Structure
Article

Recognizing Protein-Ligand Binding Sites by Global Structural Alignment and Local Geometry Refinement

Ambrish Roy[1,3] and Yang Zhang[1,2,3,*]
[1]Department of Computational Medicine and Bioinformatics
[2]Department of Biological Chemistry
University of Michigan, Ann Arbor, MI 48109-2218, USA
[3]Center for Bioinformatics, University of Kansas, Lawrence, KS 66047, USA
*Correspondence: zhng@umich.edu
DOI 10.1016/j.str.2012.03.009

SUMMARY

Proteins perform functions through interacting with other molecules. However, structural details for most of the protein-ligand interactions are unknown. We present a comparative approach (COFACTOR) to recognize functional sites of protein-ligand interactions using low-resolution protein structural models, based on a global-to-local sequence and structural comparison algorithm. COFACTOR was tested on 501 proteins, which harbor 582 natural and drug-like ligand molecules. Starting from I-TASSER structure predictions, the method successfully identifies ligand-binding pocket locations for 65% of apo receptors with an average distance error 2 Å. The average precision of binding-residue assignments is 46% and 137% higher than that by FINDSITE and ConCavity. In CASP9, COFACTOR achieved a binding-site prediction precision 72% and Matthews correlation coefficient 0.69 for 31 blind test proteins, which was significantly higher than all other participating methods. These data demonstrate the power of structure-based approaches to protein-ligand interaction predictions applicable for genome-wide structural and functional annotations.

INTRODUCTION

Proteins bind with other molecules to bolster or inhibit biological functions. The binding partner, commonly referred to as ligand, can be metal ions, small organic/inorganic molecules, or macromolecules like proteins or nucleic acids. In all these protein-ligand interactions, only a few key residues are involved in the partner recognitions and for the affinity that tethers the ligand to its receptor molecule. Identification of these key residues is imperative for understanding protein's function, analyzing molecular interactions and guiding further experimental procedures (Rausell et al., 2010). Although the experimental determination provides the most accurate assignment of the binding locations, the procedure can be time- and labor-intensive.

Computational approaches to recognize these functional sites in proteins are generally classified into sequence- and structure-based methods. Most of the sequence-based approaches (Capra and Singh, 2007; Pei and Grishin, 2001; Valdar, 2002; Wang et al., 2008) are based on the presumption that functionally important residues are preferentially conserved during the evolution, because natural selection acts on function. In many cases, however, the sequence or evolutionary conservation of residues does not necessarily translate into their involvement in ligand binding, as these residues may play a structural role in maintaining the global scaffold. Nevertheless, the advantage of sequence-based methods is that 3D structure is not a prerequisite and they require negligible time to generate predictions.

Structure-based methods for ligand binding-site identification start with the 3D structure of protein molecules. Most of the early approaches followed the Emil Fisher's assumption that ligand binding in proteins is like "an insertion of key into a lock" (Fischer, 1894); hence shape and physiochemical complementarity are often used to detect concave pockets on proteins surface (Brady and Stouten, 2000; Hendlich et al., 1997; Huang and Schroeder, 2006; Laskowski, 1995; Le Guilloux et al., 2009; Levitt and Banaszak, 1992; Weisel et al., 2007). There are other methods that use calculated interaction energies (Goodford, 1985; Laurie and Jackson, 2005; Wade et al., 1993) or protein structure dynamics (Landon et al., 2008; Lin et al., 2002) to examine the click of "lock and key." With recent increase in number of known protein-ligand complexes in Protein Data Bank (PDB) (Rose et al., 2011), it is becoming evident that homologous proteins with similar global topology often bind similar ligands using a conserved set of residues (Russell et al., 1998). Accordingly, many contemporary methods utilize both geometric match and evolutionary information to identify binding site pockets and residues. Some of them use known protein-ligand complexes as templates (Brylinski and Skolnick, 2008; Glaser et al., 2003; Oh et al., 2009; Tseng and Li, 2011; Wass et al., 2010; Xie and Bourne, 2008), whereas others utilize purely sequence-based homology information (Capra et al., 2009; Huang and Schroeder, 2006; Laskowski, 1995).

Following the sequence-to-structure-to-function paradigm, here we develop a hierarchical approach, COFACTOR, which uses structure modeling and a combined global-and-local

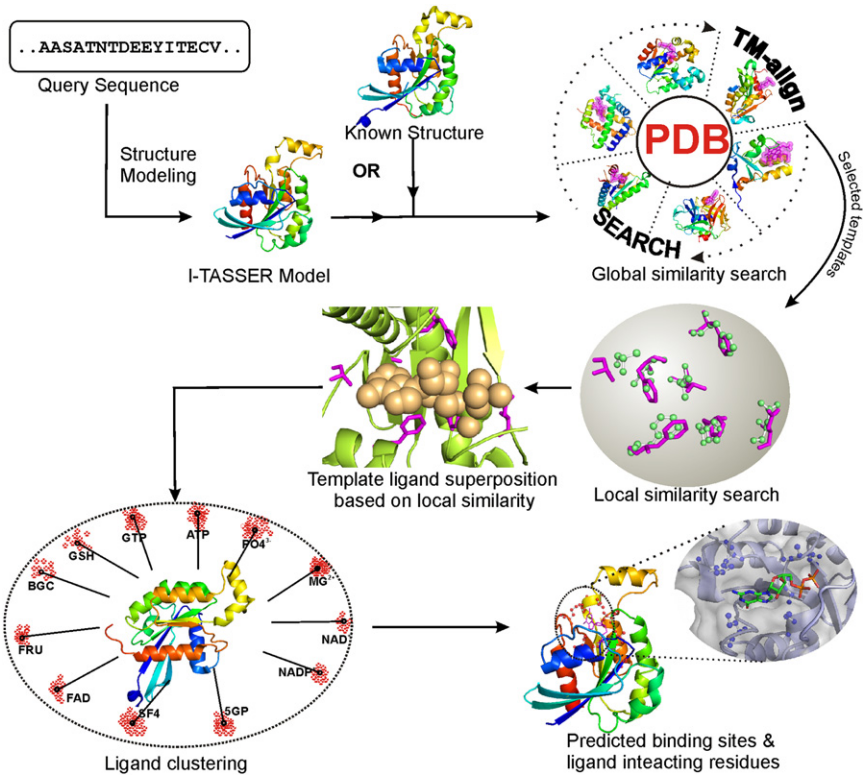

Figure 1. COFACTOR Protocol for Ligand Binding Site Prediction
See also Figure S1.

similarity search scheme to identify binding pockets and ligand-interacting residues in query (target) protein. Figure 1 shows a schematic diagram describing the procedure of COFACTOR algorithm. Starting from the query sequence, the 3D structure model is first generated using the I-TASSER fragment assembly simulations (Roy et al., 2010; Zhang, 2008). Experimental structure can also be used in the following steps. Template proteins with bound ligands in the PDB library are collected based on their global structural similarity to query protein, using the TM-align structure alignment program (Zhang and Skolnick, 2005). Meanwhile, to examine the ligand-binding details, the binding pockets of templates are scanned through the target structure to identify the best local geometric and sequence matches. The binding pose of the ligand in the target structure is predicted based on the local alignment of predicted and template binding site residues. Finally, superposed ligands from multiple templates are clustered to procure the ligand-binding predictions.

The algorithm is evaluated using both the I-TASSER models and the experimental structures of query proteins. Large-scale benchmarking results show that COFACTOR can correctly identify ligand-binding locations for 65%–69% test cases and interacting residues with MCC of 0.55–0.58, for both natural and drug-like molecules. The algorithm was also tested in the recent community-wide CASP9 experiments, where the method outperformed all other participating methods in recognizing ligand binding residues for both metal and nonmetal ligands. The results highlight the potential applicability of the method for genome-scale functional annotations.

RESULTS

Benchmarking of Binding Site Predictions

The performance of protein-ligand binding predictions can be evaluated based on their ability to detect the spatial location of ligand binding pocket and the competency to delineate protein residues that interact with the ligand. In the first evaluation, the prediction errors are evaluated by measuring the spatial distance between the center of the predicted binding pocket and the ligand in experimental structure, whereas in the second one evaluates the assignment accuracy of ligand-interacting residues in the protein sequence. Here, we evaluate COFACTOR on both criteria. The results are controlled by two recently developed structure-based methods, FINDSITE (Brylinski and Skolnick, 2008) and ConCavity (Capra et al., 2009). FINDSITE predicts binding sites by matching the target structure with template proteins identified by threading (Brylinski and Skolnick, 2008), whereas ConCavity assigns binding residues as those closest to the spatial cavities surrounding the protein surface (Capra et al., 2009).

Ligand-Binding Pocket Predictions

The ability of the algorithms to identify ligand-binding pocket is tested on 501 benchmarking proteins, collected from three previous experiments (Dessailly et al., 2008; Hartshorn et al., 2007; Perola et al., 2004), which harbor 582 ligands. The experimental structure of the protein-ligand complexes were collected from the PDB library (Berman et al., 2000).

Figure 2 shows the cumulative fraction of predicted binding pockets as a function of distance between the center of mass of the native ligand and the center of the predicted binding pocket. If we make a cutoff at the pocket distance <4.5 Å, which is close to the average radius of gyration of all ligands in the benchmark set (4.41 Å), the binding pocket predictions by COFACTOR are correct in 65% cases when the low-resolution I-TASSER structure models were used. The control methods FINDSITE and ConCavity correctly predicted binding pocket for 56% and 34% cases, respectively. These differences are statistically significant, where the p value of paired Student's t test for the COFACTOR prediction is 2.3e-6 to FINDSITE and 3.2e-12 to ConCavity results.

Compared to ConCavity, both COFACTOR and FINDSITE are not very sensitive to the accuracy of the protein structure predictions, as long as the global topology of the target model is correct. When the apo-form experimental structures of the target proteins were used, the accuracy of the binding pocket

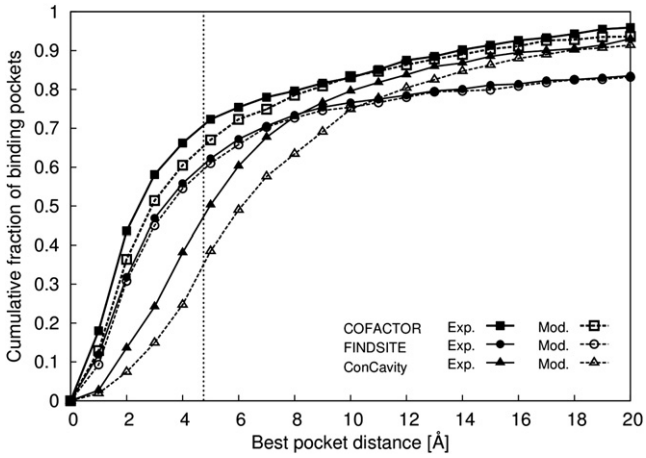

Figure 2. Comparison of Different Methods in Identifying Ligand Binding Pocket Using Either I-TASSER Models or Experimental Structures

Results are presented as the cumulative fraction of predicted binding site pockets versus distance between the center of the native ligand position and the center of the best in top five predicted ligand-binding poses.
See also Table S1.

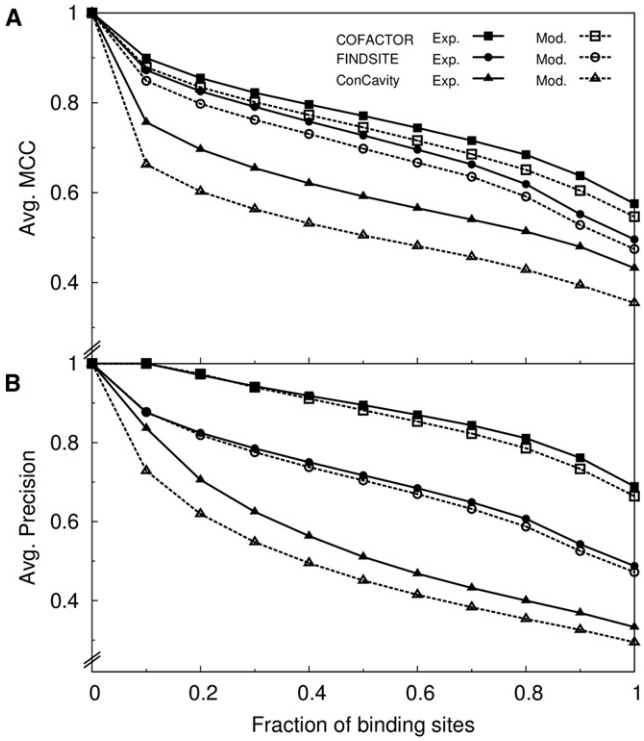

Figure 3. Performance of Different Methods in Detecting Ligand Interacting Residues

(A and B) The data are shown using cumulative data of average Matthews's correlation coefficient (MCC) (A), and average precision of predicted binding sites (B).

predictions by COFACTOR and FINDSITE was only marginally increased to 69% and 59%, respectively, where that of ConCavity was significantly changed from 34% to 45%. This difference in structural sensitivity is probably due to the fact that the cavity-based methods such at ConCavity are sensitive to the local geometry of the target structures, whereas the template-based methods rely more on the global similarity of the target-template topologies. Although homologous templates have been excluded from the I-TASSER template library, the majority of the I-TASSER models (91%) have a correct topology with TM-score >0.5, which explains the independence of the average performance of COFACTOR and FINDSITE on the models chosen of the target structures.

We observed that in 9% of the cases FINDSITE didn't generate any pocket predictions, due to lack of good threading templates in its binding-site library. As a result, ConCavity shows an improved performance over FINDSITE in difficult cases, i.e., ConCavity outperforms FINDSITE in cumulative fraction of binding pocket when the pocket distance increases. If we consider only 447 proteins (with 516 binding sites) where all the three methods successfully generated a prediction, the average binding-pocket distance of the best in top-five predictions by COFACTOR, FINDSITE, and ConCavity using I-TASSER models are 4.7 Å, 5.4 Å, and 7.4 Å, respectively. When the experimental structures are used, the average distance errors are reduced to 4.5 Å, 5.0 Å, and 7.0 Å, respectively. This data shows that for both easy and hard targets the binding pockets identified by COFACTOR are on average closer to the actual binding pocket.

Ligand Binding-Site Residue Assignments

To evaluate the ability of COFACTOR to detect the binding site residues, in Figure 3 we plotted the cumulative data of the average Matthews correlation coefficient (MCC) and precision of the predicted binding residues as a function of the coverage of the predicted binding residues under consideration, where MCC and binding precision were defined in Equations S6 and S7 available online, respectively.

When using the I-TASSER predicted models, COFACTOR can identify binding-site residues for 90% of the targets with an average MCC of 0.60. The average MCC for all targets is 0.55. The average precision of the binding residue prediction is 73% (69%) for 90% (all) targets. Compared to the control methods (FINDSITE and ConCavity), COFACTOR shows an overall improvement of 17%–57% on MCC (Figure 3A), and 46%–137% improvement on the prediction precision (Figure 3B). The reason for the obviously low precision and MCC for ConCavity is that the algorithms defines all the conserved residues lining with the predicted pockets as potential ligand interacting residue, which although increases the recall values (Table 1) but also considerably increases the rate of false positive prediction and results in the low MCC and precision. When using experimental structure, the MCC and precision of the binding site residues by COFACTOR slightly improve to 0.64 (0.58) and 76% (71%), for 90% (all) targets (Figure 3).

This improvement in MCC by COFACTOR is not due to the possible enrichment of analogous structural similarities. Even if we remove the targets for which predictions were generated from templates with a TM-score >0.7, the MCC of COFACTOR prediction is still 26% and 30% higher than that of FINDSITE and ConCavity, respectively.

Table 1. Average MCC, Pre, and Rec of Ligand-Binding Residue Predictions by ConCavity, FINDSITE, and COFACTOR Using I-TASSER Models and Experimental apo Structures as Receptor Structure

			First Prediction			Best in Top Five		
Protein Structure	Ligands (n)	Methods	MCC	Pre	Rec	MCC	Pre	Rec
I-TASSER models	Natural (382)	ConCavity	0.33	0.27	0.58	0.35	0.34	0.62
		FINDSITE	0.42	0.43	0.45	0.47	0.55	0.53
		COFACTOR	0.47	0.58	0.42	0.55	0.70	0.52
	Drug-like (200)	ConCavity	0.32	0.25	0.56	0.34	0.27	0.59
		FINDSITE	0.39	0.37	0.44	0.44	0.41	0.49
		COFACTOR	0.45	0.51	0.40	0.54	0.68	0.50
	Overall (582)	ConCavity	0.33	0.27	0.57	0.35	0.29	0.61
		FINDSITE	0.41	0.41	0.45	0.47	0.47	0.52
		COFACTOR	0.46	0.56	0.41	0.55	0.69	0.51
Experimental structures	Natural (382)	ConCavity	0.40	0.31	0.69	0.43	0.34	0.73
		FINDSITE	0.44	0.44	0.47	0.51	0.51	0.54
		COFACTOR	0.48	0.58	0.43	0.57	0.71	0.54
	Drug-like (200)	ConCavity	0.40	0.30	0.69	0.43	0.32	0.73
		FINDSITE	0.42	0.38	0.47	0.47	0.43	0.52
		COFACTOR	0.47	0.54	0.43	0.58	0.72	0.55
	Overall (582)	ConCavity	0.40	0.30	0.69	0.43	0.33	0.73
		FINDSITE	0.43	0.42	0.47	0.50	0.49	0.54
		COFACTOR	0.48	0.57	0.42	0.58	0.71	0.54

MCC, Matthews's correlation coefficient; Pre, precision; Rec, recall.

Drug-Like versus Natural Ligands

If we define biomolecules binding to enzyme active and allosteric sites as "natural" ligands and artificially designed molecules as "drug-like" ones, 382 out of 582 ligands are classified as natural ligands, whereas the remaining 200 are drug-like in our benchmark set. Based on the results shown in Figure 4A, we find that there is little difference in the average MCC of predicted binding site residues for the different ligand types. The difference becomes notable for prediction precision (Table 1), where ligand interacting residues for natural ligands were predicted with 5%–8% higher precision than for drug-like compounds.

In Figure 4B, we further analyzed the chemical similarity between the predicted ligands by COFACTOR and the native ligands in experimental structure, measured by the Tanimoto coefficient (TC). It is appealing to observe that for ~70% of the proteins with bound natural ligands, the predicted ligands by COFACTOR shared an average chemical similarity (TC) of 0.74, and can therefore be used for a more detailed level elucidation of protein function. For the targets with bound drug-like molecules, even though the predicted residues had an overall high average MCC (54%), close to that of the natural counterpart, the predicted and solved ligands were chemically similar in only 8% cases. This observation recapitulates the fact that the majority of these drug-like molecules are targeted near the active/allosteric sites, where even though they are chemically dissimilar to the substrate molecules, they are tethered by similar set of binding residues. These high accuracy predicted binding site residues by COFACTOR therefore can also be used for creating binding-site based 3D-pharmacophore models for ligand-screening and structure-based drug design even for proteins with unknown structure.

Ligand Shape Comparison

In Table S1 (available online), we compare the shape of the predicted binding pocket/ligand with that of the native ligands (average volume 743 $Å^3$) bound in the experimental structure, as an assessment of predicted ligand conformation. Predicted ligands by COFACTOR, FINDSITE, and ConCavity using the I-TASSER model (experimental structure) have an average Jaccard coefficient (JC) of 0.33 (0.37), 0.27 (0.29), and 0.19 (0.24), respectively, whereas the average volume of ligand/pocket predicted by the three methods are 932 (952), 964 (962), and 2,208 (2,307), respectively. The result demonstrates that although the volume of predicted ligands by COFACTOR are on average smaller, the shape of the predicted ligands matches the best with the native ligands, which is important for shape similarity based studies such as docking and ligand screening (Giganti et al., 2010). Moreover, the average numbers of non-physical protein-ligand clashes are generally fewer in complexes generated by COFACTOR (Table S1).

Confidence Score of Prediction

An estimation of the accuracy of the predictions is important for blind predictions where the answer is unknown, because the accuracy of the predictions essentially decides how the biologist users will use the predictions. The confidence of the predictions in COFACTOR is measured by the C-$score_{LB}$ (see Equation 2 in Experimental Procedures). To examine the correlation of C-$score_{LB}$ with the experimental results, we plot in Figure 5 the average MCC data versus C-$score_{LB}$. The overall Pearson correlation coefficient between C-score and MCC is 0.62. If we use a cutoff of C-score >0.25 and assign MCC >0.5 as correct prediction, the average false positive and false negative rates are 18% and 20%, respectively.

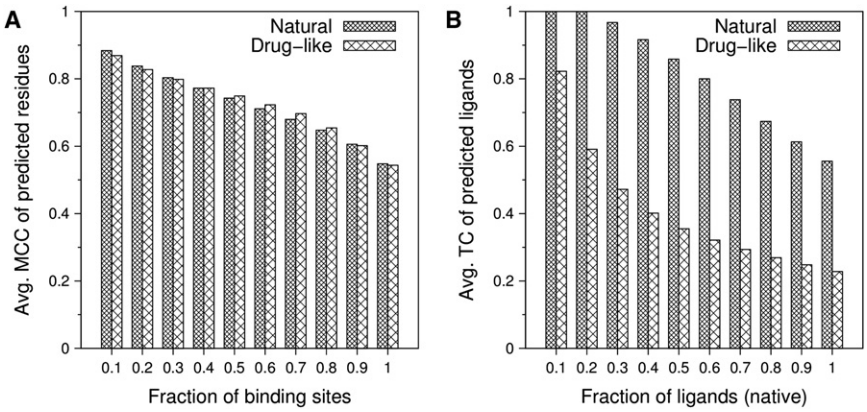

Figure 4. COFACTOR Ligand-Binding Predictions for Natural Ligands and Drug-like Compounds

(A) Cumulative data of Matthews's correlation coefficient (MCC) of predicted ligand interacting residues as a function of the fraction of binding sites.

(B) Chemical similarity between the native bound ligands and the predicted ligands assessed using cumulative average of Tanimoto coefficient (TC). For both analyses, I-TASSER models are used as the apo receptor structure.

As a control, we also present the data of FINDSITE that uses the fraction of templates sharing the same pocket as the confidence score (Brylinski and Skolnick, 2008), where the correlation is 0.21. Apparently, the combination of the global and local similarities based on both sequence and structure comparisons help increase the sensitivity of C-score$_{LB}$ to the quality of the predictions.

Blind Test of COFACTOR in CASP9

The ninth community-wide critical assessment of techniques for protein structure prediction (CASP9) released 129 target protein sequences for blind test of protein structure and function prediction methods. The function prediction section was focused on evaluating the ligand binding-site predictions, where the predictors were asked to identify ligand-interacting residues in the provided protein sequence.

During CASP9, we first generated the 3D structural models using I-TASSER and the structure-based ligand binding site predictions were generated using the COFACTOR algorithm. Although we generated predictions for all the 129 targets, only 31 proteins were solved in their holo form and were used in the official assessment (Schmidt et al., 2011). The definition of the binding site residues in our analysis follows the CASP9 assessor's rendition. The COFACTOR prediction results on the 31 proteins are listed in Table 2. Overall, the models by COFACTOR

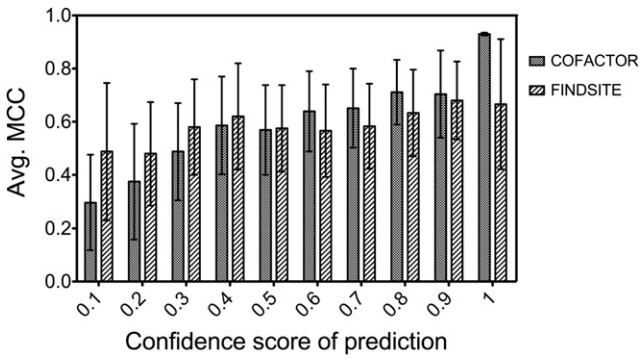

Figure 5. Distribution of MCC of Predicted Binding Residues as a Function of Prediction Confidence Score

The histogram shows the mean MCC of predicted binding site residues and the error bar represents SD.

(named "I-TASSER_FN" in the server section and "Zhang" in the human section) were ranked at the top two positions based on the mean MCC *Z*-scores with and without bootstrapping experiment (Figure S3). As CASP9 assessors concluded, among all 33 participant groups "Two groups (FN096, Zhang; FN339, I-TASSER_FUNCTION) performed better than the rest, while the following ten prediction groups performed comparably well." (Schmidt et al., 2011).

Overall, for the 31 evaluated proteins, the binding-site residues were predicted with an average MCC of 69%, which is slightly higher than the above benchmark test because CASP9 has more easy targets (Schmidt et al., 2011). For the best 24 proteins, more than 50% ligand interacting residues were correctly identified. We observed that most of the high precision predictions are for binding-sites harboring nonmetal ligands (average precision of 75.5%), whereas the binding-site residues for metal ions have a slightly lower average precision 69.8%. The metal ion binding residues also show large variations in their prediction recall. One of the major reasons for the moderate metal-involved predictions is the relatively lower quality of receptor models. The average TM-score is 0.66 ± 0.21 for the metal-bound proteins whereas that for nonmetal proteins is 0.82 ± 0.12. Also, in some of these metal-binding proteins COFACTOR additionally predicted nonmetal ligand binding sites (for example $PO4^{3-}$ in T0635) and was the source of overprediction. Nevertheless, similar to observations in the benchmarking analysis, in most of the cases, the predicted and native ligands are highly similar, implying the applicability of COFACTOR for a more detailed elucidation of protein function.

Figure 6 shows two representative examples of easy and hard test cases, T0609 and T0518, for which COFACTOR predictions significantly outperformed other groups. Target T0609 (PDB ID: 3os7) is a putative galactose mutarotase crystallized with tartaric acid. Although the crystal structure was solved without the native ligand, the CASP9 assessors inferred that the protein binds β-D-galactose (GAL) in the same binding cleft as the crystallized tartaric acid. Figure 6A shows the successful prediction (MCC = 0.82, accuracy = 0.75) by COFACTOR for this target, where four of the five binding site residues were correctly identified (shown in green). This prediction was deduced from a distant homolog protein of Gal10 bifunctional protein (PDB ID: 1z45) from *Saccharomyces cerevisiae*, which also binds GAL. Most groups in CASP9 missed the prediction because the template by threading has

Table 2. Binding Site Predictions by COFACTOR for 31 CASP9 Targets

Target	TM-Score[a]	Native Ligand(s)	Predicted Ligand(s)	C-score$_{LB}$	MCC	Pre	Rec
T0515[b]	0.89	PLP, LYS	ORX, PLP	0.61, 0.45	0.68	0.64	0.75
T0516	0.89	PF1	PF1, HMH	0.79, 0.88	0.84	0.85	0.85
T0518	0.80	NA	CA, MN	0.41, 0.39	0.38	0.38	0.43
T0521	0.52	2 CA	4 CA	0.67, 0.76, 0.66, 0.60	0.08	0.10	0.22
T0524[b]	0.87	GAL	GAL	0.75	0.66	0.73	0.62
T0526[b]	0.88	GLA	GAL	0.55	0.46	0.42	0.56
T0529	0.23	MN	ZN, AMP	0.72, 0.23	0.55	0.31	1.00
T0533	0.79	PHE	2 PHE	0.88, 0.09	0.88	1.00	0.79
T0539	0.64	ZN, ZN	ZN, ZN	0.85, 0.77	1.00	1.00	1.00
T0547[b]	0.71	PLP, LYS	PLP, LYS, AZ1, ORX, P3T	0.61, 0.61, 0.61, 0.54,0.54	0.77	0.74	0.82
T0548	0.56	ZN	SAL, ZN	0.21, 0.67	0.69	0.50	1.00
T0565[b]	0.74	DGL, ALA	DLG, ALA, UNL	0.88, 0.50, 0.52	0.86	1.00	0.75
T0570	0.88	MG, GOL	CA, GOL, PO4	0.83, 0.21, 0.34	0.87	0.88	0.88
T0582	0.85	ZN	ZN	0.64	1.00	1.00	1.00
T0584[b]	0.83	IPR, DST	IPR, RIS, MG, MG, PO4	0.51, 0.25, 0.68, 0.75, 0.58	0.75	0.63	0.92
T0585	0.78	ZN	ZN	0.85	0.77	1.00	0.60
T0591	0.89	LLP	PLP, PLP	0.83, 0.81	0.76	0.65	0.91
T0597	0.86	ANP	MG, ATP, AMP	0.93, 0.83,0.80	0.70	0.80	0.63
T0599[b]	0.95	ISC	MG, ISC	0.88, 0.83	0.83	0.75	0.92
T0604	0.41	FAD	FAD	0.72	0.45	0.54	0.42
T0607[b]	0.86	ZN, ZN, BES	MN, MN, BIB	0.93, 0.83, 0.68	0.50	0.71	0.36
T0609[b]	0.78	GAL	GAL	0.74	0.82	0.75	0.90
T0613[b]	0.96	GAR, NHS	UNL, THH	0.48, 0.58	0.70	0.77	0.67
T0615[b]	0.71	MN, GPX	MN, PO4	0.83, 0.77	0.50	0.83	0.33
T0622[b]	0.69	NAD	NAD,ATP	0.66, 0.71	0.76	0.67	0.93
T0625	0.74	ZN	ZN	0.79	1.00	1.00	1.00
T0629	0.34	6 FE	ZN	0.45	0.37	1.00	0.14
T0632	0.74	COA	COA, PHB	0.68,0.60	0.46	0.67	0.38
T0635	0.91	CA	MG, PO4	0.96, 0.90	0.60	0.38	1.00
T0636[b]	0.93	HAS, PLP	HAS, PMP, PMP	0.51, 0.32, 0.51	0.79	0.78	0.82
T0641	0.91	STE	PLM	0.82	0.83	0.80	0.89
Average					0.69	0.72	0.72

MCC, Matthews's correlation coefficient; PDB, Protein Data Bank; Pre, precision; Rec, recall.
[a]TM-score of I-TASSER models for the target protein.
[b]Holo structure of these proteins was solved with nonnative ligand and the native ligand binding information was inferred by CASP9 assessors from homologous PDB structures.

a poor alignment quality; while COFACTOR used the I-TASSER full-length models (TM-score = 0.78), which correctly detected the template with correct alignment by TM-align. This is an example showing the advantage of COFACTOR by using a better quality of receptor models by I-TASSER.

T0518 (PDB ID: 3nmb) is a putative sugar hydrolase crystallized with sodium ion. Although the receptor was an easy target for structure modeling (TM-score of I-TASSER model is 0.80) and a close homolog (PDB ID: 3imm) had a very similar Na^+ binding site, most predictors in CASP9 failed to predict the binding site because Na^+ was considered a crystallization artifact. The COFACTOR template library also missed this template protein. However, a local similarity was detected between the I-TASSER model and peanut-lectin (PDB IDs: 2dv9 and 2tep). Two binding sites for Mn^{2+} and Ca^{2+} were then predicted by COFACTOR although with a low confidence score in the same binding cleft. Out of the seven native ligand-binding residues (Figure 6B), three residues were correctly identified (shown in green). Five were incorrectly annotated as binding residues (shown in red), whereas four correct residues (shown in yellow) were missed during the prediction. Nonetheless, T0518 represents a typical successful example, where although a close template was not present in the template library, COFACTOR correctly identified a remote homolog of the protein using local comparisons and provided a reasonable prediction that could be useful for understanding the function.

Why Does COFACTOR Work?

An important question is: why COFACTOR outperforms most of the state-of-the-art methods in the overall binding site prediction

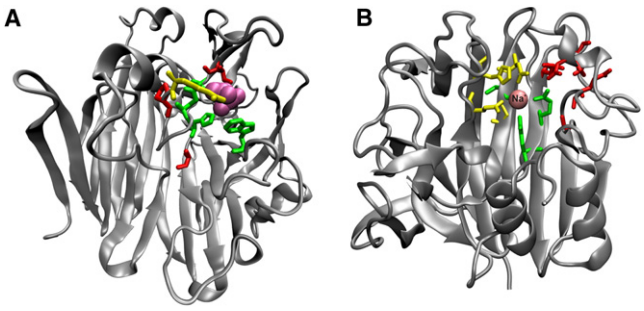

Figure 6. Examples of Successful Predictions by COFACTOR in CASP9

Models in (A) and (B) are from T0609 and T0518, respectively. Correctly predicted residues are shown in green (true positive), false positive predictions highlighted in red, and false negatives residues shown in yellow. The overall ranking results of all targets in CASP9 can be seen in Figure S3.

accuracy, although both COFACTOR and these other methods have exploited the sequence and structural information in their predictions?

In Figure 7A, we analyzed the dependence of binding pocket predictions by COFACTOR and the two control methods (FINDSITE and ConCavity) on the accuracy of predicted receptor structure. For clearness, the data set in Figure 7 includes only those proteins on which the three methods perform differently. A more complete version of the data is presented in Figure S4 that contains all protein targets, including those on which the three methods are all successful and failed. The local structure quality of predicted receptors is evaluated by the root-mean-square deviation (RMSD) of known ligand binding residues, whereas that of global structure is measured by the RMSD of full-length receptor models. For targets with approximately correct global topology (RMSD <8 Å), all three methods have a reasonable ability to predict the ligand binding pocket. Nevertheless, COFACTOR generates 15% and 92% more correct (distance error <4.5 Å) binding pocket predictions than FINDSITE and ConCavity (Figure 7A, inset), respectively. Moreover, in these correct predictions, the average distance error of pocket prediction by COFACTOR is lower (1.9 Å), compared to that by FINDSITE (2.1 Å) and ConCavity (3.0 Å), which highlights the fact that a combination of local and global structural alignment improves the accuracy of binding site predictions for easy modeling proteins.

Even for the harder cases, when the global topology of the receptor models is incorrect (global RMSD >8 Å) but the ligand binding pocket is correctly formed (local RMSD <8 Å), COFACTOR had 13% and 94% more correct predictions, compared to the control methods (lower-right area of Figure 7A), respectively. Because the topology of the receptor models is incorrect, methods that rely only on global comparisons will have difficulty to identify the correct template, which was improved in COFACTOR by using local structural comparisons.

In Figure 7B, we analyzed the performance of COFACTOR in relation to global and local similarity between target and template structures. When target and template proteins have a similar fold (TM-score >0.5) and the local match near the binding pockets are significant (BS-score >1.0), i.e., upper-right region of Figure 7B, in 80% cases the predictions generated by COFACTOR were correct and the average distance error was 1.81 Å. Conversely, for protein that use template proteins of the same fold but the local match was relatively poorer (BS-score <1.0, the lower-right region of Figure 7B), the prediction accuracy rapidly decreased to 53% and ligand distance error increased to 2.3 Å. This highlights the sensitivity of local structural comparisons for selecting templates in template-based

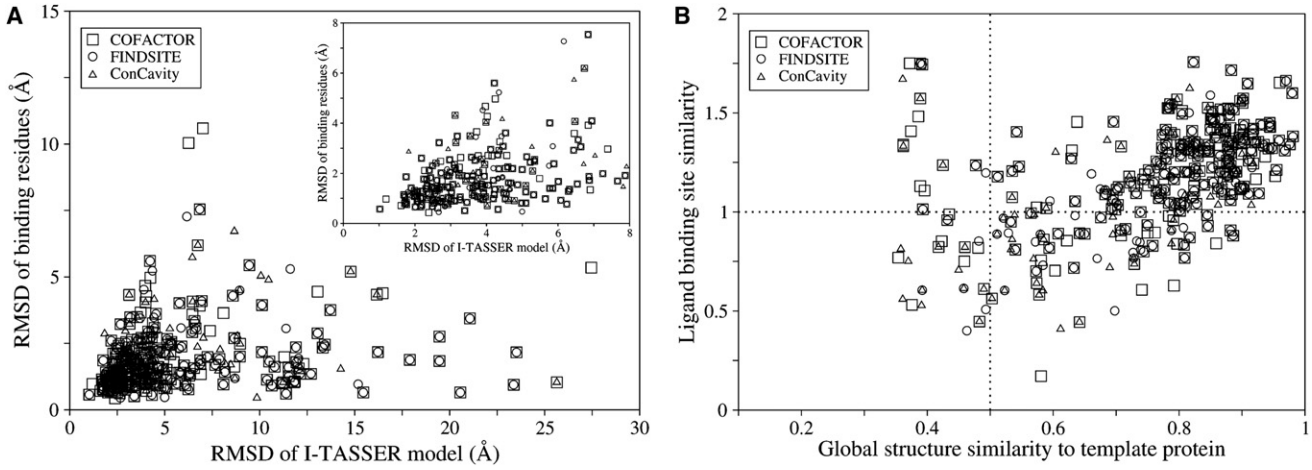

Figure 7. Influence of Local and Global Protein Structure Modeling on the Accuracy of Ligand Binding Site Predictions

(A) Structural accuracy of ligand binding residues versus the accuracy of full-length receptor models. Ligand binding pocket predictions using higher resolution receptor models are shown in the inset.

(B) Local versus global similarity of template to target structures. The local similarity is evaluated by BS-score (Equation 1), whereas global structural similarity is measured by TM-score of template and the I-TASSER model. In both the plots, the correct predictions with a distance error <4.5 Å by different methods are represented by different symbols. For clarity, data points of binding pockets for which either all the methods correctly identified the pocket (128 cases) or all the methods failed to identify the pocket (147cases) have been omitted.

See also Figures S2 and S4.

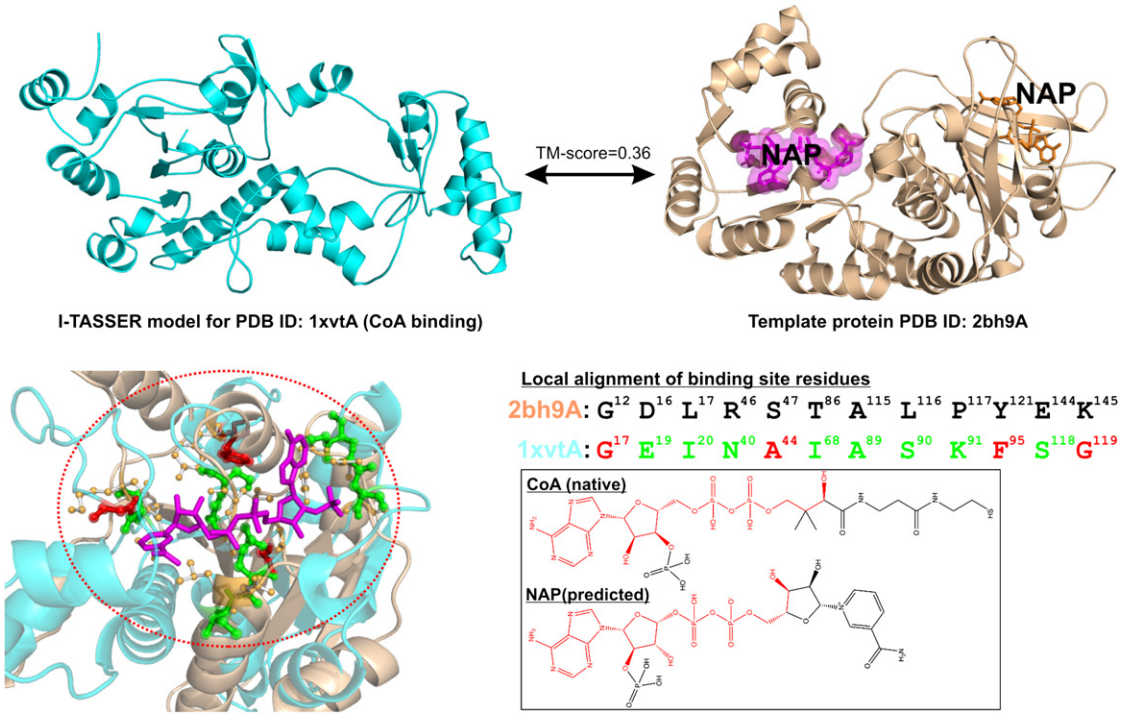

Figure 8. A Representative Example of COFACTOR Binding-Site Prediction Based on Local Structural Comparisons

Binding site residues of the carnitine CoA-transferase (PDB ID: 1xvtA) was detected using glucose-6-phosphate dehydrogenase (PDB ID: 2bh9A) as template with MCC of 56% and precision of 75%. The NAP binding site in N-terminal domain of 2bh9A (ligand shown in magenta) was used for the prediction. The overall TM-score of two structures is 0.36 (TM-score = 0.24 if only the binding domain of 1xvtA (4–330) and 2bh9A (27–199) is considered). The true positive residues are shown in green and false positive ones are in red. Inset shows that CoA (native ligand) and NAP (predicted ligand) have similar chemical structure (adenine and ribo-phosphate moiety shown in red). No local similarity was detected using the C-terminal NAP (shown in orange) binding site of template.

binding site prediction methods in addition to the global structural similarity. Nevertheless, if we completely ignore the global similarity (TM-score and ID_{str}) from C-$score_{LB}$, the percentage of the correctly predicted binding pocket is reduced from 65% to 59% with the average distance error increasing from 1.9 Å to 2.06 Å. Similarly, if we completely ignore the local similarity search and use TM-align alignment for binding pocket prediction, the percentage of correct predictions decreases to 48% and the average distance error increases to 2.72 Å. Thus, both the global and local comparisons are important in binding-site recognitions.

We further examine cases in the upper left region of Figure 7B that is most interesting because the templates used by COFACTOR have a different fold from the query model (TM-score <0.5). When a good local match near the binding pocket is identified (i.e., BS-score >1), the binding pocket prediction is correct in 75% cases, which is 88% and 67% higher than the control methods FINDSITE and ConCavity, respectively, in the same region. Apparently the advantage of algorithm on the proteins in this category contributes the most to competition of COFACTOR to these two methods.

A further analysis of all the predictions based on templates of different folds reveals that the average sequence similarity between the target and template binding site residues is 56 ± 27% for the correctly predicted targets, whereas that for the failed predictions is only 35 ± 19%. The average structural similarity (measured using left-hand term in Equation 1) of the local binding motifs for the correctly predicted cases are relatively more conserved (0.66 ± 0.21), than for incorrect predictions (0.45 ± 0.20). These data suggest that both ligand binding residues and the spatial position of the residues have been highly preserved in functional sites during evolution, even though the overall structural similarity has dwindled. Therefore, a combination of both structural and sequence similarity in the local pocket comparison is essential.

In Figure 8, we show a successful example from carnitine CoA-transferase (PDB ID: 1xvtA), which demonstrates the strength of local structural matches. In this example, the correct template protein is from the glucose-6-phosphate dehydrogenase (PDB ID: 2bh9A) that has, however, a completely different overall fold with a TM-score to the target 0.36 (Figure 8). Nevertheless, the structure of both template and target contains a pocket with three-layer (aba) sandwich architecture in their N-terminal region, which forms a NADP+ (bound NAP in 2bh9A) binding site in glucose-6-phosphate dehydrogenase and a CoA binding site in carnitine CoA-transferase. Although there is no global structural similarity, COFACTOR identifies this local structural similarity of the two proteins with a high BS-score, which results in predicted ligand-binding residues with an MCC of 56% and precision of 75%. The predicted ligand (NAP) for the query contains the same adenine and ribo-phosphate moiety as "native" ligand (bound CoA in 1xvtA).

All the data of COFACTOR ligand binding prediction presented in Figures 2, 3, 7, and 8 using the I-TASSER models, as well as the template distributions for each entry, are listed on our web page at http://zhanglab.ccmb.med.umich.edu/COFACTOR/benchmark.

DISCUSSION

A hierarchical approach, COFACTOR, for high accuracy prediction of protein-ligand interaction has been developed. Anatomy of results obtained on a large-scale data set containing functionally diverse proteins, shows that the algorithm could accurately identify binding pockets in 65% of cases with an average error of 2 Å, when predicted protein structures were used and homologous templates were completely excluded from both structure and protein-ligand template libraries. In 90% of the cases, without knowing the ligand a priori, the ligand interacting residues were assigned with an average Matthews correlation coefficient of 60% and precision of 73%.

We have analyzed the predicted binding sites for both natural and drug-like molecules, but no significant difference was observed between the predictions for the two classes of molecules. In particular, for 70% of the proteins with bound natural ligand, the predicted ligand shared a high chemical similarity to the bound ligand in native state, which suggests a potential application of the method for a more elaborate functional elucidation of uncharacterized proteins. Successful predictions were also observed for drug-like compounds, which open up the possibility for structure-based drug design even for proteins that have no structural information.

We have compared our benchmarking results with two recently developed structure-based methods (FINDSITE and ConCavity). Starting from the same set of structural models, the MCC of ligand-binding residue predicted by COFACTOR is 17% and 57% higher than that by FINDSITE and ConCavity, respectively, whereas the distance error in locating ligand-binding pocket by COFACTOR is 0.7 Å and 2.7 Å lower than that by the aforementioned two control methods. In the recent community-wide CASP9 experiment (Schmidt et al., 2011), COFACTOR achieved an average MCC 0.69 and precision 0.72, which significantly outperforms all other methods from 33 participating groups (Figure S3).

The major advantage of COFACTOR over the existing methods is the optimal combination of global and local structural comparisons for identifying ligand-binding sites. First, it outperforms the popular cavity-based methods (Capra et al., 2009; Laskowski et al., 2005; Sael and Kihara, 2010) in the cases when only low-resolution protein models are available, because global topology comparisons can reliably identify the correct functional templates as their accuracy is not sensitive to the local structural errors. Second, for proteins that have functional templates with different global topology but similar conserved binding pockets, local structural comparisons help COFACTOR to correctly recognize the ligand-binding residues, which cannot be achieved by the purely global structural comparison methods (Brylinski and Skolnick, 2008; Oh et al., 2009; Wass et al., 2010).

The latter advantage of local structural comparison is particularly important for functional annotations of proteins in the so-called "twilight-zone" regions, where the protein structure prediction methods often have difficulties in generating correct global fold due to the lack of appropriate templates. However, many methods, including I-TASSER (Roy et al., 2010; Zhang, 2007), can almost always generate models with correct super-secondary structures (Ben-David et al., 2009; Jauch et al., 2007), especially in the functionally conserved regions, which provide important insight for local-structure based functional inferences. Thus, combining the presented method with the state-of-the-art protein structure predictions represents an automated and optimal method for genome-wide structural and functional annotations for the majority of the proteins that lack experimental structures.

A couple of improvements are planned for further development of COFACTOR algorithm. First, the algorithm currently uses Needleman-Wunsch (NW) dynamic programming (Needleman and Wunsch, 1970) as the search engine to identify the best local match between target and the template proteins. Because the NW alignment is sequence-order dependent, it may limit the applicability of the algorithm to the broader range of functional sites because the spatial order of ligand-binding residues is often different from the sequential order. Developing a sequence-order independent search engine will help identify these cases. Second, the current COFACTOR prediction is based on the comparison analysis of monomer chains, although in many cases active/binding sites are located at protein-protein interfaces. Although all the ligand-binding templates (regardless of their interaction status) are included as monomers in the COFACTOR library and the ligand-binding from protein-protein interactions can be in principle predicted by the current algorithm if monomer similarity is sufficiently high, the inclusion of the complex structures in the comparisons may further improve the precision and recall of the algorithm.

EXPERIMENTAL PROCEDURES

For a target protein, the structure models are first generated by the automated I-TASSER structural assembly method (Roy et al., 2010; Wu et al., 2007). The ligand binding information is then derived from the known proteins (templates) in a comprehensive protein-ligand complex library, where the best templates are identified using both global and local structure comparisons between the target and template proteins. In the benchmarking test, to exclude the contamination of homologous proteins, all templates having a sequence identity >30% to the target, were removed from both our structure and function libraries. A flowchart of the COFACTOR algorithm is shown in Figure 1, where a detailed description is provided in Figure S1 and the related discussions in Supplemental Information.

The global structure match is performed by TM-align (Zhang and Skolnick, 2005), which identifies the best alignment between the target and template structures by a heuristic dynamic programming iteration using TM-score rotation matrix. A TM-score (Zhang and Skolnick, 2004), with the value in [0, 1], is reported to assess the global structural similarity. All template proteins with a nonrandom structural similarity (i.e., TM-score >0.3) to the target structure (Xu and Zhang, 2010) (or up to top 100 templates if less than 100 templates have such TM-score, which rarely happen) are selected for further processing.

The local match between the target and template proteins is conducted in two steps (Figure S1). The first step is to identify a set of conserved residues in target that are used as the seed of local structure comparisons. For this purpose, multiple sequence alignment (MSA) of the query target sequence is constructed by PSI-BLAST (Altschul et al., 1997) through the NCBI nonredundant (NR) sequence database. Conserved residues in query sequence are then identified from the MSA based on their Jensen-Shannon divergence score (Capra and Singh, 2007). Triplets of these conserved residues (noted

as *a*, *b*, *c*), along with their two flanking residues, are used for generating initial candidate binding-site motifs. This is based on the fact that residues lining the ligand binding pocket are evolutionarily more conserved than the rest of the sequence (Valdar, 2002); therefore by generating the motifs using only evolutionarily conserved residues, the search space is largely reduced. Similarly, for any given template protein (*t*) with known binding site (*b*), motifs are generated by selecting ligand-interacting residue triplets (l_{tb}, m_{tb}, n_{tb}, see Figure S2).

In the second step, the structure of each of the candidate binding site motifs (*a*, *b*, *c*) is superposed on the template motif (l_{tb}, m_{tb}, n_{tb}). The rotation and translation matrix acquired from this local superimposition is used to bring the complete structure of query and template proteins together. A sphere of radius *r* is then defined around the geometric center (C_{tb}) of template motif, where *r* is the maximum distance of template binding site residues from C_{tb} (Figure S2A). The sphere here defines a local environment, under which the compatibility of query and template to bind similar ligand is compared, based on the sequence and structure similarity of residues lining the pocket. The query-template alignment within the selected sphere area provides an initial seed alignment, which is refined further using a iterative NW dynamic programming (Gotoh, 1982). The alignment score S_{ij} during this iteration is given by

$$s_{ij} = \frac{1}{1+\left(\frac{d_{ij}}{d_0}\right)^2} + M_{ij},$$

where d_{ij} is the C_α distance between *i*th residue in the query and *j*th residue in the template, $d_0 = 3$ Å is the distance scaling factor, and M_{ij} is the substitution score between *i*th and *j*th residues taken from BLOSUM62 matrix. For each alignment, a raw alignment score is defined for evaluating the binding site similarity (BS-score), given by

$$\mathrm{BS-score} = \frac{1}{N_t}\sum_{i=1}^{N_{ali}} \frac{1}{1+\left(\frac{d_{ii}}{d_0}\right)^2} + \frac{1}{N_t}\sum_{i=1}^{N_{ali}} M_{ii}, \quad (1)$$

where N_t represents the total number of template residues in the binding site sphere and N_{ali} is the number of aligned residue pairs in the sphere. This procedure is repeated until the final alignment is converged. This local search procedure is performed for all possible candidate binding site motifs (*a*, *b*, *c*) and known binding site residues triplets (l_{tb}, m_{tb}, n_{tb}). It should be noted that the first step PSI-BLAST based conservation analysis was used only to generate initial candidate motifs and the final binding sites can be completely different from the initial assignment dependent on the local structure comparisons.

For each template binding site (*b*), the region that gives the highest BS-score is recorded as the corresponding predicted binding site in the query, and the residues aligned with known binding site residues in the template are assigned as the binding site residues in target. As the ligand copied directly from the template may have overlaps with the target structure, a quick Metropolis Monte-Carlo simulation is performed for each inferred ligand to improve the local geometry by maximizing the number of contacts between ligand and predicted residues, meanwhile minimizing the protein-ligand overlaps.

The predicted ligand conformations from all the templates are clustered based on their spatial proximity with a distance cutoff 8 Å. If a binding pocket binds multiple ligands (e.g., an ATP binding pocket may also bind MG, $PO4^{3-}$, and ADP), ligands within the same pocket were clustered further based on their chemical similarity using Tanimoto coefficient. Finally, the model with highest ligand-binding confidence score (C-$score_{LB}$) among all the clusters is selected, which is defined as:

$$\mathrm{C-score_{LB}} = \frac{2}{1+e^{-\left(\frac{N}{N_{tot}}\times\left(0.25\mathrm{BS-score}+\mathrm{TM-score}+2.5\mathrm{ID_{Str}}+\frac{2}{1+\langle D\rangle}\right)\right)}} - 1, \quad (2)$$

where N is the multiplicity of ligand decoys in the cluster and N_{tot} is the total number of predicted ligands using the templates. BS-score defined in Equation 1 and TM-score measure local and global similarity of the target to the template protein, respectively. ID_{str} is sequence identity between the target and the template in the structurally aligned region. <*D*> is the average distance of the predicted ligand to all other predicted ligands in the same cluster. C-$score_{LB}$ represents a combined score of the cluster size, and local and global similarities of sequence and structure between target and functional templates.

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, one table, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.str.2012.03.009.

ACKNOWLEDGMENTS

The project is supported in part by NSF Career Award (DBI 1027394), and National Institute of General Medical Sciences (GM083107 and GM084222).

Received: November 5, 2011
Revised: March 20, 2012
Accepted: March 26, 2012
Published online: May 3, 2012

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Ben-David, M., Noivirt-Brik, O., Paz, A., Prilusky, J., Sussman, J.L., and Levy, Y. (2009). Assessment of CASP8 structure predictions for template free targets. Proteins *77* (*Suppl 9*), 50–65.

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

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

Brylinski, M., and Skolnick, J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc. Natl. Acad. Sci. USA *105*, 129–134.

Capra, J.A., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. Bioinformatics *23*, 1875–1882.

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*, e1000585.

Dessailly, B.H., Lensink, M.F., Orengo, C.A., and Wodak, S.J. (2008). LigASite—a database of biologically relevant binding sites in proteins with known apo-structures. Nucleic Acids Res. *36* (Database issue), D667–D673.

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. Berichte der deutschen chemischen Gesellschaft *27*, 2985–2993.

Giganti, D., Guillemain, H., Spadoni, J.L., Nilges, M., Zagury, J.F., and Montes, M. (2010). Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. J. Chem. Inf. Model. *50*, 992–1004.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics *19*, 163–164.

Goodford, P.J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J. Med. Chem. *28*, 849–857.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. J. Mol. Biol. *162*, 705–708.

Hartshorn, M.J., Verdonk, M.L., Chessari, G., Brewerton, S.C., Mooij, W.T., Mortenson, P.N., and Murray, C.W. (2007). Diverse, high-quality test set for the validation of protein-ligand docking performance. J. Med. Chem. *50*, 726–741.

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*, 359–363, 389.

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

Jauch, R., Yeo, H.C., Kolatkar, P.R., and Clarke, N.D. (2007). Assessment of CASP7 structure predictions for template free targets. Proteins *69* (*Suppl 8*), 57–67.

Landon, M.R., Amaro, R.E., Baron, R., Ngan, C.H., Ozonoff, D., McCammon, J.A., and Vajda, S. (2008). Novel druggable hot spots in avian influenza neuraminidase H5N1 revealed by computational solvent mapping of a reduced and representative receptor ensemble. Chem. Biol. Drug Des. *71*, 106–116.

Laskowski, R.A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J. Mol. Graph. *13*, 323–330, 307–328.

Laskowski, R.A., Watson, J.D., and Thornton, J.M. (2005). ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res. *33* (Web Server issue), W89–W93.

Laurie, A.T., and Jackson, R.M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics *21*, 1908–1916.

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

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

Lin, J.H., Perryman, A.L., Schames, J.R., and McCammon, J.A. (2002). Computational drug design accommodating receptor flexibility: the relaxed complex scheme. J. Am. Chem. Soc. *124*, 5632–5633.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

Oh, M., Joo, K., and Lee, J. (2009). Protein-binding site prediction based on three-dimensional protein modeling. Proteins *77* (*Suppl 9*), 152–156.

Pei, J., and Grishin, N.V. (2001). AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics *17*, 700–712.

Perola, E., Walters, W.P., and Charifson, P.S. (2004). A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins *56*, 235–249.

Rausell, A., Juan, D., Pazos, F., and Valencia, A. (2010). Protein interactions and ligand binding: from protein subfamilies to functional specificity. Proc. Natl. Acad. Sci. USA *107*, 1995–2000.

Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Westbrook, J.D., et al. (2011). The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res. *39*, 392–401.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. *5*, 725–738.

Russell, R.B., Sasieni, P.D., and Sternberg, M.J. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. J. Mol. Biol. *282*, 903–918.

Sael, L., and Kihara, D. (2010). Binding ligand prediction for proteins using partial matching of local surface patches. Int. J. Mol. Sci. *11*, 5009–5026.

Schmidt, T., Haas, J., Cassarino, T.G., and Schwede, T. (2011). Assessment of ligand binding residue predictions in CASP9. Proteins *79* (*Suppl 10*), 126–136.

Tseng, Y.Y., and Li, W.H. (2011). Evolutionary approach to predicting the binding site residues of a protein from its primary sequence. Proc. Natl. Acad. Sci. USA *108*, 5313–5318.

Valdar, W.S. (2002). Scoring residue conservation. Proteins *48*, 227–241.

Wade, R.C., Clark, K.J., and Goodford, P.J. (1993). Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. J. Med. Chem. *36*, 140–147.

Wang, K., Horst, J.A., Cheng, G., Nickle, D.C., and Samudrala, R. (2008). Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information. PLoS Comput. Biol. *4*, e1000181.

Wass, M.N., Kelley, L.A., and Sternberg, M.J. (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res. *38* (*Web Server issue*), W469–W473.

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

Wu, S., Skolnick, J., and Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. *5*, 17.

Xie, L., and Bourne, P.E. (2008). Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. Proc. Natl. Acad. Sci. USA *105*, 5441–5446.

Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics *26*, 889–895.

Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. Proteins *69* (*Suppl 8*), 108–117.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. BMC Bioinformatics *9*, 40.

Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins *57*, 702–710.

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302–2309.