SUPPLEMENTAL INFORMATION

Supplemental Data

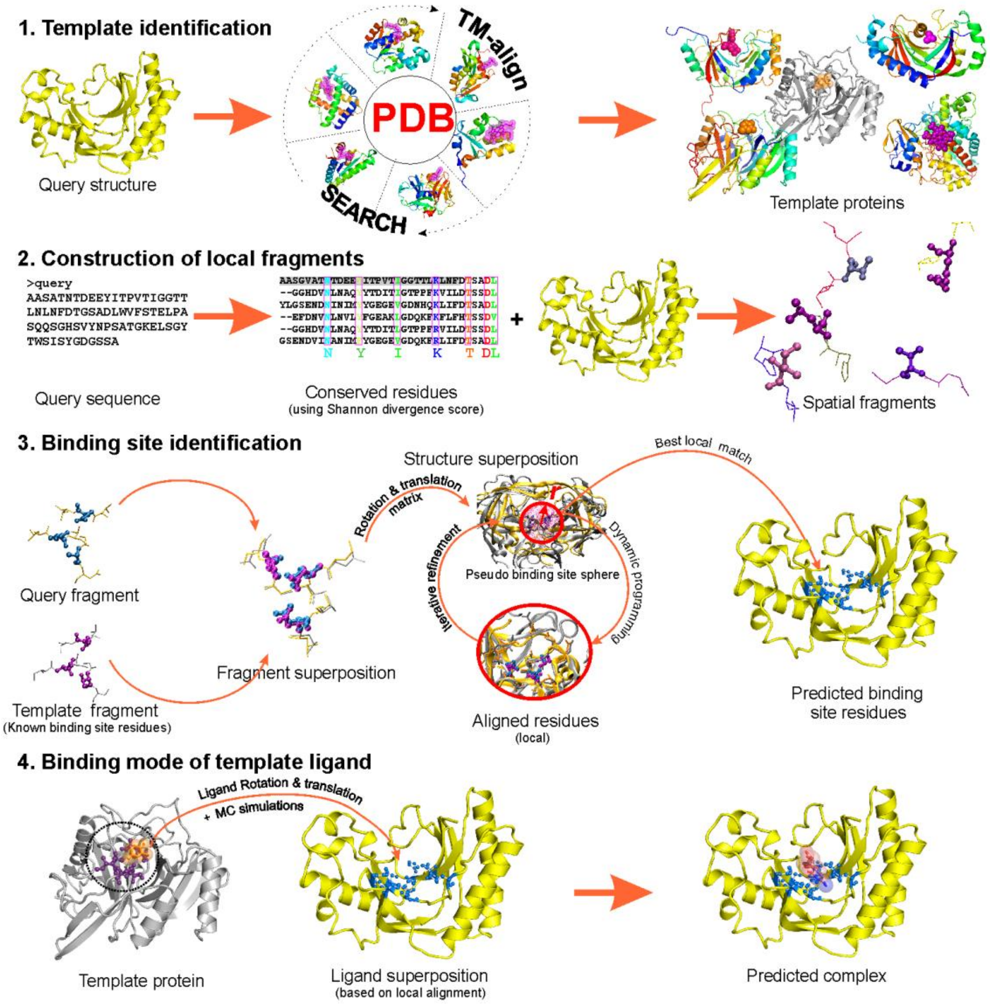
1. Template identification
Query structure
TM-align
PDB
SEARCH
Template proteins
2. Construction of local fragments
>query
AASATNTDEEYITPVTIGGTT
LNLNFDTGSADLWVFSTELPA
SQQSGHSVYNPSATGKELSGY
TWSISYGDGSSA
Query sequence
AASGVATNTDEEYITPVTIGGTTLKLNFDTSADL
--GGHDVNLNAQYYTDITIGTPPFKVILDTSSDL
YLGSENDNINIMYYGEGEVGDNHQKLIFDTSADL
--EFDNVNLNVLYFGEAKIGDQKFKFLFHTSSDV
--GGHDVNLNAQYYTDITLGTPPFRVILDTSSDL
GSENDVINANIMYYGEGEVGDQKFRLIFDTSADL
N Y I K T DL
Conserved residues
(using Shannon divergence score)
+
Spatial fragments
3. Binding site identification
Query fragment
Template fragment
(Known binding site residues)
Fragment superposition
Rotation & translation matrix
Structure superposition
Best local match
Iterative refinement
Pseudo binding site sphere
Dynamic programming
Aligned residues
(local)
Predicted binding site residues
4. Binding mode of template ligand
Ligand Rotation & translation
+ MC simulations
Template protein
Ligand superposition
(based on local alignment)
Predicted complex

Figure S1. Outline of COFACTOR protocol for protein-ligand binding site predictions. (1) Template proteins from the ligand-binding library are identified using TM-align global similarity search. (2) Conserved residues in query sequence are identified based on Shannon diverge score which are used to glean local 3D-fragments from the query structure. (3) Each local 3D-motif of query is iteratively aligned with known binding site residues fragments from template where the binding pocket similarity between query and template is evaluated using BS-score. (4) The template ligand is transferred onto the query structure which is refined by a short Monte Carlo (MC) simulation to improve the local geometry.

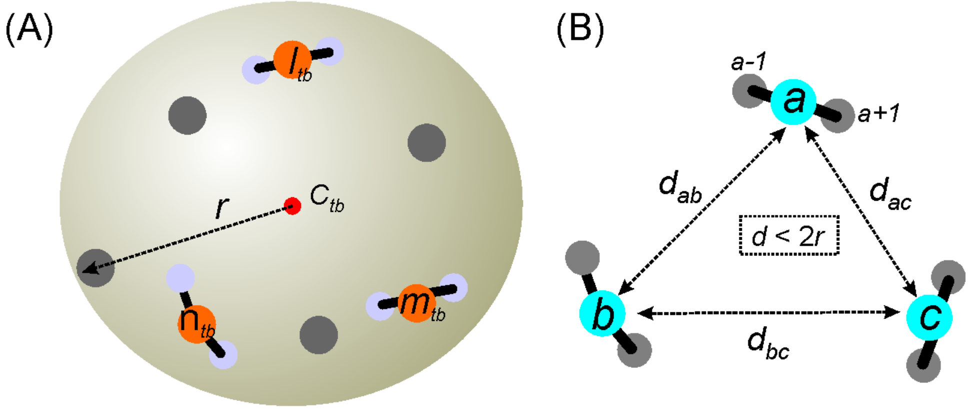

Figure S2. Schematic diagram showing putative binding residues in template and query sequence. (A) template binding site (*tb*) defined using a sphere of radius *r* from geometric centroid of binding site. The selected binding site residue triplets (*l*, *m*, *n*) are highlighted in orange. (B) conserved residues triplets (*a*, *b*, *c*) of query protein with inter-residue distance (*d*) < 2*r*. In both query and template, for any residue *i*, two flanking residues *i*-1 and *i*+1 are also selected.

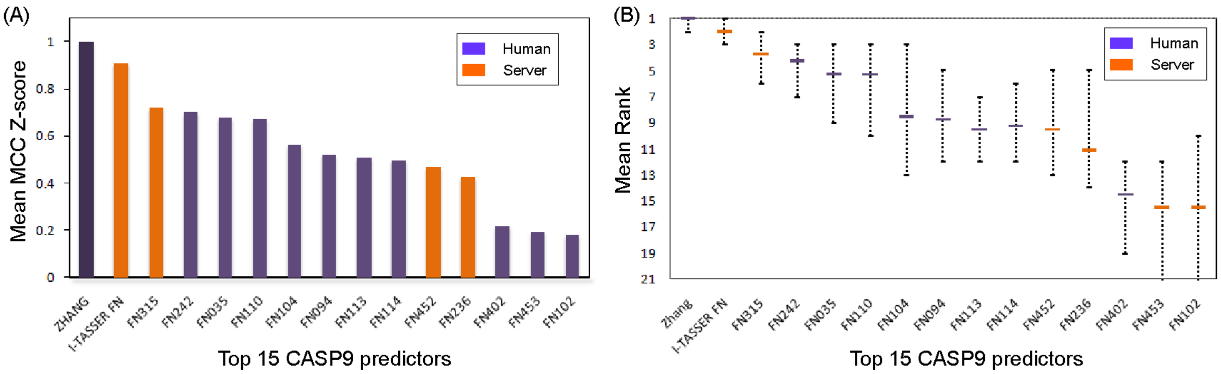

Figure S3. Performance of the top 15 groups in the binding site prediction category in CASP9 with data taken from the CASP9 assessors (Schmidt et al., 2011). (A) Mean MCC Z-score. (B) Mean Rank of CASP9 predictors based on bootstrapping experiment. The top two groups ('Zhang' as human group and 'I-TASSER FN' as automated server group) used COFACTOR to predict the binding site residues in protein structures obtained from I-TASSER predictions.

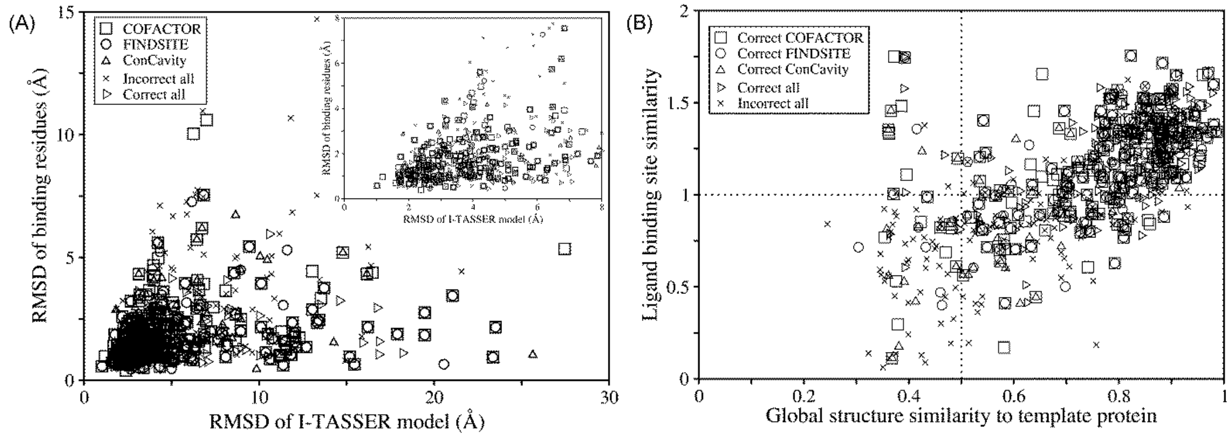

Figure S4. (A) Structural accuracy of ligand binding residues versus the accuracy of full-length receptor models. Ligand binding pocket predictions using higher resolution receptor models are shown in the inset. (B) Local versus global similarity of template to target structures. The local similarity is evaluated by BS-score (Eq. 1), while 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. This plot is complementary to Figure 7, where Figure 7 includes only the proteins in which three methods perform differently, but this plot includes all proteins in the test set.

Table S1. Comparison between predicted and bound ligands.

Protein structure	Methods	Exp. ligand volume (Å^3)	Prediction volume (Å^3)	Jaccard Coefficient	Average # of clashes
I-TASSER models	ConCavity	743	2208	0.19	282
	FINDSITE		964	0.27	63
	COFACTOR		**932**	**0.33**	**32**
Experimental structures	ConCavity	743	2307	0.24	287
	FINDSITE		962	0.29	49
	COFACTOR		**952**	**0.37**	**29**

Supplemental Experimental Procedure

Construction of benchmarking dataset and training

The benchmarking proteins were collected from ligAsite benchmark set (v9.1) (Dessailly et al., 2008), which contains 364 protein chains bound to small molecule ligands, including 63 “drug-like” and 382 “natural” ligands. To increase the sample size of drug-like compounds we further added 137 proteins bound with drug-like molecules from references (Hartshorn et al., 2007) and (Perola et al., 2004). Metal ions were filtered out from this analysis, as the control methods (FINDSITE and ConCavity) could not predict binding sites for metal ions. We also excluded ligands bound at the interface of protein chains, since the current I-TASSER protein structure modeling could be performed only for single protein chains and both COFACTOR and FINDSITE do not incorporate oligomeric state of the protein.

The parameters of the COFACTOR algorithm include the three weights that balance the sequence and structure similarities in Eq. 2, and the distance cutoffs used in the local structural search and final ligand clustering. They were optimized on the 48 protein targets from the CASP7 and CASP8 dataset and 300 protein-ligand complexes collected from the PDB. These training proteins are non-homologous to the testing set. The list of both training and testing proteins are available at http://zhanglab.ccmb.med.umich.edu/COFACTOR/benchmark.

Binding site template library

Constructing a binding site library containing biologically relevant ligands in a non-trivial problem, considering the large number of crystallization artifacts in the solved structures in PDB (Berman et al., 2000). Using homology information provides some respite to this problem, as in most cases homologous proteins bind ligands near similar locations (Brylinski and Skolnick, 2009).

To construct a comprehensive library with biologically relevant ligands, all protein chains with ligand interacting residues were first screened through the PDB library. Commonly used crystallization buffers, non-biological ions and heavy metal were pre-filtered. Protein sequence was extracted from the co-ordinates file of filtered complexes, while translating modified amino acids to their parent amino acid. Thereupon, sequences were clustered using CD-HIT (Li and Godzik, 2006) at 40% sequence identity cutoff, with the purpose of grouping them into homologous families. For orphan protein chains that formed single entry cluster, we tried to identify its homologous cluster by first performing a PSI-BLAST (Altschul et al., 1997) search against already clustered proteins, else proteins with similar structure were identified using TM-align (Zhang and Skolnick, 2005) search (TM-score >0.7).

The longest protein in each cluster was selected as the cluster representative, and all the cluster members were structurally superposed on the cluster representative using TM-align. Pair-wise distance between center of mass of ligands in superposed complexes was calculated. To judge whether a ligand is biologically relevant or not, we implemented the following filtering criteria: (a) the ligand should either have at least one ligand present in a superposed homologous structure (sequence identity <90%) within 5Å; (b) if the ligand is metal or inorganic ions, it should have at least 3 binding site residues; (c) if it is a non-metal ligand, 5 or more binding site residues is a prerequisite. Complexes that satisfied any of these three criteria were re-clustered based on ligand type and redundant binding site were removed by comparing binding site residues at 90% sequence identity cutoff. We also consulted Binding MOAD database (Hu et al., 2005), which contains both drug and natural ligands, to check for ligands that may have been missed during this automated procedure.

At present, the binding site library contains 45,381 entries, containing 13,763 metal ligands, 1,417 biopolymers and 30,201 monomeric ligands that include both drug-like and natural ligands. The library is freely available at http://zhanglab.ccmb.med.umich.edu/COFACTOR/library.

I-TASSER pipeline for automated protein structure prediction

I-TASSER is a hierarchical approach for protein structure and function predictions. A detailed description of the pipeline has been given in previous publications (Roy et al., 2010; Wu et al., 2007; Zhang, 2007, 2009). Here, we briefly outline the structure prediction methodology of I-TASSER as used for the benchmarking proteins.

For a given query protein, I-TASSER first threads the sequence through a representative PDB structure library using LOMETS (Wu and Zhang, 2007) threading programs, where homologous proteins with sequence identity >30% to the query protein were excluded from the threading library. The aim of the threading procedure is to identify possible folds or super-secondary structures which can be similar to the query. Continuous fragments (>5 residues) in threading aligned regions are excised and used for assembling full-length models, while the unaligned regions are built by *ab initio* modeling. The protein conformation during I-TASSER simulations is represented by C_α atoms and the side-chain centers of mass. The structure assembly is conducted by a modified replica-exchange Monte Carlo algorithm (Zhang et al., 2002). Low temperature structure trajectories are clustered by SPICKER (Zhang and Skolnick, 2004b) and cluster centroids are obtained by averaging the C_α coordinates of all clustered structures. Since the averaging of coordinates may create multiple steric clashes and secondary structure distortions in cluster centroids, further structure assembly simulations are conducted to remove the clashes and refine the global structure, where spatial restraints are taken from the centroid structures and from similar structures in PDB, as identified by TM-align. Finally, the lowest energy structure is selected and full atomic refined models are generated using REMO (Li and Zhang, 2009) through optimization of the hydrogen-bonding network.

COFACTOR algorithm

The COFACTOR algorithm consists of four major steps (see Figure S1**.** First, structural analogs of the query protein are identified by performing a global structure similarity search using TM-align (Zhang and Skolnick, 2005), where the structural analogs are ranked based on TM-score (Zhang and Skolnick, 2004a). The underlying hypothesis is that proteins with similar structure usually have similar function, and hence they may bind similar ligands. However, this is not always true since many observations have demonstrated that proteins with similar functions can have different global topology. This necessitates local structural comparisons since similar ligands often have similar binding pockets, which is the goal of the next steps.

In the second step, homologs of query sequence are identified by performing a PSI-BLAST (Altschul et al., 1997) search through the NCBI non-redundant (NR) sequence database. The sequences obtained during the third iteration are re-aligned by the “alignhits.pl” program (included in HHsearch (Soding et al., 2005) package) for generating the multiple sequence alignments (MSAs). Conserved residues in query sequence are then identified from the MSA, based on their Jensen–Shannon divergence score (Capra and Singh, 2007). These residues mark potential binding site locations in query structure. The structures of all combined sets of these marked residues will be used as candidate binding site motifs.

In the third step, for any given template (t) with known binding site (b), residue triplets (l_{tb}, m_{tb}, n_{tb}) are selected from binding site residues (see Figure S2A). Similarly, conserved residues triplets (a, b, c) are selected from query as candidate binding site motif (Figure S2B). The structure of these candidate sites (a, b, c) is superposed on the known binding site residues (l_{tb}, m_{tb}, n_{tb}). As a pre-filter, we discard any candidate binding site motif for which a pair-wise residue distance (d_{ab}, d_{bc} or d_{ac}) > 2r, where r is the maximum distance of any template binding site residues from the geometric center (C_{tb}) of template binding site. Furthermore, to increase the reliability of the structure superimposition, for each residue i, the coordinates of C_α atom and side-chain center of mass of two neighboring residues, i.e. the i-1th and i+1th residues, in both template and query are also included in the superposition. If the residues have glycine involved or side-chains conformations are missed, only backbone atoms of the corresponding residue-pairs are used for fragment superposition.

To account for the similar local environment in query and template, a requirement for accommodating similar ligand molecules, the entire structure of the query is superposed onto the entire structure of the template based on the rotation matrix acquired from the superposition of the candidate binding site motifs and template residues. A sphere of radius r is then defined around the geometric center (C_{tb}). The sphere here represents a probable binding pocket, under which the sequence and structural similarity of query and template are compared. Because a sphere comprising of very small number of residues can easily generate false positive hits, when the defined binding site region on the template is small (i.e. the number of residues within the sphere is less than 15), r is gradually incremented by 0.5 Å, until the number of residues inside the sphere is larger than 15.

A heuristic procedure, similar to that used in TM-align (Zhang and Skolnick, 2005), is then used to refine the local match between the query and template structures, inside the sphere. Starting from the initial superposition of query and template protein structures based on the candidate motif, a Needleman-Wunsch dynamic programming (Gotoh, 1982) is performed to generate a new alignment within the selected sphere areas of query and template, where the alignment score S_{ij} for aligning ith residue in query and jth residue in template is given by

$$S_{ij} = \left[\frac{1}{1+\left(\frac{d_{ij}}{d_0}\right)^2} + M_{ij} \right]. \qquad \text{(S1)}$$

Here, d_{ij} is the C_α distance between ith residue in the query and jth residue in the template, d_0 is the distance scale chosen to be 3.0 Å, M_{ij} is the substitution scores between the ith and jth residues taken from the BLOSUM62 mutation matrix. The element value in the BLOSUM62 matrix was normalized in between [0, 1], in order to keep both the distance and mutation scores in Eq. S1 in the same scale. Gap penalty is empirically set as -1. Based on the initial seed alignment, the areas within the spheres are re-superimposed and a new scoring matrix S_{ij} is then constructed, which will result in a newer alignment from dynamic programming. This procedure is repeated until the final alignment is converged. For each alignment, a raw alignment score is defined for evaluating the binding site similarity (BS-score):

$$\mathrm{BS-score} = \frac{1}{N_t} \sum_{i=1}^{N_{a\ l\ \ i}} \frac{1}{1+\left(\frac{d_{i\ i}}{d_0}\right)^2} + \frac{1}{N_t} \sum_{i=1}^{N_{a\ l\ \ i}} M_i \qquad \text{(S2)}$$

where N_t represents the number of residues within the binding site sphere of the template, N_{ali} is the number of aligned residue pairs. The procedure is repeated for all possible candidate binding site motifs (*a*, *b*, *c*) and known binding site residues triplets (l_{tb}, m_{tb}, n_{tb}) in this template binding site. Finally, BS-score, that determines the best local match between query and the known template binding site, is obtained:

$$\mathrm{BS-score}_{\max} = \max_{\forall abc \to lmn} \left[\frac{1}{N_t} \sum_{i=1}^{N_{ali}} \frac{1}{1+\left(\frac{d_{ii}}{d_0}\right)^2} + \frac{1}{N_t} \sum_{i=1}^{N_{ali}} M_{ii} \right] \qquad \text{(S3)}$$

This marks a binding site prediction from a known binding pocket of one template that was scanned for all conserved motifs in the query.

The ligand pose will be copied from the template structure by superposing the structure of the known binding site residues in the template onto the structure of the predicted binding site residues in the query. To remove the overlap between the ligand and the query protein structures, a quick Metropolis Monte Carlo simulation is conducted to improve the local geometry of ligand-binding, in which the energy is defined as the sum of the number of contacts made by ligand with predicted binding site residues, the reciprocal of the number of ligand-protein clashes, and the contact distance error which is calculated as difference between inter-atomic ligand-protein contact distance in template and that in query model.

The last step of the procedure is to rank the predicted binding sites based on multiple templates. To do so, all the locally superposed ligands on the query structure are clustered based on their spatial proximity (distance cutoff = 8Å). The binding pockets with larger cluster size are supposed to have higher chance to be correct. As each binding pocket can bind multiple ligands (for example, an ATP binding pocket in enzymes can also bind MG, $\mathrm{PO4}^{3-}$ and ADP), ligands within the same pocket are clustered again based on their chemical similarity (Tanimoto coefficient cutoff = 0.7) using the average linkage clustering procedure. From each ligand-specific cluster, a confidence score is defined as:

$$\mathrm{C-score_{LB}} = \frac{2}{1+e^{-\left(\frac{\mathrm{N}}{\mathrm{N_{tot}}}\times\left(0.25\mathrm{BS-score}+\mathrm{TM-score}+2.5\mathrm{ID_{Str}}+\frac{2}{1+\langle D\rangle}\right)\right)}} - 1, \qquad \text{(S4)}$$

where N is the multiplicity of ligand decoys in the cluster and $\mathrm{N_{tot}}$ is the total number of predicted ligands using the templates, BS-score and TM-score measure local and global similarity of the query to the template, $\mathrm{ID_{str}}$ is sequence identity between the query and the template in the structurally aligned region, and <D> is the average distance of the predicted ligand to all other predicted ligands in the same cluster.

The C-score definition of Eq. S4 thus represents a combination of the cluster size and the structural and sequence similarities of target and template proteins. The parameters have been chosen to keep $\mathrm{C-score_{LB}}$ in [0, 1]. The ligand binding prediction with the highest $\mathrm{C-score_{LB}}$ is finally selected.

Control methods: ConCavity and FINDSITE

Two recently developed, structure-based methods have been used as the control in our benchmarking experiments.

ConCavity (Capra et al., 2009) was designed to identify solvent-accessible pockets formed by surface residues. The identified pockets are ranked based on sequence conservation of residues associated with the pocket. Residues in the predicted pockets are smudged using a Gaussian filter to identify potential ligand interacting residues. ConCavity program was used to detect ligand binding sites in I-TASSER models and experimentally determined structures using default parameters and by providing evolutionary sequence conservation information estimated based on Jensen–Shannon divergence (JSD) score (Capra and Singh, 2007) of residues. JSD scores for each residue were computed using multiple sequence alignment of query protein with identified homologues in NR sequence database using PSI-BLAST (Altschul et al., 1997). A predicted pocket by ConCavity is represented by a set of 3D grid points; hence we used geometric center of predicted grid points as the location of predicted binding pocket.

FINDSITE (Brylinski and Skolnick, 2009) is a template-based method that first uses PROSPECTOR (Skolnick and Kihara, 2001) to identify the threading template proteins in the PDB library. Homologous template proteins of the identified threading templates are then collected from the FINDSITE binding site library and superposed on the query structure (I-TASSER models or experimental solutions) using Fr-TM-align (Pandit and Skolnick, 2008). FINDSITE predicts binding pocket as a single point, calculated as the center of mass of all the threading template ligands superposed on query structure. Binding site residues are also predicted based on concurrence of residues that make contact with ligands in the cluster.

Evaluation of ligand-binding predictions

The binding pocket predictions are evaluated by calculating the distance between the center of mass of the bound ligand in the experimental structure and the center of the predicted binding pocket in the query. We used 4.5Å as a cutoff to evaluate correct binding pocket predictions, which is close to the average radius of gyration of the 582 experimental ligands in the benchmark set.

The shape similarity between the predicted ligand and the bound ligand in experimental structure is evaluated based on the volume overlap, measured as Jaccard Coefficient (JC):

$$\mathrm{JC} = \left|\frac{\text{predicted vol.} \ \cap \ \text{native vol.}}{\text{predicted vol.} \ \cup \ \text{native vol.}}\right| \tag{S5}$$

For FINDSITE, we selected the ligand from template amongst the clustered templates that had highest sequence identity to query protein and best-predicted pocket; while for ConCavity, predicted pocket grid points were presumed as H-atom of ligand. Volume calculation is done on 3-D grid of 1Å spacing.

The ligand-binding residue predictions are evaluated mainly by the Matthews Correlation Coefficient (MCC) between the predicted and experimental binding residues:

$$\mathrm{MCC} = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{S6}$$

where *TP*, *TN*, *FP* and *FP* are abbreviations for true positive, true negative, false positive, and false negative binding residue predictions. MCC ranges between 1 and −1, where a MCC of 1 indicates a prefect prediction, 0 a random prediction, and -1 an inverse prediction. We also define the precision of the binding site prediction as

$$\text{Precision} = TP/(TP+FP) \tag{S7}$$

which measures the ratio of the correctly predicted binding residues over the total number of predicted residues. The recall is defined as

$$\text{Recall} = TP/(TP+FN) \tag{S8}$$

which measures the portion of the correctly predicted binding residues over the total number of binding residues in the experimental structure. Here, true binding site residues are defined as those that have any heavy atom within a distance of 0.5Å plus the sum of the van der Waals radius of protein atom and ligand atoms in the experimental structure.

Chemical similarity between two compounds (*A* and *B*) is evaluated using the Tanimoto coefficient:

$$T_{AB} = \frac{N_A+N_B-N_{AB}}{N_{AB}} \tag{S9}$$

where N_A and N_B are the number of chemical and structural features that are present in each ligand and N_{AB} is the number of common features between *A* and *B*. Features of all the ligands were defined using Open Babel package (Guha et al., 2006).

Supplemental References

Altschul, S.F., Madden, T.L., Schaffer, 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.

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.

Brylinski, M., and Skolnick, J. (2009). FINDSITE: a threading-based approach to ligand homology modeling. PLoS Comput Biol *5*, e1000405.

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.

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

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*, D667-673.

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

Guha, R., Howard, M.T., Hutchison, G.R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J., and Willighagen, E.L. (2006). The Blue Obelisk-interoperability in chemical informatics. J Chem Inf Model *46*, 991-998.

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.

Hu, L., Benson, M.L., Smith, R.D., Lerner, M.G., and Carlson, H.A. (2005). Binding MOAD (Mother Of All Databases). Proteins *60*, 333-340.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.

Li, Y., and Zhang, Y. (2009). REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins *76*, 665-676.

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.

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.

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.

Schmidt, T., Haas, J., Gallo Cassarino, T., and Schwede, T. (2011). Assessment of ligand binding residue predictions in CASP9. Proteins, in press.

Skolnick, J., and Kihara, D. (2001). Defrosting the frozen approximation: PROSPECTOR--a new approach to threading. Proteins *42*, 319-331.

Soding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res *33*, W244-248.

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

Wu, S., and Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res *35*, 3375-3382.

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

Zhang, Y. (2009). I-TASSER: Fully automated protein structure prediction in CASP8. Proteins *77*, 100-113.

Zhang, Y., Kihara, D., and Skolnick, J. (2002). Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins *48*, 192-201.

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

Zhang, Y., and Skolnick, J. (2004b). SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem *25*, 865-871.

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