W400–W408 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 22 May 2012*
doi:10.1093/nar/gks421

PocketAnnotate: towards site-based function annotation

Praveen Anand, Kalidas Yeturu and Nagasuma Chandra*

Department of Biochemistry, Indian Institute of Science, Bangalore 560012, Karnataka, India

Received February 24, 2012; Revised April 15, 2012; Accepted April 23, 2012

ABSTRACT

A computational pipeline PocketAnnotate for functional annotation of proteins at the level of binding sites has been proposed in this study. The pipeline integrates three in-house algorithms for site-based function annotation: PocketDepth, for prediction of binding sites in protein structures; PocketMatch, for rapid comparison of binding sites and PocketAlign, to obtain detailed alignment between pair of binding sites. A novel scheme has been developed to rapidly generate a database of non-redundant binding sites. For a given input protein structure, putative ligand-binding sites are identified, matched in real time against the database and the query substructure aligned with the promising hits, to obtain a set of possible ligands that the given protein could bind to. The input can be either whole protein structures or merely the substructures corresponding to possible binding sites. Structure-based function annotation at the level of binding sites thus achieved could prove very useful for cases where no obvious functional inference can be obtained based purely on sequence or fold-level analyses. An attempt has also been made to analyse proteins of no known function from Protein Data Bank. PocketAnnotate would be a valuable tool for the scientific community and contribute towards structure-based functional inference. The web server can be freely accessed at http://proline.biochem.iisc.ernet.in/pocketannotate/.

INTRODUCTION

In the post-genomic era, the number of completely sequenced genomes has increased tremendously widening the sequence-structure gap. Various structural genomics consortia have been setup to determine the structures of those unique sequences to bridge this sequence-structure gap. As an obvious outcome, there is significant increase in the number of proteins whose functions have not been determined (1,2). Annotation can be of several types—from a crude association of a biological process to detailed mechanistic view of the underlying biological reaction. Typically for homologous proteins, sequence-to-function models suffice in obtaining broad associations, based on the premise that when two sequences are similar to each other, their structures and hence function is also likely to be similar. In cases where structure may be conserved but sequences are not homologous, structure to function models are used, as it is well understood that two proteins, that share no significant sequence similarities, could still adopt the same fold and hence exhibit the same function. Although useful in many cases, fold-level descriptions are not always sufficient as structures having the same fold can exhibit different functions and vice versa (3–5).

The rationale behind the biological function of any protein molecule is its ability to specifically recognize the ligand. The binding site(s) of protein must, therefore, necessarily hold clues about the function of the protein. Algorithms such as CE (6), DALI (7), VAST (8), SSM (9) and STAMP (10) capture the similarity in overall structure of the protein and also detect remote homologs that can lead to functional inference but fail to detect functional equivalence between proteins of different fold. The local structure-based methods with varying scope such as FFF (11), SPASM (12), Cavbase (13), ASSAM (14), PINTS (15), JESS (16), ef-Site (17), Query3d (18), SATABSEARCH (19), CMASA (20) and ProBis (21) can in some cases capture functional similarities that are missed by global structure comparisons. 3D-Ligandsite (22), I-tasser (23) and firestar (24) are examples of tools that use both sequence and fold-level analyses to identify important residues that could take part in catalysis or ligand binding. To our knowledge, there is no explicit tool to aid in annotation of a protein's function based on identification of binding sites and the ligands that they could bind to. In this study, we report PocketAnnotate web server that integrates three algorithms (25–27) to guide the user through an interactive web interface during each step of the annotation process from choosing the substructure, to a database search and finally detailed alignment, visual inspection and identification of possible ligands.

*To whom correspondence should be addressed. Tel: +91 80 22932892; Fax: +91 80 23600814; Email: nchandra@biochem.iisc.ernet.in

WORKFLOW

PocketAnnotate workflow consists of three different algorithms that are integrated together to form a pipeline. The workflow of PocketAnnotate (Figure 1) is as follows:

(i) The user input protein structure is first subjected to PocketDepth (25), a geometry-based algorithm to detect the putative binding sites in the given protein structure through a depth-based clustering algorithm. The predicted pockets are ranked using various schemes such as polarity, size, surface atom count, depth factor and hydrophobicity. The user can interactively analyse the output produced on the web server through a *jmol* applet and select an appropriate pocket.

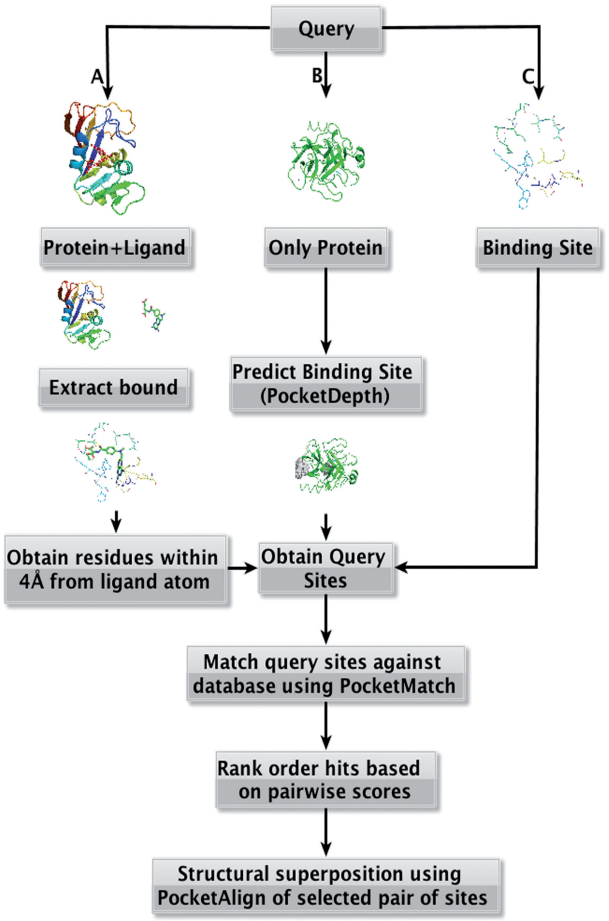

Figure 1. PocketAnnotate workflow. All the steps involved in the PocketAnnotate workflow for three different type of queries has been illustrated. **(A)** If a protein–ligand complex is chosen, the 4 Å region around the ligand is extracted and considered to be binding site. This extracted binding site is then compared against the database using PocketMatch, and the highest scoring similar pockets can be visualized using PocketAlign. **(B)** If Only protein option is chosen, the pockets are predicted for the protein using PocketDepth. The suitable pocket can then be compared with pockets in the database using PocketMatch, and relevant hits can be subjected to PocketAlign to obtain the annotation for the pocket of interest. **(C)** Only substructure of a protein can also be given as input and can be compared against the database to obtain the similarity with other binding sites using PocketMatch and PocketAlig*n*.

(ii) The selected pocket is then analysed by PocketMatch algorithm (26) to extract the distance elements between all atoms and compared against a database of known binding sites. The algorithm reports two scores PMS_{MIN} (local similarity score) and PMS_{MAX} (global similarity score) for all the binding sites compared in the database.

(iii) This step involves selection of a pair of high-scoring pockets by the user to analyse the similarities using PocketAlign (27) that gives a detailed alignment including residue correspondences. PocketAlign produces alignment according to four different schemes, and each of them can be visualized on the web server using *jmol* applet.

DATABASE OF BINDING SITES

The reliability of the hits and time taken for comparison against the database is directly dependent on the quality of the database used. A non-redundant binding site database of biologically relevant ligands was derived out of Protein Data Bank (PDB) (28). The detailed description of various steps followed in creation of database is illustrated in Figure 2. Approximately 49 866 protein–ligand complexes (excluding deoxyribonucleic acid/ribonucleic acid and only X-ray crystal structures as on 24 October 2011) were downloaded from PDB. The set of residues whose atoms lie within 4 Å radius of any ligand atom is considered as the binding site. The binding sites of metal ions, covalently bound ligands and crystallization components were excluded (list provided on the web server). Modified residues were also filtered out, as these would sometimes be represented as HETATM in the PDB. Altogether there were around 311 ligands removed including 67 metal ions and 485 modified residues. A distance cutoff of 2 Å between protein atom and ligand atom was used to identify and exclude covalently bound ligands in the data set through ligand protein contact (LPC) (29).

Non-redundancy of the binding sites is essential to speed up the process of database comparison. To generate a non-redundant set of binding sites, a novel scheme of associating a site with fold information in terms of structural classification of proteins (SCOP) (30) identifiers of its constituent chain(s) has been proposed. The SCOP code consists of four fields, each representing a unique Class, Fold, Superfamily and Family, respectively. All the SCOP identifiers at the zone of binding sites are considered as strings in alphabetical order and concatenated to result in BScID's, a unique structural identifier for the binding site. For example, the BScID for adenosine triphosphate (ATP) bound to PDB:1A49 is b.58.1.1–c.1.12.1, implying that the residues contributing to the binding site comes from two different folds—Pyruvate kinase (PK) β-barrel domain-like (b.58) and Triosephosphate isomerase (TIM) β/α barrel (c.1). In PDB, there are 12 such binding sites of ATP, all with the same BScID. To remove redundancy, only one best site per BScID is chosen by considering the resolution of the

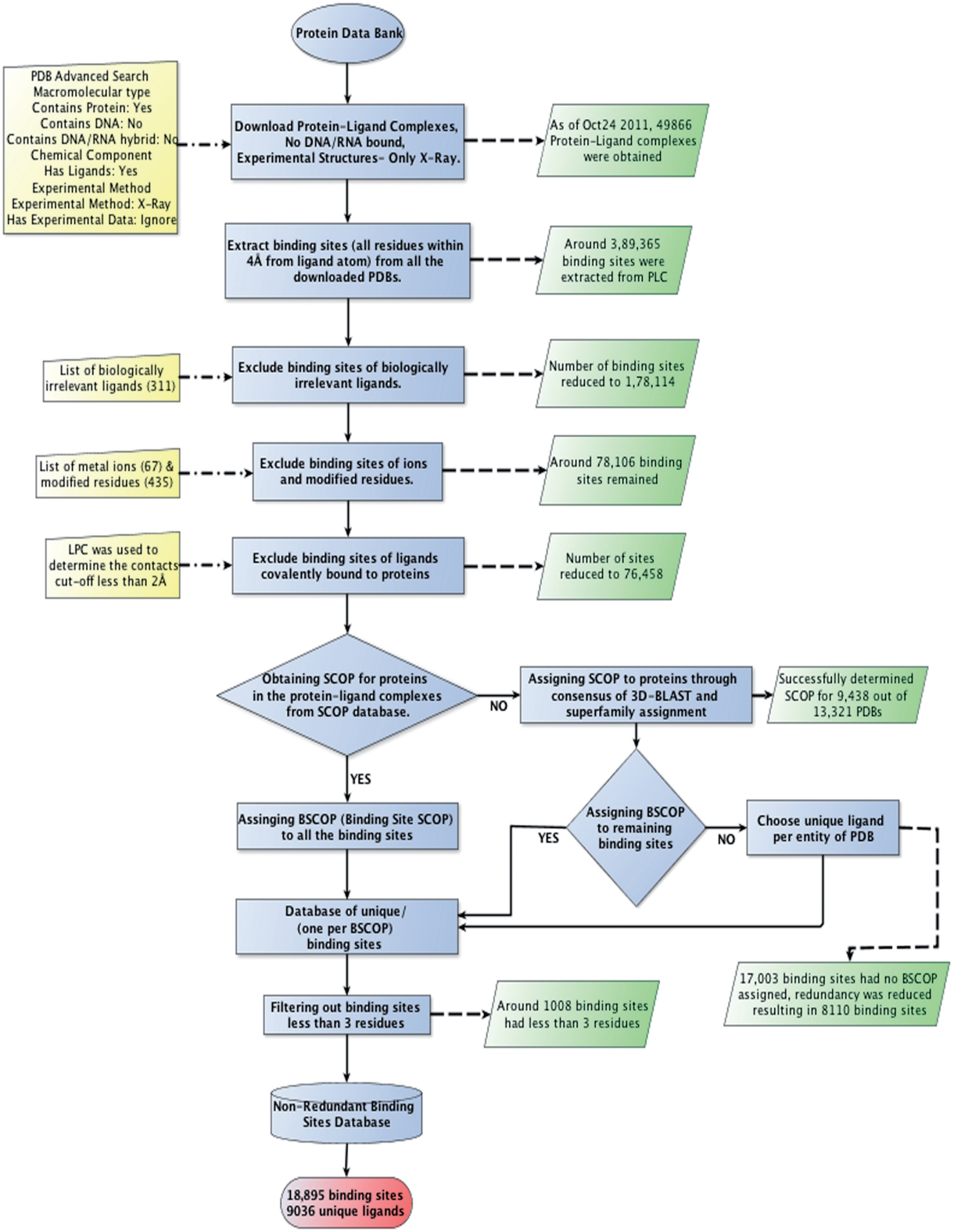

Figure 2. Non-redundant binding site database. All the steps involved in generation of non-redundant binding sites based on SCOP association has been illustrated.

structures. In this case, the ATP binding site of PDB: 1A49 was considered because of its lowest resolution of 2.1 Å, among the 12. In a similar way, for each ligand type (defined by three-letter HETATM codes), one highest resolution site per BScID is chosen. Although there is a possibility of ligand orientation being slightly different even when the fold is the same, we observed that high similarity exists in significant portions of the

corresponding pockets, as judged by a thorough analysis of comparison of all the binding sites belonging to same BScID. It resulted in average PMAX score of 0.59 and PMIN score of 0.63 (http://proline.biochem.iisc.ernet.in/pocketannotate/info/Cluster_similarity.htm). The results of the comparison are available at http://proline.biochem.iisc.ernet.in/pocketannotate/info/cluster_results/. Thus, the method used by us to pick non-redundant sites seems reasonable, even when ligand orientation differs within the same fold. As not all PDB entries of the protein–ligand complex have the SCOP codes associated with it, consensus of superfamily assignment and 3D-BLAST (31) were used to assign SCOP identifiers to the chain. Approximately 9438 structures were assigned SCOP through this method. For the remaining structures that could not be associated with any SCOP, unique binding site per entity (unique chain) in the structure was chosen. Finally, 18 895 binding sites corresponding to 9036 unique ligands were derived to form the non-redundant binding site database.

Analysis of the database indicates that heme (HEM) ranked as the ligand with most number of different binding sites in the database (Supplementary Figure S1), which was followed by other biologically important co-factors such as ADP, ATP, ANP (Phosphoaminophosphonic acid-adenylate ester), Flavin-adenine dinucleotide (FAD) and S-adenosyl-L-homocysteine (SAH). The fold that shows the highest number of unique ligand associations was found to be TIM beta/alpha barrel, which was followed by Protein-kinase like, Acid proteases, Trypsin-like serine protease and Rossmann-fold. The database of binding sites is available for download from the website (Supplementary Figure S2).

WEB SERVER IMPLEMENTATION

The entire web server has been implemented using Hypertext Preprocessor (PHP). The algorithms used, PocketDepth, PocketMatch and PocketAlign, have been coded in 'C' and bash. A typical database comparison for a query-binding site of 23 residues (ADP) takes ~3 min to search the entire database of 18 895 sites using PocketMatch on Intel 2.83 GHz quad-core system running 32 bit Linux vs2 2.6.31-19-generic (Ubuntu). A comprehensible help section is added to help the user understand each of the steps involved and infer the output of the results produced in the workflow. A screen video is also provided as tutorial.

Input

The input required is a protein structure in PDB format. Three types of query can be provided as input to PocketAnnotate. An apo protein structure in PDB format can either be uploaded or the PDB code along with chain ID can be entered. The second type of query is supported by an option of uploading or entering PDB code of protein bound with a ligand. In such a scenario, all the residues within 4 Å radius of any atom of the ligand are extracted as the binding site, which can then be queried against the database, to identify similar binding sites with known ligands. The third type of query type would allow user to submit his/her own sub-structure (any zone of continuous/discontinuous residues in structure) for comparison.

Output

The results produced by all three algorithms are displayed on separate html pages during the pipeline (Supplementary Figure S3), wherein the user can interactively analyse the results. PocketDepth predicts the putative binding sites, which can be visualized on the server through a *jmol* applet. There are various ranking schemes for the predicted clusters/pockets and the user can visualize the pockets through all the ranking schemes and choose the appropriate one. PocketMatch results are displayed in a scroll bar along with the scores in a separate html page, wherein the user can select the high-scoring pockets for PocketAlign. PocketAlign results are displayed in a separate interactive window, with both the selected sites superposed and report the correspondences between the residues in the pair aligned with RMSD scores.

VALIDATION

PocketAnnotate pipeline uses three different algorithms; a brief description of validation protocols that were used for each of the algorithms is listed on the web server in the validation section. PocketAnnotate was validated for three different data sets (i) protein–ligand complexes across different folds, (ii) apo-holo structures and (iii) homologous pairs of ligand bound protein structures.

A set of 537 protein–ligand complexes such that it contained one ligand per fold type was prepared from our database. The correct pocket was detected in 409 of such cases automatically with default parameters when compared with the crystallographically determined pockets in the corresponding structures. In others, a part of the site was often identified, as with the case of most site prediction algorithms. A total of 329 of the predicted sites gave significant hits against its native ligand-binding sites with PMAX > 0.25, PMIN > 0.4 and at least 200 matching distance elements (http://proline.biochem.iisc.ernet.in/pocketannotate/php/Across_different_structures.htm). Similar analysis was carried out on a data set of 124 pairs of apo-holo structures that identified correct hits in 75 cases (http://proline.biochem.iisc.ernet.in/pocketannotate/info/apo_holo.htm) and 6166 pairs of homologous apo-holo protein structures that identified 4400 pairs with correct ligand having <60% identity obtained through BUDDY-system (32) and BindingDB (33) database (http://proline.biochem.iisc.ernet.in/pocketannotate/info/homologous_pairs.htm). With a careful examination of the matches and with altered cutoffs or manual inspection, many of the remaining sites could also be annotated in all the data sets.

STATISTICAL SIGNIFICANCE

P value

The database of binding sites obtained as described earlier was non-redundant in terms of ligand association, but to evaluate statistical significance for PocketMatch scores, we needed a database of non-redundant pockets (irrespective of ligand association). An all versus all comparison of the pockets in our database of 18 895 sites was performed using PocketMatch and later filtered to obtain unique binding sites by neglecting the pairs that gave a perfect score of 1. The distribution of scores (~178 million) thus obtained (both PMIN and PMAX) was carefully analysed. The distribution was continuous and supported on a bounded interval between 0 and 1 as PocketMatch scores are reported with this range. The probability density function (pdf) and cumulative distribution function (cdf) are shown in Figure 3. Cullen and Frey analysis performed on the scores suggested that it follows β distribution (Figure 4a and Supplementary Figure S4). The distribution of both PMIN and PMAX scores could then be represented in the mathematical form $f(x;\alpha,\beta) = \Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta)x^{\alpha-1})(1-x)^{\beta-1}$ -(pdf) and (cdf)-$F(x;\alpha,\beta) = B_x(\alpha,\beta)/B(\alpha,\beta) = I_x(\alpha,\beta)$, where Γ is the gamma function, x is the pocketmatch score, B is the beta function, $I_x(\alpha, \beta)$ is regularized incomplete beta function for α and β, which are the shape parameters that should be estimated from the distribution. Matching moment method was used to estimate the shape parameter, and the goodness of fit is shown in Figure 4b. The goodness of fit was also evaluated by Kolmogorov-Smirnov test that converged with the statistical nearness value of 0.017, resulting in almost the same value for shape parameters. In this scenario, our null hypothesis would state that a score reported is from random match (from the unique binding site database match), and the P value can be evaluated as the area under the curve (AUC) given by $1-\text{cdf}$. A $P > 0.01$ would mean acceptance of the null hypothesis of random match. The score of 0.4 has been reported to

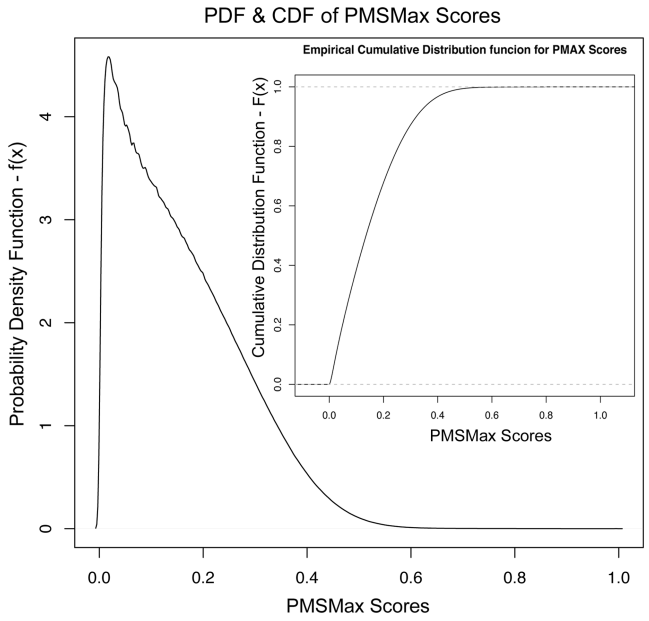

Figure 3. Distribution of PMS_{MAX} scores. Probability density function of the PMS_{MAX} distribution of all versus all pairwise comparisons in the data set. Starting from a frequency histogram of PMS scores in data set comparison, the probability density graph is obtained by using a kernel density estimation method with smoothing parameter (bandwidth) of 0.002311. The inset shows the cumulative distribution for the same data. (Refer text for mathematical description).

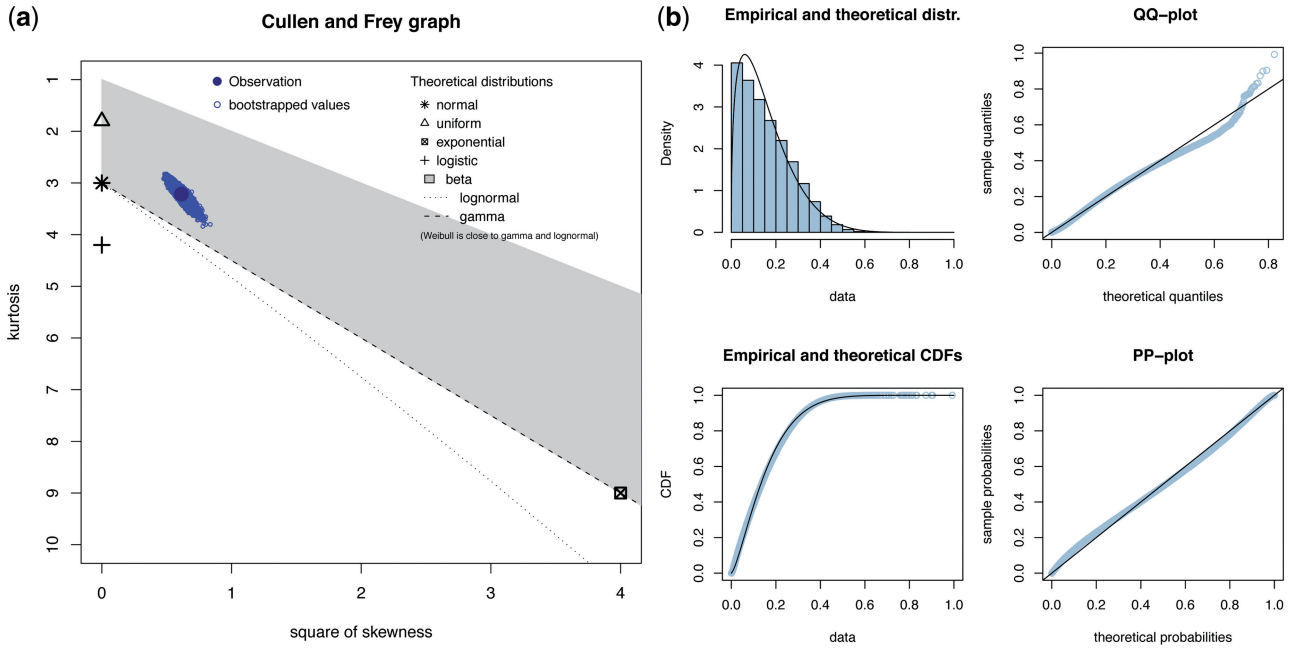

Figure 4. Statistical analysis and distribution fitting of all versus all PMS_{MAX} scores from pocket comparisons in the data set. (**a**) Cullen and Frey analysis suggesting that data points observed follows beta distribution. [**b**(i–iv)] Goodness of fit to the estimated beta distribution.

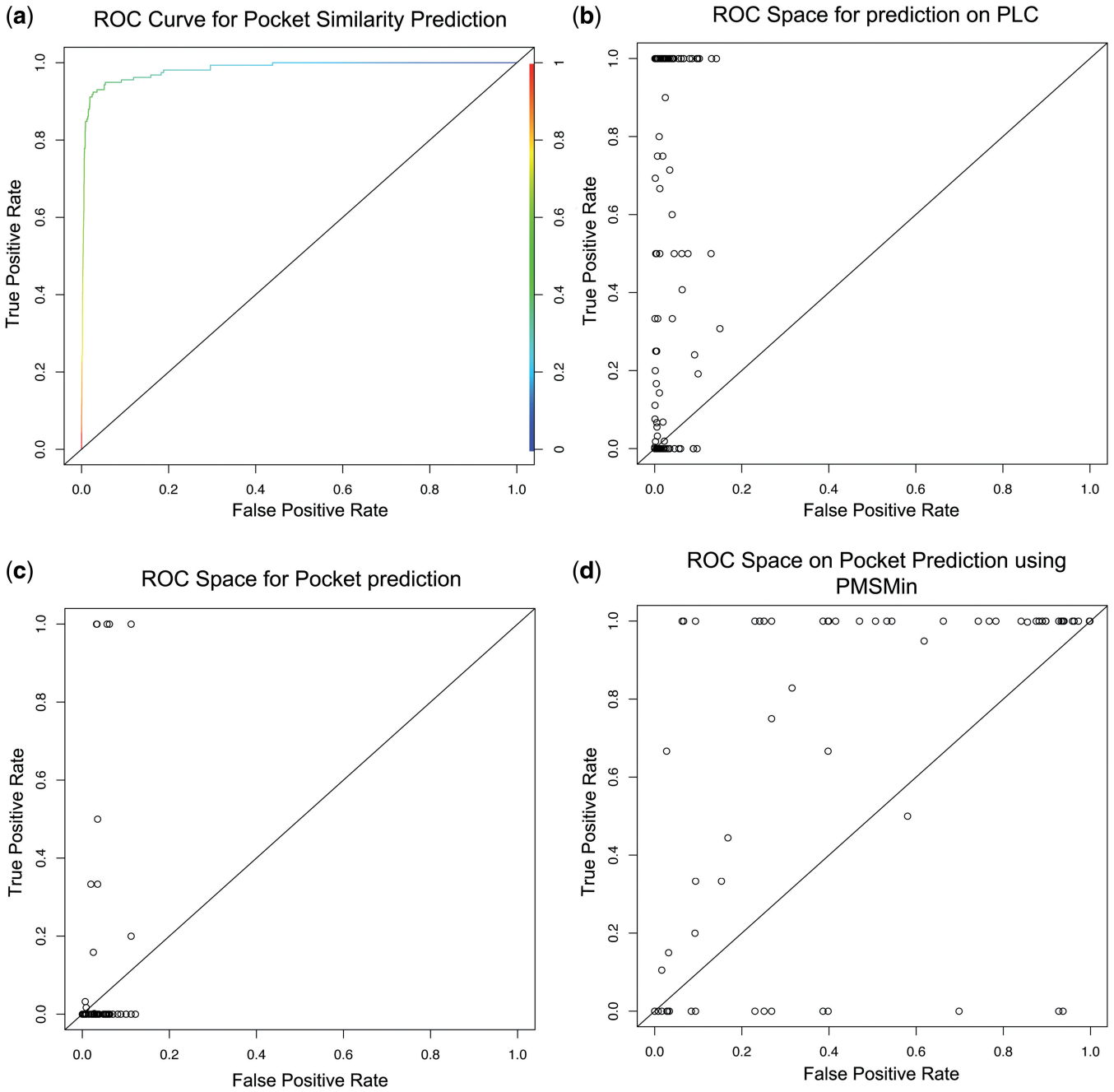

Figure 5. Sensitivity and specificity analyses. (**a**) ROC curve obtained for pocket similarity prediction of sites extracted from 98 protein–ligand complexes (holo form). (**b**) ROC space of each of prediction made for the hit against the database by considering each binding site of protein–ligand complex (holo form). (**c**) ROC space for the pockets predicted by the PocketDepth on corresponding set of apo proteins. (**d**) ROC space for the pockets by considering PMS_{MIN} scores for the PocketDepth predictions on apo-form of proteins.

be significant earlier by the analysis performed on pockets obtained from same SCOP fold that corresponds to statistical P value of 4×10^{-2}. Thus, a $P < 4 \times 10^{-2}$ obtained for a comparison of binding sites can be considered to be significant.

Sensitivity and specificity

The data set of protein–ligand complexes (98 complexes) that had representative binding sites in our database was obtained from the apo-holo data set used earlier for the validation in the same study. An all versus all comparison of the binding sites was performed using PocketMatch. This data set included diverse (completely different binding sites) and analogues or inhibitors (same binding site) of the ligands, and hence, a classifier (0 for different site and 1 for same site) was manually generated to test the specificity and sensitivity of the PMS_{MAX} scores obtained (http://proline.biochem.iisc.ernet.in/pocketannotate/info/classifier.htm). The ROC curve (Figure 5a) was obtained

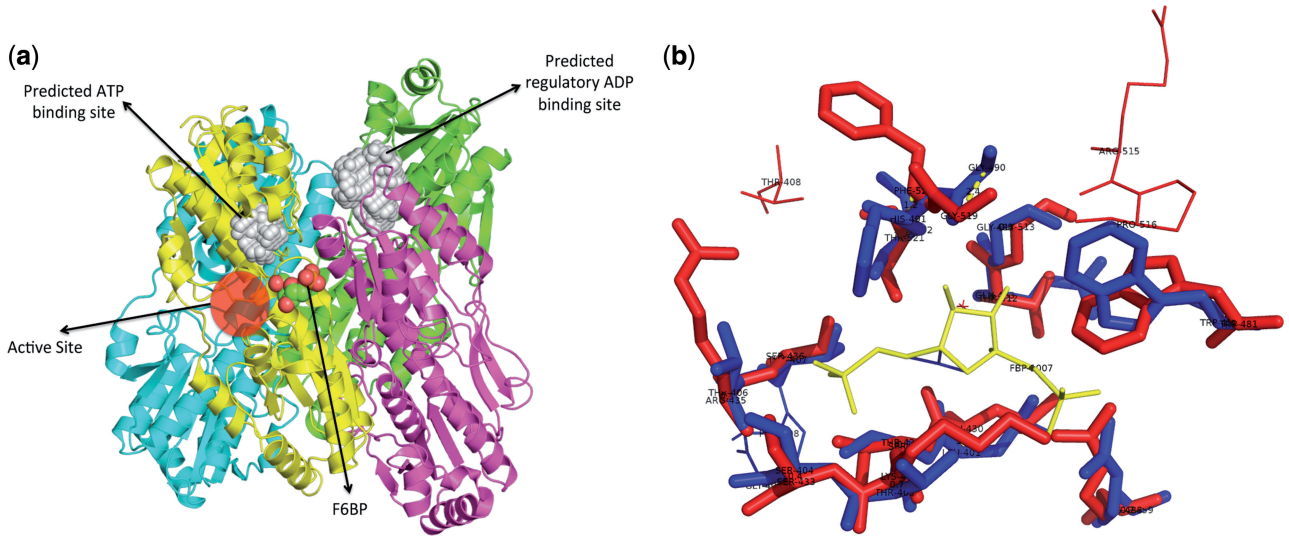

Figure 6. Examples for exploring protein–ligand interaction space. (**a**) Predicted binding sites in phosphofructokinase. Both the active site of the enzyme in its monomeric subunit along with regulatory binding site of ADP at the interface of subunits has been rightly picked up. (**b**) Superposition of the residues in the predicted binding site of PKM2 (pyruvate kinase) with the topmost hit of fructose-biphosphate (FBP) binding site. FBP is known to be an allosteric activator of PKM.

(AUC = 0.983) with maximum threshold at 0.4. Accuracy calculated at this threshold resulted in sensitivity of 93% and specificity of 95%. Each of the ligand-binding sites obtained were compared with the binding sites in database, and true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) were analysed for each prediction to understand the receiver operating characteristic (ROC) space occupied by them. Obtaining the right definitions of these terms was a non-trivial task. An assumption of similar ligands binding to similar pockets was made for the sake of simplicity. Tanimoto scores (Lig score) obtained between the ligands through FP2 fingerprint was used to assess the ligand similarity. A comparison was considered to be TP if the $PMS_{MAX} > 0.4$ and Lig score > 0.9, TN if $PMS_{MAX} < 0.4$ and Lig score < 0.9, FP if $PMS_{MAX} > 0.4$ and Lig score < 0.9 and FN if $PMS_{MAX} < 0.4$ and Lig score > 0.9. Although there is a possibility of same ligand binding to dissimilar sites and different ligands binding to same pocket, however rare, only FP's and FN's would change under this assumption, leading to underestimation of performance rather than an overestimation. The ROC space indicates good predictions for sites obtained directly from protein–ligand complexes (Figure 5b) but not for predicted pockets from corresponding apo-forms of the protein (Figure 5c). This would be expected because of the boundaries of the pockets not being defined exactly, which is a general limitation with any binding-site prediction tool. However, the ROC space improves significantly when PMS_{MIN} scores are considered, because, in most cases the actual pocket is a subset of the predicted pocket (Figure 5d). We have, therefore, included an additional ranking scheme based on PMS_{MIN} scores along with PMS_{MAX}.

The 'fitdistrplus' and 'ROCR' packages under R statistical programming language was used to carry out all the analysis mentioned earlier.

EXAMPLES

PocketAnnotate pipeline can prove to be extremely useful in cases where one has to explore alternate binding sites on the protein surface. As an example, phosphofructokinase (PFK) enzyme was explored to probe into regulatory sites along with the catalytic site using PocketAnnotate pipeline. PFK is a major enzyme that takes part in glycolysis by converting fructose-6-phosphate to fructose-1, 6-biphosphate. PFK is known to be regulated by many small molecules such as ATP, ADP, AMP, citrate and fructose-6-bisphosphate. Prokaryotic PFK (PDB: 1MTO), known to be a homotetramer in its biologically functional form, was subjected to PocketDepth analysis (only protein query), as the regulatory sites could also be found at the interfaces. The pipeline correctly predicted all the active sites of the individual monomers and the regulatory sites through which ADP modulates the function at the interface of the subunits also. One such predicted regulatory site known to bind to ADP along with active site was correctly identified is shown in Figure 6(a). Pyruvate kinase (PDB:1ZJH, unbound form) was also analysed for the presence of any allosteric sites, and the pipeline identified the right allosteric site modulated by fructose-bi-phosphate (Figure 6b). The details of both the examples can be obtained from the example section of the web server.

An attempt was also made to analyse the protein structure with unknown function. An advance text word search of 'unknown function' in PDB yielded ~3333 (as of November 2011) structures. A filter of 70% identity cutoff was used, and only the unique entities from each PDB resulting in 2155 chains were subjected to PocketAnnotate workflow till PocketMatch stage. Approximately 1837 proteins were found to have significant hits with the cutoff mentioned earlier (http://proline.biochem.iisc.ernet.in/pocketannotate/info/unknown_

function.htm). Selected examples are included in the example section of the web server. The results of this analysis are also available on the website. Few of the positive hits from each of the data set mentioned earlier and validation are made available for visualization in the example section of web server.

CONCLUSIONS

In this study, we present a pipeline for site-based function annotation of a protein structure through in-house algorithms. We believe that such a pipeline would allow us to explore protein–ligand interaction space. Such binding site-based structure annotation can also be used for genome scale structure annotations to gain useful information (34). With general advances in site prediction and scoring methods, as well as an increase in the experimental protein–ligand structures, we expect that this approach will become more accurate and grow in its scope for functional annotation of proteins.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–4.

ACKNOWLEDGEMENTS

The authors thank Sumanta Mukherjee for his help and useful suggestions on setup, maintenance and improvement of the server. They thank Raghu Bhagavat for his efforts on setting up the binding site database and Deepesh Nagarajan for his help on faster implementation of all versus all comparisons of pockets. They appreciate all the members of the laboratory for their valuable suggestions, feedbacks and inputs on the development of the web server. The computational facilities at Supercomputer Education Research Centre (SERC) are greatly acknowledged.

FUNDING

Department of Biotechnology (DBT); Government of India. Funding for open access charge: DBT.

Conflict of interest statement. None declared.

REFERENCES

1. Chandonia,J.M. and Brenner,S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.
2. Jaroszewski,L., Li,Z., Krishna,S.S., Bakolitsa,C., Wooley,J., Deacon,A.M., Wilson,I.A. and Godzik,A. (2009) Exploration of uncharted regions of the protein universe. *PLoS Biol.*, **7**, e1000205.
3. Redfern,O.C., Dessailly,B. and Orengo,C.A. (2008) Exploring the structure and function paradigm. *Curr. Opin. Struct. Biol.*, **18**, 394–402.
4. Kosloff,M. and Kolodny,R. (2008) Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins*, **71**, 891–902.
5. Gan,H.H., Perlow,R.A., Roy,S., Ko,J., Wu,M., Huang,J., Yan,S., Nicoletta,A., Vafai,J., Sun,D. *et al.* (2002) Analysis of protein sequence/structure similarity relationships. *Biophys. J.*, **83**, 2781–2791.
6. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein. Eng.*, **11**, 739–747.
7. Holm,L. and Rosenstrom,P. (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res.*, **38**, W545–W549.
8. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
9. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta. Crystallogr. D Biol. Crystallogr.*, **60**, 2256–2268.
10. Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.
11. Fetrow,J.S. and Skolnick,J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.*, **281**, 949–968.
12. Kleywegt,G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, **285**, 1887–1897.
13. Schmitt,S., Kuhn,D. and Klebe,G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, **323**, 387–406.
14. Spriggs,R.V., Artymiuk,P.J. and Willett,P. (2003) Searching for patterns of amino acids in 3D protein structures. *J. Chem. Inf. Comput. Sci.*, **43**, 412–421.
15. Stark,A. and Russell,R.B. (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Res.*, **31**, 3341–3344.
16. Barker,J.A. and Thornton,J.M. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, **19**, 1644–1649.
17. Kinoshita,K. and Nakamura,H. (2003) Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein. Sci.*, **12**, 1589–1595.
18. Ausiello,G., Via,A. and Helmer-Citterich,M. (2005) Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinformatics*, **6(Suppl. 4)**, S5.
19. Stivala,A.D., Stuckey,P.J. and Wirth,A.I. (2010) Fast and accurate protein substructure searching with simulated annealing and GPUs. *BMC Bioinformatics*, **11**, 446.
20. Li,G.H. and Huang,J.F. (2010) CMASA: an accurate algorithm for detecting local protein structural similarity and its application to enzyme catalytic site annotation. *BMC Bioinformatics*, **11**, 439.
21. Konc,J. and Janezic,D. (2010) ProBiS: a web server for detection of structurally similar protein binding sites. *Nucleic Acids Res.*, **38**, W436–W440.
22. Wass,M.N., Kelley,L.A. and Sternberg,M.J. (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38**, W469–W473.
23. Roy,A., Kucukural,A. and Zhang,Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.*, **5**, 725–738.
24. Lopez,G., Maietta,P., Rodriguez,J.M., Valencia,A. and Tress,M.L. (2011) firestar–advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39**, W235–W241.
25. Kalidas,Y. and Chandra,N. (2008) PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. *J. Struct. Biol.*, **161**, 31–42.
26. Yeturu,K. and Chandra,N. (2008) PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinformatics*, **9**, 543.
27. Yeturu,K. and Chandra,N. (2011) PocketAlign a novel algorithm for aligning binding sites in protein structures. *J. Chem. Inf. Model.*, **51**, 1725–1736.
28. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
29. Sobolev,V., Eyal,E., Gerzon,S., Potapov,V., Babor,M., Prilusky,J. and Edelman,M. (2005) SPACE: a suite of tools for protein

structure prediction and analysis based on complementarity and environment. *Nucleic Acids Res.*, **33**, W39–W43.
30. Lo Conte,L., Ailey,B., Hubbard,T.J., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.
31. Yang,J.M. and Tung,C.H. (2006) Protein structure database search and evolutionary classification. *Nucleic Acids Res.*, **34**, 3646–3659.
32. Morita,M., Terada,T., Nakamura,S. and Shimizu,K. (2011) BUDDY-system: a web site for constructing a dataset of protein pairs between ligand-bound and unbound states. *BMC Res. Notes*, **4**, 143.
33. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
34. Anand,P., Sankaran,S., Mukherjee,S., Yeturu,K., Laskowski,R., Bhardwaj,A., Bhagavat,R., Consortium,O., Brahmachari,S.K. and Chandra,N. (2011) Structural annotation of Mycobacterium tuberculosis proteome. *PLoS One*, **6**, e27044.