PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Comparison of structure-based and threading-based approaches to protein functional annotation

Michal Brylinski and Jeffrey Skolnick*

Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30318

ABSTRACT

To exploit the vast amount of sequence information provided by the Genomic revolution, the biological function of these sequences must be identified. As a practical matter, this is often accomplished by functional inference. Purely sequence-based approaches, particularly in the "twilight zone" of low sequence similarity levels, are complicated by many factors. For proteins, structure-based techniques aim to overcome these problems; however, most require high-quality crystal structures and suffer from complex and equivocal relations between protein fold and function. In this study, in extensive benchmarking, we consider a number of aspects of structure-based functional annotation: binding pocket detection, molecular function assignment and ligand-based virtual screening. We demonstrate that protein threading driven by a strong sequence profile component greatly improves the quality of purely structure-based functional annotation in the "twilight zone." By detecting evolutionarily related proteins, it considerably reduces the high false positive rate of function inference derived on the basis of global structure similarity alone. Combined evolution/structure-based function assignment emerges as a powerful technique that can make a significant contribution to comprehensive proteome annotation.

Proteins 2010; 78:118–134.

Key words: binding pocket detection; gene ontology molecular function; protein function annotation; protein threading; sequence-based methods; structure-based methods; virtual screening.

INTRODUCTION

In the postgenomic era, the rapid accumulation of proteins whose functions have not yet been experimentally characterized has created a great demand for automated computational tools that can provide insights into their function.[1,2] Many methods have been developed to address this issue; they can be roughly divided into sequence-based and structure-based approaches (for reviews see Ref. 3–5). The simplest approaches infer function from close homologues as detected by sequence similarity.[6–9] However, the functional divergence observed at high levels of sequence identity (60–70%) significantly complicates annotation transfer by homology.[10,11] To address this problem, some sequence-based techniques exploit family specific sequence identity thresholds,[11] increase their accuracy by detecting the presence of functionally discriminating residues[12] or by identifying small sequence signatures and functional motifs.[13–16] Nevertheless, purely sequence-based approaches are in general limited to higher levels of sequence identity; predictions in the "twilight zone" of sequence similarity[17] may be inaccurate. When accuracy is maintained, it is often at the expense of adequate coverage.[18]

To extend functional inference approaches to low levels of sequence identity, a number of structure-based methods have been developed.[19–22] Template-free methods rely on the purely structural properties of the target protein of interest. They analyze the geometrical and physicochemical features of a protein surface in order to detect functionally important sites. Most of these techniques focus on the detection of clefts and cavities that likely bind ligands.[23–29] Other methods consider blind docking of small molecules to the protein's structure,[30] scanning of the protein surface with chemical probes,[31,32] or they provide functional information by examining various physicochemical properties of the protein residues to infer binding sites.[19] These include the degree of surface residue conservation,[33,34] the electrostatic potential,[35,36] the hydrophobicity distribution,[37] perturbed p*Ka* values[38] or the destabilizing effect of local surface residues on the protein's structure.[39,40]

Similar to function annotation approaches based on short sequence motifs, local structural signatures are also widely used to identify functionally important sites in proteins.[41–43] Here, the library of predefined three-

The authors state no conflict of interest.
Grant sponsor: NIH; Grant number: GM-48835.
*Correspondence to: Jeffrey Skolnick, Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, 250 14th Street NW, Atlanta, GA 30318. E-mail: skolnick@gatech.edu
Received 24 June 2009; Accepted 22 July 2009
Published online 5 August 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22566

dimensional arrangements of a small set of key residues is used to screen a target protein structure in order to identify similar motifs that often indicate a common function.[44–46] In general, methods based on structure calculations, as well as local structure comparison approaches are successful when applied to high-resolution structures; their performance typically drops off when approximate protein models, particularly these modeled using remote protein homology,[47,48] are used as the target structures.[49–51] Given the current state-of-the-art in protein structure prediction,[52–55] powerful structure-based methods that effectively utilize low-to-moderate quality protein models for function assignment would be of considerable practical assistance in proteome-scale function prediction. However, all of this is moot until one ascertains what precisely are the limits of functional inference given exact experimental structures. This will constitute the upper bound that any approach using predicted structures could achieve.

It is well known that within a protein family, the global fold is more strongly conserved than the protein's sequence.[56] Hence, at low sequence identity levels, structure-based identification of remote homology and functional relationships inevitably outperforms sequence-based methods.[57–60] Examination of known protein structures in the SCOP database[61] reveals the tendency of certain protein folds to bind substrates at a similar location, suggesting that very distantly homologous proteins often have common binding sites.[62] This observation forms the basis for FINDSITE, a structure/evolution-based approach for ligand-binding site prediction and function annotation.[49] However, one should bear in mind that divergent and convergent evolution results in a non-unique relationship between protein fold and protein function.[63,64] Therefore, template-based function inference using solely global structure similarity might lead to a high false positive rate.

While a variety of purely structure-based approaches to functional inference have been developed,[36,60,65,66] their precision, sensitivity, and specificity have not been assessed in a large-scale benchmark. To address this issue, in this article, we present the results of a large-scale benchmark comparison of structure-based and threading-based approaches to the inference of protein function, given the experimental structure of the protein of interest. The simplest structure-based approach for functional inference merely requires significant structural similarity between a pair of proteins. As shown below, to achieve a low false positive rate, using structure alone requires a high structure similarity threshold, which results in very low coverage. This problem can be addressed by introducing various filters. Here, we demonstrate that the use of threading[47,67] identified templates that share a common binding site greatly reduces the high false positive rate in template-based function annotation by detecting evolutionarily related homologues. Furthermore, rather similar ligands tend to bind at a given common location in the protein's structure; this emphasizes the importance of a local component, such as spatial ligand clustering, in ligand selection for virtual screening. We compare the set of templates selected on the basis of significant structure similarity to these identified from protein threading with respect to the conservation of ligand-binding sites and chemical properties of bound ligands. In the "twilight zone" of sequence identity, the accuracy of function assignment is assessed at the level of binding site prediction, molecular function transfer and the construction of ligand templates for use in virtual screening.

MATERIALS AND METHODS

Dataset

We consider a representative dataset of 901 nonhomologous protein-ligand complexes that cover the PDB at 35% sequence identity, where the lengths of the proteins are between 50 and 400 residues and the minimum and maximum number of ligand atoms ranges from 6 to 100.[49] Of these, we selected 842 targets for which at least one weakly homologous (less than 35% sequence identity) template with a Z-score[47] $\geq$4 and a TM-score[68] $\geq$0.4 can be identified by threading. Thus, all templates with a sequence identity >35% to the target are excluded from all aspects of the analysis. Moreover, a subset of 710 targets for which a gene ontology (GO) annotation is provided by Gene Ontology[69] or UniProt[70] was used to assess the performance of the molecular function transfer. The datasets are available at http://cssb.biology.gatech.edu/skolnick/files/FINDSITE.

Overview of structure-based and threading-based approaches to function assignment

To assess the importance of protein threading for template-based function assignment (our variant of which is the FINDSITE algorithm[4,49]), as shown in Figure 1, we apply four procedures for template selection. For a given target, template structures are selected from the template library either by structure alignment to the native target structure (assumed to be apo throughout this analysis) using a purely structure-based approach or threading, (threading-based approach). For the case of functional inference, we can simply collect the GO terms for templates above a structural similarity threshold, (left hand pane) structure-based approach (no local filtering). We can also estimate an upper bound for the performance of a purely structure-based approach by using only those templates that in addition to the significant global structure similarity to their targets also have ligand-binding sites in similar locations to the target structure (structure-based approach, correct pocket localization). Here, we consider template structures whose binding pockets

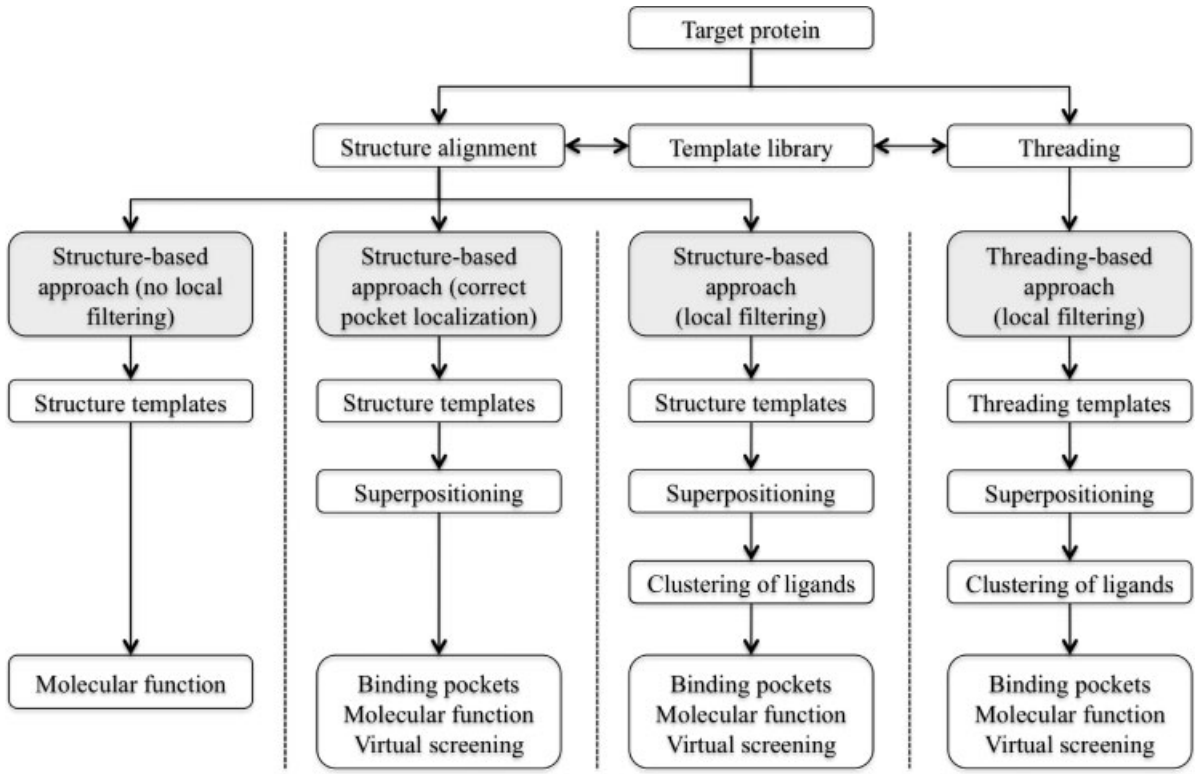

Figure 1

Flowchart of structure- and threading-based approaches to function inference. Details are given in the text.

are within a distance of 4 Å from the target pockets upon structure alignment. If no such templates can be identified for a given target protein, the distance is gradually increased by 1 Å until at least one template is found.

For binding site based functional inference, both structure-based and threading-based approaches, follow the same procedure to predict binding pockets and to assign the function (Fig. 1, two right panes). Ligand-bound template structures are superimposed onto the target's structure using the TM-align structure alignment algorithm[71]; see below. Then, binding pockets are identified by the spatial clustering of the centers of mass of template-bound ligands by an average linkage clustering procedure and ranked by the number of binding ligands. This step is termed "local filtering." Thus, this is a binding site matching approach based on the location and frequency of bound ligands. It is not based on identifying clefts present in the protein structure. The simulation time depends on the number and size of the identified template proteins and varies from minutes to hours on a single state of the art processor core. For template-based binding site prediction, the fraction of templates that share a common top-ranked binding site is used to construct a primary confidence index that classifies the reliability of the pocket prediction as easy, medium, or hard.[48] We demonstrate below that the overall accuracy of binding site prediction is well correlated with this classification. In all cases, the performance of structure- and threading-based approaches is compared with randomly selected patches on the target protein surface.

Structure-based template selection

Given a native structure, the structure-based approach uses structural alignments to identify the set of relevant templates. In general, protein structure alignment approaches attempt to establish equivalences between a pair of structures based on their three-dimensional conformation where the equivalent residues are not a priori given.[71–73] Here, we use the TM-align structure alignment algorithm[71] that combines the TM-score[68] rotation matrix and Dynamic Programming to identify the "best" structural alignment. By weighting small inter-structural distances stronger than large distances, the TM-score rotation matrix is more sensitive to the global topology than the traditionally used global root-mean-square-deviation[74] (RMSD). Moreover, the statistical significance of the alignment for a given TM-score is protein length independent. For a pair of randomly

related protein structures, their average TM-score is 0.30, with a standard deviation of 0.01. For each target protein, structurally similar templates (with <35% sequence identity to the target protein) are selected from the template library based on the TM-score reported by TM-align. Here, we used the TM-score threshold of 0.4, which is indicative of highly significant structural similarity.[75] A detailed comparison of the performance of TM-align with other algorithms has been done elsewhere.[71,76,77]

Threading-based template selection

Protein threading was developed to match target sequences to proteins adopting very similar structures.[67] In practice, threading that employs a strong sequence profile component[47] works by detecting evolutionary related proteins.[49] For a given sequence, template structures are identified from a nonredundant fold library by threading the target sequence through the template structures and selecting the best alignment by a scoring function. Score significance is evaluated by a *Z*-score (score in standard deviation units relative to the mean of the structure template library) of the sequence mounted in a given template structure using the best alignment given by dynamic programming. For threading-based template selection, we used the PROSPECTOR_3 program,[47] but in principle any-state-of-the-art algorithm can be used with comparable results (unpublished results). From the threading templates provided by PROSPECTOR_3, we used only those templates with <35% sequence identity to the target protein, a *Z*-score ≥4 and a TM-score ≥0.4 between the template and the experimental structure.

Template selection by a sequence profile–profile algorithm

Ligand-binding site prediction using the set of templates selected by threading is compared to those identified by a sequence profile-profile alignment algorithm. Here, we use HHpred 1.5.0.1, which is based on the pairwise comparison of profile hidden Markov models (HMMs).[78] For a given target sequence, the HMM profile is constructed from a nonredundant sequence library and the secondary structure is predicted by PSIPRED 2.61.[79] Subsequently, each query HHM is calibrated on a nonredundant SCOP[61] library. Remote homologues (<35% sequence identity to the target and a TM-score ≥0.4) are selected from the template library using an estimated probability of 0.5 for a template to be evolutionarily related to the target sequence. If no hits are detected at the 0.5 threshold for a given target protein, a probability of 0.3 is used. The set of templates selected by HHpred are used to replace those identified by threading or structure alignment in ligand binding site prediction by FINDSITE (see two right panes in Fig. 1).

Template-free pocket prediction methods

The results of ligand-binding site prediction using the template-based approach (FINDSITE[49]) were compared to those obtained using geometric template-free algorithms: Ligsite^{CS}34 and Fpocket.[26] Ligsite^{CS}, an extension of Ligsite,[24] uses the Connolly molecular surface (BALL implementation[80]), which is a combination of the van der Waals surface of the protein and the probe sphere surface to detect putative binding sites. Fpocket employs Voronoi tessellation (Qhull implementation[81]) and evaluates the identified binding sites using a pocket score that considers several pocket descriptors: the number of alpha spheres, the cavity density, a polarity score, a mean local hydrophobic density and the ratio of apolar/polar alpha spheres. For both programs, the default set of parameters was used.

Inference of molecular function

Each target protein is annotated with a set of GO terms[69] extracted from the template proteins. Molecular function is transferred from the templates to the target protein with a probability that corresponds to the fraction of templates annotated with a particular GO term. GO parent nodes are traced to explore the more general ontology classes. Function transferability is investigated for an increasing probability threshold from 0.0 (all template GO terms are transferred to the target) to 0.95 (only highly conserved GO terms that are common for 95% of the templates are transferred). Of course, such an approach has all the disadvantages and advantages of the GO description of molecular function. Function annotation using the sets of threading and structure templates is compared with randomly assigned molecular function according to the frequencies of GO terms in the dataset. The results are assessed by Precision-Recall analysis[82] with the precision and recall defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{2}$$

where TP, FP, and FN denote true positives, false positives and false negatives, respectively.

In addition to the molecular function annotation, the conservation of GO terms with respect to the TM-score is evaluated for all target-template pairs. Here, we use Matthew's correlation coefficient (MCC) to quantify the functional similarity between a template and its target:

$$\text{MCC} = \frac{\text{TP}\times\text{TN} - \text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})\times(\text{TP}+\text{FN})\times(\text{TN}+\text{FP})\times(\text{TN}+\text{FN})}} \tag{3}$$

where TP, TN, FP, and FN denote respectively: true positives (number of GO terms common for both the target and its template), true negatives (number of GO terms absent in the template as well as in its target), false positives (number of GO terms specific only for the template), and false negatives (number of GO terms specific only for the target).

Virtual screening

As in the case of FINDSITE,[4,49] for the purely structure-based approach, we can also exploit information on the chemical properties of the binding ligands to construct ligand templates for ligand-based virtual screening. For each predicted binding pocket, the bound ligands are extracted from the template complex structures, converted into 1024-bit SMILES strings[83] and clustered using a Tanimoto coefficient[84] of 0.7. Subsequently, representative molecules selected from the clusters are used to rank the screening library using a weighted Tanimoto coefficient ($m\mathrm{TC}^{\mathrm{ave}}$):

$$m\mathrm{TC}^{\mathrm{ave}} = \sum_{i=1}^{n} w_i \mathrm{TC}_i^{\mathrm{ave}} \quad (4)$$

where n is the number of ligand clusters, w_i is the fraction of ligands that belong to cluster i, and $\mathrm{TC}_i^{\mathrm{ave}}$ is the average TC ($\mathrm{TC}^{\mathrm{ave}}$) calculated for the representative ligand from cluster i and a library compound. The overlap between two fingerprints is measured by the average Tanimoto coefficient, $\mathrm{TC}^{\mathrm{ave}}$, defined as[84–86]:

$$\mathrm{TC}^{\mathrm{ave}} = (\mathrm{TC} + \mathrm{TC}')/2 \quad (5)$$

where TC′ is the TC calculated for bit positions set to zero rather than to one as in the traditional TC.[84]

As a background library in ligand-based virtual screening, we use the KEGG compound library[88] that consists of 12,158 chemically diverse molecules. The performance of threading and structure based template selection is assessed based on the ranks assigned to the compounds complexed with the target proteins in the crystal structure with respect to the background molecules and compared to random ligand selection. Finally, similar to the primary confidence index for pocket detection, we demonstrate that the relative size of the largest cluster of ligands extracted from the predicted binding sites (with a minimum of five ligands) can be used as a secondary confidence index to assess the reliability of ligand ranking.

RESULTS

Functional and structural relationships between templates and their targets

First, we analyze the conservation of ligand binding features in the templates selected by threading and structure alignment with respect to the target crystal structures. Figure 2 shows the fraction of templates whose binding pocket center is within 4 and 8 Å from the target's pocket center as a function of the TM-score. We again note that a TM-score ≥0.4 indicates significant structural similarity. Clearly, threading effectively detects and eliminates evolutionarily unrelated proteins with different binding site localization, particularly for a TM-score <0.7. For example, only 6% (18%) of the templates selected based on structure similarity alone and having a TM-score between 0.4 and 0.5 bind ligands within a distance of 4 Å (8 Å) from the target bound ligand [Fig. 2(A)]. Using threading, this fraction increases to 34% (49%) [Fig. 2(B)]. As shown in subsequent analysis, the higher fraction of templates that bind ligands in similar locations greatly improves the accuracy of the pocket prediction and the ranking capability in particular.

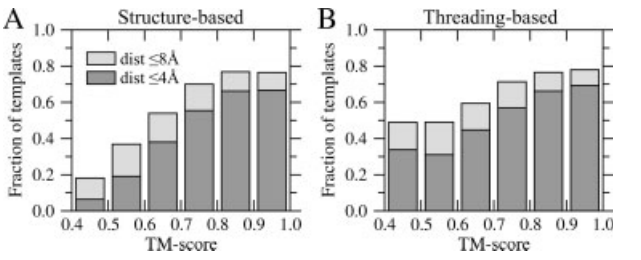

Figure 2

Fraction of templates selected by (**A**) structure alignment and (**B**) threading that have a binding pocket center within 4 and 8 Å from the target pocket center as a function of the TM-score.

Next, the conservation of molecular function according to the gene ontology classification, one of the most common classification systems for proteins that provides the functional description for both enzymes and nonenzymes,[69] is presented for all target-template pairs in Figure 3. Here, we use GO molecular functions, which typically describe molecular events such as catalytic or binding activities that can be directly linked to the active or binding sites. A relatively low functional similarity between structure templates selected based on TM-score alone and their targets is observed for a TM-score <0.7 [Fig. 3(A)]. However, the number of templates annotated with similar GO terms increases if threading filtered templates are considered [Fig. 3(B)]. For a TM-score of 0.4–0.5, the fraction of templates that have similar molecular functions, as assessed by a MCC ≥0.7 (≥0.5), is 3% (9%) and 17% (34%) for the structure-based and threading-based set of templates, respectively. Moreover, as shown in Figure 3(C), the fraction of templates having the same gene ontology classification as the target significantly improves when local filtering by the common binding site localization is applied. Here, for a TM-score of 0.4–0.5, the fraction

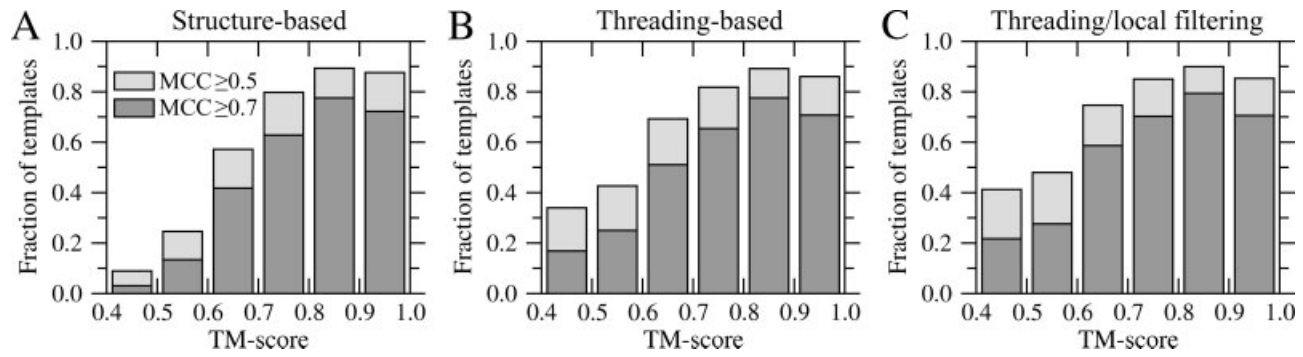

Figure 3

Fraction of templates annotated with similar GO terms as their targets as selected by (**A**) structure alignment and (**B**, **C**) threading as a function of the TM-score. Functional similarity is assessed by a Matthew's correlation coefficient (MCC) ≥0.7 and ≥0.5. In C, templates that bind ligands with a distance from the target-bound ligand >8 Å are excluded.

of threading templates with a MCC ≥0.7 (≥0.5) to their targets is 22% (41%). Furthermore, as it is evident in Figure 4, protein threading also tends to detect templates that bind similar ligands to the target-bound molecules. For a TM-score of 0.4–0.5, 8% (14%) of the structure-based templates bind ligands whose Tanimoto coefficient to the native ligand is ≥0.7 (≥0.5) [Fig. 4(A)]. This fraction increases to 28% (39%) if the templates identified by threading are used [Fig. 4(B)]. Additionally, local component, filtering by the common localization of the binding pockets, promotes the selection of ligands with even higher chemical similarity to the target-bound compounds for all values of the TM-score above 0.4, This is shown in Figure 4(C), where only threading filtered templates that have binding sites within 8 Å from the target-bound ligands are considered. Here, 50% of ligands have a Tanimoto coefficient ≥0.7 in the ranges of TM-scores between 0.4 and 0.5. This effect is of particular importance in virtual screening, where the screening library is ranked by fingerprint-based ligand profiles constructed from the compounds extracted from the template complex structures.

These results suggest that in practice, a high structural similarity cutoff should be used for the template selection in purely structure-based function assignment. The disadvantage of such an approach is that this structurally discriminative threshold eliminates most functionally related templates; thus the number of suitable targets for template-based function annotation would be limited. As shown in Figure 5, 90% of the templates in our dataset have a TM-score <0.7 to their targets. In the evolutionarily related set of templates, as provided by threading, 60% of the templates have a TM-score <0.7; these would be undetected if one applied a discriminative TM-score cutoff of 0.7. Hence, protein threading appears as a more functionally oriented filter that allows using more permissive structural similarity cutoffs with the false positive cases eliminated by evolutionary restraints.

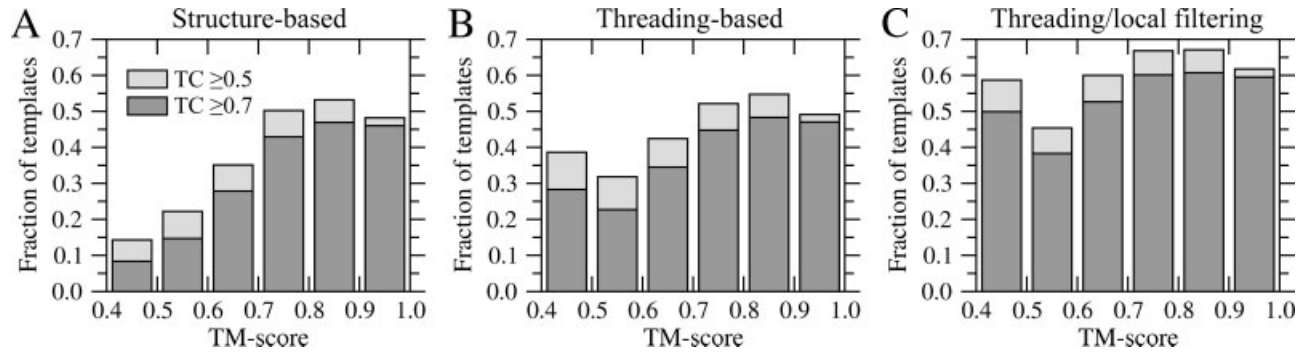

Figure 4

Fraction of templates selected by (**A**) structure alignment and (**B**, **C**) threading that bind similar ligands to the target-bound molecules as a function of the TM-score. Chemical similarity of ligands is assessed by a Tanimoto coefficient (TC) ≥0.7 and ≥0.5. In C, templates that bind ligands with a distance from the target-bound ligand >8 Å are excluded.

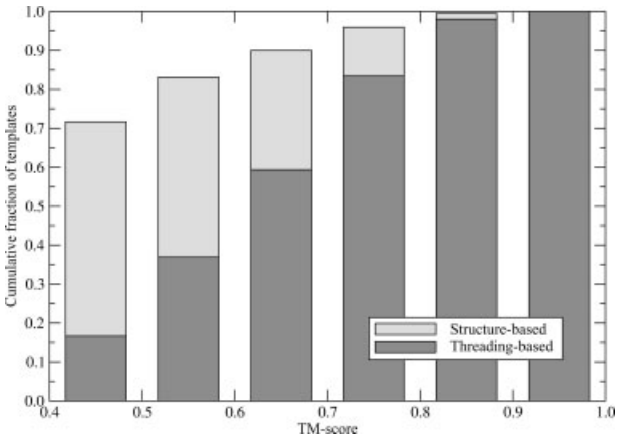

Figure 5

Cumulative fraction of template proteins selected by structure alignment and threading that have a TM-score to the target crystal structure ≤ the value on the *x*-axis to the right of the corresponding bar graph.

Template detection by threading

Finally, for each target protein, we identify the largest set of templates with similar binding pockets and assess the recall and precision of their detection by threading. These templates are selected based on significant global structure similarity (TM-score ≥0.4), similar ligand-binding site localization (the distance between target-bound and template-bound ligands upon the structure alignment of the proteins ≤4 Å) and the chemical properties of the bound ligands (Tanimoto coefficient between target-bound and template-bound ligands ≥0.7). Figure 6 presents the recall and precision of the template identification by threading with respect to the global target-template structure similarity. Above a TM-score of 0.5, the recall of templates with similar pocket localization is >0.66. When the chemical similarity of the bound ligands is also taken into account, the recall of the templates increases to >0.79 for a TM-score ≥0.5. This clearly demonstrates that protein threading effectively detects template structures that bind chemically similar ligands in similar locations. However, the substantially lower precision values suggest that the threading-identified set of templates also contain many false positives, that is, proteins which, despite their global structure similarity to the target, bind ligands in different locations or tend to bind chemically unrelated (Tanimoto coefficient <0.7) molecules in similar locations. As we show in the function annotation benchmarks (see below), this false positive rate can be considerably reduced using subsequent filtration by the subset of templates that share the most frequent binding site.

Function annotation benchmarks

Next, we assess the performance of purely structure-based and threading-based templates, in comprehensive function annotation in the "twilight zone" of sequence similarity. Protein function has many facets, ranging from biochemical to cellular to phenotypical.[3,88] In this work, we focus on catalytic and binding activities that involve direct interactions with small molecules and report the results of functional annotation at the level of binding pocket detection, molecular function assignment and ligand-based virtual screening.

Primary confidence index for pocket detection

Binding pockets are identified by the spatial clustering of the center of mass of template-bound ligands aligned to the target crystal structure and ranked by the number of binding ligands. Figure 7(A,B) show for purely structure similarity based approach and the threading-based approach the fraction of targets for which the binding pocket center can be predicted within a distance of 4 and 8 Å as a function of the fraction of templates that share a common top-ranked binding site, with a minimum of five templates identified. Quite similar behavior for structure-based and threading-based approaches is seen. High accuracy in binding pocket prediction typically requires a relatively high fraction (≥0.4) of the templates that have a common pocket. If this fraction drops below 0.2, the chances that the top-ranked binding site is predicted within 4 or 8 Å are rather low (~20% using 4 Å as a hit criterion). We use this observation to construct a primary confidence index that classifies targets as Easy (≥0.4), Medium (<0.4 and ≥0.2) and Hard (<0.2 or <5 templates) for binding pocket prediction.

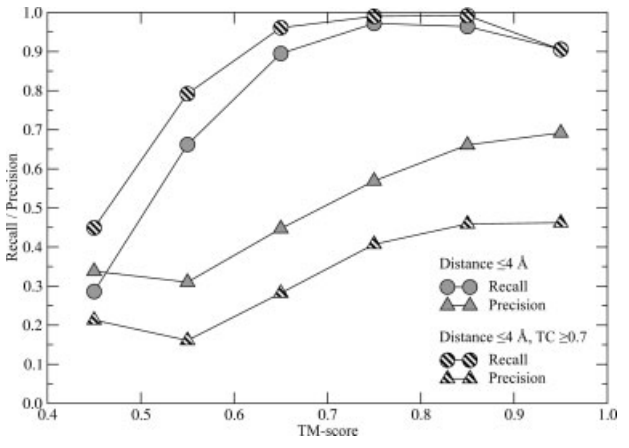

Figure 6

Recall and precision of template detection by threading as a function of the target-template global structure similarity (assessed by the TM-score, *x*-axis).

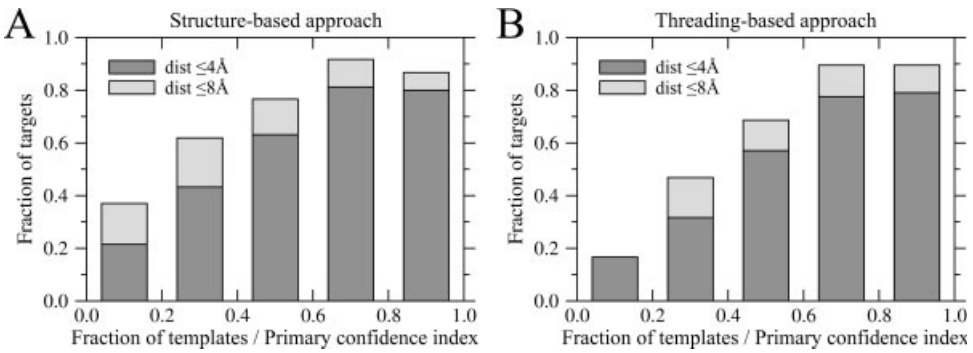

Figure 7

Fraction of targets for which the top-ranked binding pocket's center was predicted within a distance of 4 and 8 Å from the center of mass of a ligand in the crystal complex as a function of the primary confidence index for (**A**) structure-based and (**B**) threading-based approach. The primary confidence index corresponds to the fraction of templates that share a common top-ranked predicted binding site.

The fraction of Easy, Medium and Hard targets in the benchmark set of 842 proteins is presented in Figure 8 for the set of templates selected by structure similarity and threading. The high content of false positives in the structure-based set of templates leads to mainly moderate confidence predictions (44.1%) [Fig. 8(A)]. In contrast, most of the proteins in the dataset appear as Easy targets if the threading filtered set of templates is used [Fig. 8(B)]. Consequently, for these targets, the threading-based template selection approach identifies binding pockets with quite high accuracy, as shown below.

Binding pocket prediction

The performance of structure-based and threading-based template identification in binding site detection and ranking is presented in Figure 9. Figure 9(A) shows the cumulative fraction of proteins for which the center of the best of top five predicted binding sites was predicted within some distance from the center of mass of a ligand in the crystal complex. As the set of structure-based templates can be considered as a superset with respect to the templates selected by threading, all binding pockets identified using threading-based templates are also detected by employing the structure-based set. This explains the relatively small difference (5%) in the pocket distance prediction using a 8 Å cutoff as a hit criterion, if the best of top five predicted binding sites is considered. However, a significant drop off in the ranking capability is observed when structure-based templates are used [Fig. 9(A), inset]. For the set of templates selected by structure similarity and threading, the best predicted binding pocket is at rank 1 in 56.3 and 78.5% of the cases, respectively. In addition, Figure 9(B) presents the ranking accuracy when the top 100 predicted binding pockets are considered. Here, the ability of the structure- and threading-based approach to assign the best pocket with rank 1 is 50.2% and 75.9%, respectively.

Using the set of structure templates with similar binding site localization (see Materials and Methods), an estimated upper bound for pocket detection accuracy is 93.9% (98.0%) for a distance threshold of 4 Å (8 Å) [Fig. 9(A), Structure/pockets curve]. To provide a better assessment for the performance of the threading-based approach with respect to the theoretical limit, we subsequently divided the dataset into two subsets. The first subset consists of 555 targets for which there is at least one template that bind ligands within a distance of 4 Å from the binding site in the target structure and whose Tanimoto coefficient to target-bound compounds is $\geq$0.7. The second subset comprises 259 targets for which no templates that bind similar ligands within 4 Å or only these that bind chemically dissimilar ligands (TC $<$ 0.7) can be identified. We note that 28 targets for which HHpred[78] failed to detect remotely related templates with $<$35% sequence identity and a TM-score of $\geq$0.4 are removed from this analysis; thus threading provides slightly higher coverage. In Figure 10, we compare the performance of the threading-based approach and the sequence profile-profile algorithm (HHpred) to the

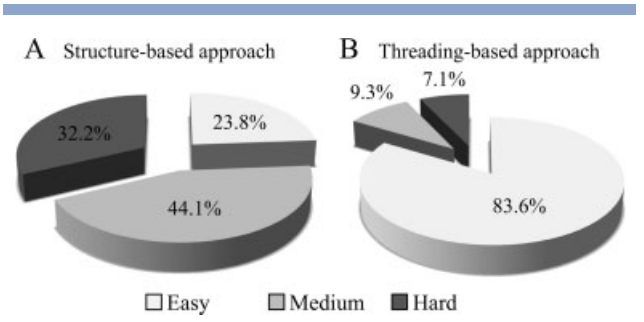

Figure 8

Primary confidence of FINDSITE predictions using the set of templates selected by structure alignment and threading.

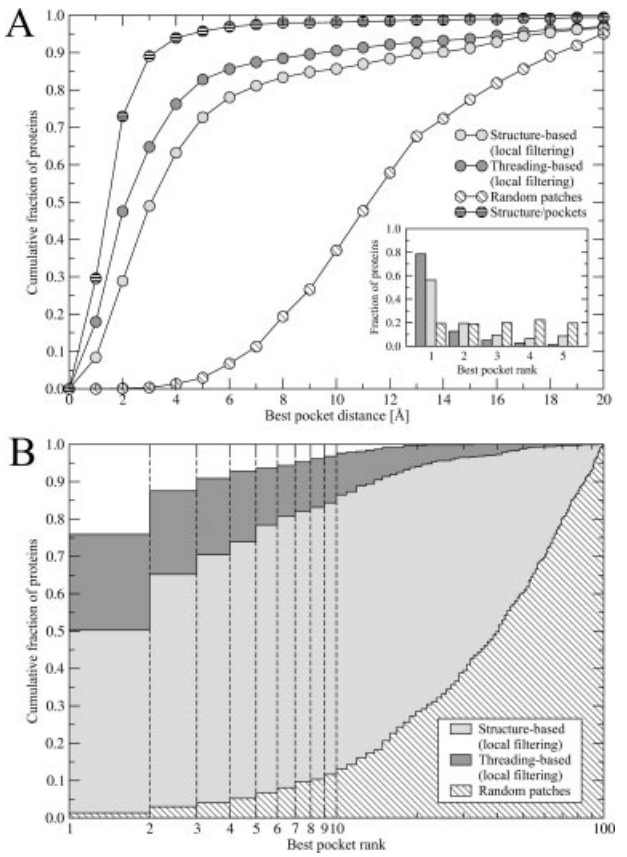

Figure 9

Performance of FINDSITE in binding site prediction using the set of templates selected by structure alignment and threading compared with randomly selected patches on the target protein surface and similar pockets identified in the template structures (Structure/pockets). (**A**) Cumulative fraction of proteins with a distance between the center of mass of a ligand in the crystal complex and the center of the best of top five predicted binding sites displayed on the *x*-axis. Inset plot shows the rank of the best of top five predicted binding sites. (**B**) Best pocket rank considering the top 100 predicted binding pockets.

cleft-based methods for binding site prediction: Ligsite^{CS}34 and Fpocket,[26] separately for each subset of targets. Figure 10(A) shows the results obtained for the first subset of 555 targets. Here, the performance of both methods that employ sequence profiles, the threading-based approach and HHpred, is very close to the theoretical limit at detecting binding sites with a similar location in the structure that are associated with binding ligands with similar chemical properties, whereas the accuracy of both cleft-based methods is significantly lower. Considering a distance cutoff of 4 Å and the best of top five predicted binding sites, the accuracy of the threading-based approach, HHpred, Ligsite^{CS}, and Fpocket for the first subset is 91.0, 90.4, 49.4, and 51.9%, respectively (98.9, 99.1, 78.6, and 86.7% for a 8 Å cutoff, respectively). The strong preference of sequence profile driven methods to detect binding sites that bind similar ligands (see Figs. 4 and 6) explains the relatively lower accuracy of the threading-based approach and HHpred in pocket detection for the second subset of 259 targets [Fig. 10(B)]. Most of the templates that bind dissimilar ligands in similar locations remain undetected by threading and HHpred; therefore their performance is well below the estimated theoretical limit. Here, the performance of the cleft-based methods is notably higher. Again, considering a distance cutoff of 4 Å and the best of top five predicted binding sites, the accuracy of the threading-based approach, HHpred, Ligsite^{CS}, and Fpocket for the second subset is 49.4, 50.2, 54.4, and 47.5%, respectively (70.3, 70.7, 82.6, and 84.2% for a 8 Å cutoff, respectively). As expected, the performance of Ligsite^{CS} and Fpocket is comparable for both subsets of targets. We also find that the deterioration in accuracy of the sequence profile driven approaches observed for the second subset is accompanied by a significantly higher fraction of Medium and Hard targets according to the primary confidence index (see above); this is shown as the inset bar plots in Figure 10.

GO function transfer

The prediction of the binding pocket localization is typically the initial step in function annotation that is followed by more detailed molecular function assignment, for example, using the GO ontology.[69] The accuracy of the function transfer from the structure- and threading-based templates to the target is evaluated by Precision-Recall graphs that have been suggested to provide an appropriate measure for skewed class distributions,[82] as such, they are frequently used for assessing GO protein function prediction methods.[90,91] Figure 11(A,B) presents the Precision-Recall graphs for the top-ranked and the best of top five predicted binding pockets, respectively. Because of the relatively high false positive rate, both precision and recall are notably lower for the top-ranked binding sites using structure-based templates [Fig. 11(A)]. This is because in almost half of the cases (49.8%), the best pocket is at a lower rank [see Fig. 9(B)]. If the best of top of five binding sites is considered, then the precision of function annotation using the templates selected by structure similarity (open squares) is comparable to the precision obtained for the threading filtered set of templates (solid circles) [Fig. 11(B)]; however, it is comparable only at significantly lower recall levels (by 20–30%). Moreover, a minor improvement is observed if a local component (spatial clustering of ligands, solid triangles) is applied in purely structure-based function assignment [Fig. 11(B)]. Interestingly, the overall accuracy of molecular function transfer using structure templates with similar binding site localization as the native structure (solid squares) is lower than that when only threading identified templates are used. That is, global structure similarity and the com-

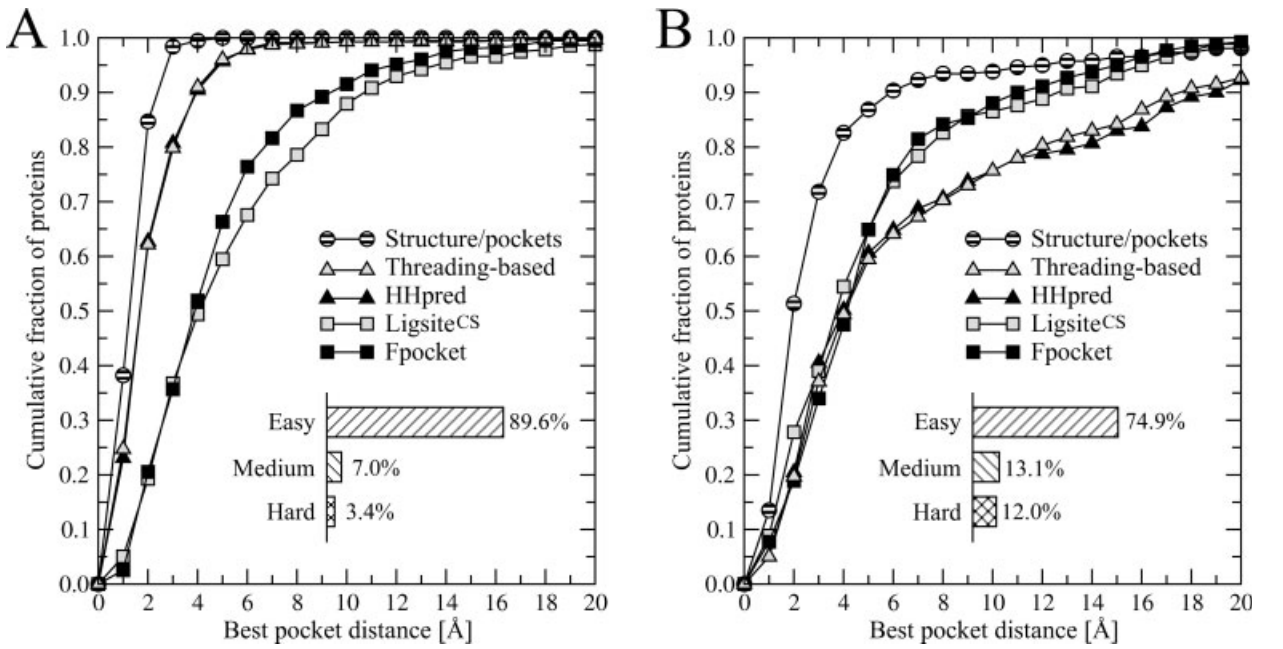

Figure 10

Performance of FINDSITE in binding site prediction using the set of templates selected by threading and HHpred compared with Ligsite^{CS}, Fpocket and similar pockets identified in the template structures (Structure/pockets). Results are presented as the cumulative fraction of proteins with a distance between the center of mass of a ligand in the crystal complex and the center of the best of top five predicted binding sites displayed on the *x*-axis. Binding site prediction accuracy is reported for (**A**) the subset of 555 proteins with at least one template structure that binds a similar ligand in the similar location and (**B**) the subset of 259 proteins with no such templates (see text for details). Inset bar plots show the fraction of easy, medium, and hard targets using the primary confidence index.

mon binding site location do not automatically imply a common molecular function. These results clearly show that threading not only eliminates false positives with respect to the binding site location, but also detects and removes proteins that, despite common binding sites, have unrelated molecular functions, and thus bind

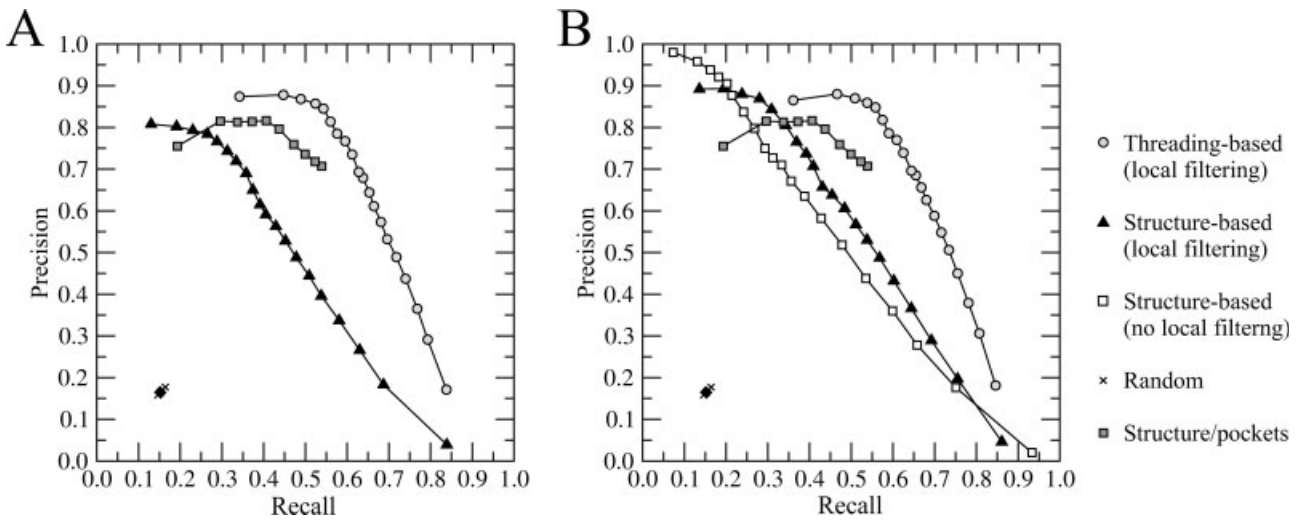

Figure 11

Precision-Recall graphs for GO molecular function prediction using the set of templates selected by structure alignment and threading, compared to random function assignment and function transfer from similar pockets identified in the template structures (Structure/pockets). The results are shown for (**A**) the top-ranked and (**B**) the best of top five predicted binding sites.

chemically dissimilar ligands. In addition to the more accurate function annotation, this important feature of protein threading can be further exploited in the construction of molecular fingerprints for virtual screening.

Ligand-based virtual screening

Finally, in the last step of the comprehensive function annotation, we carry out virtual screening simulations for the 842 proteins present in the benchmark dataset. Here, we apply a simple fingerprint-based method to rank a diverse and relatively large screening library (see Materials and Methods) and assess the rank of the natively bound ligand, that is, the compound co-crystallized with the target protein. Figure 12(A,B) present the results for the top-ranked and the best of top five predicted binding sites compared to random ligand selection. Similar to the case of molecular function transfer, the performance in virtual screening is significantly better when threading-based set of templates is used. Native ligands are found in the top 1% (10%) of the ranked library in 28.7% (52.1%) and 50.2% (71.4%) of the cases, when the scoring fingerprints are constructed using ligands extracted from top-ranked binding sites predicted from the structure- and threading-based set of templates, respectively [Fig. 12(A)]. Of course, this difference can be trivially explained by the ineffective ranking of the binding sites caused by the high false positive content in the set of templates selected by structure similarity alone; the molecular fingerprints are constructed from ligands that occupy incorrectly predicted pockets. Nevertheless, even when the best of top five binding sites is used in virtual screening [Fig. 12(B)], the accuracy when threading templates are employed is still notably higher than that obtained using the structure-based set of templates. The native ligand is ranked within the top 1% (10%) in 54.4% (75.3%) of threading-based and 37.6% (62.6%) of structure-based, cases respectively. Furthermore, ligand fingerprints provided by protein threading perform comparably (top ranked binding pockets) or better (the best of top five predicted pockets) in ligand-based virtual screening than these extracted from the binding pockets with similar localization in the template and native structures (Fig. 12, Structure/pockets). These results demonstrate that threading identifies functionally related templates bound to chemically similar ligands.

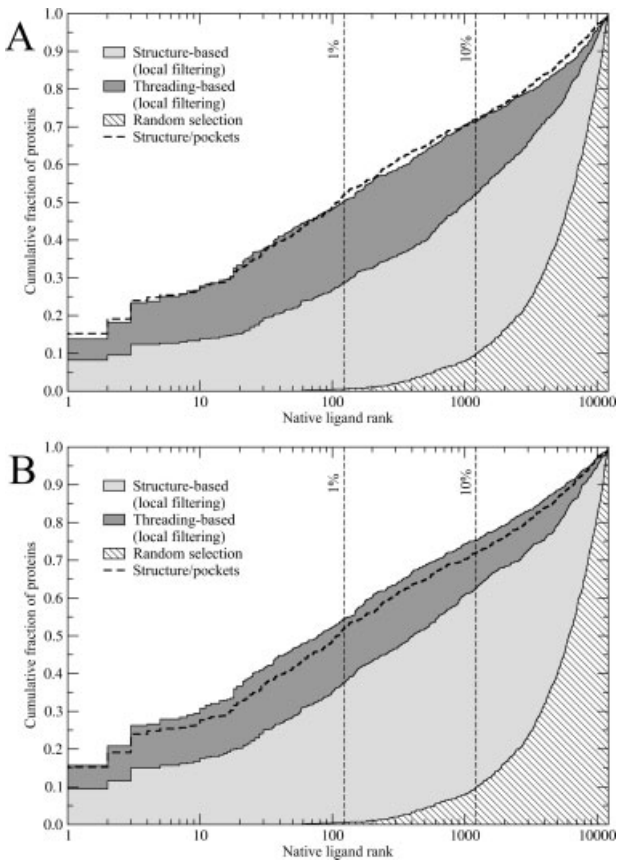

Figure 12

Performance of ligand-based virtual screening to identify the native bound ligand using the set of templates selected by structure alignment and threading. The native ligand ranking accuracy using ligand templates extracted from (**A**) the top-ranked and (**B**) the best of top five predicted binding sites is compared to random ligand selection and ligand ranking using spatially similar pockets identified in the template structures (Structure/pockets). Dashed lines delineate the top 1% and 10% of the ranked screening library.

Secondary confidence index for ligand ranking

The fraction of templates that share a common top-ranked predicted binding site is used to construct a primary confidence index, which is well correlated with the overall accuracy of pocket detection (see Fig. 7). However, as we show in this study, the common localization of ligand-binding sites does not always imply a common molecular function. Therefore, ligand-based virtual screening that follows the prediction of binding pockets requires a separate confidence index. Here, we use the relative size of the largest cluster of ligands extracted from the predicted binding sites (see Materials and Methods) to construct a secondary confidence index for assessing the reliability of ligand ranking. Figure 13 presents the fraction of targets for which the native ligand was ranked within the top 1 and 10% of the screening library using both structure-based [Fig. 13(A)] and threading-based [Fig. 13(B)] approaches with respect to the secondary confidence index. Effective ligand-based virtual screening requires relatively a high fraction ($\geq$0.4) of the ligands that form the largest cluster of chemically similar molecules. If this criterion is satisfied, one can expect the native ligands to be ranked within the top 1%

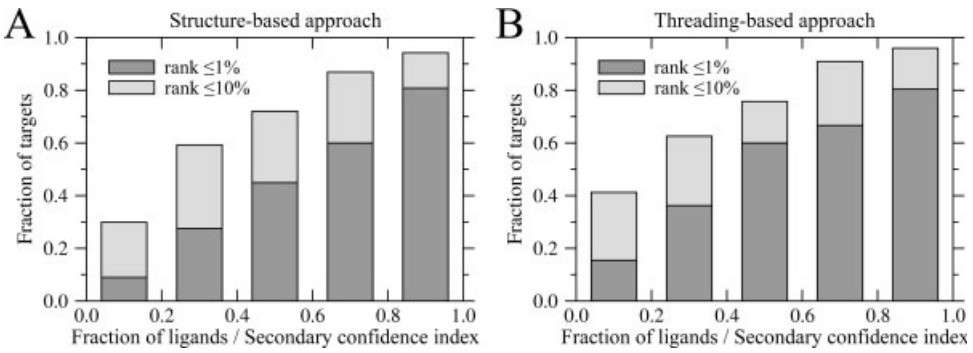

Figure 13

Fraction of targets for which the native ligand was ranked within the top 1% and 10% of the screening library as a function of the secondary confidence index for (**A**) structure- and (**B**) threading-based approach. The secondary confidence index corresponds to the relative size of the largest cluster of ligands extracted from the best of top five predicted binding sites.

(10%) of the screening library by the structure- and threading-based approach in at least 44.9% (72.0%) and 60.0% (75.7%) of the cases, respectively. The accuracy of ligand ranking drops dramatically if the fraction of ligands in the most populated cluster is <0.2, which indicates that the template proteins bind a collection of chemically diverse ligands. It is noteworthy that when the threading-based set of templates is used, the fraction of highly confident targets according to the secondary confidence index is higher than for the set templates selected by structure similarity; the fraction of ligands that form the largest cluster is ≥ 0.4 for 64.4 and 46.8% of the cases for the threading-based and structure-based approach, respectively.

Performance of the threading-based approach (FINDSITE) in CASP8

In addition to the results of ligand-binding site prediction and virtual screening reported in this study for a representative dataset of protein-ligand complexes, we assess the performance of the threading-based approach, FINDSITE,[49] in CASP8. FINDSITE was an integral part of the SiteHunter server (group number 163) in the function prediction category. The predicted binding sites for small ligands are presented in Table I for 13 CASP8 targets (metal ion binding proteins are excluded from this analysis since the version of FINDSITE used in CASP8 was strictly designed to predict binding sites for organic molecules[49]). In 11 cases, the binding pocket center was predicted within a distance of 4 Å from the center of mass of a ligand in the crystal complex. Moreover, for all targets, the best predicted binding pocket is on rank 1. Using binding residues identified by LPC,[92] the median Matthew's correlation coefficient for residues predicted by FINDSITE to bind a ligand is 0.73. In CASP8, we also carried out a ligand-based virtual screening against the KEGG compound library[88] (12,158 diverse molecules) in order to predict binding ligands. The results were included in the "Remark" section of the submission files (available from CASP8 website). Here, we include the native ligand (co-crystallized with the target protein) and re-rank the screening library using ligand fingerprint profiles constructed by FINDSITE in CASP8. The native ligand is ranked within the top 1% (10%) of the screening library for 8 (11) targets (Table I). For targets T0431, T0485, and T0508, the native ligand is found at rank 1.

DISCUSSION

Over the past years, a number of protein function prediction techniques have been developed to facilitate the functional annotation of the sequenced genomes. Powerful structure/template-based methods[41,45,49,59,60,93] are particularly well-suited for practical applications in the "twilight zone" of sequence similarity, which roughly covers ~2/3 of known protein sequences.[94] However, due to the complex and equivocal relations between protein fold and function, very conservative similarity thresholds or target-specific thresholds are requisite for the effective function transfer by structure similarity alone.[95] An alternate way to significantly reduce the relatively high false positive rate caused by using more permissive cutoffs is to introduce evolutionary restraints. The evolutionary similarities of the protein functional sites provided by the Evolutionary Trace method have been successfully exploited to improve the accuracy of transfer of functional annotations.[96] Sequence profile–profile algorithms that amplify the patterns defining protein families have been demonstrated to detect remote homologies with strong functional similarities.[78,97] Several other methods for function prediction integrate sequence-based searches to establish functional relationships.[98,99]

Table I
Performance of the SiteHunter Server in CASP8

Target protein			TASSER model accuracy			FINDSITE (threading-based approach)			
			Global		Local	Pocket prediction			Virtual screening
CASP-ID	PDB-ID	Native ligand	RMSD[a]	TM-score	RMSD[b]	Rank[c]	Distance[d]	MCC for binding residues[e]	Native ligand rank[f]
T0422	3d8b	ADP	8.34	0.83	1.73	1	0.91	0.82	4
T0426	3da2	4MD	0.92	0.98	1.90	1	1.15	0.63	7
T0429	3db3	ARG-M3L-SER	11.39	0.23	9.81	1	7.96	0.11	7014
T0430	3dlz	AMP	13.48	0.46	7.97	1	2.96	0.56	97
T0431	3dax	HEM	3.63	0.75	3.36	1	1.28	0.41	1
T0445	3dao	1PE	1.56	0.92	3.68	1	3.90	0.56	4167
T0450	3da1	FAD	2.65	0.92	1.73	1	0.51	0.74	124
T0477	3dkp	ADP	4.54	0.85	2.88	1	7.54	0.63	15
T0483	3dls	ADP	4.89	0.85	2.55	1	1.40	0.85	134
T0485	3dlc	SAM	5.15	0.79	2.40	1	0.91	0.74	1
T0490	3dme	FAD	2.87	0.88	2.04	1	0.70	0.73	124
T0494	2vx3	D15	3.96	0.88	1.77	1	0.77	0.84	35
T0508	3dou	SAM	1.62	0.92	2.47	1	1.27	0.84	1

SiteHunter employs a threading-based approach (FINDSITE) to predict binding site for small organic molecules in protein models generated by TASSER. The accuracy of TASSER models used in ligand-binding site detection is also reported.
[a]Cα RMSD [Å].
[b]All-atom RMSD[Å] calculated over the binding residues.
[c]Rank of the best predicted binding site.
[d]Distance between the center of mass of a ligand in the crystal complex and the center of the predicted binding sites.
[e]Matthew's correlation coefficient calculated for the predicted binding residues.
[f]Predicted rank of the native ligand in the KEGG compound library of 12,158 molecules.

In this work, we demonstrate that threading, which employs a strong sequence-profile component,[47] plays an important role as a selective evolutionary filter in the template-based function annotation of proteins. It considerably enriches the set of selected templates with those that have similar binding pockets, molecular functions and bind chemically similar ligands. Thus, it significantly improves the efficiency and confidence of the function assignment. Conceptually, similar are the AnnoLite and AnnoLyze programs that combine sequence and structure similarity for the comparative protein annotation.[93] In these algorithms, a significant template-target structural similarity is established when at least 75% of the template's Cα atoms can be aligned to the target structure within 4 Å RMSD. This criterion roughly corresponds to a TM-score of 0.5; thus, as shown here, it is too promiscuous for the efficient function transfer by structure similarity. Therefore, to avoid the high false positive rate, a decreasing cutoff for sequence identity is applied. Because closely homologous templates with the sequence identity to the target up to 90% were included in their benchmark set, it is rather difficult to assess the performance of the AnnoLite/AnnoLyze in the "twilight zone" of the sequence identity. Furthermore, a very similar approach was used to assign new GO annotations to PDB sequences,[100] with the results suggesting that a permissive structural similarity threshold of 5 Å RMSD for the aligned region typically requires a high sequence identity of the aligned residues, while in the low sequence identity regime, a more selective cutoff of 2 Å RMSD should be used.[66] Here, we show that in the "twilight zone" of the sequence identity, sequence profile driven threading provides more sensitive means for detecting evolutionarily related proteins; thus, it maximizes the set of templates and the number of suitable targets by relaxing the structure similarity criteria.

This important feature also allows for the accommodation of the structural imperfections of the theoretical protein models used as targets for template-based function annotation. In this work, we focused on the crystal structures of the target proteins; however, as we reported previously,[4,49] template-based methods that employ protein threading[47] and global structure comparisons[71] are generally more suitable than binding pocket detection by local geometry for practical applications using protein models. The deformation of binding regions in predicted target structures was estimated for weakly homologous models to be ~2 Å RMSD[4]; this may critically affect the accuracy of the approaches employing the local structure similarity. For example, local structure matching using automated functional templates (AFT) identifies binding sites in protein models that have a RMSD from the crystal structure <3 Å,[44] whereas FINDSITE (which is a threading-based template selection algorithm) tolerates global structural inaccuracies in protein models to the RMSD from the crystal structure up to 8–10 Å.[49]

The ultimate aim of computational function annotation is to discover lead compounds with preferred pharmacological properties or those that can be further subjected to chemical modifications to attain a desired

activity toward a given protein target. Many virtual screening techniques have been developed to achieve this goal.[101,102] One approach is ligand-based virtual screening that typically requires a collection of already known active compounds.[103] In contrast, for a target protein given just the protein's sequence, in addition to the binding pocket localization and molecular function, another approach provides a set of molecules extracted from the template-ligand complex structures.[49] These can be used as the template ligands when no other information concerning potential binders is available. It has been already reported for α-helical proteins that related proteins tend to bind similar ligands.[104] Here, we show that this observation is in fact more general and applies to all evolutionarily related proteins. Evolution tends to conserve not only the functionally important region in the protein structure but also conserves a subset of ligand binding features. Thus, protein threading, when used as the evolutionary component in the template-based function assignment, typically detects templates that bind compounds with distinct chemical relationships to the target-bound molecules. Furthermore, the chemical similarity of ligands is higher if they bind to common binding sites; this emphasizes the importance of a local component in structure based functional inference algorithms, which in practice corresponds to the spatial clustering of the template-bound ligands.

Paradoxically, from the point of view of protein structure prediction, an ideal threading algorithm whose performance is comparable to that of structure alignment, would not improve the inference of molecular function as it would reduce to the purely structure based approach examined here with its demonstrated poorer results. This dictates two independent directions for the further development of threading algorithms. For the purpose of protein structure prediction, structure-based threading with the capabilities to detect structurally related templates should be pursued to detect those unrecognized templates with a similar fold. On the other hand, effective function inference requires an evolutionary-oriented version of threading that employs a strong sequence profile component. Interestingly, the variant of threading used here as well as HHpred, a sequence profile–profile method, already performs very close to the estimated theoretical limit for template-based function inference. Rather, it is the absence of structurally and functionally related templates that is the major limiting factor. The growing number of protein crystal structures solved in the complexed state will expand the pool of suitable targets for sequence profile driven template-based annotation of proteins. Thus, the combined evolution/structure-based function assignment emerges as a powerful technique to assist in comprehensive and fully automated proteome annotation.

ACKNOWLEDGMENTS

This article is dedicated to the memory of Dr. Angel Ortiz, a very talented scientist and fine human being who passed away long before his time.

REFERENCES

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. Science 2001;291:1304–1351.
2. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D,

Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES. Initial sequencing and comparative analysis of the mouse genome. Nature 2002;420:520–562.

3. Punta M, Ofran Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comput Biol 2008;4:e1000160.
4. Skolnick J, Brylinski M. FINDSITE: a combined evolution/structure-based approach to protein function prediction. Brief Bioinform 2009;10:378–391.
5. Wolfson HJ, Shatsky M, Schneidman-Duhovny D, Dror O, Shulman-Peleg A, Ma B, Nussinov R. From structure to function: methods and applications. Curr Protein Pept Sci 2005;6:171–183.
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
7. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
8. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science 1985;227:1435–1441.
9. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 1998;26:320–322.
10. Rost B. Enzyme function less conserved than anticipated. J Mol Biol 2002;318:595–608.
11. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 2003;333:863–882.
12. Tian W, Arakaki AK, Skolnick J. EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res 2004;32:6226–6239.
13. Henikoff JG, Henikoff S. Blocks database and its applications. Methods Enzymol 1996;266:88–105.
14. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. The PROSITE database. Nucleic Acids Res 2006;34 (Database issue):D227–D230.
15. Neduva V, Russell RB. DILIMOT: discovery of linear motifs in proteins. Nucleic Acids Res 2006;34 (Web Server issue):W350–W355.
16. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Supertifurga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 2003;31:3625–3630.
17. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
18. Arakaki AK, Huang Y, Skolnick J. EFICAz2: Enzyme function inference by a combined approach enhanced by machine learning. BMC Bioinformatics 2009;10:107.
19. Gherardini PF, Helmer-Citterich M. Structure-based function prediction: approaches and applications. Brief Funct Genomic Proteomic 2008;7:291–302.
20. Hawkins T, Kihara D. Function prediction of uncharacterized proteins. J Bioinform Comput Biol 2007;5:1–30.
21. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. Curr Opin Struct Biol 2005;15:275–284.
22. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys 2003;36:307–340.
23. Brady GP, Jr., Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 2000;14:383–401.
24. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 1997;15:359–363, 389.
25. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 1995;13:323–330, 307–328.
26. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics 2009;10:168.
27. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 1998;7:1884–1897.
28. Petrek M, Otyepka M, Banas P, Kosinova P, Koca J, Damborsky J. CAVER: a new tool to explore routes from protein clefts, pockets and cavities. BMC Bioinformatics 2006;7:316.
29. Xie L, Bourne PE. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. BMC Bioinformatics 2007;8 (Suppl 4):S9.
30. Hetenyi C, van der Spoel D. Blind docking of drug-sized compounds to proteins with up to a thousand residues. FEBS Lett 2006;580:1447–1450.
31. Ruppert J, Welch W, Jain AN. Automatic identification and representation of protein binding sites for molecular docking. Protein Sci 1997;6:524–533.
32. Silberstein M, Dennis S, Brown L, Kortvelyesi T, Clodfelter K, Vajda S. Identification of substrate binding sites in enzymes by computational solvent mapping. J Mol Biol 2003;332:1095–1113.
33. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307:447–463.
34. Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 2006;6:19.
35. Jones S, Shanahan HP, Berman HM, Thornton JM. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 2003;31:7189–7198.
36. Petrey D, Honig B. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. Methods Enzymol 2003;374:492–509.
37. Brylinski M, Prymula K, Jurkowski W, Kochanczyk M, Stawowczyk E, Konieczny L, Roterman I. Prediction of functional sites based on the fuzzy oil drop model. PLo S Comput Biol 2007;3:e94.
38. Ondrechen MJ, Clifton JG, Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure. Proc Natl Acad Sci USA 2001;98:12473–12478.
39. Dessailly BH, Lensink MF, Wodak SJ. Relating destabilizing regions to known functional sites in proteins. BMC Bioinformatics 2007;8:141.

40. Elcock AH. Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 2001;312:885–896.
41. Ausiello G, Gherardini PF, Marcatili P, Tramontano A, Via A, Helmer-Citterich M. FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures. BMC Bioinformatics 2008;9 (Suppl 2):S2.
42. Kleywegt GJ. Recognition of spatial motifs in protein structures. J Mol Biol 1999;285:1887–1897.
43. Stark A, Sunyaev S, Russell RB. A model for statistical significance of local similarities in structure. J Mol Biol 2003;326:1307–1316.
44. Arakaki AK, Zhang Y, Skolnick J. Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. Bioinformatics 2004;20:1087–1096.
45. Laskowski RA, Watson JD, Thornton JM. Protein function prediction using local 3D templates. J Mol Biol 2005;351:614–626.
46. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997;6:2308–2323.
47. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins 2004;56:502–518.
48. Zhang Y, Arakaki AK, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins 2005;61 (Suppl 7):91–98.
49. Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci USA 2008;105:129–134.
50. Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. J Med Chem 2004;47: 45–55.
51. McGovern SL, Shoichet BK. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J Med Chem 2003;46:2895–2907.
52. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T. Automated server predictions in CASP7. Proteins 2007;69 (Suppl 8):68–82.
53. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69 (Suppl 8):38–56.
54. Read RJ, Chavali G. Assessment of CASP7 predictions in the high accuracy template-based modeling category. Proteins 2007;69 (Suppl 8):27–37.
55. Zhou H, Pandit SB, Lee SY, Borreguero J, Chen H, Wroblewska L, Skolnick J. Analysis of TASSER-based CASP7 protein structure prediction results. Proteins 2007;69 (Suppl 8):90–97.
56. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
57. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
58. Irving JA, Whisstock JC, Lesk AM. Protein structural alignments and functional genomics. Proteins 2001;42:378–382.
59. Kristensen DM, Ward RM, Lisewski AM, Erdin S, Chen BY, Fofanov VY, Kimmel M, Kavraki LE, Lichtarge O. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. BMC Bioinformatics 2008;9:17.
60. Lisewski AM, Lichtarge O. Rapid detection of similarity in protein structure and function through contact metric distances. Nucleic Acids Res 2006;34:e152.
61. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
62. Russell RB, Sasieni PD, Sternberg MJ. Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol 1998;282:903–918.
63. Gerlt JA, Babbitt PC. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem 2001;70:209–246.
64. Hegyi H, Gerstein M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol 1999;288:147–164.
65. Jambon M, Andrieu O, Combet C, Deleage G, Delfaud F, Geourjon C. The SuMo server: 3D search for protein functional sites. Bioinformatics 2005;21:3929–3930.
66. Ponomarenko JV, Bourne PE, Shindyalov IN. Assigning new GO annotations to protein data bank sequences by combining structure and sequence homology. Proteins 2005;58:855–865.
67. Jones DT, Hadley C. Threading methods for protein structure prediction. In: Higgins D, Taylor WR, editors. Bioinformatics: sequence, structure and databanks. Heidelberg: Springer-Verlag; 2000. pp 1–13.
68. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
69. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–29.
70. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 2006;34 (Database issue):D187–D191.
71. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
72. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA 1998;95:5913–5920.
73. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
74. Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr A 1978;34:827–828.
75. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci USA 2006;103:2605–2610.
76. Pandit SB, Skolnick J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. BMC Bioinformatics 2008;9:531.
77. Teichert F, Bastolla U, Porto M. SABERTOOTH: protein structural alignment based on a vectorial structure representation. BMC Bioinformatics 2007;8:425.
78. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–960.
79. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
80. Kohlbacher O, Lenhof HP. BALL–rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library. Bioinformatics 2000;16:815–824.
81. Barber CB, Dobkin DP, Huhdanpaa HT. The Quickhull algorithm for convex hulls. ACM Trans on Mathematical Software 1996;22:469–483.
82. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA. 2006. pp 233–240.
83. Daylight Theory Manual. 4.9. Aliso Viejo, CA: Daylight Chemical Information Systems, Inc.; 2007.
84. Tanimoto TT. An elementary mathematical theory of classification and prediction. New York: IBM; 1958.

85. Xue L, Godden JW, Stahura FL, Bajorath J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. J Chem Inf Comput Sci 2003;43:1151–1157.
86. Xue L, Godden JW, Stahura FL, Bajorath J. Similarity search profiles as a diagnostic tool for the analysis of virtual screening calculations. J Chem Inf Comput Sci 2004;44:1275–1281.
87. Xue L, Stahura FL, Bajorath J. Similarity search profiling reveals effects of fingerprint scaling in virtual screening. J Chem Inf Comput Sci 2004;44:2032–2039.
88. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.
89. Skolnick J, Fetrow JS. From genes to protein structure and function: novel applications of computational approaches in the genomic era. Trends Biotechnol 2000;18:34–39.
90. Chua HN, Sung WK, Wong L. Using indirect protein interactions for the prediction of Gene Ontology functions. BMC Bioinformatics 2007;8 (Suppl 4):S8.
91. Wass MN, Sternberg MJ. ConFunc–functional annotation in the twilight zone. Bioinformatics 2008;24:798–806.
92. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. Automated analysis of interatomic contacts in proteins. Bioinformatics 1999;15:327–332.
93. Marti-Renom MA, Rossi A, Al-Shahrour F, Davis FP, Pieper U, Dopazo J, Sali A. The AnnoLite and AnnoLyze programs for comparative annotation of protein structures. BMC Bioinformatics 2007;8 (Suppl 4):S4.
94. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
95. von Grotthuss M, Plewczynski D, Vriend G, Rychlewski L. 3D-Fun: predicting enzyme function from structure. Nucleic Acids Res 2008;36(Web Server issue):W303–W307.
96. Ward RM, Erdin S, Tran TA, Kristensen DM, Lisewski AM, Lichtarge O. De-orphaning the structural proteome through reciprocal comparison of evolutionarily important structural features. PLoS ONE 2008;3:e2136.
97. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
98. Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 2005;33(Web Server issue):W89–W93.
99. Lopez G, Valencia A, Tress ML. Firestar–prediction of functionally important residues using structural templates and alignment reliability. Nucleic Acids Res 2007;35(Web Server issue):W573–W577.
100. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
101. Muegge I, Oloff S. Advances in virtual screening. Drug Discov Today 2006;3:405–411.
102. Walters WP, Stahl MT, Murcko MA. Virtual screening—an overview. Drug Discov Today 1998;3:160–178.
103. Stahura FL, Bajorath J. New methodologies for ligand-based virtual screening. Curr Pharm Des 2005;11:1189–1202.
104. Mitchell JB. The relationship between the sequence identities of alpha helical proteins in the PDB and the molecular similarities of their ligands. J Chem Inf Comput Sci 2001;41:1617–1622.