BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 20 2013, pages 2588–2595
doi:10.1093/bioinformatics/btt447

Structural bioinformatics

Advance Access publication August 23, 2013

# Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment

Jianyi Yang[1], Ambrish Roy[1] and Yang Zhang[1,2,*]

[1]Department of Computational Medicine and Bioinformatics and [2]Department of Biological Chemistry, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA

Associate Editor: Anna Tramontano

**ABSTRACT**

**Motivation:** Identification of protein–ligand binding sites is critical to protein function annotation and drug discovery. However, there is no method that could generate optimal binding site prediction for different protein types. Combination of complementary predictions is probably the most reliable solution to the problem.

**Results:** We develop two new methods, one based on binding-specific substructure comparison (TM-SITE) and another on sequence profile alignment (S-SITE), for complementary binding site predictions. The methods are tested on a set of 500 non-redundant proteins harboring 814 natural, drug-like and metal ion molecules. Starting from low-resolution protein structure predictions, the methods successfully recognize >51% of binding residues with average Matthews correlation coefficient (MCC) significantly higher (with *P*-value $<10^{-9}$ in student *t*-test) than other state-of-the-art methods, including COFACTOR, FINDSITE and ConCavity. When combining TM-SITE and S-SITE with other structure-based programs, a consensus approach (COACH) can increase MCC by 15% over the best individual predictions. COACH was examined in the recent community-wide COMEO experiment and consistently ranked as the best method in last 22 individual datasets with the Area Under the Curve score 22.5% higher than the second best method. These data demonstrate a new robust approach to protein–ligand binding site recognition, which is ready for genome-wide structure-based function annotations.

**Availability:** http://zhanglab.ccmb.med.umich.edu/COACH/

**Contact:** zhng@umich.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on May 22, 2013; revised on July 19, 2013; accepted on July 30, 2013

## 1 INTRODUCTION

Proteins perform the biological functions through interactions with other molecules (called ligands). The identification of specific ligand-binding site (LBS) on proteins is often the first important step toward understanding the function of protein molecules, or for rational design of new therapeutic compounds to modulate the protein functions (Greer *et al*., 1994; Hubbard, 2006). Due to the technical difficulties and high cost associated with experimental determination, however, the structural details for protein–ligand interactions are unknown for most proteins. Even for the proteins with experimentally solved 3D structure, there are nearly 40% of proteins (i.e. 35 633 out of 90 424) in the PDB which lack biologically relevant ligand-binding information, as shown in BioLiP (Yang *et al*., 2013). Accurate prediction of ligand–protein binding is therefore required for both biological and therapeutic studies.

A variety of methods have been developed for computational prediction of protein LBSs; these methods can be generally categorized into two groups, i.e. sequence-based (Capra and Singh, 2007; Fischer *et al*., 2008; Lopez *et al*., 2011; Rausell *et al*., 2010) or 3D structure-based (An *et al*., 2005; Brylinski and Skolnick, 2008; Capra *et al*., 2009; Hendlich *et al*., 1997; Laskowski, 1995; Roche *et al*., 2011; Roy *et al*., 2012; Roy and Zhang, 2012; Wass *et al*., 2010) methods. Most of the sequence-based methods rely on the residue conservation analyses under the assumption that the ligand-binding residues are functionally important and therefore conserved in evolutionary process (Capra and Singh, 2007). This approach has the advantage of generating prediction from sequence alone but the precision of predictions is low (typically around 35% at 20% recall) because many non-binding residues can also be of high degree of conservation due to the diverse roles (e.g. to keep the fold stable).

For the structure-based methods, two different approaches prevail. In the first approach, the ligand-binding pocket is identified by recognizing the surface cavities on the 3D structural model of the target protein (An *et al*., 2005; Capra *et al*., 2009; Hendlich *et al*., 1997; Laskowski, 1995). It has the advantages of *ab initio* modeling of the LBSs as predictions are made without using templates, but the false positive rate can be high, especially for the low-resolution models generated from protein structure predictions. The second structure-based approach is to infer ligand-binding information from the known template proteins, which have similar global and/or local structure to the query (Brylinski and Skolnick, 2008; Roche *et al*., 2011; Roy *et al*., 2012; Roy and Zhang, 2012; Wass *et al*., 2010). As shown in the recent community-wide CASP experiments (Schmidt *et al*., 2011), this type of template-based approaches represent by far the most accurate methods, especially for targets which have close homologs in the ligand–protein complex structure databases. Nevertheless, no individual methods can generate sufficiently accurate predictions for different targets. For instance, the template-based methods do not outperform *ab initio* pocket identification methods for distant-homologous targets; and the

*To whom correspondence should be addressed.

performance of pocket-based methods can be significantly degraded in the targets without high-resolution models, where sequence-based methods may have their advantage.

In this work, we aim to develop a reliable approach which could generate highly accurate ligand-binding predictions for different categories of protein targets. We first design a new structure-based algorithm (TM-SITE) which derives LBSs from structure-related templates with the alignments built on binding-specific substructures matches. Second, a new binding-specific sequence profile alignment method (S-SITE) is developed for evolution-based LBS recognition. Finally, we design COACH to combine the prediction results of TM-SITE and S-SITE with other available LBS tools by the support vector machine (SVM) training. Because one of the major objectives in this work is for genome-wide function annotation following the sequence-to-structure-to-function paradigm, we will examine and test our methods on low-resolution structure models generated by the state-of-the-art protein structure predictions (Roy *et al.*, 2010; Zhang, 2007; Zhang, 2008). All the algorithms and data developed in this work are freely accessible at our website.

## 2 MATERIALS AND METHODS

### 2.1 Datasets

Our benchmark testing dataset consists of 500 non-redundant proteins that harbor 814 ligands (410 natural ligand, 238 drug-like ligand and 164 metal ions). This set of proteins are taken from the previous benchmark study (Roy *et al.*, 2012; Roy and Zhang, 2012), but one protein (PDB ID: 3cq3) was removed because its ligand (GOL) is not biologically relevant as evaluated in BioLiP (Yang *et al.*, 2013). Ligands in this benchmark are more comprehensive, including metal ions and other ligands from the BioLiP database. In addition, a set of 400 non-redundant proteins with at least 5 residues bound to known ligands are collected from BioLiP as the training set to train our methods. None of proteins in the training set has a sequence identity >30% to the proteins in the test set.

For each protein, the structural models are generated by the standard I-TASSER pipeline (Roy *et al.*, 2010) where all homologous templates with a sequence identity >30% to the query sequence are excluded from the template library. This sequence identity cutoff is also used to filter out all ligand-binding templates when generating the LBS predictions. The two datasets of proteins, together with the binding-ligands and the I-TASSER models can be downloaded at http://zhanglab.ccmb.med.umich.edu/COACH/benchmark.

### 2.2 TM-SITE

TM-SITE is designed to derive the LBSs by structurally comparing the query with the proteins of known LBSs (called template). There are generally two types of structural comparisons. The global comparison aligns the entire structure of two proteins which is most robust to recognize the similarity of two folds; but the alignment on the ligand-binding regions, which is important for LBS predictions, can be distracted by the structural variations in the regions far away from the binding pockets. On the other hand, the local comparison based on the binding residue structures is more sensitive to detect the similarity of specific binding pockets; but it has a high false positive rate due to the too small number of residues involved in the comparisons. TM-SITE takes an intermediate approach balancing the two comparisons, i.e. to compare the structures of a subsequence from the first binding residue to the last binding residue (called SSFL) on the query and template proteins. A flowchart of TM-SITE is depicted in Figure 1A, which consists of three steps of SSFL generation, binding template identification and LBS selection.

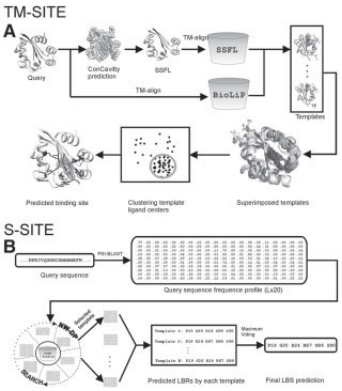


**Fig. 1.** Flowchart of (**A**) TM-SITE and (**B**) S-SITE for protein–ligand binding site prediction

*2.2.1 SSFL generation* For templates, an SSFL library is pre-calculated for all proteins in the BioLiP database by collecting the residues associated with the known ligands. For the query, various surface cavities are first identified by ConCavity (Capra *et al.*, 2009) which detects the cavities based on the shape of the structural surface. For each identified cavity, residues that have any heavy atoms within 2.5 Å to the cavity's mesh points are considered as putative binding residues. The cavities with >15 putative binding residues are defined as putative binding pockets which are then used to determine the SSFL of the query. If there is no cavity with >15 binding residues, the largest cavity is used for the query SSFL definition.

*2.2.2 Binding template identification* For a given query protein, we use TM-align (Zhang and Skolnick, 2005) to thread the query SSFL structure through the SSFL library by optimal structural alignments. The match between each pair of query and template SSFL structures is evaluated by a composite scoring function which counts for both global and local, structural and sequence similarities, i.e.

$$q_{\mathrm{str}} = \frac{2}{1 + e^{-\left[L_c\left(0.4L_g + 0.3L_s + 0.2\mathrm{JSD}\right) + \mathrm{TM}\right]^2}} - 1, \qquad (1)$$

where $L_c$ is the fraction of template binding residues that are aligned to the query structure by TM-align. $L_g = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\left(d_i/d_0\right)^2}$ accounts for the local structure similarity between the binding pockets of query and template proteins, where $n$ is the number of the aligned residue pairs associated with the binding pockets of the template and $d_i$ is the distance of $i$-th residue pair. $L_s = \frac{1}{n}\sum_{i=1}^{n} B(R_i^q, R_i^t)$ measures the evolutionary relation between the aligned binding residue pairs where $B(R_i^q, R_i^t)$ is the normalized BLOSUM62 score for the $i$-th aligned residue pair. JSD is an evolutionary conservation index defined as the average Jensen–Shannon divergence score over the predicted binding residues, which is calculated from multiple sequence alignments (see Supplementary Materials for detail). Here, JSD should not be considered redundant to $L_s$, since $L_s$ is defined for the pairwise conservation between the query and the structure template from BioLiP, but JSD is for the conservation between query and all other homologous sequences detected from sequence database by PSI-BLAST (Altschul et al., 1997). Finally, $\mathrm{TM} = 2\frac{\mathrm{TM}_t * \mathrm{TM}_q}{\mathrm{TM}_t + \mathrm{TM}_q}$ is the harmonic average of the two TM-scores returned by TM-align when aligning the query and template SSFLs, where $\mathrm{TM}_t$ and $\mathrm{TM}_q$ are TM-scores normalized by query and template lengths, respectively. The format of Equation (1) and the weighting parameters have been determined by extensive trial and error on the 400 training proteins, which results in the best performance on the ligand-bind site recognitions according to the average Matthews's correlation coefficient (MCC) values.

All proteins in the template SSFL library with a $q_{str}$ score >0.65 are collected as putative templates; if no proteins have a $q_{str} > 0.65$, the protein with the highest $q_{str}$ will be selected. Meanwhile, a second scan using the entire query structure as a probe is conducted through the BioLiP library based on TM-align (Fig. 1A). All proteins, which are missed by the first SSFL-based scan but hit by the whole-chain based scan, are added to the putative template pool. Since the whole-structure based scan has lower specificity, a lower cutoff, $q_{str} > 0.57$, is used in the second structure scan.

*2.2.3 LBS clustering and selection* To select the ligand-binding residues on the query, all ligands bound with the proteins in the putative template pool are projected to the query structure based on the corresponding alignments from TM-align. These ligands are then clustered based on the spatial distance between their geometric centers, where an average linkage clustering algorithm is conducted with a distance cutoff 4 Å. At the beginning, each ligand sample is treated as a cluster. Two clusters are merged into one when their distance is below the cutoff. For clusters containing multiple ligands, the cluster distance is computed as the average over the pair-wise distances of the geometric centers between all the ligands from the first cluster and all the ligands from the second cluster. The cluster merging iteration starts from the cluster pairs of the closest distance, and stops when no cluster pairs have the distance smaller than the distance cutoff (4 Å).

For each cluster, a set of consensus binding residues, which usually correspond to one binding pocket, are deduced from all the ligands in the cluster based on the maximum voting. The residues receiving >25% votes, i.e. more than a quarter of templates in the cluster have the same residues as the binding site, are considered as the final predicted LBS residues in the binding pocket. This clustering procedure and cutoff selections are similar to the ones used by previous methods, including 3DLigandSite (Wass *et al.*, 2010), COFACTOR (Roy *et al.*, 2012; Roy and Zhang, 2012), FINDSITE (Brylinski and Skolnick, 2008) and FunFOLD (Roche *et al.*, 2011). A confidence score of the predicted binding residues, associated with specific ligand-binding clusters, is defined by

$$\mathrm{CS}_t = \frac{2}{1 + e^{-\left[\frac{m}{M} q_{\mathrm{str}}^{\max} + 0.2\ln(1+\sqrt[3]{m}) + 0.2\mathrm{JSD}_{\mathrm{Ta}}\right]}} - 1, \quad (2)$$

where $m$ is the number of template ligands in the cluster and $M$ is the total number of templates selected. $q_{str}^{max}$ is the maximum of $q_{str}$ score from the templates in the cluster as calculated by Equation (1). $JSD_{Ta}$ is the average JDS score for the predicted LBS residues from the ligand cluster. Again, this form of scoring function was decided by trial and error on our training set proteins. In the final TM-SITE predictions, the binding pockets are selected and ranked based on the $CS_t$ score, with the binding residues sorted by the number of votes in each binding pocket. In our training data, a prediction with $CS_t > 0.35$ has average false positive and false negative rates below 0.16 and 0.13, respectively.

## 2.3 S-SITE

S-SITE is another template-based method, which detects protein templates and the LBSs using binding site specific, sequence profile–profile comparisons. The procedure of S-SITE is illustrated in Figure 1B, which consists of three steps of sequence profile generation, template identification and LBS selection.

*2.3.1 Profile generation* To obtain the profile of the query protein, PSI-BLAST is used to thread the query sequence through the NCBI sequence database for constructing multiple sequence alignments. A position-specific frequency matrix (PSFM) is then computed from the multiple sequence alignments. Similarly, the template profiles, which are represented by the position-specific scoring matrices (PSSM), are pre-constructed by the PSI-BLAST searches for all proteins in the BioLiP library.

*2.3.2 Template identification* To detect homologous templates from BioLiP, the query profile PSFM is compared with the template profile PSSMs in the library using the Needleman–Wunsch dynamic programming algorithm (Needleman and Wunsch, 1970). The score for aligning the $i$-th residue in the query to the $j$-th residue in template is defined as

$$S_{i,j} = \sum_{k=1}^{20} F_{i,k}^{q} P_{j,k}^{t} + \delta\left(s_i^q, s_j^t\right) + 2b_j^t B\left(R_i^q, R_j^t\right), \quad (3)$$

where $F_{i,k}^{q}$ (or $P_{j,k}^{t}$) is the $i,k$-th ($j,k$-th) element in the query PSFM (or the template PSSM); $s_i^q$ (or $s_j^t$) $\in \{H, E, C\}$ is the three-state secondary structure ('H' = alpha helix, 'E' = beta strand and 'C' = random coil) of $i$-th (or $j$-th) residue in the query (or template); the secondary structure for templates are assigned by STRIDE (Heinig and Frishman, 2004) using the PDB structures. For query, its secondary structure is predicted by PSIPRED (Jones, 1999). $\delta(x, y)$ equals to 1 if $x = y$, or 0 otherwise; $b_j^t = 1$ if the $j$-th residue is at the binding site in the template, or $b_j^t = 0$ otherwise; $B(R_i^q, R_j^t)$ is the normalized BLOSUM62 similarity score for residues $R_i^q$ in query and $R_j^t$ in template with value in [0,1] (see Supplementary Materials). Overall, the first term in Equation (3) accounts for the query-to-template profile alignments, the second for the secondary structure match, and the third for evolutionary relation of residues in the LBSs.

The quality of a template match is estimated by

$$q_{\mathrm{seq}} = \frac{2}{1 + e^{-(0.5A_S + 0.5L_c L_s + 0.2\mathrm{JSD})}} - 1, \quad (4)$$

where $A_S = \frac{1}{L}\sum_{i=1}^{L_{\mathrm{ali}}} S_{i,i}$ is the profile-alignment score normalized by the query sequence length $L$ following Equation (3). $L_c$, $L_s$ and JSD are similar to that defined in Equation (1) but with the alignments generated from the ligand-binding specific profile–profile comparisons. All proteins in BioLiP with a $q_{seq}$ score above 0.5 are selected as the putative templates. If the number of putative templates is below 10, the top 10 templates with the highest $q_{seq}$ score will be returned for the next step of LBS selection analysis.

*2.3.3 LBS selection by maximum voting* The residues on the query, which are aligned with the binding residues on the templates following the sequence profile–profile alignments, are assigned as putative binding residues in the S-SITE prediction. Since the binding sites of different templates will match with different query residues, a consensus voting scheme is applied to select the most consensus binding residues. The residues receiving >25% votes are considered as the final binding residues by S-SITE.

A confidence score $CS_s$ is defined for the binding residues:

$$\mathrm{CS}_s = \frac{2}{1 + e^{-\left[q_{\mathrm{seq}}^{\max} + 0.1\ln(1+\sqrt[3]{N}) + 0.2\mathrm{JSD}_{\mathrm{Sa}}\right]}} - 1, \quad (5)$$

where $q_{seq}^{max}$ is the maximum value of $q_{seq}$ among all the putative templates, $N$ is the number of the selected templates, $JSD_{Sa}$ is the average Jensen–Shannon divergence score of all the predicted LBS residues. Here, the confidence score $CS_s$ is in a similar format as that of TM-SITE ($CS_t$) but no clustering is conducted in S-SITE since the templates detected from the sequence profile comparison are converged in most cases. Based on the training data, a prediction with $CS_s > 0.25$ has average false positive and false negative rates below 0.24 and 0.21, respectively.

## 2.4 Control programs for LBS prediction

Three methods representing different prediction principles are used as control in this study. First, ConCavity (Capra *et al.*, 2009) is an *ab initio* prediction method which identifies LBSs from the surface cavity of the target structure with score combining residue conservation information. FINDSITE (Brylinski and Skolnick, 2008) is a template-based approach which derives the query binding sites from template proteins identified by the threading programs (Skolnick *et al.*, 2004). COFACTOR (Roy *et al.*, 2012; Roy and Zhang, 2012) is a structural comparison approach which identifies ligand-binding residues by global and local matches of the query

and template structures. The ConCavity and FINDSITE are publically available programs and COFACTOR was developed in our lab. All the methods are run using default parameters.

## 2.5 COACH

COACH is a consensus approach to LBS prediction that combines the multiple prediction results of algorithms from TM-SITE, S-SITE, COFACTOR, FINDSITE and ConCavity, with the architecture presented in Supplementary Figure S1. To generate a prediction, the query sequence along with the structure are provided as input and fed into the individual programs. The top-scoring predictions from each of the programs are combined using a linear SVM as implemented by the software SVM-light (Joachims, 2006).

The probability of a residue to be a binding residue is calculated from individual methods, which are used as the feature vectors for the residue. For TM-SITE and S-SITE, the probabilities are computed as the confidence score of the ligand cluster or templates (i.e. $CS_t$ and $CS_s$) multiplied by the ratio of votes on the residue. The probabilities for FINDSITE and COFACTOR are taken from the default confidence scores. For ConCavity, the probability is calculated by a linear combination of the residue conservation score and the distance of the residue to the predicted binding pockets. Finally, all feature vectors are fed into SVM to make consensus prediction, with classifiers trained on the 400 non-redundant training proteins, which have sequence identity <30% to the proteins in the test set.

The linear kernel in SVM-light was used with the optimal value of the cost parameter $C$ selected based on an exhaustive grid search. To avoid over-training, a 10-fold cross-validation procedure was applied as follows. The training set was randomly divided into 10 subsets of equal size, where 9 subsets are used to train the SVM and the remaining subset was used as validation to calculate the average MCC. For each parameter $C$ in the grid space, such random sample division was repeated by 10 times and an overall MCC was calculated as the mean of the 10 average MCCs. The parameter $C$ with the highest overall MCC was finally selected for SVM training.

## 2.6 Evaluation

The LBS prediction results are mainly evaluated by the Matthews correlation coefficient (MCC), precision and recall:

$$\begin{cases} \text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \\ \text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \\ \text{recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \end{cases}, \quad (6)$$

where TP (FP) is the number of true (false) binding residues in the prediction, and TN (FN) is the number of true (false) non-binding residues. In general, MCC represents a score combining both the accuracy and coverage of the prediction which has a better balance of both aspects than the individual precision and recall values. The MCC equation was also used in the CASP official evaluation for protein–ligand binding site prediction (Schmidt *et al.*, 2011).

Another measure to combine precision and recall is the Area Under the Curve (AUC) value of the true positive rate versus false positive rate plot, or the ROC (Receiver Operating Characteristic) plot. This measurement has been used in another community-wide blind experiment, CAMEO (http://www.cameo3d.org), for accuracy evaluation. Since MCC and AUC contain similar information, we will mainly use MCC for our benchmark/training data, and AUC for CAMEO experiment in our data analysis.

# 3 RESULTS

## 3.1 Accuracy of the I-TASSER structural models

Because the quality of the receptor models has impact on the structure-based methods for LBSs prediction, we computed TM-score and RMSD of the I-TASSER models for the 500 testing proteins. In Supplementary Figure S2, we present the histogram distribution of the TM-score and RMSD of the first I-TASSER models. It is shown that the majority of the proteins (=90%) can be modeled with a correct fold (TM-score > 0.5) and 65% have a RMSD below 6 Å, although all close homologous templates were excluded in the model generations. The average TM-score and RMSD for the proteins in the test dataset is 0.77 and 6.4 Å, respectively. Nevertheless, there are still 38 cases which have incorrect folds with a TM-score below 0.5, and 176 cases with RMSD higher than 6 Å. Even in those with a correct fold, the local structure associated with the LBSs have structural deviations with an average RMSD = 2.8 Å to the native holo-structures. These represent a set of models in a typical range of accuracy of the template-based protein structure predictions (Zhang, 2009).

## 3.2 Summary of individual methods

The MCCs for TM-SITE and S-SITE are summarized in first two columns of Table 1. The average MCC score for TM-SITE is 0.48, with precision 0.57 and recall 0.49, respectively. The MCC and precision values of TM-SITE are higher than that of S-SITE (MCC = 0.45, precision = 0.45 and recall = 0.58), although TM-SITE has a slightly lower recall. The difference in the MCC values is statistically significant which has a $P$-value $<10^{-3}$ in the Student's $t$-test (Supplementary Table S1). This data demonstrate the usefulness of structural information in I-TASSER models that were taken by TM-SITE for template detection, while S-SITE is based on the sequence profile comparison. The higher value of recall in the S-SITE predictions is partly due to the fact that many S-SITE predictions contain residues from multiple ligand-binding pockets, since no clustering was performed in S-SITE. The clustering procedure in TM-SITE increases the specificity and thus MCC of the predictions. In S-SITE, since the templates are mostly converged, we found that the clustering on the S-SITE predictions did not increase the overall MCC score but there was a slight increase in precision (0.52) and reduction in recall (0.49).

The TM-SITE predictions also outperform the three other control methods, with the average MCC value 14% higher than FINDSITE and COFACTOR, and 84.6% higher than ConCavity. The $P$-value in the Student's $t$-test is below $10^{-9}$ in all the comparisons (Supplementary Table S1). The precision of TM-SITE and COFACTOR is similar while the recall of TM-SITE is higher, which is partly due to the fact that the TM-SITE prediction is made based on a combination of multiple templates, while COFACTOR prediction is on a single top-scoring template. ConCavity's recall (0.51) is slightly higher than TM-SITE (0.49), while its precision is very low due to over-prediction, resulting in a low MCC value 0.26.

In Supplementary Figure S3, we present a head-to-head comparison of TM-SITE versus the four other methods based on MCC. Out of the 500 test proteins, there are 345, 354, 347 and 400 cases where TM-SITE has equal or higher MCC than S-SITE, COFACTOR, FINDSITE and ConCavity, respectively. Interestingly, although S-SITE has a higher average MCC score than COFACTOR and FINDSITE, it has similar number of cases in which it outperforms TM-SITE. This indicates that the S-SITE algorithm is probably less complementary to

**Table 1.** LBS predictions by different programs on the 500 test proteins

| | | TM-S | S-SI | COF | FIN | Con | COA |
|---|---|---|---|---|---|---|---|
| ITA | MCC | 0.48 | 0.45 | 0.42 | 0.42 | 0.26 | **0.54** |
| | Precision | **0.57** | 0.45 | 0.56 | 0.44 | 0.23 | 0.54 |
| | Recall | 0.49 | 0.58 | 0.39 | 0.49 | 0.51 | **0.63** |
| EXP | MCC | 0.51 | 0.45 | 0.46 | 0.44 | 0.33 | **0.60** |
| | Precision | 0.59 | 0.45 | **0.61** | 0.45 | 0.26 | 0.59 |
| | Recall | 0.51 | 0.58 | 0.41 | 0.51 | 0.62 | **0.70** |

TM-S, TM-SITE; S-SI, S-SITE; COF, COFACTOR; FIN, FINDSITE; Con, ConCavity; COA, COACH; ITA, I-TASSER models; EXP, experimental structures.
Bold values denote the best performance in each category.

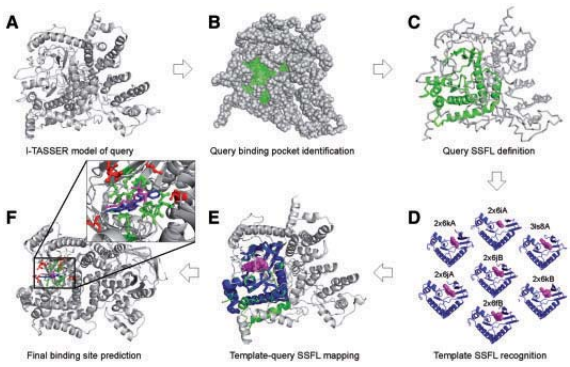


**Fig. 2.** An illustrative example of TM-SITE binding site prediction on the phosphatidylinositol 4,5-bisphosphate 3-kinase protein. (**A**) I-TASSER model with TM-score = 0.74 and global/local-binding RMSD = 19.8/1.5 Å. (**B**) Binding pocket identification by ConCavity (green mesh). (**C**) Query SSFL definition based on predicted binding pockets for the query protein (green cartoon) which has a RMSD to the native 3.8 Å. (**D**) Template SSFLs recognized by TM-align. (**E**) Superposition of all template SSFLs on the query structure. (**F**) Final model of the predicted binding sites: native/predicted ligands are shown in magenta/blue sticks; true/false positive binding sites are highlighted in green/red ball-sticks

TM-SITE compared to other methods, since several common terms (including the binding residue coverage and the JSD score) have been used in both methods for template recognitions. In addition, the Pearson's correlation coefficients (PCC) between TM-SITE and other methods are reported in Supplementary Figure S3. TM-SITE has the highest PCC with COFACTOR, probably because both methods use structural alignment and the same template library. As expected, ConCavity has the lowest PCC with TM-SITE as it is a template-free method.

In the lower part of Table 1, we also list the MCC results of the programs when the experimental structure of the receptor proteins is used. As expected, the performance of all the structure-based methods is improved due to the increase of structural accuracy of the receptor proteins. S-SITE has no change since it only uses the sequence profile information. Nevertheless, it has a comparable or superior result than other structure-based methods from FINDSITE, COFACTOR and ConCavity.

### 3.3 Why TM-SITE/S-SITE work better than other methods?

The method with the closest overall performance to TM-SITE is COFACTOR, but TM-SITE still has an MCC value 14% (using I-TASSER models) and 10.9% (using experimental structure) higher than COFACTOR on the 500 tested proteins.

A major advantage of the TM-SITE is the use of the SSFL residues for structural alignments, which helps increase the sensitivity and specificity of the template detections, compared to either the global or local structural comparison taken by COFACTOR. Figure 2 shows an illustrative example from the phosphatidylinositol 4,5-bisphosphate 3-kinase protein (PDB ID: 2chzA), which has the ligand molecule N-(5-(4-Chloro-3-(2-Hydroxy-Ethylsulfamoyl)-Phenylthiazole-2-Yl)-Acetamide bound with Residues S567, W573, I592, Y628, I640, E641, I642, V643, A646, N711, M713, F721, I723 and D724. We first run I-TASSER which generates the first model with a TM-score = 0.74 and RMSD = 19.8 Å. Based on the I-TASSER model, the ConCavity program detects 3 putative binding pockets, each with more than 15 associated binding residues, which result in an SSFL definition from residue M565 to residue P787 (Fig. 2B). Despite the high RMSD of the global I-TASSER model, the SSFL region has high accuracy with a $RMSD_{SSFL}$ = 3.8 Å, which provides an opportunity for the accurate SSFL structure matches to the structures in BioLiP. Following the SSFL-based TM-align alignment, seven templates, including 2x6kA, 2x6iA, 2x6jA, 2x6jB, 2x6kB, 3ls8A and 2x6fB, have a $q_{str}$ score above 0.65, where 16 binding residues, including M565, A566, S567, P571, I592, K594, D602, Y628, I640, E641, I642, V643, D645, M713, I723 and D724, are transferred as putative LBS residues to the query (Fig. 2E). After the ligand clustering and the maximum vote, 14 residues receiving >25% votes (M565, A566, I592, K594, D602, Y628, I640, E641, I642, V643, D645, M713, I723 and D724) are eventually selected, which have a $CS_t$ score = 0.57, significantly higher than the default confident $CS_t$ cutoff (0.35). This prediction results an MCC score = 0.64 with precision = 0.64 and recall = 0.64, respectively.

COFACTOR uses either the local-binding site or the entire structure comparison. The former comparison does not result in significant hits since a large number of possible 3D motifs have been aligned in the query structure which generates similar scores. The global structure comparison picks up two templates from 1n38A and 1n35A, which have a completely different binding pocket although they have a similar fold to the query protein. These hits result in an LBS prediction with an MCC = –0.01 for COFACTOR, in this example.

Accordingly, ConCavity has selected 76 residues as the binding residues which also results in a low MCC = 0.32, due to the low precision value (=0.14) although 11 out of the 14 native LBS residues are included in the ConCavity prediction. FINDSITE failed to identify the correct binding residues since most of the templates have only a weak homology to the query protein and the threading alignment by FINDSITE selected an incorrect template (with a TM-score = 0.26).

The second advantage of TM-SITE is the composite scoring function balancing both structural and sequence similarities in the SSFL region which is essential for recognizing the correct templates and binding sites. One of the major differences between the scoring functions of TM-SITE and FINDSITE is

that the evolutionary information and local structure similarity of binding site are not considered in FINDSITE. These two terms are calculated by $L_g$ and $L_S$ in TM-SITE (Equation 1). If we turned off these two terms, the performance of TM-SITE would be significantly degraded with the average MCC value reduced from 0.48 to 0.43. Similarly, the JSD score was not considered in score-function of CAFACTOR although it was used for initial binding residue screening. If we turned off the JSD term in Equation 1, the MCC value is also reduced.

In the example of Figure 2, although S-SITE takes a similar threading approach to FINDSITE, by using the profile–profile alignment for template screening, it recognizes a set of binding residues, including M565, A566, S567, P571, I592, K594, D602, Y628, I640, E641, I642, V643, D645, A650, M713, I723 and D724, which has an MCC score (0.64) much higher than FINDSITE (0.0). A detailed analysis shows that the binding-specific conservation scores, including $L_s$ and JSD in Equations (1) and (4), dominate the profile–profile comparisons and therefore plays an important role in the query-template alignment selections. We also tried to replace Equations (1) and (4) by the global structure similarity TM-score and the threading $Z$-score, respectively. The performances of TM-SITE and S-SITE are dramatically reduced by 21% and 59%, respectively. These data demonstrate the extreme importance of a balanced scoring function combining different effects from global and local, structural and evolutionary information of ligand–protein binding.

### 3.4 Combination of different methods by COACH

In the last column of Table 1, we list the prediction results of COACH, which combines the predictions from the five individual algorithms using the SVM. The average precision and recall of the COACH predictions are 0.54 and 0.63, respectively, which result in an overall MCC = 0.54. This MCC value is 12.5% higher than the best individual prediction from TM-SITE or 107.7% higher than the prediction by ConCavity. The improvements made by COACH are mainly attributed to the complementary property of the individual component predictors, as highlighted by the head-to-head comparison between TM-SITE and other methods shown in Supplementary Figure S3. Although the best individual prediction by TM-SITE outperforms that by other methods for most proteins (see the number of data points located in the upper triangle), there are still a considerable number of proteins where TM-SITE performs worse than others. Thus, a combination of the results provides the opportunity to pick up the cases that are incorrectly predicted (mostly having weak scores) by one program but correctly predicted (mostly having high scores) by others.

Supplementary Figure S4 presents head-to-head comparisons of COACH versus all individual programs. Indeed, COACH has dominantly more number of proteins that have a higher MCC than any of the individual programs. We note that the consensus approach in COACH is different from most of the meta-server approaches in the protein structure prediction which are designed to *select* the models from individual programs (Ginalski *et al*., 2003; Wu and Zhang, 2007). Therefore, the final model in the meta-server predictions for a specific protein should be identical to that by some individual program. In COACH, however, the LBS predictions are combined from different programs. Thus, none of the COACH predictions is identical to that of the individual programs, which can be seen in Supplementary Figure S4 whereby there are almost no points on the diagonal lines of the comparisons. The PCC values between COACH and other methods are listed in Supplementary Figure S4, which are approximately associated with the contribution of the component methods to COACH. As shown in the figure, TM-SITE has the highest PCC with COACH, meaning that it contributes the most to COACH, followed by COFACTOR, S-SITE, FINDSITE and ConCavity, respectively.

As shown in Supplementary Table S1, the $P$-values in the Student's $t$-test between COACH and other methods are all $<10^{-14}$, demonstrating that the improvement from consensus is statistically significant.

### 3.5 Test of COACH in blind experiments

CAMEO (Continuous Automated Model EvaluatiOn) is a community-wide ligand-binding experiment, which was designed to evaluate computational methods in a continuous base. Every week, a set of pre-released sequences from the PDB are collected and sent to the online service systems with the LBS prediction generated *before* the experimental structure is released. The prediction results are evaluated based on the subsequently released PDB structures. The CAMEO experiment is complementary to the CASP experiment but has the advantage to assess the participating methods on a large number of targets and in a continuous base, whereby the CASP experiment has often too few targets to draw a reliable conclusion on the ligand-binding prediction (Schmidt *et al*., 2011). Instead of the MCC score, the AUC score of the ROC plot has been used by CAMEO for the evaluation of the LBS predictions.

COACH participated in CAMEO since December 7, 2012. Figure 3 summarizes the official assessment results of the LBS predictions by COACH in the last 22 weeks, together with other 7 predictors, which contains results on 1203 released proteins. The detailed AUC values are listed in Supplementary Table S2, taken from the official CAMEO website http://www.cameo3d.org/ligand_binding/weekly_summary.html. Because the number of targets modeled by different predictors can be different (Supplementary Table S2), we use the 'average accuracy' for the comparison in Figure 3, whereby COACH has generated

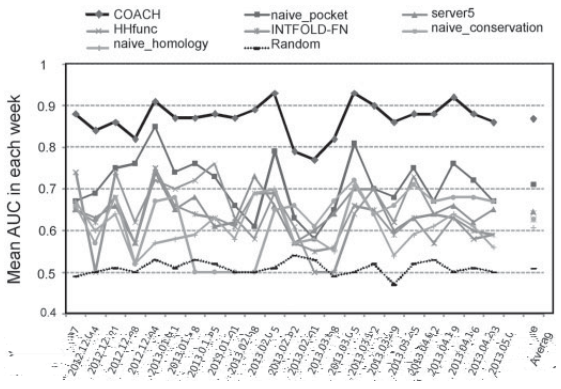


**Fig. 3.** Summary of blind LBS predictions in the COMEO experiment during last 22 weeks. The mean AUC was computed over all targets tested in the corresponding weeks

predictions for all the targets. As shown in Figure 3, the COACH predictions achieve the highest AUC score among all the predictors in each of the 22 weeks. Overall, the mean AUC of COACH is 0.87, which is 22.5% higher than the second best predictor ('naive_pocket'). We have also made an internal assessment of TM-SITE and S-SITE on the same set of proteins, which achieves an AUC score 0.82 and 0.78, which are 15.5% and 10% higher than the 'naive_pocket' program, respectively. Again, TM-SITE outperforms S-SITE in these sets, which demonstrates the advantage of using structure information over sequence profiles. The AUC of TM-SITE is 10.4%, 12.7% and 64.3% higher than that of COFACTOR, FINDSITE and ConCavity, respectively, which are consistent with the data in our benchmark tests although the assessment here is based on AUC whereby the benchmark analysis was on the MCC values.

In Figure 4, we present four examples of the successful COACH predictions in the CAMEO experiment for the proteins bound with four typical categories of ligands (poly-nucleotide, peptide, organic and ion-like). These examples are from (A) DNA binding ETS domain of the human protein FEV in complex with DNA (PDBID: 3zp5A); (B) beta'-COP/Insig-2 complex (4j82A); (C) Glutathione transferase homolog from Lodderomyces elongisporus bound with CIT (4ivfA); and (D) R39-imipenem Acyl-enzyme bound with $Mg^{2+}$ ion (4benA). The I-TASSER models for the receptors all have correct fold with a TM-score >0.5 in these examples (Xu and Zhang, 2010). The BioLiP library contains biologically relevant ligands of both small molecules (e.g. metal ions and organic molecules) and big molecules (e.g. nucleotides and peptides); these provide the opportunity to predict resides bound with various ligand types through the template-based modeling. As a result, the final models of binding sites have the AUC = 0.9, 0.93, 0.97 and 0.97, respectively, which are 32.4%, 52.4%, 27.6% and 73.2% higher than the best predictions from the other server groups.

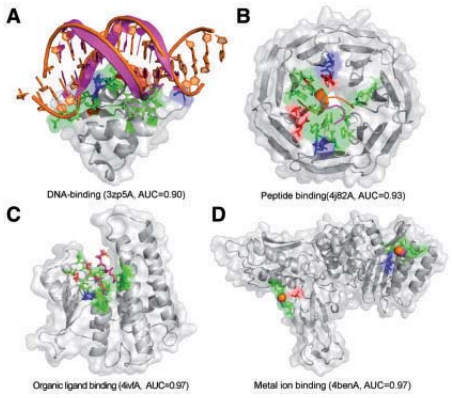


**Fig. 4.** Illustrative examples of successful predictions by COACH in the CAMEO. The receptor structures are shown in gray cartoon buried in transparent surface. The native and predicted ligands are in magenta and orange colors, respectively. The true positive, false positive and false negative predictions of the ligand-binding residues by COACH are highlighted in green, red and blue sticks, respectively. (**A**) DNA binding ETS domain of FEV in complex with DNA. (**B**) beta'-COP/Insig-2 complex. (**C**) Glutathione transferase homolog from Lodderomyces elongisporus bound with CIT. (**D**) R39-imipenem Acyl-enzyme bound with $Mg^{2+}$ ion

### 3.6 Setting up of online COACH server

An online COACH server is set up and made freely available at http://zhanglab.ccmb.med.umich.edu/COACH/. To use the server, users can provide either sequence or 3D structure of the query protein. If the query sequence is provided, I-TASSER (Roy *et al*., 2010) will be used to construct 3D models for the query, which are then used by the individual COACH programs for the structure-based LBS predictions. The final LBS models will be created by the SVM-based combinations. If a 3D model of the target is provided, the structure will be directly used for the LBS predictions.

Starting from a given 3D structural model, the COACH prediction typically takes ~1–10 h depending on the size of the query proteins. An additional time will be needed for the I-TASSER structure prediction (~5–20 h), if the users only provide the sequence. After the prediction is completed, an email alert is sent to user with instruction to access the results, which will be kept on the COACH website for 3 months. For each target, the top 10 COACH predictions are listed, together with the confidence score, ligand cluster size, the representative ligand–protein templates, and the consensus LBS residues. Supplementary Figure S5 shows an illustration of the COACH results, taken from a snapshot of http://zhanglab.ccmb.med.umich.edu/COACH/CH000001/. Except for the consensus LBS predictions by COACH, up to top 5 predictions for the component predictors are also summarized in the same web page.

## 4 CONCLUSION

Accurate identification of LBSs is essential to protein function annotation and drug discovery. Inspired by the fact that no individual methods can generate the optimal prediction for all proteins, we have developed two complementary algorithms, TM-SITE and S-SITE, for protein–ligand binding site predictions. TM-SITE is built on the structural comparison of a subset of continuously distributed residues associated with the binding pockets in the query and template proteins, while S-SITE is based on the binding-specific sequence profile–profile alignments. Starting from the low-resolution 3D models generated by protein structure predictions, the LBS prediction methods are tested on a set of 500 non-redundant proteins harboring 814 natural, drug-like and metal ion ligands, where the correct LBSs with an MCC above 0.5 are generated for 302 and 243 cases, by TM-SITE and S-SITE, respectively. Among the successful predicted cases, 125 cases are predicted by either TM-SITE or S-SITE which demonstrate the complementary feature of the two algorithms.

The methods are controlled with three state-of-the-art methods from COFACTOR (Roy and Zhang, 2012), FINDSITE (Brylinski and Skolnick, 2008) and ConCavity (Capra *et al*., 2009). The TM-SITE predictions have an average MCC 0.48 which is 14% higher than COFACTOR and FINDSITE, and 84.6% higher than ConCavity; the differences correspond to *P*-values $<10^{-11}$, $<10^{-9}$, $<10^{-53}$, respectively, in Student's *t*-test. The detailed data analysis showed that the major advantages of the methods are attributed to the binding-specific substructure alignment search for the template recognition, and the composite potential appropriately balancing multiple scoring matrices from both global and local, structural and sequence comparisons.

Due to the complementarity of the LBS prediction methods, the best predictions are shown to be generated by a consensus of different programs, as collected by a new meta-LBS predictor COACH which outputs the residue-specific ligand-binding probability by the SVM technique. In the same testing set of 500 non-redundant proteins, COACH generates predictions with the MCC value 12.5% higher than the best individual algorithms. COACH is also tested in the community-wide CAMEO experiments, which was ranked as the best method in each of the past 22 weeks with an overall AUC score of 0.87, 22.5% higher than the second best method from other predictors.

An online server for COACH has been set up and made freely