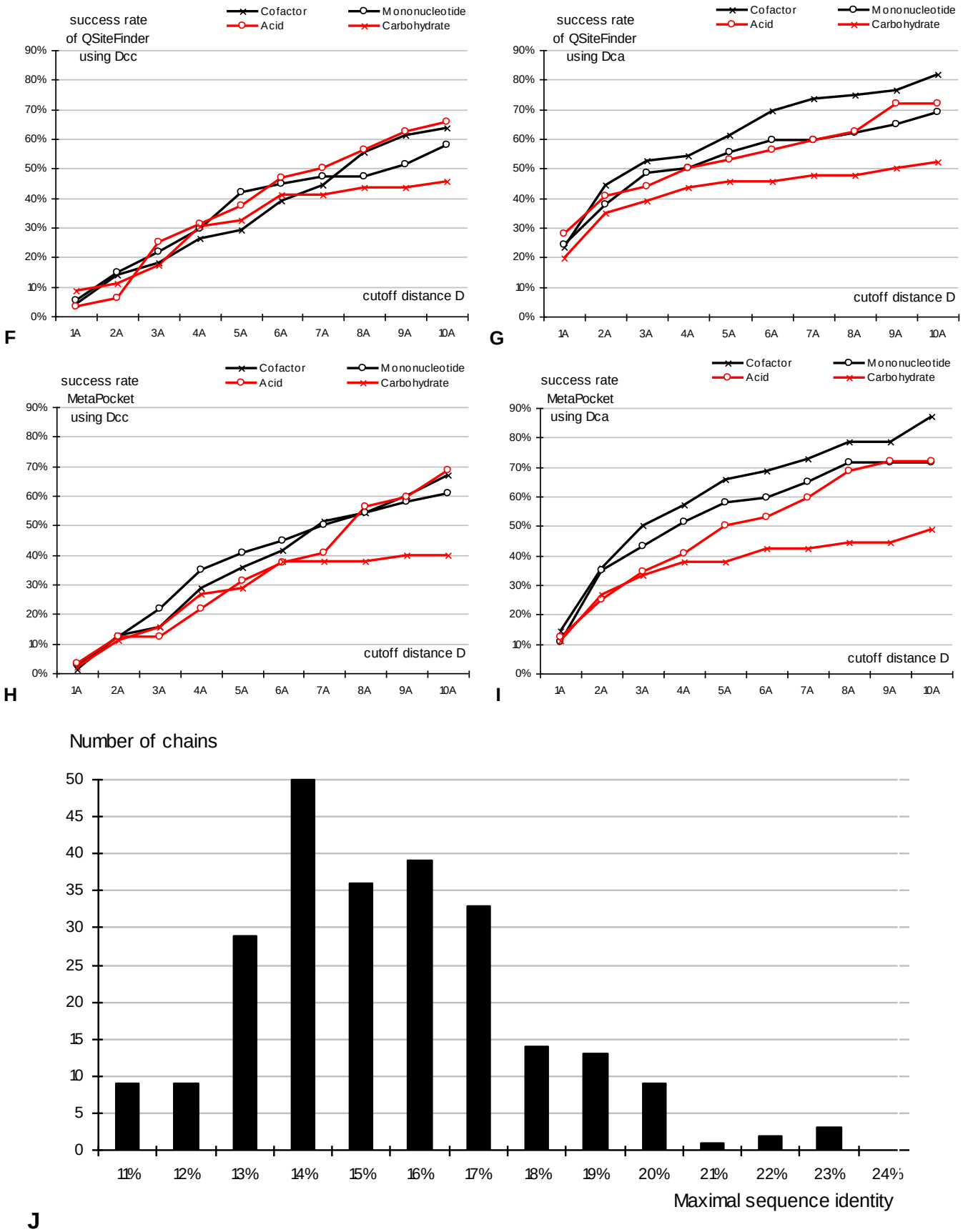
C


Cofactor
Mononucleotide
Acid
Carbohydrate
success rate
of ConCavity
using Dcc
80%
70%
60%
50%
40%
30%
20%
10%
0%
cutoff distance D
1A
2A
3A
4A
5A
6A
7A
8A
9A
10A

D


Cofactor
Mononucleotide
Acid
Carbohydrate
success rate
of ConCavity
using Dca
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
cutoff distance D
1A
2A
3A
4A
5A
6A
7A
8A
9A
10A

E

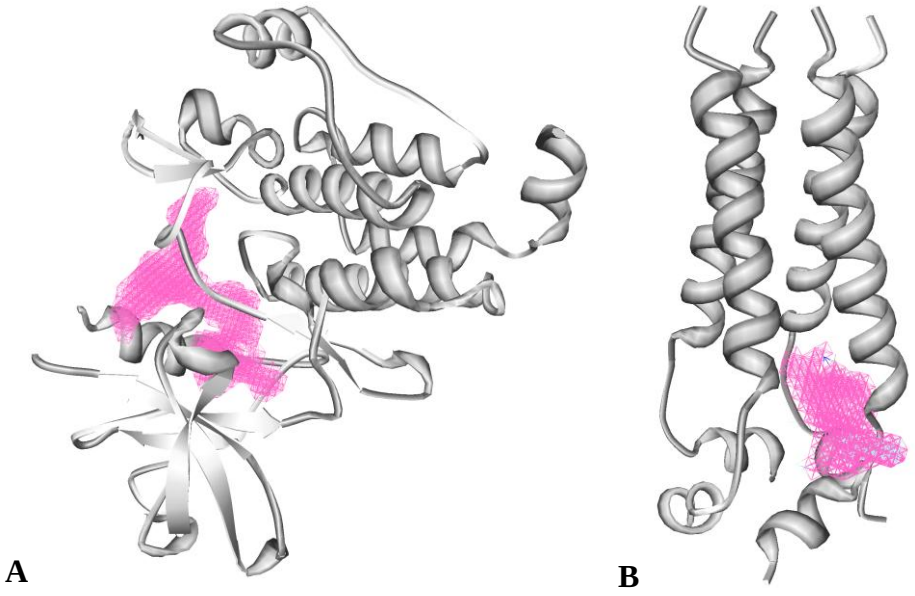


Supplementary Figure 3.

(A) The rate of occurrence of the four major ligand groups, which include nucleotides, cofactors, carbohydrates and acids, in the benchmark dataset. These four groups cover 46% of all ligands in the dataset.

(B to I) Comparison of the success rates (*y*-axis) for prediction of binding sites for four categories of ligands including acids, carbohydrates, mononucleotides and cofactors measured using D_{CC} (panels on the left) and D_{CA} (panels on the right). The *x*-axis shows the cutoff distance *D* used to calculate the

success rates. B) results of Findsite measured using D_{CC}; C) results of Findsite measured using D_{CA}; D) results of ConCavity measured using D_{CC}; E) results of ConCavity measured using D_{CA}; F) Results of Q-SiteFinder measured using D_{CC}; G) results of Q-SiteFinder measured using D_{CA}; H) results of MetaPocket measured using D_{CC}; I) results of MetaPocket measured using D_{CA}.
(J) Distribution of maximal pairwise sequence similarities (*x*-axis) between a given chain and any other chain in the benchmark dataset. The *y*-axis shows the count of chains that have a given maximal pairwise identity.



Supplementary Figure 4. The pockets identified by the ConCavity, denoted by a pink mesh, for (A) chain A of the Bcr-Abl protein; (B) M2 proton channel.