Quantum Bio-Informatics III
eds. L. Accardi, W. Freudenberg & M. Ohya
© 2010 World Scientific Publishing (pp. 327–336)

NOVEL COMPUTATIONAL APPROACHES TO DRUG DISCOVERY

JEFFREY SKOLNICK[†]

*Center for the Study of Systems Biology, Georgia Institute of Technology,
250 14th Street NW, Atlanta, GA, 30076 USA*

MICHAL BRYLINSKI

*Center for the Study of Systems Biology, Georgia Institute of Technology,
250 14th Street NW, Atlanta, GA, 30076 USA*

New approaches to protein functional inference based on protein structure and evolution are described. First, FINDSITE, a threading based approach to protein function prediction, is summarized. Then, the results of large scale benchmarking of ligand binding site prediction, ligand screening, including applications to HIV protease, and GO molecular functional inference are presented. A key advantage of FINDSITE is its ability to use low resolution, predicted structures as well as high resolution experimental structures. Then, an extension of FINDSITE to ligand screening in GPCRs using predicted GPCR structures, FINDSITE/QDOCKX, is presented. This is a particularly difficult case as there are few experimentally solved GPCR structures. Thus, we first train on a subset of known binding ligands for a set of GPCRs; this is then followed by benchmarking against a large ligand library. For the virtual ligand screening of a number of Dopamine receptors, encouraging results are seen, with significant enrichment in identified ligands over those found in the training set. Thus, FINDSITE and its extensions represent a powerful approach to the successful prediction of a variety of molecular functions.

1. Introduction

One of the key goals of Systems Biology is to understand the function of all molecules in a cell and how they interact on a system-wide level [1]. In that respect, we specifically focus on computational tools designed to elucidate biochemical function. By detecting evolutionary relationships between proteins, sequence-based methods can provide insights into the function of about 50% of the ORFs in a given proteome [2], with the remainder being too evolutionarily distant to accurately infer function based on sequence information alone [3]. Thus, the prediction of the function of these unannotated ORFs is a significant challenge. However, since protein structure is more conserved than protein sequence [4], it can play an essential role in annotating genomes [5], including lead compound identification for subsequent use in drug discovery [6]. Of course, the key question is whether one can use low-to-moderate resolution predicted structures or if high-resolution experimental structures are required [7, 8]. This issue also has implications for the requisite scope of structural genomics that aims for high-throughput protein structure determination [9]. If low-to-moderate resolution models were to prove useful for functional inference, then the value of contemporary protein structure prediction approaches would be significantly enhanced [10].

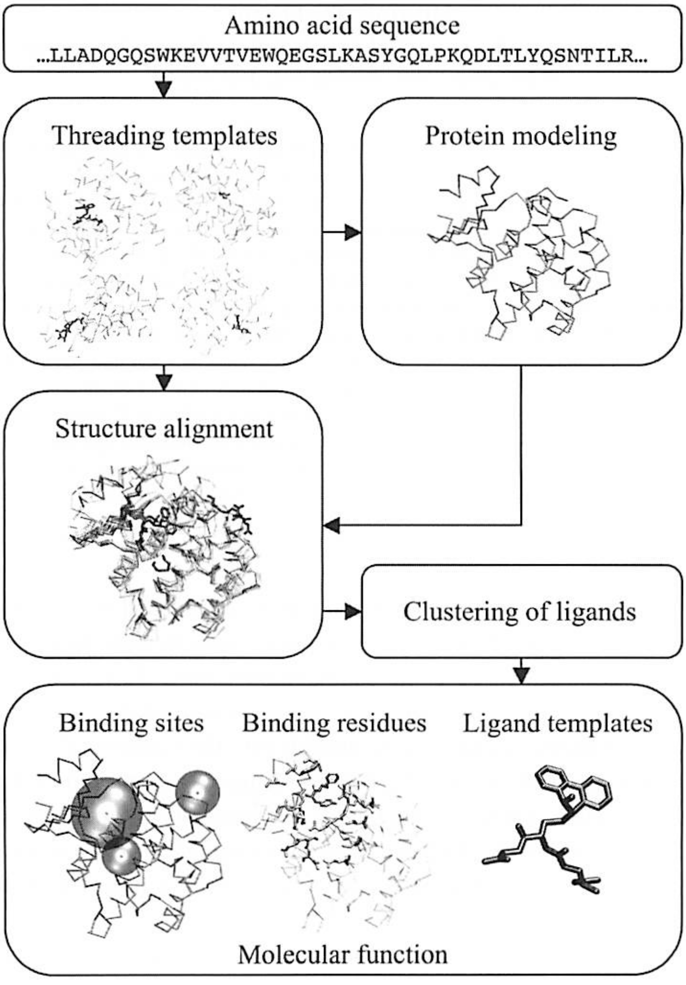

Figure 1. Overview of the FINDSITE threading-based protein functional inference algorithm.

2. FINDSITE: A threading based approach to protein functional inference

A systematic analysis of known protein structures grouped according to SCOP [11] reveals a general tendency of certain protein folds to bind substrates at a similar structural location, suggesting that evolution tends to conserve the functionally important region and a subset of ligand binding features [12]. If so, it should be possible to develop an approach for ligand binding site identification that is less sensitive than pocket-detection methods to distortions in the modeled structure. Thus, we developed FINDSITE [7], a method for the prediction of ligand-binding sites and protein functional annotation based on binding site similarity among superimposed groups of template structures identified from threading [13]. A schematic overview of the methodology is shown in Figure 1. For a given target protein, PROSPECTOR_3.5 [14] identifies ligand-bound structure templates. Then, holo-templates are superimposed onto the predicted (or experimental, if available) target protein structure using the structural alignment algorithm TM-align [15]. Upon superimposition, the clustered centers of mass of the ligands bound to the threading templates identify putative binding sites, and the predicted sites are ranked according to the number of templates that share a common binding pocket. FINDSITE also specifies the chemical properties of the ligands that likely occupy the binding site. To assess its validity, we employed a representative set of 901 proteins with < 35% sequence identity to their templates and generated models using the structure prediction algorithm TASSER [16].

2.1. *Ligand binding site prediction*

We evaluated the performance of both the LIGSITE^{CSC} [17] pocket-detection and FINDSITE threading-based approaches on a non-redundant benchmark set of 901 proteins in terms of the accuracy of ligand-binding site prediction and the ability to correctly rank identified pockets in both crystal structures and protein models. The results of ligand binding site prediction are shown in Figure 2. In Figure 2A, we employ the target protein's crystal structure. For LIGSITE^{CSC}, the native structure is scanned to identify binding pockets. For FINDSITE, predicted template models (with sequence identity <35% to the target) are superimposed onto their crystal structure and the most populated binding site is selected. FINDSITE performs better than the pocket-detection method in both overall accuracy and ranking ability of identified pockets. When the native crystal structure is used, using the best of top five identified binding pockets, the success rate is 70.9% and 51.3% for FINDSITE and LIGSITE^{CSC}, respectively.

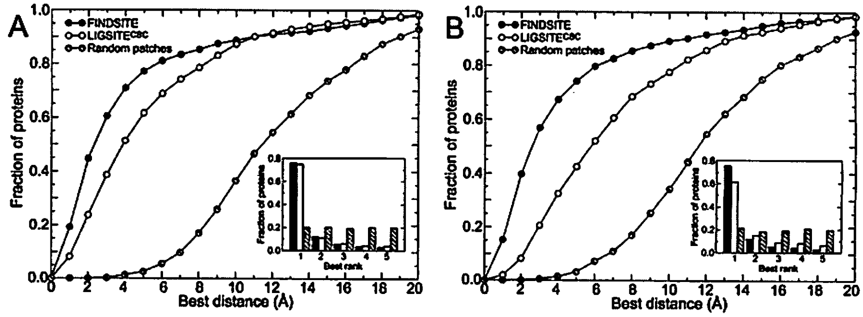

Figure 2. Performance of FINDSITE and LIGSITE^{CSC} compared to randomly selected patches on a target protein surface using target crystal structures (A) and TASSER models (B). Results are presented as the cumulative fraction of proteins with a distance between the center of mass of a ligand in the native complex and the center of the best of top five predicted binding sites ≤ the distance displayed on the x-axis and the rank of the best pocket selected from the top five predictions (inset). The results are much worse when random patches are chosen.

As shown in Figure 2B, using LIGSITE^{CSC}, the prediction accuracy falls off considerably when modeled protein structures rather than the experimental structures are used. LIGSITE^{CSC}'s success rate decreases from 51.3%, when crystal structures are used to 27.2% and 32.5% for PROSPECTOR_3 template structures and protein models generated by TASSER, respectively. This decrease is accompanied by deterioration in the ability to correctly rank the binding site. For the top-ranked threading templates and TASSER models, only 59.1% and 61.4% of the best pockets are assigned rank 1 by LIGSITE^{CSC}. In contrast, with FINDSITE both the high accuracy of ligand binding site prediction and the ability to correctly rank the identified binding sites are sustained if models instead of native structures are used as reference structures for holo-template superposition. The success rate is 67.0% and 67.3% for the top-ranked PROSPECTOR_3 templates and TASSER models, respectively, with a corresponding top ranking accuracy of 75.9%, 75.5% and 75.7%. In practice, FINDSITE tolerates structures with a global RMSD from native of 8-10 Å, provided that the local binding site has a backbone RMSD from native in the range of 2-3 Å.

2.2. *Ligand screening*

FINDSITE also provides information on the chemical properties of the binding ligands, termed here "template ligands". Subsequently, these molecules are used as ligand templates by fingerprint-based similarity searching [18] in a simple ligand-based virtual screening experiment against the KEGG compound library that contains 12,478 compounds [19]. We note that for a given target protein, template ligands can be selected even when the crystal structure is unavailable and its molecular function is unknown. Figure 3 presents the cumulative distribution of enrichment factors calculated for the 901 representative target proteins that have <35% sequence identity to the closest template protein. For accurately predicted binding sites (whose center of mass is ≤ 4 Å from the experimental one; this holds for 70.9% of the target proteins), in 78% of the cases, FINDSITE performs better than random. The ideal enrichment factor (all native-like compounds in the top 1% of the ranked library [7]) was observed for 50% of target proteins. For less accurately predicted binding pockets, ligand selection is notably worse (the ideal enrichment factor was obtained for 12% of the cases and is better than random for 34%). Finally, a case study examined the performance of FINDSITE in virtual screening for 895 active *HIV-1 protease inhibitors* in a 123,331 compound library. Again, if only templates with <35% sequence identity to the target are used, the enrichment factor of the top 1% of compounds is 40.

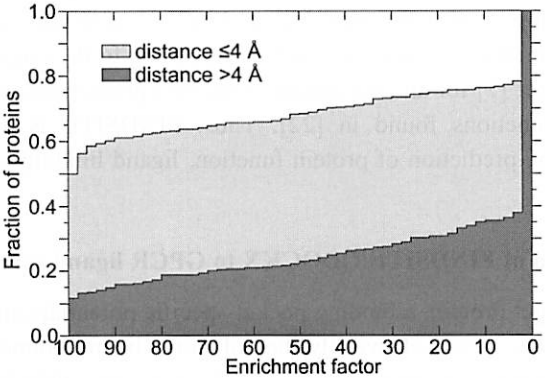

Figure 3. Cumulative distribution of enrichment factors resulting from the ligand-based virtual screening against the KEGG compound library using ligand templates selected by FINDSITE. Target proteins are divided into the two subsets with respect to binding pocket prediction accuracy (the distance between the top-ranked pocket and the center of mass of the native ligand ≤4 Å and >4 Å).

2.3. *Molecular Function Inference*

The relatively high accuracy of the ligand selection procedure encouraged us to investigate the transferability of specific functions of the threading templates to the target. We use Gene Ontology [20], GO, to describe molecular protein function. From the benchmark set, we selected 753 proteins for which a GO annotation is provided by Gene Ontology [20] or UniProt [21]. The procedure for molecular function prediction employs the superimposed group of holo-templates selected by threading as previously used for binding site detection and ligand selection. For benchmarking purposes, only predicted threading templates with <35% sequence identity to a target protein are used. For each target protein, all GO annotations are identified for the threading templates that share the top-ranked predicted binding site using the GO and UniProt databases. Then, the target protein is assigned a function with a probability that corresponds to the fraction of threading templates annotated with that molecular function. For a probability threshold of 0.5 (i.e. at least one half of the threading holo-templates must be annotated with the same GO term to transfer it to the target protein), the maximal Matthew's correlation coefficient of 0.64 is found. This corresponds to a precision (True Positives / (True Positives + False Positives)) of 0.76 with a recall (True Positives / (True Positives + False Negatives)) of 0.54. In addition, we calculated predictive metrics with respect to individual GO identifiers. When the closest template has <35% sequence identity, FINDSITE distinguishes between enzymatic and non-enzymatic function, with a precision and sensitivity of 0.93 and 0.89, respectively. Moreover, many molecular functions that cover a broad spectrum of molecular events including both enzymatic and binding activities are accurately transferable from the templates selected by threading to the target proteins. See SI, Table 1 of ref [7] for an assessment of the best predictions, with the full set of predicted functions found in [22]. Thus, FINDSITE is an encouraging approach for the prediction of protein function, ligand binding sites and ligand screening.

3. Application of FINDSITE/QDOCKX to GPCR ligand screening

For a given target protein, a binding pocket-specific potential can be derived by FINDSITE from a set of weakly homologous ligand-bound templates to improve the ligand docking accuracy [7, 23]. For most globular proteins, the requisite evolutionary related template structures can be identified by a threading approach that employs a strong sequence profile component [24]. For GPCRs, the limited number of experimentally solved structures requires construction of a synthetic protein structure/ligand library compiled from GPCR models complexed with small molecules as predicted by ligand docking. For selected GPCRs, known binders extracted from ligand databases (Drug Bank [25] and the MDL Drug Data Report [26]) are used to model the receptor-ligand complexes by low-resolution ligand docking using the Q-Dock ligand docking/screening algorithm [27]. The first pass of the docking simulations employs a generic force-field for protein-ligand interactions derived from the non-redundant dataset of ligand-bound proteins from the Protein Data Bank, PDB [28]. To improve binding mode prediction, the generic contact potential is amplified at highly conserved amino acid positions (taken from GPCR class-specific multiple sequence alignments provided by GPCRDB (information system for G-protein-coupled receptors) [29]) to enforce the experimentally known interactions with ligands. Subsequently, the known bioactive molecules are clustered using a SIMCOMP (a chemical compound-matching algorithm that provides atom equivalences) [30] similarity cutoff of 0.7. For each cluster of similar compounds, a common molecular structure, "the anchor", is then identified. The anchor substructure is defined as a maximum set of functional groups present in at least 90% of the ligands from a single cluster. Having a well-defined anchor substructure, seed molecules are extracted from ligands docked into the receptor binding pocket. Seed molecules comprise the largest set of compounds that have their common substructures docked within a 4 Å RMSD from each other. The consensus-binding mode is derived by averaging the anchor substructure pose in the seed molecules. Finally, in the second pass of docking simulations, the known active molecules are re-docked into the receptor binding pocket, but now with harmonic RMSD restraints imposed on the consensus binding mode of the anchor substructure. The synthetic GPCR-ligand library consists of TASSER-generated protein models with the bioactive molecules placed into the binding pockets during the second pass of constrained docking. Instead of the library of ligand-bound crystal structures, the synthetic library is then used by FINDSITE to derive pocket-specific potentials for high-throughput ligand docking/screening.

3.1. *Virtual Screening Experiments*

For a preliminary assessment of the ranking capability of our approach, using the identical synthetic protein/ligand library protocol as above, we performed a simplified virtual screening experiment against relatively small ligand libraries for the following three dopamine receptors: D_2, D_3, and D_4. For each target, known bioactive molecules were extracted from the MDL Drug Data Report [26] and divided into a training and benchmark subsets. The training subset was used to construct 1024-bit Daylight fingerprints [18] for ligand-based virtual screening as well as to compile the synthetic GPCR library in order to derive pocket-specific potentials and the consensus binding modes of the common ligand substructures. In ligand-based virtual screening, we followed the protocol previously applied in virtual screening of HIV-1 protease [7]. For each active compound from the benchmark set, 10 background compounds were randomly selected from the Asinex compound libraries [31]. The goal of this study was to examine whether virtual screening is capable of ranking the known bioactive molecules over the background ligands. We carried out ligand- and structure-based screening separately for each target and then we combined the results using data fusion techniques [32]. The results were assessed in terms of the enrichment behavior, i.e. the fraction of known active compounds recovered in the top-ranked fraction of the library.

We present the results for dopamine D_2, D_3, and D_4 receptor screening in Figure 4. For these GPCR targets, virtual screening performed significantly better than random ligand selection. Moreover, we observed that the enrichment calculated for the combined ranks from ligand- and receptor-based screenings is typically higher than that calculated for each method alone. These results suggest that the proposed protocol for virtual screening GPCRs may be useful in the discovery of new biopharmaceuticals.

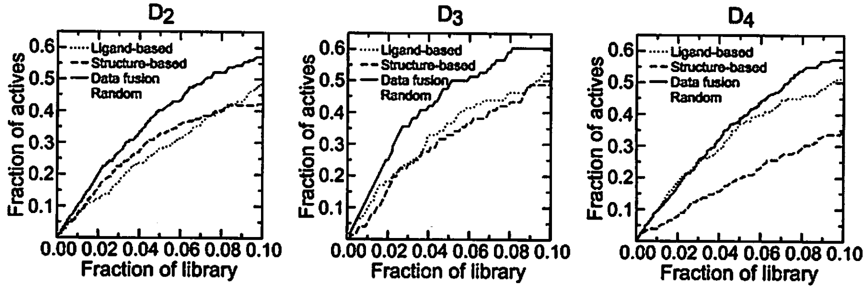

Figure 4. Results of ligand library screening of three dopamine receptors, D_2, D_3, and D_4, using known binders to construct the ligand fingerprint (dotted), structure based (dash), Data fusion (solid) and random (gray) approaches.

4. Discussion

Many methods for protein function prediction employ functional inference by homology [33, 34]; that is, if two proteins are evolutionary related, then the function of the protein of known function is assigned to that of unknown function. However, for enzymes [2], because the extent of function conservation depends on the protein family, care must be taken if the goal is both high accuracy and coverage of sequence space. To address this issue, a number of structure-based approaches based on three-dimensional geometric descriptors of enzymatic function have been developed [5]; In practice, such approaches have been restricted to enzymes and require extensive manual intervention. To eliminate these limitations, we developed the FINDSITE algorithm [7] which exploits the fact that binding sites are strongly conserved among evolutionary distant proteins. The conservation of binding sites among threading identified templates can be used to predict the target binding site, the ligands that bind to this site and consensus GO molecular functions [20]. Since these observations hold for evolutionary quite distant proteins (well below 35% sequence identity), this has profound implications as to how protein molecular function evolved.

Not only is the protein's structure conserved, but the chemical features of the ligands that bind to the protein are also conserved. This provides a type of signal averaging that can be exploited in ligand screening.

A clear advantage of FINDSITE is that predicted as well as experimental structures can be used with comparable results. This is important because state-of-the-art approaches provide useful predicted structures for > 2/3 of protein domains in a given proteome [16, 35]. It is quite likely that there are other functional properties that can be detected by extensions of the FINDSITE approach, one of which, the extension to GPCR virtual ligand screening, was described here with encouraging preliminary results. The idea is to identify distantly related structures with common functional features and then transfer these features to the protein of interest. Thus, this promising avenue of investigation should help extend the range of applicability of structure-based approaches to protein function prediction.

Acknowledgments

This work is supported by NIH grant Nos. GM-48835, GM-37408, GM-084222 and RR-12255.

References

1. L. Hood *et al.*, Science **306**, 640 (2004).
2. W. Tian, and J. Skolnick, J Mol Biol **333**, 863 (2003).
3. R. Kolodny, P. Koehl, and M. Levitt, J Mol Biol **346**, 1173 (2005).
4. A. Andreeva *et al.*, Nucleic Acids Res **36**, D419 (2008).
5. J. S. Fetrow, and J. Skolnick, J Mol Biol **281**, 949 (1998).
6. D. M. Schnur, Curr Opin Drug Discov Devel **11**, 375 (2008).
7. M. Brylinski, and J. Skolnick, Proc Natl Acad Sci U S A **105**, 129 (2008).
8. M. Brylinski, and J. Skolnick, Proteins **70**, 363 (2008).
9. M. Gerstein *et al.*, Science **299**, 1663 (2003).
10. A. Kryshtafovych, K. Fidelis, and J. Moult, Proteins 69 Suppl **8**, 194 (2007).
11. A. G. Murzin *et al.*, J Mol Biol **247**, 536 (1995).
12. R. B. Russell, J Mol Biol **279**, 1211 (1998).
13. D. T. Jones, and C. Hadley, in Bioinformatics: Sequence, structure and databanks, edited by D. Higgins, and W. R. Taylor (Springer-Verlag, Heidelberg, 2000), pp. 1.
14. S. Y. Lee, and J. Skolnick, Biophys J (2008).
15. Y. Zhang, and J. Skolnick, Nucleic Acids Res **33**, 2302 (2005).
16. Y. Zhang, and J. Skolnick, Proc Natl Acad Sci U S A **101**, 7594 (2004).
17. B. Huang, and M. Schroeder, BMC Struct Biol **6**, 19 (2006).
18. in Daylight Theory Manual, (Daylight Chemical Information Systems, Inc., Aliso Viejo, CA, 2007).

19. M. Kanehisa, and S. Goto, Nucleic Acids Res **28**, 27 (2000).
20. M. Ashburner *et al.*, Nat Genet **25**, 25 (2000).
21. C. H. Wu *et al.*, Nucleic Acids Res **34**, D187 (2006).
22. M. Brylinski, and J. Skolnick, 2008).
23. M. Brylinski, and J. Skolnick, J Comput Chem **29**, 1574 (2008).
24. J. Skolnick, and D. Kihara, Proteins **42**, 319 (2001).
25. D. S. Wishart *et al.*, Nucleic Acids Res **34**, D668 (2006).
26. A. E. Fournier, T. GrandPre, and S. M. Strittmatter, Nature **409**, 341 (2001).
27. M. Brylinski, and J. Skolnick, J Comput Chem (2008).
28. H. M. Berman *et al.*, Nucleic Acids Res **28**, 235 (2000).
29. F. Horn *et al.*, Nucleic Acids Res **31**, 294 (2003).
30. M. Hattori *et al.*, Genome Inform **14**, 144 (2003).
31. A. Gaulton, and T. K. Attwood, Nucleic Acids Res **31**, 3333 (2003).
32. M. Whittle *et al.*, J Chem Inf Model **46**, 2193 (2006).
33. D. Groth, H. Lehrach, and S. Hennig, Nucleic Acids Res **32**, W313 (2004).
34. G. Zehetner, Nucleic Acids Res **31**, 3799 (2003).
35. M. A. Marti-Reno*m et al.*, Annu Rev Biophys Biomol Struct **29**, 291 (2000).