6-1-2013

EFindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands

Michal Brylinski
Louisiana State University

Wei P. Feinstein
Louisiana State University

Recommended Citation

Brylinski, M., & Feinstein, W. (2013). EFindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *Journal of Computer-Aided Molecular Design, 27* (6), 551-567. https://doi.org/10.1007/s10822-013-9663-5

*e*FindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands

Michal Brylinski · Wei P. Feinstein

Received: 6 April 2013 / Accepted: 1 July 2013 / Published online: 10 July 2013

Abstract Molecular structures and functions of the majority of proteins across different species are yet to be identified. Much needed functional annotation of these gene products often benefits from the knowledge of protein–ligand interactions. Towards this goal, we developed *e*FindSite, an improved version of FINDSITE, designed to more efficiently identify ligand binding sites and residues using only weakly homologous templates. It employs a collection of effective algorithms, including highly sensitive meta-threading approaches, improved clustering techniques, advanced machine learning methods and reliable confidence estimation systems. Depending on the quality of target protein structures, *e*FindSite outperforms geometric pocket detection algorithms by 15–40 % in binding site detection and by 5–35 % in binding residue prediction. Moreover, compared to FINDSITE, it identifies 14 % more binding residues in the most difficult cases. When multiple putative binding pockets are identified, the ranking accuracy is 75–78 %, which can be further improved by 3–4 % by including auxiliary information on binding ligands extracted from biomedical literature. As a first across-genome application, we describe structure modeling and binding site prediction for the entire proteome of *Escherichia coli*. Carefully calibrated confidence estimates strongly indicate that highly reliable ligand binding predictions are made for the majority of gene products, thus *e*FindSite holds a significant promise for large-scale genome annotation and drug development projects. *e*FindSite is freely available to the academic community at http://www.brylinski.org/efindsite.

Keywords Ligand binding site prediction · Binding residue prediction · Protein threading · Ligand virtual screening · Machine learning · Support vector machines

M. Brylinski (✉) · W. P. Feinstein
Department of Biological Sciences, Louisiana State University,
Baton Rouge, LA 70803, USA
e-mail: michal@brylinski.org

M. Brylinski
Center for Computation and Technology, Louisiana State
University, Baton Rouge, LA 70803, USA

Introduction

Proteins carry diverse molecular functions mainly through their ability to bind other molecular species present in a cell. Interactions between proteins and other molecules are critical to numerous biological processes, e.g. signal transduction, protein transport and folding, DNA replication and repair, and cell division, just to mention a few examples. To comprehend the immense repertoire of molecular functions and to describe key domains of molecular biology, a number of controlled, structured vocabularies, known as ontologies, have been developed [1, 2]. One of the most widely used public resources is Gene Ontology (GO), which provides precisely defined annotation standards and hierarchical classifications for describing the roles of genes and gene products in any organism [3, 4]. As of June 2013, a simple keyword search at the Gene Ontology website using the word “binding” returns 1,791 GO terms, with 1,655 under the molecular function category. Furthermore, it reports 220,312 gene products annotated with the “binding” term, defined by GO as “the selective, non-covalent, often stoichiometric, interaction of a molecule with one or more specific sites on another molecule”. This demonstrates how diverse, prevalent and important binding interactions are for cellular processes.

Proteins bind to a broad spectrum of molecular species in a cell including small organic molecules, nucleic acids, inorganic clusters, metal ions, peptides as well as other proteins. Binding is facilitated by the tertiary structure of a protein, with the region responsible for interacting with other molecules known as the binding site, which often forms a depression on the protein surface. The identification of a binding site and the corresponding binding residues is typically a first step in the comprehensive functional annotation of a gene product. A wide range of experimental techniques have been used to characterize binding events. X-ray crystallography and NMR can provide the detailed atomic structures of molecular complexes; however, experimental structure determination typically requires considerable efforts and time, therefore the structures of most complexes will not be solved in the near future. Other experimental techniques, such as site-directed mutagenesis, provide indirect structural information; however, these methods are most effective when supported by high-resolution data from X-ray or spectroscopic studies [5]. On that account, the identification of binding sites in proteins is strongly supported by computational approaches. As a matter of fact, due to the advances in genome sequencing technologies [6, 7], these methods represent the only practical strategy to keep up with the rapid accumulation of sequence information [8, 9].

Over the past years, a number of algorithms for binding site prediction have been developed. The simplest sequence-based methods build on homology, i.e. they transfer binding sites and residues from already annotated proteins. For example, methods based on position specific scoring matrices were successfully applied to find DNA binding sites in proteins [10, 11]. These methods often integrate machine learning, which can increase the predictive power, as demonstrated for the prediction of protein–protein interactions and interfacial residues [12, 13]. Furthermore, they frequently employ sensitive sequence search techniques, e.g. based on hidden Markov models [14] to increase the sensitivity by extracting the functional information from remotely related proteins [15]. Nevertheless, homology-based transfer is complicated by ambiguous relationships between protein sequence and function [16], thus it typically requires rather high sequence similarity thresholds to reduce the considerable risk of misannotation [17].

To overcome these limitations, alternative methods exploit purely structural information and attempt to capture a causal relation between protein structure and function. For instance, structure-based methods predict protein–protein interfaces from structural neighbors [18], identify ligand binding sites using hydrophobicity profiling [19, 20] and use short structural motifs as signatures for e.g. metal binding locations [21]. Many geometric methods take advantage of the fact that ligand binding events often occur inside cavities and depressions on the protein surface. Consequently, the detection of deep pockets has become a popular technique to predict ligand binding sites [22–24]. Depending on benchmarking datasets, prediction procedures and evaluation criteria, the accuracies of 60–69 % for LIGSITE^{CS} [25], 67–83 % for Fpocket [26] and 75–77 % for MSPocket [27] have been reported. Furthermore, consensus methods such as MetaPocket gain additional ~5 % over single methods by combining results obtained from individual predictors [28]. If the best of top three predictions is considered, the accuracy of geometry-based methods reaches 90–95 %. Many of these techniques, however, require experimentally solved structures, preferably in the “bound” conformational state, to achieve a high accuracy [29].

A new class of evolution/structure-based methods has emerged recently; these powerful techniques incorporate both sequence and structure components and cover many aspects of protein molecular function including interactions with small organic compounds [30–32], metal ions [33], nucleic acids [34] and other proteins [35, 36]. For drug discovery and development, of particular interest are interactions between proteins and small organic compounds, which are typical candidates for drugs. One of the earliest approaches to evolution/structure-based ligand binding prediction, FINDSITE [31], employs protein threading to detect weakly homologous templates, which are subsequently superposed onto the target structure using TM-align [37]. Upon the global superposition, putative binding sites are identified by the average linkage clustering of the geometrical centers of template-bound ligands. 3DLigandSite [30] is a similar method, which first identifies significant structural matches to the target protein using MAMMOTH [38]. Next, template ligands are extracted and a single linkage clustering is performed to detect ligand binding sites in the target structure. Residue conservation mapped onto the target structure serves as a sequence component to improve the accuracy of 3DLigandSite. Another algorithm, FunFOLD [32], is similar in concept to the abovementioned methods; however, it uses a novel automated method for ligand clustering and the identification of binding residues. FunFOLD assigns ligands to clusters using an agglomerative hierarchical clustering algorithm that accounts for a continuous mass of contacting ligands; this is followed by binding residue prediction by an optimized residue voting system.

At the conceptual level, all these methods capitalize on a general tendency of certain protein families to bind small molecules at similar locations [39]. Using structure information helps overcome the limitations of purely sequence-based methods effectively exploiting very remote evolutionary relationships in the “twilight zone” of sequence identity. The sequence component relaxes structure similarity criteria without increasing the false positive rate [40] thus allows for using modeled target structures instead of those solved experimentally. As a consequence, evolution/structure-based approaches provide a viable strategy for proteome-wide functional annotation. Across-proteome ligand binding site prediction may not only discover new target sites for pharmacotherapy [41], but also can help identify off-targets for existing drugs to support rational drug repositioning [42, 43].

Here, we describe the development and benchmarking of *e*FindSite, a new method for ligand binding site and residue prediction, which includes a series of important improvements over its predecessor, FINDSITE [31]. It employs a highly sensitive meta-threading procedure optimized specifically towards the identification of functionally related ligand-bound template structures. Moreover, it uses an improved clustering algorithm [44] to exploit both template-target as well as pairwise template structure similarities and includes a fine-tuned template weighting scheme. *e*FindSite extensively uses various machine learning techniques to efficiently integrate structural and evolutionary information and provide a reliable system for confidence estimation. As an additional feature, we include the possibility to support binding site prediction by using those ligands known to bind to target proteins. In large-scale benchmarks against crystal structures as well as different quality protein models we demonstrate that *e*FindSite outperforms FINDSITE and other methods for ligand binding site and residue prediction.

As a first genome-scale application of *e*FindSite, we describe the results obtained for the entire proteome of *Escherichia coli* comprising 4,552 gene products, whose crystal structures are unknown. Using protein models constructed by *e*Thread, a meta-threading protein structure modeling pipeline [45, 46], we predict ligand binding pockets and residues by *e*FindSite for the majority of *E. coli* proteins. The results are encouraging and hold a significant promise for the application of *e*FindSite in large-scale genome annotation and drug development projects.

Materials and methods

Ligand-bound template library and benchmarking dataset

The set of protein–ligand complexes used in this study as a template library was obtained from Protein Small Molecule Database [47]. The redundancy was removed at the 40 % pairwise sequence identity by PISCES [48]. However, two proteins that share more than 40 % sequence identity were included in the library if they bind ligands in different locations, i.e. the distance between ligand geometric centers upon the global structure alignment is >8 Å. Ligands are small organic compounds that have 6–100 heavy atoms non-covalently bound to the receptor proteins. The complete *e*FindSite template library consists of 15,285 proteins complexed with 20,215 ligands.

From the template library, we selected target proteins 50–600 residues in length, for which at least three weakly homologous (<40 % sequence identity) ligand-bound templates can be identified using meta-threading as described below. Moreover, we require templates to structurally align onto the target with a statistically significant TM-score of ≥0.4 [49]; structure alignments are generated by fr-TM-align [37]. This resulted in a non-redundant dataset of 5,784 protein–ligand complexes, which were used for the derivation of *e*FindSite parameters and machine learning models. For benchmarking purposes, we identified a subset of 3,659 complexes, in which receptor proteins have a single pocket, i.e. bind ligands in approximately the same location (within 8 Å radius) according to the Protein Data Bank [50] (PDB). We note that all benchmarks are carried out using twofold cross validation, randomly splitting the dataset to avoid memorization issues. Moreover, those templates that have >40 % sequence identity to the target are excluded from benchmarking calculations.

Target protein structures

Target crystal structures were obtained from PDB [50]. In addition, for each target structure, we generated two protein models: high- and moderate-quality. Structure models were constructed by *e*Thread, a recently developed method for template-based protein structure modeling [45, 46]. For each target, we generated up to 20 weakly homologous models: 10 using *e*Thread/Modeller and 10 using *e*Thread/TASSER-Lite. Two randomly selected models, one with a TM-score to native of >0.7 and one with a TM-score within 0.4–0.7 were included in the high- and moderate-quality set, respectively. When the modeling procedure did not provide models of appropriate quality, we artificially distorted the crystal structure to a desired resolution using a simple Monte Carlo procedure [51].

Selection of functional templates by meta-threading

To identify ligand-bound templates, we use *e*Thread that integrates ten state-of-the-art protein threading/fold recognition algorithms: CSI-BLAST [52], COMPASS [53], HHpred [14], HMMER [54], pfTools [55], pGenThreader [56], SAM-T2 K [57], SP3 [58], SPARKS2 [58] and Threader [59]. Originally, *e*Thread was designed to select structural templates using machine learning and a set of feature vectors composed of individual threading scores. Here, we extend this functionality to include the estimates of ligand binding probability by constructing two additional machine learning models for the selection of functional templates. The first model assesses whether a particular template binds its ligands in similar locations as the target. Note that a similar ligand binding location for a given target-template pair means that they bind ligands within a distance of 8 Å upon the global alignment of their structures. The second model estimates a probability that template ligands are chemically similar to the target bound molecule, where similar ligands are defined by using a Tanimoto coefficient [60] threshold of >0.5 for 1024-bit Daylight fingerprints calculated by OpenBabel [61]. Both classifiers use a naïve Bayes algorithm to combine individual threading scoring functions into a single probabilistic score. Here, the real-value attributes are modeled from a Gaussian distribution, i.e. the classifier first estimates a normal distribution for each threading component by computing the mean and standard deviation of the training data in that class, which is then used to estimate the posterior probabilities during classification [62]. Furthermore, since *e*Thread uses two template libraries: chain and domain, we constructed separate machine learning models for each library. Both *e*Thread structure libraries are mapped to the *e*FindSite ligand-bound template library using global sequence identity calculated by a Needleman-Wunsch algorithm [63]. The accuracy of template selection is assessed using twofold cross validation excluding those templates, whose sequence identity to target is >40 %; note that this sequence identity cutoff is also applied in all subsequent modeling steps.

*e*FindSite engine

*e*FindSite builds upon the original FINDSITE algorithm, which was one of the first of its kind in evolution/structure-based ligand binding site prediction. *e*FindSite significantly extends its functionality and includes a series of major improvements over the original implementation to provide higher coverage, significantly lower false positive rate and better tolerance to structural errors in protein models. It is specifically tuned to exploit structural as well as functional information on ligand binding extracted from threading templates using machine learning. This optimized procedure allows us to use more distantly, yet functionally related templates at a reduced risk of predicting false positives.

A typical evolution/structure-based algorithm for binding site prediction superimposes a set of evolutionarily related templates complexed with ligands onto the target structure. Then, the centers of mass of bound ligands are clustered and the resulting clusters are used to identify putative binding sites in the target protein structure [30, 31, 64]. Here, we developed a slightly different approach. We use structure alignments constructed by fr-TM-align to calculate all-against-all binding site distances between templates. This matrix is subsequently used to identify clusters of template-bound molecules by Affinity Propagation (AP) [44], a recently developed clustering algorithm. As input, AP takes a matrix of similarities and exchanges real-valued messages between data points in order to identify a high-quality set of exemplars and the corresponding cluster members. It was demonstrated to uniformly detect clusters with much lower error rates compared to other methods. Finally, the identified template clusters are structurally aligned onto the target to mark the locations of putative binding sites. By design, this procedure is less sensitive to the quality of the target structure than a traditional clustering in Cartesian space upon the superposition of templates onto the target. To speed up calculations, we pre-computed pairwise similarities within the template library to compose a lookup table. Furthermore, to each template, we assign a weight that corresponds to the probability of having a binding site in similar location as the target. These probabilities are provided by machine learning models implemented in *e*Thread. In doing so, templates predicted to have similar binding sites give a stronger contribution to the pocket location prediction than those with lower probability.

Binding residue prediction

For each putative binding pocket, binding residues are predicted using machine learning and a set of the following features: sequence and secondary structure profiles, a distance from the predicted pocket center, standard deviation for distances between the pocket center and the centers of mass of template-bound ligands, the fraction of templates that have a residue in structurally aligned position in contact with a ligand and the average molecular weight of template bound ligands. Sequence-based features as well as geometric characteristics ensure a proper structural and chemical environment at the predicted binding sites for binding ligand molecules. We also impose a requirement of a minimum number of three confidently predicted binding residues to designate a site as ligand binding; this further reduces the false positive rate particularly for function annotations using low-homology templates. At last, a 2-class (binding/non-binding) Support Vector Classification (SVC) model is constructed to assign a given residue in the target structure a ligand binding probability.

Pocket ranking and confidence estimates

Similar methods commonly use majority voting to rank predicted binding sites. While it works well for relatively easy targets, it may encounter some problems in the case of medium difficulty targets, for which a couple of largest clusters often have comparable multiplicities. To address this issue, we developed a machine learning protocol to rank the predicted sites and estimate the corresponding ranking confidence. It employs a vector of the following features: the fraction of templates that share a particular site, the number of templates (cluster multiplicity), the average TM-score of the templates to the target, the number and the average confidence of predicted binding residues, and a protein–ligand binding index [65] calculated over the predicted binding residues. Similar to binding residue prediction, a 2-class SVC model is constructed to estimate whether a given site center is predicted within 8 Å from the geometrical center of a natively bound ligand. This confidence is then used to rank all putative binding sites predicted for a given target.

Auxiliary ligands

In many cases, the chemical identity of binding ligands is known, for instance, can be found in biomedical literature. Therefore as an option, we incorporate this data to enrich binding site information in *e*FindSite. If such auxiliary ligand is provided, each predicted binding site is assigned a probability to bind this compound by an SVC model, which assesses a physicochemical match to the template-bound molecules. Here, we calculate a classical (*TC*), an average (*aveTC*) as well as a continuous Tanimoto coefficient (*conTC*) between the auxiliary compound and template-bound ligands identified for a given binding site; see Appendix. The TC scores are calculated using two popular chemical fingerprints, 1024-bit Daylight (Daylight Chemical Information Systems Inc.: http://www.daylight.com) and 166-bit MACCS (Symyx Software: MACCS structural keys. San Ramon, CA), which give 6 features. The remaining 5 features comprise the following physicochemical properties: molecular weight (*MW*), octanol/water partition coefficient (*logP*), polar surface area (*PSA*), and the number of hydrogen bond donors (*HBD*) and acceptors (*HBA*). The calculations of 1024-bit Daylight fingerprints, *MW*, *logP* and *PSA* are conducted by OpenBabel [61] and 166-bit MACCS fingerprints, *HBD* and *HBA* by MayaChemTools (http://www.mayachemtools.org/). For each property, we first calculate the average value and the corresponding standard deviation for the set of template-bound ligands that are used to identify a given binding site in the target structure. Then we use a single Gaussian restraint *R* (Eq. 1) to evaluate how well the auxiliary compound *i* matches the putative binding site with respect to a particular molecular property, e.g. *MW*:

$$R_i^{MW} = 0.5 \times \left(\frac{MW_i - \langle MW \rangle}{\sigma}\right)^2 - \ln\frac{1}{\sigma\sqrt{2\pi}} \qquad (1)$$

where R_i^{MW} is the molecular weight restraint for compound i, $\langle MW \rangle$ is the average molecular weight of template-bound molecules and σ is the standard deviation. The restraints for the remaining properties are calculated in a similar way.

Finally, an SVC classifier was developed to estimate the posterior probability of an auxiliary compound binding to each identified site. If such compound is provided, we also include this probability estimate as an additional feature for binding site ranking and confidence estimation.

Other methods for binding pocket prediction

Binding site prediction by *e*FindSite is compared to several other methods. First, we consider two nearest-neighbor approaches: sequence- and structure-based. In the sequence-based variant, binding site location is directly transferred from a template protein with the highest global sequence similarity to the target; sequence similarity is calculated by Needleman-Wunsch dynamic programming [63]. In the structure-based approach, the closest template is identified based on the lowest *E*-value reported by MAMMOTH [38], which indicates the highest global structure similarity. We note that MAMMOTH is a frequently used structure alignment algorithm in template-based ligand binding site prediction [30, 66, 67]. To maintain the consistency of both nearest-neighbor approaches with other benchmarks reported in this study, closely homologous templates with >40 % sequence identity to the target are excluded.

In addition to these nearest-neighbor methods, we compare the performance of *e*FindSite to FINDSITE [31], Fpocket [26], ghecom [68], LIGSITE^{CS} [25], MSPocket [27] and MetaPocket [28, 69]. FINDSITE is one of the first approaches that integrate evolutionary information with structure-based annotation of ligand binding sites in proteins. Here, simulations were carried out as described in the original publication, except for the template-target sequence identity threshold, which was set to 40 % instead of 35 %. Fpocket, ghecom, LIGSITE^{CS} and MSPocket are purely structure-based binding pocket predictors; for each algorithm we used the default set of parameters. Except for MetaPocket benchmarked against different datasets as described below, the performance of all pocket prediction algorithms is evaluated using the complete dataset of 3,659 protein–ligand complexes, where each target protein exists in three different conformations: experimental structure, high- and moderate-quality protein model.

MetaPocket represents a majority-voting meta-method that effectively combines the results of several individual algorithms to significantly improve the prediction accuracy.

The comparison of *e*FindSite to MetaPocket 1.0 and 2.0 is conducted using three datasets previously selected for MetaPocket benchmarking [69]: 48 unbound/bound structures (MPK-48), 210 bound structures (MPK-210) and a non-redundant dataset of 198 drug-target complexes (MPK-198). Note that in these benchmarks, only the crystal structures of target proteins are used. For *e*FindSite, we follow the standard procedure excluding closely related templates, whose sequence identity to the target is >40 %. Furthermore, meta-threading failed to identify structurally related ligand-bound templates, whose TM-score to the target is ≥0.4, for 6, 7 and 33 target proteins in the MPK-48, MPK-210 and MPK-198 dataset, respectively. To keep *e*FindSite results consistent with those previously obtained for MetaPocket [69], in these cases, we use all identified templates regardless of the structure alignment quality.

Structure modeling of *E. coli* proteins

For genome-scale protein structure modeling and ligand binding pocket prediction, we selected *E. coli* K12 strain [70], which is routinely used in bioengineering and molecular biology research. First, we used *e*Thread to construct structural models for 4,552 gene products 50–600 residues in length. Full-length models were assembled using either Modeller [71] or TASSER-Lite [72]. Our benchmarking calculations indicate that Modeller constructs higher quality models for easy cases, whereas TASSER-Lite more effectively handles difficult cases providing better coverage [45]. Therefore for each gene product, we built an initial model using *e*Thread/Modeller; when the estimated TM-score was <0.5, indicating difficult structure modeling, we constructed another model using *e*Thread/TASSER-Lite. In these cases, a model with the higher estimated TM-score is designated as the final structure. We note that estimated TM-score values correlate well with the real ones with Pearson correlation coefficient of 0.89 and 0.81 for *e*Thread/Modeller and *e*Thread/TASSER-Lite, respectively [45].

Results and discussion

Structural characteristics of benchmarking proteins

The primary application of *e*FindSite is high-throughput ligand binding site prediction using modeled protein structures. The quality of protein models can vary and strongly depends on the availability of structurally related templates detectable by threading. Therefore, in addition to target crystal structures, we benchmark *e*FindSite against two sets of protein models with high- and moderate-quality structures. Table 1 shows that models constructed either by *e*Thread/Modeller or *e*Thread/TASSER-Lite were included in the high- and moderate-quality dataset for 79.8–95.6 % of the target proteins, respectively. Owing to the fact that no models with a TM-score to native of >0.7 (0.4–0.7) have been constructed for 741 (159) targets, both datasets are enriched with a proportionate number of structures distorted to a desired resolution. Note that our structure modeling procedure employed only weakly homologous templates with a sequence identity to the target of at most 40 %. Overall, as shown in Table 1, the high-quality set consists of models, whose average TM-score to native is 0.81. The average backbone Cα-RMSD (root-mean-square deviation) is below 5 Å with ligand binding regions fairly well preserved to an average all-atom RMSD of 2.3 Å. The moderate-quality set comprises significantly less accurate structures. Here, the average TM-score and Cα-RMSD is 0.55 and 11.7 Å, respectively. Furthermore, the binding sites are severely distorted with an average all-atom RMSD of 5.7 Å. These models certainly represent a considerable challenge for pocket detection and binding residue prediction algorithms.

Template selection for binding site prediction

*e*FindSite employs meta-threading and two machine learning classifiers to select ligand-bound template proteins, which are subsequently used in binding site prediction. The cross-validated accuracy of template selection is shown in Fig. 1. The first classifier assigns to each threading template a probability of having a ligand binding site in a similar location as the target. Here, the performance in selecting good templates is quite high with the true and false positive rate of 76–44 %, respectively (Fig. 1a). Interestingly, despite the fact that closely homologous templates with >40 % sequence identity to the target are excluded from the benchmarks, the second classifiers also performs fairly well in selecting these templates that bind chemically similar molecules. This is shown in Fig. 1b; here, the true and false positive rate is 78 and 54 %, respectively. As we demonstrate below, this information can be advantageously exploited to detect binding sites with high accuracy.

Binding site clustering by affinity propagation

*e*FindSite uses Affinity Propagation [44] to identify clusters of similar binding sites across a set of identified templates. AP method requires a preference factor, which controls how many data points are selected as exemplars. In Fig. 2, we show how the preference factor affects the clustering outcome for our dataset (here we use target crystal structures). As expected, low preferences lead to a large number of small clusters, with the one closest to the natively bound ligand assigned a high rank, thus using too low preference

Table 1 Composition and structure quality of two datasets of protein models used in addition to crystal structures as targets for ligand binding site prediction

Dataset	Composition			TM-score	MaxSub	GDT	Cα-RMSD [Å]	Pocket RMSD[a] [Å]
	Modeller (%)	TASSER (%)	Distorted (%)					
High-quality	30.9	48.9	20.2	0.81 ± 0.07	0.64 ± 0.11	0.67 ± 0.10	4.82 ± 2.65	2.29 ± 1.91
Moderate-quality	23.4	72.2	4.3	0.55 ± 0.09	0.35 ± 0.12	0.41 ± 0.11	11.71 ± 4.48	5.73 ± 3.74

[a] All-atom RMSD

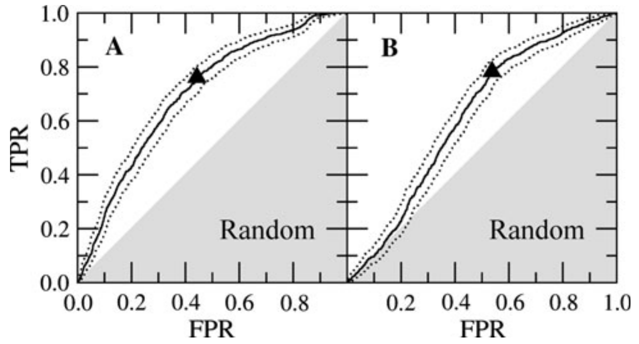

Fig. 1 ROC plots for ligand-bound template selection by *e*Thread. **a** Positives are defined as those templates that bind ligands in similar locations; **b** Bound ligands that are chemically similar to the target compound are considered positives. TPR—true positive rate, FPR—false positive rate, *black triangles* depict the maximum Matthew's correlation coefficient, *dotted lines* represent 95 % confidence bounds, and gray area corresponds to accuracy no better than random

factors would result in poor ranking abilities. High preferences produce a small number of larger clusters, which are, however, further away from natively bound ligands. We selected 8 Å as an optimal preference factor, which assigns reasonably low ranks and results in an average distance from the native ligand of ~2 Å.

Accuracy of binding residue prediction

Instead of a simple majority voting, *e*FindSite employs machine learning using SVC for binding residue prediction. The cross-validated accuracy of this model is presented in Fig. 3 for three sets of target structures with different quality. Here, we use only those targets, for which the best binding site is predicted within 8 Å from their native ligands. A posterior probability threshold of 0.25 maximizes Matthew's correlation coefficient (MCC) to 0.53, 0.50 and 0.48 for crystal structures, high- and moderate-quality protein models, respectively. Moreover, inset plots in Fig. 3 show that *e*FindSite correctly identifies 66, 67 and 66 % of ligand binding residues at the expense of 13, 15 and 15 % false positive rate, respectively. These sensitivity values correspond to a precision of 65, 58 and 56 %, respectively. Given the average distortion of ligand binding regions of almost 6 Å across the moderate-quality set, the overall accuracy of binding residue prediction is actually not only very high, but also quite insensitive to the quality of target receptor structures.

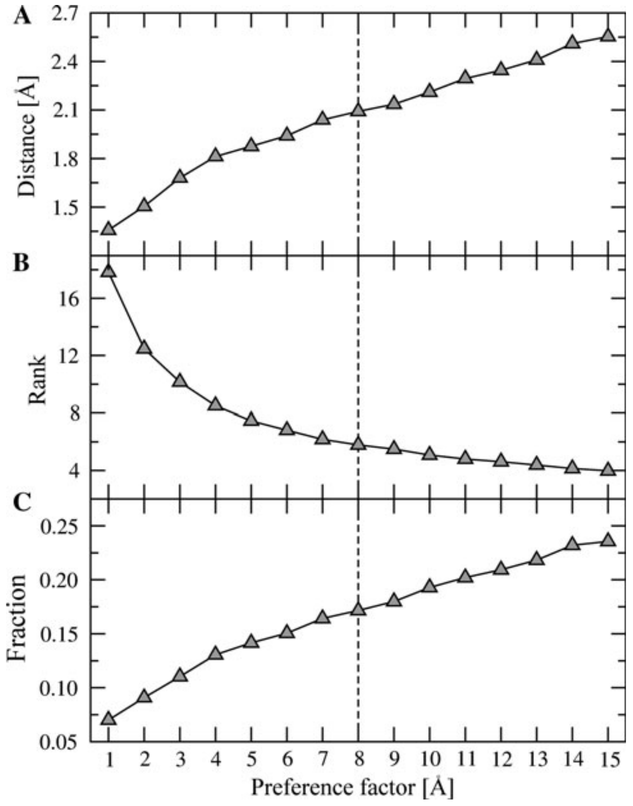

Fig. 2 Optimization of the preference factor for Affinity Propagation clustering. For a given preference factor, changing from 1 to 15 Å with 1 Å step, the geometric centers of template bound ligands are clustered and the partitioning results are assessed by the following metrics: **a** distance of the closest cluster from the ligand geometric center, **b** rank of the closest cluster and **c** fraction of templates that belong to the closest cluster. *Dashed line* depicts a preference factor of 8 Å selected to balance these three quantities

Binding site ranking

Particularly using weakly homologous templates, whose function may have diverged from that of the target, typically results in multiple putative binding sites. Therefore, a pivotal component of the prediction algorithm is a reliable system for pocket ranking. This is especially important for medium difficulty targets, for which a couple of largest clusters often have a comparable number of binding ligands. To deal with this problem, we developed a new method for pocket ranking that uses machine learning. Figure 4 demonstrates that high ranking capabilities of *e*FindSite are fairly independent on the quality of target structures. For crystal structures, high- and moderate-quality protein models, the best pocket rank is at rank 1 in 78, 76 and 75 % of the cases, respectively. This corresponds to ~3 % improvement over majority voting by cluster fraction. Furthermore, for as many as 92, 91 and 90 % of target proteins, the best pocket is found at most within the top two ranks.

Binding site prediction compared to nearest-neighbor approaches

Nearest-neighbor methods are the simplest template-based techniques that can also characterize the relationships between target proteins to the background knowledge present in the template library. For a given target protein, the prediction is made from a single template that is identified based on either global sequence or global structure similarity. Figure 5 shows the improvement of *e*FindSite over both nearest-neighbor methods for crystal structures as well as for different quality protein models. Since our benchmarks are specifically constructed to exclude closely homologous templates, the accuracy of sequence-based approach is quite low; the median distance between experimental and predicted pocket center is 15.1, 15.0 and 14.6 Å for crystal structures, high- and moderate-quality protein models, respectively. As expected, the structure-based nearest-neighbor method is generally more accurate; here, the median distance is 4.6, 5.0 and 5.9 Å, respectively. Note that in benchmarks against protein models, the modeled structure is used to identify the closest structural match in the template library, thus sequence neighbors are always the same across the three datasets, whereas structure neighbors may be different. The performance of *e*FindSite using the top-ranked predicted pockets is clearly better than both nearest-neighbor approaches with a median distance of 3.8, 3.8 and 4.3 Å for crystal structures, high- and moderate-quality models, respectively. Importantly, it is also less sensitive to distortions in the modeled structures. When moving from crystal structures to moderate-quality models, the accuracy of *e*FindSite drops off only by 0.5 Å compared to 1.3 Å for the structure-based approach. These results also concur with previous studies showing that a combined evolution/structure-based approach provides higher accuracy than function inference derived on the basis of global structure similarity alone even in the low-sequence identity regime [40]. It should be pointed out that simple nearest-neighbor techniques are computationally much less expensive; however, this analysis perspicuously demonstrates the superior performance of *e*FindSite and justifies its higher demands for computing resources.

Binding site prediction compared to other methods

In Fig. 6, the accuracy of *e*FindSite in ligand binding site prediction is compared to that of several other commonly

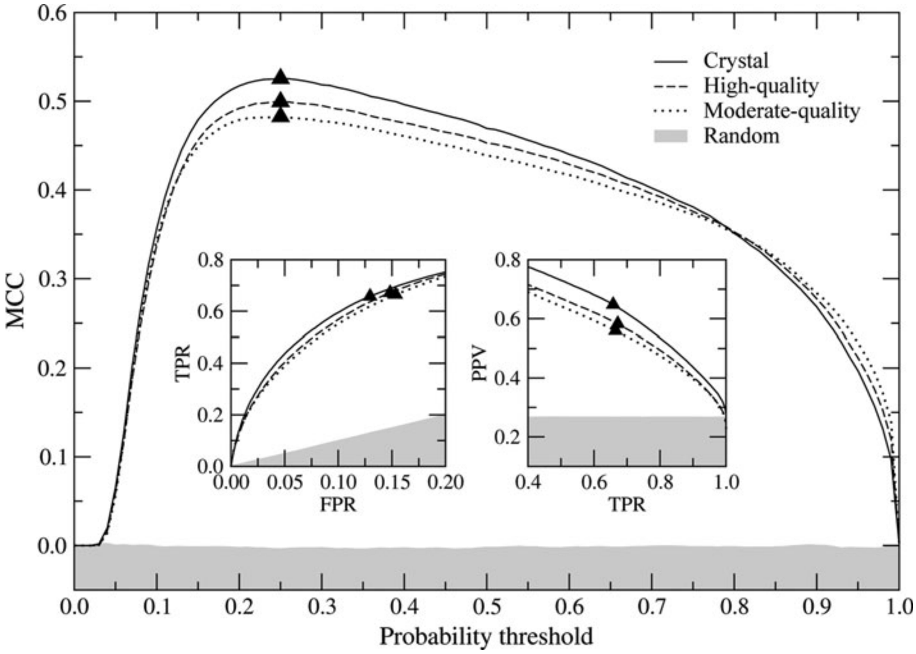

Fig. 3 Assessment of binding residue prediction using machine learning. MCC for predicted versus experimental binding residues is plotted as a function of probability estimates calculated by SVC for crystal, high- and moderate-quality target structures. *Insets*: (*left*) ROC plot and (*right*) sensitivity-precision plot; TPR—true positive rate, FPR—false positive rate, PPV—precision. *Black triangles* show the best performance in binding residue prediction, whereas *gray areas* delineate predictions no better than random

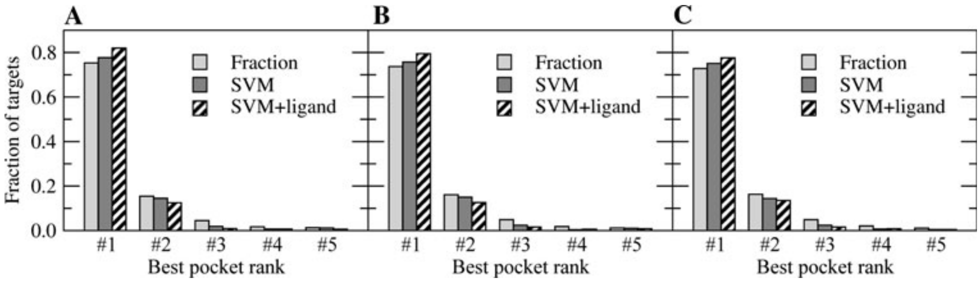

Fig. 4 Pocket ranking accuracy for different quality target structures: **a** crystal, **b** high- and **c** moderate-quality. Ranking accuracy is assessed by the fraction of targets, for which the best pocket is found at a particular rank shown on the *x*-axis. Three ranking protocols are evaluated: by cluster fraction (Fraction), machine learning (SVM) and machine learning that also considers chemical properties of native ligand (SVM + ligand)

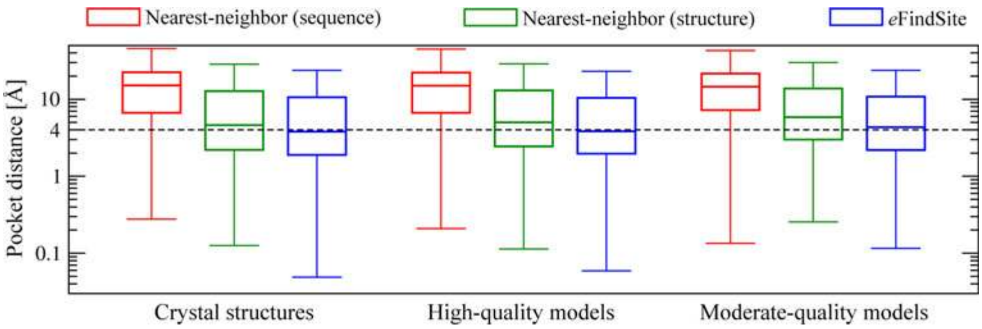

Fig. 5 Comparison of *e*FindSite to sequence- and structure-based nearest-neighbor approaches using different quality target structures. For each method, the distribution of distances between predicted pocket centers and the corresponding native ligand geometric centers across the benchmarking complexes is shown on the *y*-axis. *Boxes* end at the quartiles Q_1 and Q_3; a *horizontal line* in a *box* is the median. Whiskers point at the farthest points that are within 3/2 times the interquartile range. Dashed line depicts a distance of 4 Å between predicted and experimental pocket centers. For *e*FindSite, only top-ranked pockets are considered

used methods. Here, we consider only the top-ranked binding sites predicted for all three sets of target structures. The accuracy is assessed by Matthew's correlation coefficient calculated for predicted binding residues as well as a distance between native ligand geometric center and the predicted pocket center. Focusing on crystal structures (Fig. 6a), *e*FindSite outperforms geometrical pocket detection algorithms by 5–10 % at MCC of 0.5 for binding residues and by 15–20 % at a distance threshold of 5 Å. More importantly, it is much less sensitive to the structural distortions in protein models, see Fig. 6b, c. Here, using high- (moderate-) quality models, the accuracy measured by the fraction of proteins with MCC of ≥0.5 and the pocket center predicted within 5 Å decreases only by 4.2 % (9.9 %) and 0.9 % (4.7 %), respectively. The falloff in performance is clearly more dramatic for all purely geometrical algorithms, for example, at MCC of 0.5 (a distance threshold of 5 Å), the performance of MSPocket decreases from 47.4 % (36.4 %) to 18.3 % (20.5 %) and 8.0 % (12.5 %) for high- and moderate-quality protein models, respectively.

In binding residue prediction, *e*FindSite is also more accurate than its predecessor, FINDSITE. For target crystal structures, both algorithms predict binding residues with MCC of ≥0.5 for 55 % of the targets (Fig. 6a). However, using high- (moderate-) quality protein models, this fraction is 51 % (46 %) and 47 % (32 %) for *e*FindSite and FINDSITE, respectively. This improved performance of *e*FindSite over FINDSITE is a result of several factors: a highly optimized template selection and weighting, new clustering scheme and the extensive use of various machine learning techniques instead of majority voting. We also note that the accuracies of both programs are considerably higher than that of all geometrical methods.

To wind up comparative benchmarks, we assess the performance of *e*FindSite with respect to MetaPocket, a consensus approach currently combining eight individual pocket detection algorithms to improve prediction accuracy. Here, we use three datasets previously compiled to benchmark MetaPocket: MPK-48, MPK-210 and MPK-198; the results for MetaPocket versions 1.0 and 2.0 are taken from the original publication [69]. Table 2 presents hit rates defined as a percentage of target proteins for which the pocket center is predicted within a distance of 4 Å from the closest ligand heavy atom. The performance of *e*FindSite on MPK-48/bound, MPK-48/unbound, MPK-210

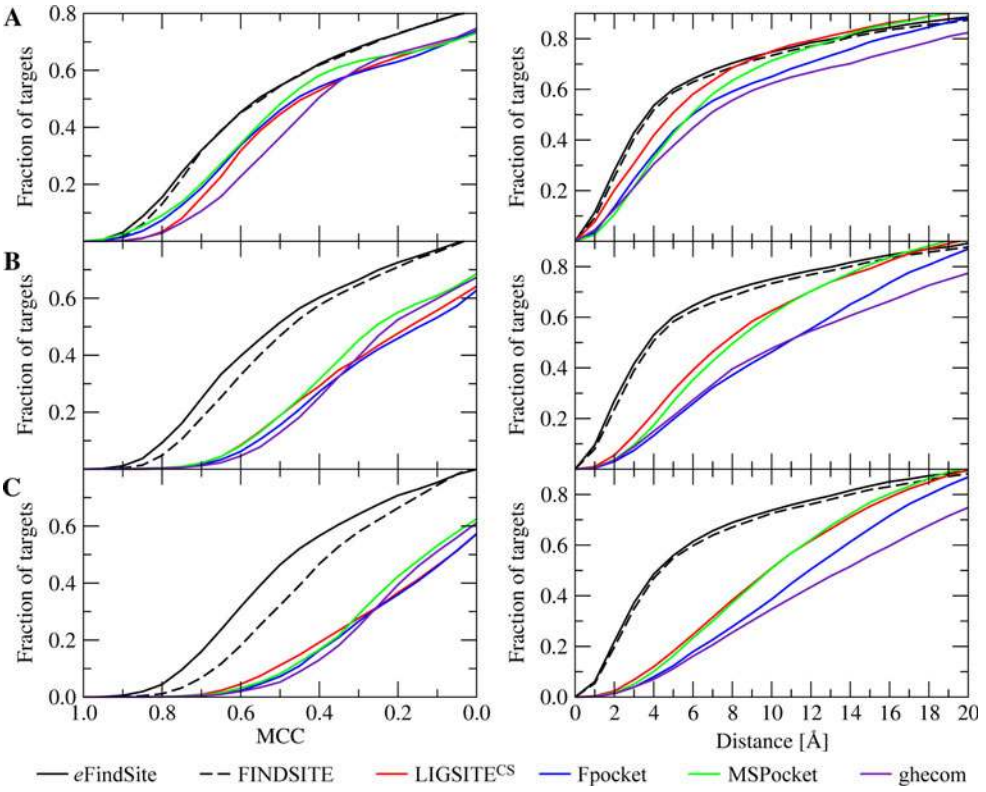

Fig. 6 Comparative assessment of ligand binding pocket prediction by *e*FindSite, FINDSITE, LIGSITE^{CS}, Fpocket, MSPocket and ghecom using different quality target structures: **a** crystal, **b** high- and **c** moderate-quality. The accuracy is assessed by (*left pane*) MCC calculated for predicted binding residues and (*right pane*) distance between native ligand geometric center and the predicted pocket center. Only top-ranked binding sites are considered

and MKP-198 is 2, 10, 7 and 5 % higher than MetaPocket 1.0, respectively. Compared to MetaPocket 2.0, *e*FindSite achieves higher hit rates for MPK-48/unbound and MPK-210 proteins. Note that *e*FindSite predictions using weakly homologous, yet structurally related templates were obtained for 42, 203 and 165 targets, which is 88, 97 and 83 % of proteins in the MPK-48, MPK-210 and MPK-198 dataset, respectively. Considering only these subsets of targets, the hit rate of *e*FindSite improves by 8 % (7 %) for MPK-48 (MPK-198). Furthermore, most unbound proteins are globally very similar to the corresponding bound forms with only local structural rearrangements of binding residues [73]. Consequently, the performance of *e*FindSite that employs global structure alignments should not depend on the functional form of MPK-48 proteins. Table 2 shows that this is indeed the case; the hit rate for both MPK-48 datasets is 85 % (93 % for the subset of 42 targets). Note that the accuracy of MetaPocket 2.0 (MetaPocket 1.0) decreases by 5 % (8 %).

Table 2 Comparison of *e*FindSite with two versions of MetaPocket

Dataset	MetaPocket 1.0 (%)	MetaPocket 2.0 (%)	*e*FindSite[a]
MPK-48 (bound)	83	85	85 % (93 %)
MPK-48 (unbound)	75	80	85 % (93 %)
MPK-210 (bound)	76	81	83 % (82 %)
MPK-198 (bound)	55	61	60 % (67 %)

The performance is assessed by hit rates for three different datasets previously used in MetaPocket benchmarking

[a] Numbers in parentheses correspond to hit rates obtained for a subset of 42, 203 and 165 proteins from MPK-48, MPK-210 and MPK-198, respectively

Improved performance by using auxiliary ligands

The results described so far were obtained using target proteins alone. Many public databases, such as BindingDB [74], PubChem [75] or DrugBank [76] provide information on binding ligands extracted from biomedical literature. For many of these compounds, the molecular target is known; however, the mode of interaction as well as specific binding sites and binding residues are undetermined. Moreover, the experimental structure of the target may not be available, which would necessitate using a protein model. A new feature of *e*FindSite is its capability of including additional information on ligands experimentally known to bind to target proteins in the prediction procedure. Here, we developed a machine learning-based model to assess how well an auxiliary ligand matches the physicochemical properties of predicted binding sites. Figure 7 shows the accuracy in recognizing correct binding sites using native ligands. At a probability threshold of 0.5, Matthew's correlation coefficient is 0.54, 0.52 and 0.52 for crystal structures, high- and moderate-quality models, respectively. This corresponds to the true/false positive rate

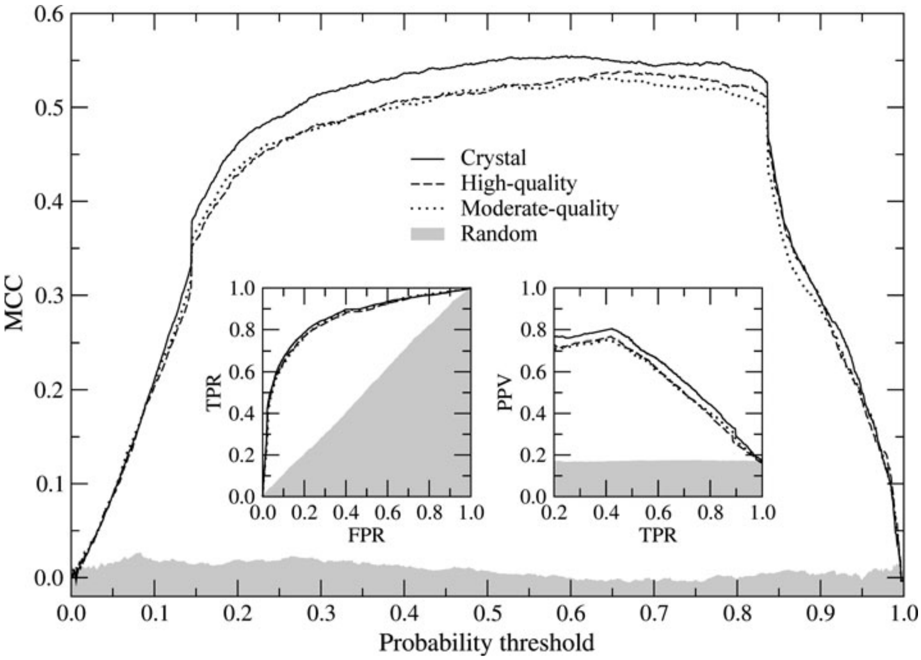

Fig. 7 Accuracy of machine learning in recognizing binding pockets using the physicochemical properties of native ligands. Matthew's correlation coefficient is plotted as a function of probability estimates calculated by SVC for different quality target structures (crystal, high- and moderate-quality). *Insets*: (*left*) ROC plot and (*right*) sensitivity-precision plot; TPR—true positive rate, FPR—false positive rate, PPV—precision. *Gray areas* correspond to predictions no better than random

of 0.63/0.08, 0.63/0.09 and 0.62/0.09, respectively (Fig. 7, inset). This high accuracy demonstrates that ligand fitness can be considered as a reliable confidence score. Moreover, Fig. 4 shows that when this information is subsequently included in the ranking procedure, it further improves the overall ranking. Now, in 82 % (95 %), 80 % (92 %) and 78 % (91 %) of the cases the best pocket is ranked 1 (at most 2) using crystal structures, high- and moderate-quality models, respectively.

As shown in Fig. 8, improved ranking leads to higher MCC values calculated for predicted binding residues. For a number of proteins highlighted by green areas, MCC for the top-ranked pocket rises above a significant threshold of 0.5. Red areas show that for notably fewer targets, additional information on ligands makes MCC scores for binding residues worse. This is caused by a very weak signal from promiscuous sites that bind to chemically diverse compounds across sets of evolutionarily weakly homologous proteins, which in turn, deteriorates ranking accuracy. Importantly, the improvement is seen not only for crystal structures, but also for both sets of protein models of high- and moderate quality.

Confidence index system for binding site prediction

Since accurate binding site predictions cannot be made for all proteins, it is critical to have a reliable confidence index system. *e*FindSite offers this functionality through posterior probabilities estimated by the SVC model for binding site ranking. In Fig. 9, we show that the confidence index correlates very well with the accuracy of binding site prediction assessed by MCC calculated for binding residues within the top-ranked pocket. Typically, accurate predictions require quite high confidence estimates of >0.8, whereas for proteins assigned a confidence of <0.2, the median MCC is close to random. Based on these results, we can categorize target proteins as "easy" (>0.8), "medium" (0.2–0.8) and "hard" (<0.2). We note that "easy" does not mean trivial; the classification simply helps estimate a level of difficulty in making an accurate prediction. Of course the overall performance of *e*FindSite is high because most of the targets in the dataset fall into the "easy" category: 74 % for target crystal structures and 69–73 % for protein models, see Table 3.

Genome-scale pocket prediction

To demonstrate the practical application of *e*FindSite in across-genome function annotation, we use it to identify putative ligand binding sites in the entire proteome of *E. coli*. First, using *e*Thread, we constructed structural models for all gene products in *E. coli* proteome. Figure 10 shows the distribution of the estimated quality of individual models generated by *e*Thread/Modeller, *e*Thread/TASSER-Lite. Since *e*Thread/TASSER-Lite was applied only to the most difficult cases, the corresponding distribution is shifted towards lower estimated TM-score values. Collecting the most confident models from both sets results in a final dataset of 4,552 structures that comprise 3,185 (70 %) and 1,367 (30 %) models constructed by *e*Thread/

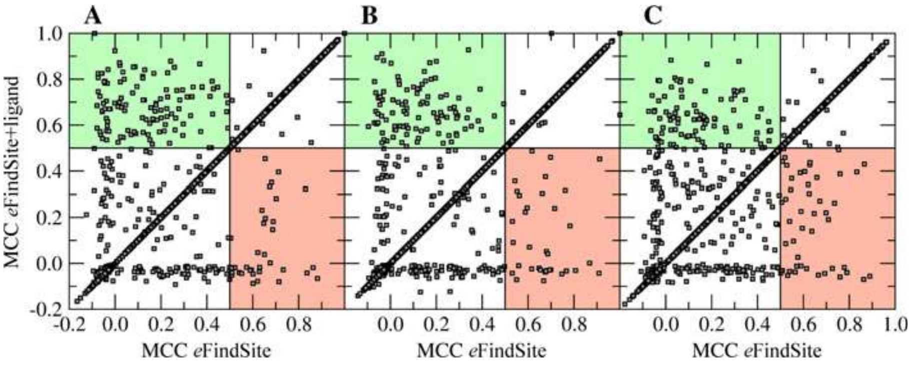

Fig. 8 Improvement of *e*FindSite + ligand over *e*FindSite for different quality target structures: **a** crystal, **b** high- and **c** moderate-quality. MCC is calculated for predicted versus experimental binding residues for the top-ranked binding pockets. *Green areas* highlight predictions significantly improved by including information on binding ligands, whereas *red areas* point out these cases, for which the performance of *e*FindSite + ligand is worse than *e*FindSite

Fig. 9 Confidence estimation system implemented in *e*FindSite. Using top-ranked binding sites, the correlation between estimated confidence and the actual accuracy of binding residue prediction is shown for **a** crystal structures, **b** high- and **c** moderate-quality protein models. The accuracy is measured by MCC calculated for predicted versus experimental binding residues

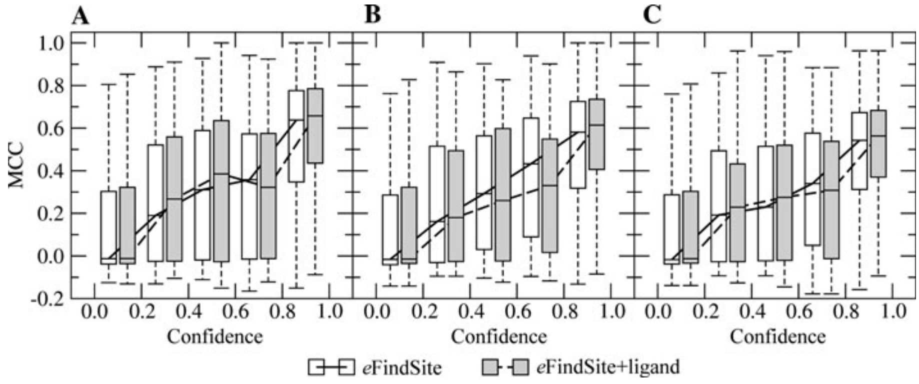

Table 3 Percentage of easy, medium and hard targets for ligand binding site prediction across three sets of different quality protein structures

Category	Confidence index	Crystal structures		High-quality models		Moderate-quality models	
		SVM (%)	SVM + ligand[a] (%)	SVM (%)	SVM + ligand[a] (%)	SVM (%)	SVM + ligand[a] (%)
Easy	>0.8	73.9	71.1	73.4	70.8	69.0	67.4
Medium	0.2–0.8	16.6	14.5	16.8	15.1	20.1	17.1
Hard	<0.2	9.5	14.4	9.8	14.1	10.9	15.5

[a] Including auxiliary ligands

Modeller and *e*Thread/TASSER-Lite, respectively. On the whole, the majority of structures are confidently predicted with an estimated TM-score of >0.7 for 1,771 (39 %) and 0.4-0.7 for 2,094 (46 %) models. Thus, ~85 % of *E. coli* proteome can be reliably moved to the structural level making these gene products promising targets for structure-based ligand binding site identification.

Using *e*FindSite, at least one ligand binding pocket is predicted for 2,828 gene products, which constitute 62 % of *E. coli* proteome. Among these, 1,300 (46 %) and 776 (27 %) are classified as “easy” and “medium” predictions, respectively, see Fig. 11. From calibration plots shown in Fig. 9, we may expect that the accuracy of identified binding residues in terms of MCC is ~0.6 and ~0.3 for “easy” and “medium” targets, respectively, indicating a fairly high precision of proteome-wide binding site prediction by *e*FindSite.

Case studies

To conclude proteome-wide binding pocket prediction for *E. coli*, we discuss a couple of representative examples that

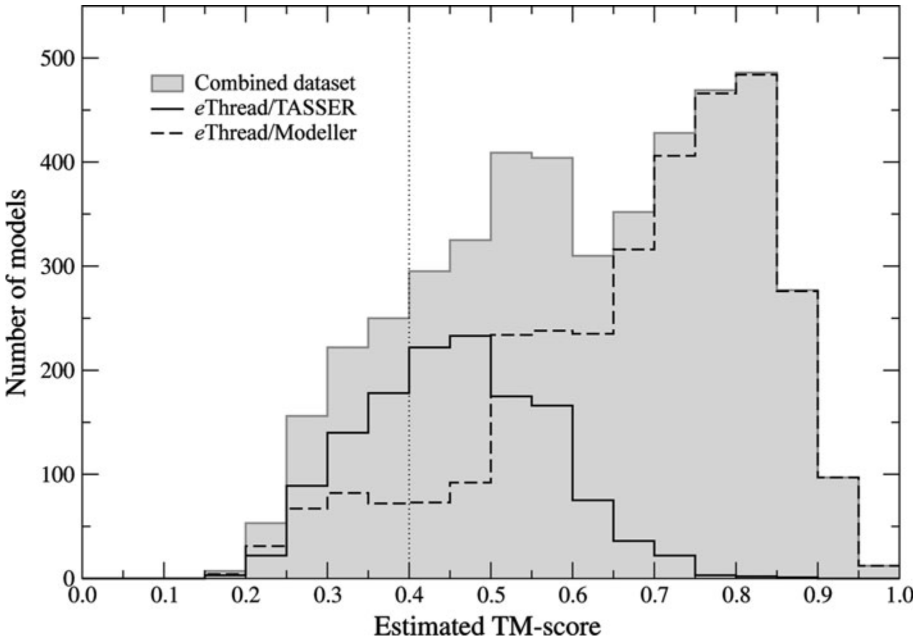

Fig. 10 Expected quality of structure models constructed for *E. coli* proteins by *e*Thread. Protein models are built using either *e*Thread/Modeller (*dashed line*) or *e*Thread/TASSER-Lite (*solid line*). Estimated TM-score is used as a quality assessment measure. The combined dataset including only the most confident models is shown in *gray*

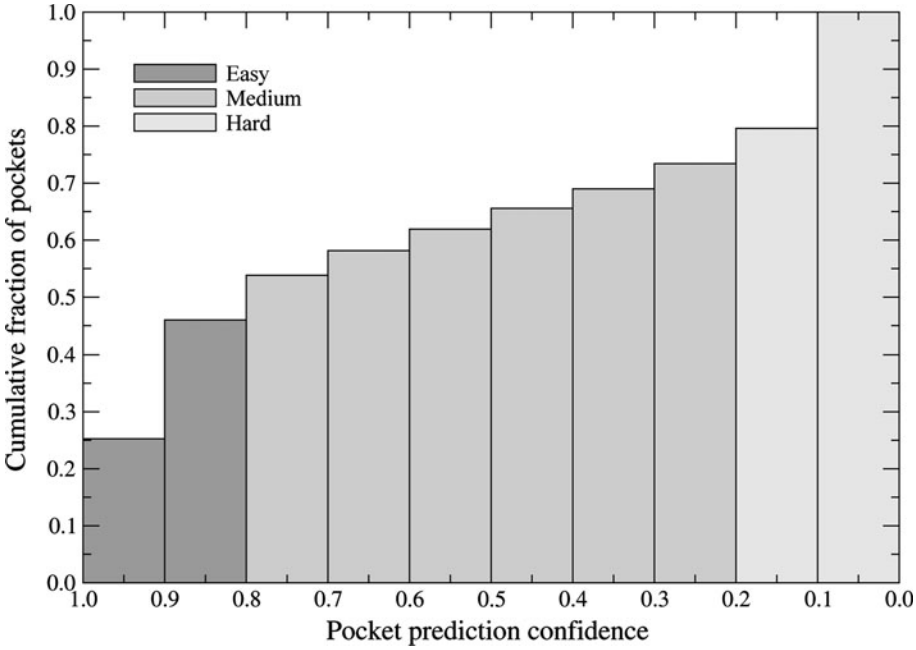

Fig. 11 Confidence of ligand binding pocket prediction across *E. coli* proteome. Confidence estimates are calculated by machine learning models calibrated on benchmarking datasets. “Easy”, “medium” and “hard” categories are shown in different shades of *gray*

demonstrate the utility of *e*FindSite in such large-scale projects. We selected two gene products, whose Ensembl IDs are EBESCP00000001015 and EBESCP00000003057. The first one is 394 amino acid long elongation factor Tu (EF-Tu; gene name: tuf) that functions as GTPase promoting the GTP-dependent binding of aminoacyl-tRNA to the A-site of ribosomes during protein biosynthesis. The estimated TM-score is 0.76 for the top-ranked model constructed by *e*Thread/Modeller from EF-Tu sequence (Fig. 12a), indicating a high accuracy of structure modeling. Next, we use *e*FindSite to predict ligand binding sites in the protein model. The top-ranked identified pocket shown in Fig. 12a has a high confidence of 0.91 and is formed by the following 13 putative binding residues: H20, V21, D22, H23, G24, K25, T26, T27, N136, K137, S174, A175, and L176. From literature, we collected experimental mutation data to validate *e*FindSite binding site prediction for EF-Tu. A single point mutation of V21 to glycine strongly reduces the GTPase activity [77]. Moreover, N136 was found essential for the correct formation of the nucleotide binding site [78]. Finally, predicted binding residues H20-K25 largely overlap with a consensus

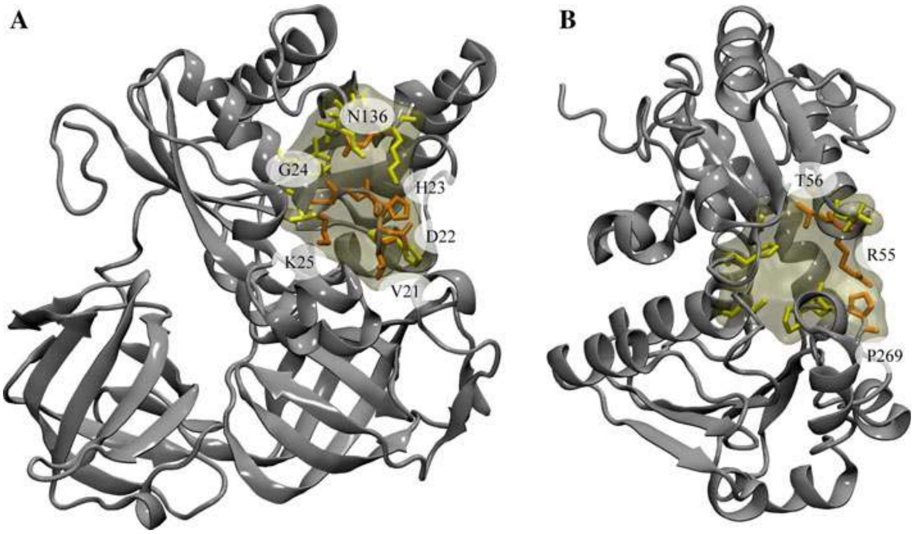

Fig. 12 Structure models constructed for **a** elongation factor Tu and **b** aspartate carbamoyltransferase from *E. coli* proteome. In each model, transparent *gray* surface shows the top-ranked binding pocket predicted by *e*FindSite with binding residues presented as sticks. Residues confirmed experimentally to bind a ligand are labeled and colored in *orange*

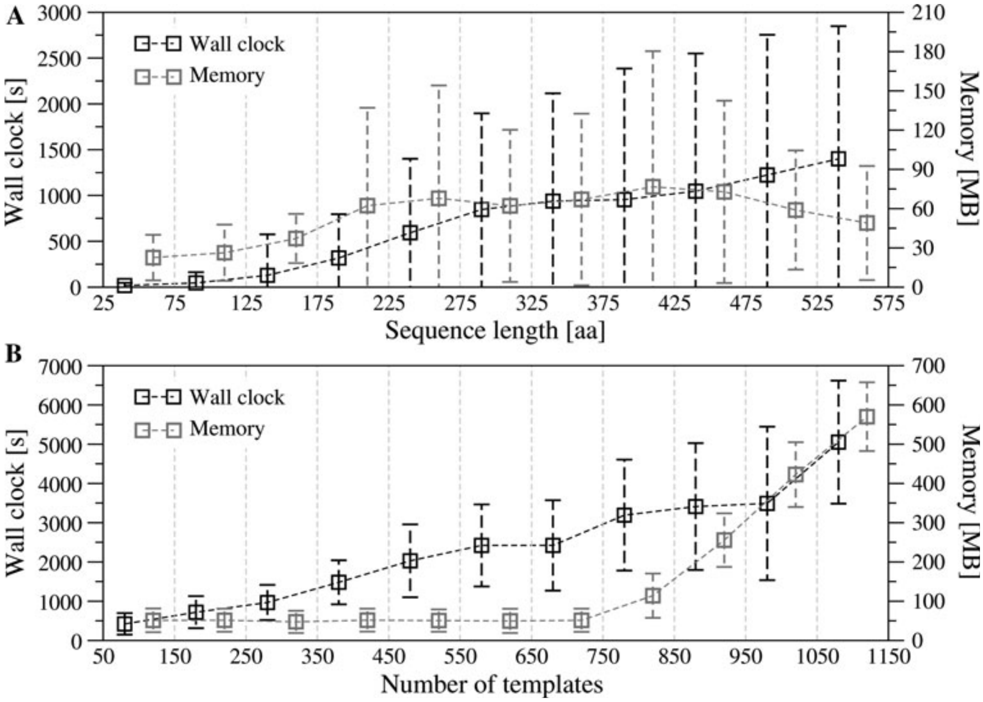

Fig. 13 Utilization of computing resources by *e*FindSite. Average ± standard deviation wall clock (*left ordinate*) and memory (*right ordinate*) is plotted as a function of **a** target protein length and **b** the number of identified template structures

sequence of residues G19-K25, which have been demonstrated to be important for interactions with GTP/GDP [79].

Our second example is 311 amino acid long aspartate carbamoyltransferase (ATCase; gene name: pyrB). Using the sequence of this protein, a highly confident model was built by *e*Thread/Modeller with an estimated TM-score of 0.73 (Fig. 12b). In this ATCase model, *e*FindSite identified a set of 10 residues that form the top-ranked predicted binding pocket, whose confidence is 0.90. These include A52, S53, R55, T56, G129, H135, T169, P267, L268 and P269. Available experimental data show that R55 as well as T56 are important for catalysis [80, 81]. Furthermore, replacing P269 with alanine dramatically decreases substrate affinity and, consequently, reduces the enzymatic activity [82]. These case studies demonstrate that predictions by *e*FindSite correlate well with site directed mutagenesis experiments.

Profiling of computational resources

Particularly for genome-scale applications that require processing a large number of jobs on high-performance computing systems, it is essential to estimate the resources needed for individual calculations. In that regard, we carry out resource profiling of *e*FindSite with respect to the CPU time and memory utilization. Figure 13 shows the average wall clock and memory usage (both ± standard deviation) for *e*FindSite. We identify two factors responsible for the resource consumption: target sequence length (Fig. 13a) and the number of template structures identified by *e*Thread (Fig. 13b). On average, *e*FindSite completes within ~30 min of CPU time and requires up to 200 MB of memory. However, larger protein targets and more template structures increase the demand for both wall clock and memory due to more intense structure alignment calculations. In a larger perspective, this comprehensive resource profiling can be used in efficient job scheduling on modern high-performance systems to maximize the utilization of computing resources and consequently, to reduce the time-to-completion of large-scale function annotation projects using *e*FindSite.

Conclusions

The knowledge of protein function needs to be continuously expanded to meet the challenges of systems biology, which is rapidly taking a center stage in biological research [83]. With the rapid accumulation of genome sequences, automated functional annotation of gene products is becoming critical. In addition to traditional experimental approaches, across-genome function inference is largely accomplished using computational techniques. Many proteins routinely interact with small molecules to regulate cellular activities and biological processes; therefore, the identification of binding sites is essential for protein function annotation. To address the limitations of purely sequence- and structure-based methods, we developed *e*FindSite, a combined evolution/structure-based approach to ligand binding prediction. A remarkable feature of *e*FindSite is its high tolerance to deformations in modeled target structures. Equally important, *e*FindSite is designed to effectively explore the "twilight zone" of sequence similarity, so that functional aspects of a target protein can be efficiently inferred from remote evolutionary relationships.

*e*FindSite employs highly sensitive meta-threading by *e*Thread [45] and the Affinity Propagation clustering algorithm [44] to optimize the selection of ligand-bound templates. This procedure is pivotal since binding site detection is essentially built upon the template selection. Furthermore, *e*FindSite extensively uses various machine learning techniques for template selection, binding residue prediction, binding site ranking and confidence estimation. Large-scale comparative benchmarks demonstrate a superior performance of *e*FindSite compared to its predecessor, FINDSITE [31], several geometrical pocket detection methods as well as binding pocket meta-predictors. A high tolerance of *e*FindSite to distortions in modeled protein structures stems from highly optimized template selection and weighting schemes, target-template as well as template–template global structure alignments, a new clustering procedure, and carefully tuned machine learning models. Interestingly, for non-native protein structures, we observe some differences in the performance of individual pocket detection algorithms depending not only on the quality of target structures, but also on the procedure used to construct these models. This will be investigated further in subsequent studies.

*e*FindSite is freely available to academic community as a user-friendly web-server as well as a well documented stand-alone software distribution at http://www.brylinski.org/efindsite; this website also provides all benchmarking datasets and results reported in this paper. Furthermore, the results of large-scale protein structure modeling and ligand binding prediction for *E. coli* proteome are freely available at http://www.brylinski.org/content/databases.

Acknowledgments This study was supported by the Louisiana Board of Regents through the Board of Regents Support Fund [contract LEQSF(2012–15)-RD-A-05] and Oak Ridge Associated Universities (ORAU) through the 2012 Ralph E. Powe Junior Faculty Enhancement Award. Portions of this research were conducted with high performance computational resources provided by Louisiana State University (http://www.hpc.lsu.edu).

Appendix

Molecular fingerprints are bit strings that represent the structural and chemical features of organic compounds (see Daylight manual for details: http://www.daylight.com/dayhtml/doc/theory/index.pdf). Tanimoto coefficient is the most popular measure to quantify the similarity of two sets of bits (e.g. molecular fingerprints). Classical Tanimoto coefficient (*TC*) [60] is defined as:

$$TC = \frac{c}{a + b + c} \qquad (2)$$

where *a* is the count of bits on in the 1st string but not in the 2nd string, *b* is the count of bits on in the 2nd string but not in the 1st string, and *c* is the count of the bits on in both strings.

In addition to the classical Tanimoto coefficient, the overlap between two molecular fingerprints can be measured by the average Tanimoto coefficient (*aveTC*) [84]:

$$aveTC = \frac{TC + TC'}{2} \qquad (3)$$

where TC' is the Tanimoto coefficient calculated for bit positions set off rather than set on.

Furthermore, a version of the Tanimoto coefficient for continuous variables (*conTC*) [85] was developed:

$$conTC = \frac{\sum x_{pi}x_{ci}}{\sum x_{pi}^2 + \sum x_{ci}^2 - \sum x_{pi}x_{ci}} \qquad (4)$$

where x_{pi} is the *i*-th descriptor of a fingerprint profile and x_{ci} is the *i*-th descriptor of a query compound. The fingerprint profile is constructed from individual fingerprints for a set of compounds, e.g. template-bound ligands that were used to identify a putative binding site in the target structure.

References

1. Hoehndorf R, Kelso J, Herre H (2009) The ontology of biological sequences. BMC Bioinformatics 10:377
2. Stevens R, Goble CA, Bechhofer S (2000) Ontology-based knowledge representation for bioinformatics. Brief Bioinformatics 1(4):398–414
3. Ashburner M et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29
4. Harris MA et al (2004) The gene ontology (GO) database and informatics resource. Nucleic Acids Res, 32(Database issue): D258–61
5. Lybrand TP (2002) In: Naray-Szabo G, Warshel A (eds) Protein-ligand interactions, in computational approaches to biochemical reactivity. Springer, Boston, pp 363–374
6. Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46
7. Zhang J et al (2011) The impact of next-generation sequencing on genomics. J Genet Genomics 38(3):95–109
8. Juncker AS et al (2009) Sequence-based feature prediction and annotation of proteins. Genome Biol 10(2):206
9. Loewenstein Y et al (2009) Protein function annotation by homology-based inference. Genome Biol 10(2):207
10. Ahmad S, Sarai A (2005) PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics 6:33
11. Hwang S, Gou Z, Kuznetsov IB (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 23(5):634–636
12. Chen P, Li J (2010) Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information. BMC Bioinformatics 11:402
13. Chen XW, Jeong JC (2009) Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics 25(5):585–591
14. Soding J (2005) Protein homology detection by HMM–HMM comparison. Bioinformatics 21(7):951–960
15. Lopez G et al (2011) Firestar—advances in the prediction of functionally important residues. Nucleic Acids Res 39(Web Server issue): W235–41
16. Lord PW et al (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19(10):1275–1283
17. Schnoes AM et al (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5(12):e1000605
18. Zhang QC et al (2011) PredUs: a web server for predicting protein interfaces using structural neighbors. Nucleic Acids Res 39(Web Server issue): W283–7
19. Brylinski M et al (2007) Prediction of functional sites based on the fuzzy oil drop model. PLoS Comput Biol 3(5):e94
20. Brylinski M et al (2007) Localization of ligand binding site in proteins identified in silico. J Mol Model 13(6–7):665–675
21. Dudev M, Lim C (2007) Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. BMC Bioinformatics 8:106
22. Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13(5):323–30, 307–8
23. Liang J, Edelsbrunner H, Woodward C (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 7(9):1884–1897
24. Levitt DG, Banaszak LJ (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 10(4):229–234
25. Huang B, Schroeder M (2006) LIGSITEcsc: predicting ligand binding sites using the connolly surface and degree of conservation. BMC Struct Biol 6:19
26. Le Guilloux V, Schmidtke P, Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics 10:168
27. Zhu H, Pisabarro MT (2011) MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets. Bioinformatics 27(3):351–358
28. Huang B (2009) MetaPocket: a meta approach to improve protein ligand binding site prediction. OMICS 13(4):325–330
29. Skolnick J, Brylinski M (2009) FINDSITE: a combined evolution/structure-based approach to protein function prediction. Brief Bioinformatics 10(4):378–391
30. Wass MN, Kelley LA, Sternberg MJ (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res 38(Web Server issue): W469–73
31. Brylinski M, Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci U S A 105(1):129–134
32. Roche DB, Tetchner SJ, McGuffin LJ (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. BMC Bioinformatics 12:160
33. Brylinski M, Skolnick J (2011) FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level. Proteins 79(3):735–751
34. Dror I et al (2011) Predicting nucleic acid binding interfaces from structural models of proteins. Proteins
35. Mukherjee S, Zhang Y (2011) Protein-protein complex structure predictions by multimeric threading and template recombination. Structure 19(7):955–966
36. Tyagi M et al (2012) Homology inference of protein–protein interactions via conserved binding sites. PLoS ONE 7(1):e28896
37. Pandit SB, Skolnick J (2008) Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. BMC Bioinformatics 9:531
38. Ortiz AR, Strauss CE, Olmea O (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 11(11):2606–2621
39. Russell RB, Sasieni PD, Sternberg MJ (1998) Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol 282(4):903–918
40. Brylinski M, Skolnick J (2010) Comparison of structure-based and threading-based approaches to protein functional annotation. Proteins 78(1):118–134
41. Laurie AT, Jackson RM (2006) Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. Curr Protein Pept Sci 7(5):395–406
42. Li YY, An J, Jones SJ (2006) A large-scale computational approach to drug repositioning. Genome Inform 17(2):239–247

43. Li YY, An J, Jones SJ (2011) A computational approach to finding novel targets for existing drugs. PLoS Comput Biol 7(9):e1002139
44. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315(5814):972–976
45. Brylinski M, Lingam D (2012) eThread: a highly optimized machine learning-based approach to meta-threading and the modeling of protein tertiary structures. PLoS ONE 7(11):e50200
46. Brylinski M, Feinstein WP (2012) Setting up a meta-threading pipeline for high-throughput structural bioinformatics: eThread software distribution, walkthrough and resource profiling. J Comput Sci Syst Biol 6(1):001–010
47. Wallach I, Lilien R (2009) The protein-small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding. Bioinformatics 25(5):615–620
48. Wang G, Dunbrack RL Jr (2003) PISCES: a protein sequence culling server. Bioinformatics 19(12):1589–1591
49. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57(4):702–710
50. Berman HM et al (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
51. Bindewald E, Skolnick J (2005) A scoring function for docking ligands to low-resolution protein structures. J Comput Chem 26(4):374–383
52. Biegert A, Soding J (2009) Sequence context-specific profiles for homology searching. Proc Natl Acad Sci USA 106(10):3770–3775
53. Sadreyev R, Grishin N (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 326(1):317–336
54. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14(9):755–763
55. Bucher P et al (1996) A flexible motif search technique based on generalized profiles. Comput Chem 20(1):3–23
56. Lobley A, Sadowski MI, Jones DT (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics 25(14):1761–1767
57. Hughey R, Krogh A (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. Comput Appl Biosci 12(2):95–107
58. Zhou H, Zhou Y (2005) SPARKS 2 and SP3 servers in CASP6. Proteins 61(Suppl 7):152–156
59. Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358(6381):86–89
60. Tanimoto TT (1958) An elementary mathematical theory of classification and prediction, in IBM Internal Report
61. Guha R et al (2006) The blue obelisk-interoperability in chemical informatics. J Chem Inf Model 46(3):991–998
62. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd ed. Morgan Kaufmann Publishers, San Francisco
63. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453
64. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5(4):725–738
65. Soga S et al (2007) Use of amino acid composition to predict ligand-binding sites. J Chem Inf Model 47(2):400–406
66. Marti-Renom MA et al (2007) The AnnoLite and AnnoLyze programs for comparative annotation of protein structures. BMC Bioinformatics 8(Suppl 4):S4
67. Liu T, Altman RB (2009) Prediction of calcium-binding sites by combining loop-modeling with machine learning. BMC Struct Biol 9:72
68. Kawabata T (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins 78(5):1195–1211
69. Zhang Z et al (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. Bioinformatics 27(15):2083–2088
70. Blattner FR et al (1997) The complete genome sequence of Escherichia coli K-12. Science 277(5331):1453–1462
71. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815
72. Pandit SB, Zhang Y, Skolnick J (2006) TASSER-Lite: an automated tool for protein comparative modeling. Biophys J 91(11):4180–4190
73. Brylinski M, Skolnick J (2007) What is the relationship between the global structures of apo and holo proteins? Proteins 70(2): 363–377
74. Chen X, Liu M, Gilson MK (2001) BindingDB: a web-accessible molecular recognition database. Comb Chem High Throughput Screen 4(8):719–725
75. Wang Y et al (2009) PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res, 37(Web Server issue): W623–33
76. Wishart DS et al (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 34(Database issue): D668–72
77. Jacquet E, Parmeggiani A (1988) Structure-function relationships in the GTP binding domain of EF-Tu: mutation of Val20, the residue homologous to position 12 in p21. EMBO J 7(9):2861–2867
78. Weijland A et al (1993) Asparagine-135 of elongation factor Tu is a crucial residue for the folding of the guanine nucleotide binding pocket. FEBS Lett 330(3):334–338
79. Gumusel F et al (1990) Mutagenesis of the NH2-terminal domain of elongation factor Tu. Biochim Biophys Acta 1050(1–3): 215–221
80. Stebbins JW et al (1992) Arginine 54 in the active site of Escherichia coli aspartate transcarbamoylase is critical for catalysis: a site-specific mutagenesis, NMR, and X-ray crystallographic study. Protein Sci 1(11):1435–1446
81. Waldrop GL et al (1992) The contribution of threonine 55 to catalysis in aspartate transcarbamoylase. Biochemistry 31(28):6592–6597
82. Jin L, Stec B, Kantrowitz ER (2000) A cis-proline to alanine mutant of *E. coli* aspartate transcarbamoylase: kinetic studies and three-dimensional crystal structures. Biochemistry 39(27):8058–8066
83. Kitano H (2002) Systems biology: a brief overview. Science 295(5560):1662–1664
84. Xue L et al (2003) Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. J Chem Inf Comput Sci 43(4):1151–1157
85. Willett P (1998) Chemical similarity searching. J Chem Inf Model 38:983–996