Published online 8 May 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W471–W477***
doi:10.1093/nar/gks372

COFACTOR: an accurate comparative algorithm for structure-based protein function annotation

Ambrish Roy, Jianyi Yang and Yang Zhang*

Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA

Received January 30, 2012; Revised March 31, 2012; Accepted April 12, 2012

ABSTRACT

We have developed a new COFACTOR webserver for automated structure-based protein function annotation. Starting from a structural model, given by either experimental determination or computational modeling, COFACTOR first identifies template proteins of similar folds and functional sites by threading the target structure through three representative template libraries that have known protein–ligand binding interactions, Enzyme Commission number or Gene Ontology terms. The biological function insights in these three aspects are then deduced from the functional templates, the confidence of which is evaluated by a scoring function that combines both global and local structural similarities. The algorithm has been extensively benchmarked by large-scale benchmarking tests and demonstrated significant advantages compared to traditional sequence-based methods. In the recent community-wide CASP9 experiment, COFACTOR was ranked as the best method for protein–ligand binding site predictions. The COFACTOR sever and the template libraries are freely available at http://zhanglab.ccmb.med.umich.edu/COFACTOR.

INTRODUCTION

The biological function of a protein molecule is decided by its 3D-shape, which eventually determines how the molecule interacts with other molecules in living cells. As such, considerable efforts have been made to determine the structure of the protein molecules and to deduce the biological functions based on their 3D-shape (1–3). One of the most common structure-based approaches in protein function annotation is to detect homologous template proteins by global structure comparisons and then transfer known functional annotations from the templates (2,4,5). However, the evidence of global structural similarity is usually insufficient for accurate functional inference, as proteins possessing similar global fold can perform different biological functions. The classic examples include the proteins with α-/β-barrel fold, which is inhabited by both enzymatic and non-enzymatic proteins (6). Accordingly, many contemporary approaches have been designed to identify local structural similarity of functionally important residues for drawing functional inferences (7,8). However, the functional annotation based on local structure alone can result in high false-positive rate, especially when the target protein has a low sequence identity to the template proteins or the target structure on its own has a low-resolution 3D structure (3,9).

In this study, we describe a newly developed COFACTOR server, which combines both global and local structural comparison algorithms to deduce the biological functions of proteins, starting from their 3D structure. The output of the server includes function annotations in three key aspects: protein–ligand binding interactions, Enzyme Commission (EC) (10) and Gene Ontology (GO) (11). Keeping in mind that high-resolution experimental structures are unavailable for most of the protein targets in genome databases, the algorithm has been extensively trained for low-resolution structures generated from computational structure predictions. Meanwhile, experimental structures undoubtedly meet the highest structural requirement and the prediction accuracy improves using these structures. In both large-scale benchmark (12) and blind experiments (2), the COFACTOR method has demonstrated significant advantages over other state-of-the-art sequence- or structure-based comparative methods.

MATERIALS AND METHODS

COFACTOR algorithm

The input to the COFACTOR server is the 3D-structure of a target protein, which can be obtained from

*To whom correspondence should be addressed. Tel: +1 734 647 1549; Fax: +1 734 615 6553; Email: zhng@umich.edu

either structure prediction or experimental determination. Figure 1 shows a general overview of the procedure followed on the COFACTOR server and the analysis done using the server, which includes detection of structural analogs in the PDB library and prediction of three different aspects of protein function, namely, EC numbers, GO terms and ligand binding sites. The structure-based function inferences are made in two steps, i.e. global structural alignment followed by local structural similarity search.

Global structural similarity identification

COFACTOR first identifies the template proteins of similar fold/topology by matching the query structure with all proteins in three newly developed representative functional libraries, which have known protein–ligand binding information, EC numbers and GO terms (J. Yang, A. Roy and Y. Zhang, submitted for publication). The global structure match is conducted by TM-align (13), a heuristic algorithm for global protein structure alignment, which starts from multiple seed alignments (gapless threading, secondary structure match and the combination of the two), followed by Needleman–Wunsch dynamic programming refinement (14). The objective function of the TM-align searching is TM-score (15):

$$\mathrm{TM-score} = \max\left[\frac{1}{L}\sum_{i=1}^{L_{\mathrm{ali}}}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \quad (1)$$

where d_i is the distance between ith pair of C_α atoms of query and template and L_{ali} is the number of aligned residue pairs identified by TM-align. d_0 is given by $d_0 = 1.24\sqrt[3]{L-15}-1.8$ and L is the length of the query protein. Since TM-score weights the short-distance residue pairs stronger than the long-distance ones, it is more sensitive to the global topology of proteins than the traditional structural similarity measurement RMSD. Meanwhile, because only the aligned residues are calculated in the summation which is normalized by the target length, TM-score in Equation (1) counts for both alignment accuracy and the alignment coverage in a single parameter. Generally, a protein pair with TM-score >0.5

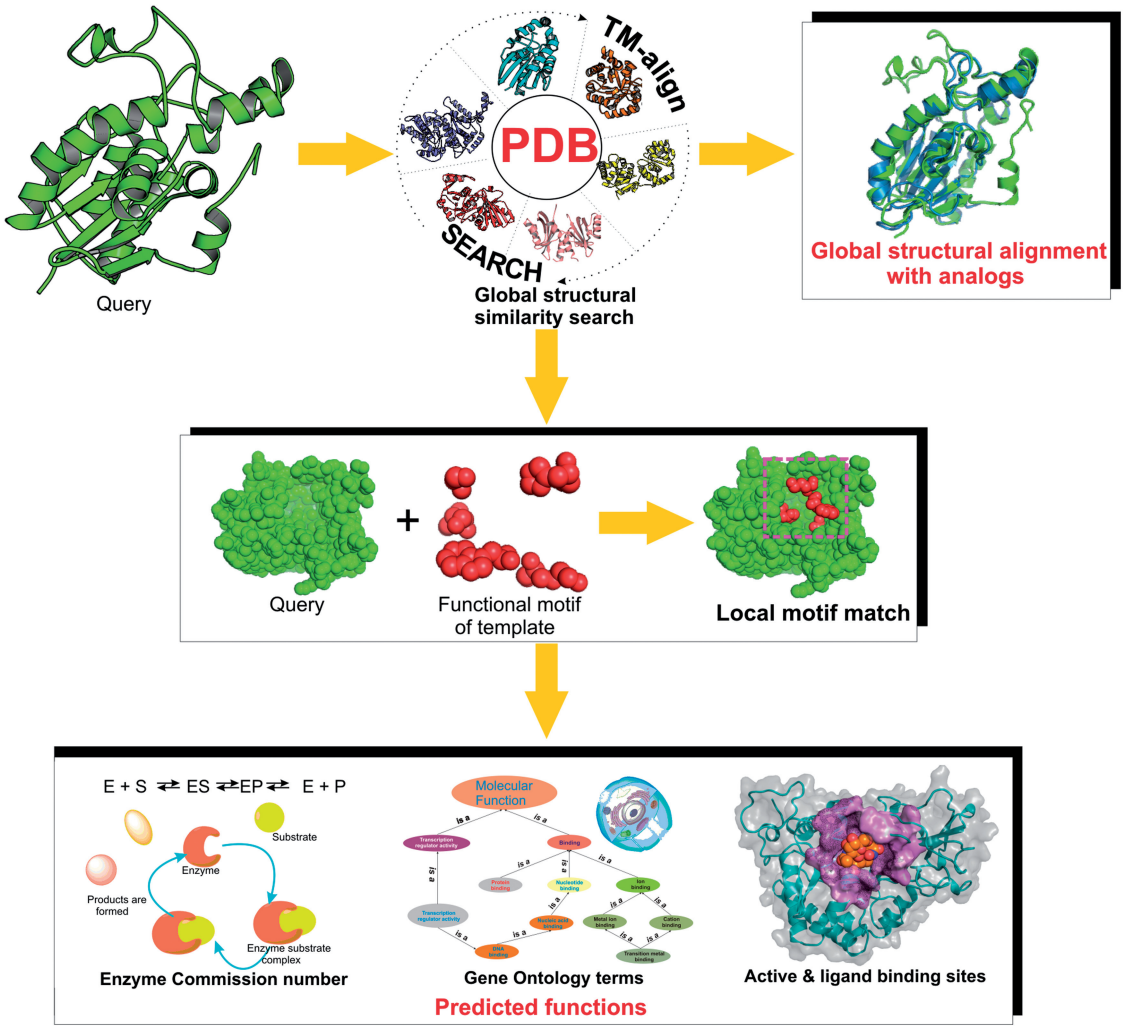


Figure 1. Illustration of structure-based function annotation by the COFACTOR server, starting from the query structure (shown in green).

indicates that they have the same fold while that with TM-score <0.3 have random structural similarity (16).

The global structural alignment between the query and template structures is useful for exploring fold/family relationships of newly solved structure or predicted structural models. However, some folds are functionally diverse and in these cases, function can be accurately predicted only by evaluating the similarity of active/binding site residues that are involved in the function. Moreover, in many cases, the functional motifs remain conserved during the evolution to maintain the function, even when the global structural similarity dwindles. Thus, a local sequence and structural comparion of functional sites may provide a more reliable way of functional annotation for the query proteins.

Accordingly, on COFACTOR server, all template proteins with a non-random structural similarity (i.e. TM-score > 0.3) (16) to the query structure (or up to 100 top templates regardless of TM-score are used if <100 non-random templates are identified) in each of the three function libraries (see below) are screened further based on their local similarity to query structure.

Local functional site identification

In the second step, a heuristic algorithm has been developed to identify the best local functional site match between the query and template structures. In Figure 2, a multiple sequence alignment is first constructed and evolutionarily conserved residues in the query sequence are identified based on their Jensen–Shannon divergence (JSD) score (17). The conformations of various triplet residues from the conserved residue pool are excised from the query structure to construct a set of local 3D-structural motifs. Each of the local query motifs is then superimposed onto the known functional site residues of the template protein.

To further refine the local structural match of the functional sites, the complete structure of the query and template proteins are brought together in the same reference frame, based on the rotation and translation matrices acquired from the initial motif superposition. A sphere of radius r is then defined around the geometric center of template motif, where r is the maximum distance of any template functional site residue from the geometric center. The residues from query and template proteins within the sphere are re-aligned by an iterative alignment procedure similar to TM-align (13), i.e. scoring matrix is repeatedly calculated from the current structural superposition and is used to generate new optimized superposition by dynamic programming, until converged. The sphere thus represents a pseudo-functional site, under which the local structural and sequence similarity (L_{sim}) between query and template proteins is evaluated by

$$L_{\mathrm{sim}} = \frac{1}{N_t}\sum_{i=1}^{i=N_{\mathrm{ali}}}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2} + \frac{1}{N_t}\sum_{i=1}^{i=N_{\mathrm{ali}}} M_{ii}, \qquad (2)$$

where N_t represents the total number of residues present within the template sphere, N_{ali} is the number of query-template aligned residue pairs within the sphere, d_i is the C_α distance between ith aligned residue pair, M_{ii} is the normalized BLOSUM62 substitution scores between ith pair of residues and d_0 is the distance cutoff chosen to be 3.0 Å. The second term in Equation (2) is to account for the evolutionary information of the functional sites. For each binding pocket on the template, this procedure is implemented for all the conserved query motifs and the one with the highest L_{sim} is recorded (Figure 2).

Functional analyses

The COFACTOR server provides a variety of available annotations for the query protein using the templates, including EC number, GO and protein–ligand binding sites. We provide a brief overview of the three aspects of predicted functions by COFACTOR server below.

Enzyme Commission number

For the purpose of classifying enzymatic proteins, all enzyme protein structures with annotated EC number(s) have been collected from the PDB library (18) with the active site residue information mapped using Catalytic Site Atlas (19). As of January 2012, this compiled enzyme template library contains 8392 protein structures.

The active site motifs of the template structures are important for the local structural comparison and the query active site identification. For the template structures where the active site residues are known, the template motifs are defined by these annotated functional sites (19). Otherwise, the algorithm uses spatially clustered and evolutionarily conserved residues for generating the template motifs (A. Roy, S. Mukherjee, P. S. Hefty and Y. Zhang, submitted for publication). For the former cases, residue correspondences from the local alignment results are mapped onto the query structure, which are used for predicting catalytic residues in the query; while in the latter, only predicted EC numbers are reported.

The confidence score for EC number prediction reflects both local and global similarities between query and template proteins and is defined as:

$$\mathrm{C-score_{EC}} = \frac{2}{1+e^{-(0.25L_{\mathrm{sim}}\mathrm{SS_{BS}}+\mathrm{TM-score}+2.5\mathrm{ID_{Str}})}} - 1, \qquad (3)$$

where L_{sim} defined in Equation (2) and TM-score in Equation (1) measure local and global similarity between query and template enzyme, respectively, IDstr is identity between query and template in structurally aligned region of TM-align alignment, and SS_{BS} is sequence similarity between predicted active site residues of query and known active site residues of template. Finally, the top five scoring hits are reported.

Gene ontology terms

The GO is a widely used machine-legible approach for automatic functional annotation. To this end, a second library of protein structures that have known GO terms was created using PDB-GO mapping taken from the Gene Ontology Annotation database (http://www.ebi.ac.uk/GOA/) and SIFTS project (http://www.ebi.ac.uk/pdbe/docs/sifts/). This library contains 24 035 non-redundant

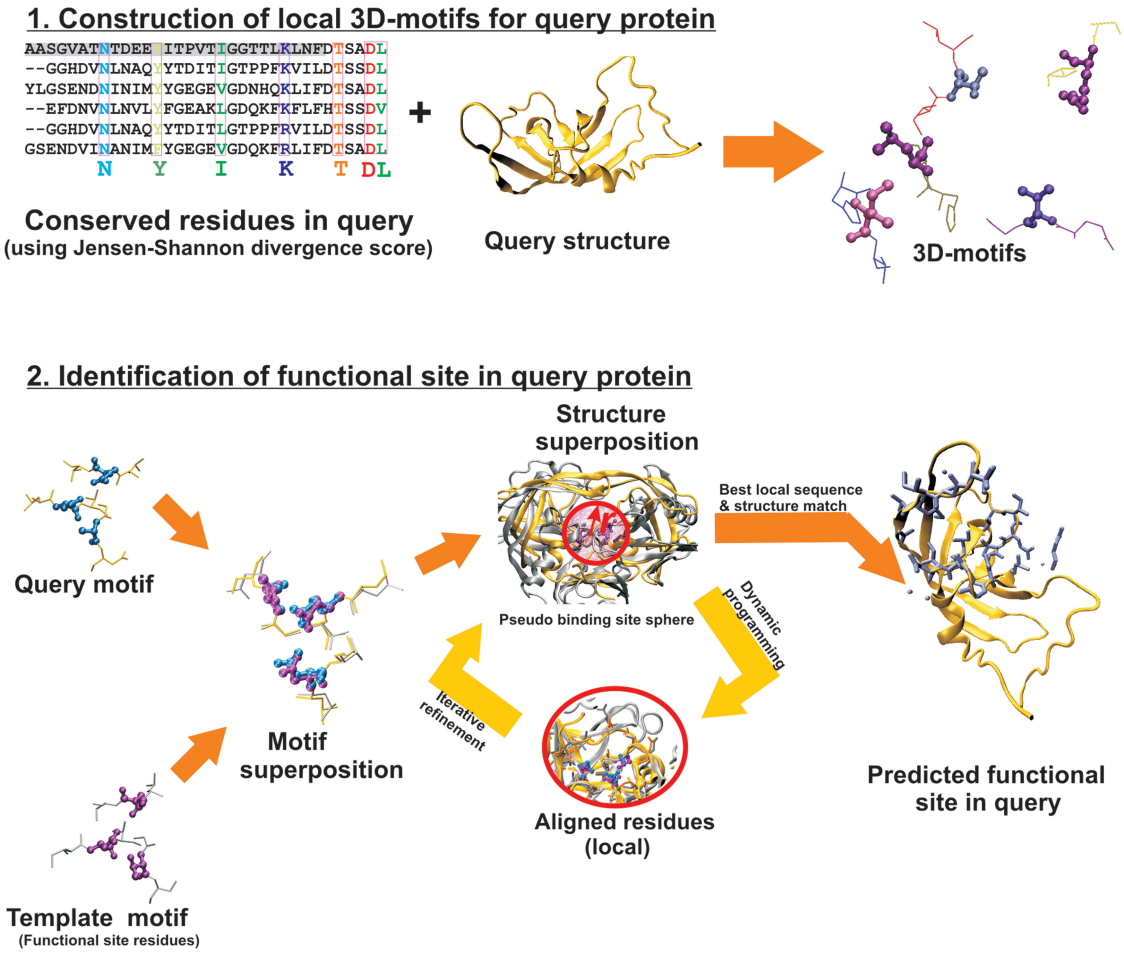


Figure 2. Flowchart of functional site identification by the COFACTOR server. (i) Conserved residues in query sequence are identified based on Jensen–Shannon diverge score, which are then used to glean local 3D-fragments from the query structure. (ii) Each local 3D-motif of query is aligned with the fragments collected from functional site of template protein and the local similarity between query and template protein is evaluated using L_{sim} Equation (2). Finally, the best match among all the probable sets with the best local match (i.e. highest L_{sim}) is selected. The residues of query protein (yellow) are shown in cyan, while those in template protein (gray) are shown in magenta.

protein chains, associated with 13 757 unique GO terms, as of January, 2012.

The procedure of identifying and scoring the identified homologs in the GO template library is similar to that used for EC number prediction, however the template motifs for the local structural comparisons are generated using both known active and ligand-binding site residues rather than active residues only. Furthermore, based on the assumption that each protein domain contributes independently to the protein function, the GO terms ascribed to the top five ranking hits are reconciled based on the PIPA algorithm (20), so that the consensus predictions identifies the intersection of functions among the top hits and provides specific annotation to the query protein.

Protein–ligand binding sites

Ligand binding pockets and ligand-interacting residues in the query protein are identified based on both global and local structural similarities to a comprehensive binding site template library, which contains 76 679 binding sites, including information on protein–protein, protein–nucleic acid, protein–lipid and protein–small molecule interactions.

The binding pose of the template ligands in the query structure is predicted based on the superposition matrix acquired from the local alignment of query and template binding site residues. A quick rigid body Metropolis Monte Carlo simulation of the superposed ligand is followed to improve the local geometry, where the energy term to guide the simulation is defined as the sum of the number of contacts made by template ligand with the predicted binding site residues, the reciprocal of the number of ligand–protein clashes, and the contact distance error which is calculated as difference between inter-atomic ligand–protein contact distance in template and that in query model. Here, contacts are those interactions that are within a distance of 0.5 Å plus the sum of the van der Waals radius of protein atom and ligand atom, while clashes are those in which the inter-atomic distance is less than sum of their van der Walls radii. The side chains of ligand binding residues are further optimized using Scwrl4 (21).

Finally, the predicted ligand conformations from all templates are clustered based on the spatial proximity with a distance cutoff 8 Å. If a binding pocket binds multiple ligands (e.g. an ATP-binding pocket may also bind MG, PO_4^{3-} and ADP), ligands within the same pocket are clustered further based on their chemical similarity (Tanimoto coefficient cutoff = 0.7) using the average linkage clustering procedure to rank the predicted binding sites.

From each cluster, the protein–ligand complex with highest ligand-binding confidence score (C-score$_{LB}$) is eventually selected as the functional site predictions for the query protein. C-score$_{LB}$ is defined as:

$$\text{C}-\text{score}_{\text{LB}} = \frac{2}{1+e^{-\left(\frac{N}{N_{\text{tot}}}\times\left(0.25\text{Lsim}+\text{TM}-\text{score}+2.5\text{ID}_{\text{Str}}+\frac{2}{1+\langle D\rangle}\right)\right)}} - 1, \quad (4)$$

where N is the number of template ligands in a cluster and N_{tot} is the total number of predicted ligands using the templates. L_{sim} defined in Equation (2) and TM-score defined in Equation (1), measuring local and global similarity of the query to the template protein, respectively. ID_{str} is sequence identity between the query and the template in the structurally aligned region. $\langle D\rangle$ is the average distance of the predicted ligand to all other predicted ligands in the same cluster.

OUTPUT

For each submitted protein, the user will be notified by email when the job is completed and the result data are reported on the COFACTOR homepage. Each of the COFACTOR result page consists of four main tables (see, e.g., http://zhanglab.ccmb.med.umich.edu/COFACTOR/example/).

In the first table, structural alignments of the query with the top 10 template proteins ranked by TM-score, identified from the PDB library, are displayed using an interactive Jmol applet (22,23). The table provides details of the structural alignment as generated by TM-align (13), including TM-score, alignment coverage (fraction of residues aligned in the query), RMSD and the sequence identity in the structurally aligned region. Each of the structural alignments can be viewed interactively in the Jmol applet by clicking the corresponding radio buttons. The links for downloading the coordinate files of superposed structures are provided in the same table.

The second table presents the top five enzyme templates ranked by confidence scores and the predicted catalytic residues in the query. These predicted catalytic residues are visually displayed using the Jmol applet in the same table.

The third table lists top scoring template proteins that are annotated with GO terms. Usually, each template protein is associated with multiple GO terms that describe different aspects of biological and cellular functions. As the template proteins have additional functional domains, rather than simply transferring GO annotation, the server presents the most frequently occurring GO terms in each of the three functional aspects (molecular function, biological process and cellular component), which are reconciled from the top five homologs. A mouse hover over each GO term provides its definition.

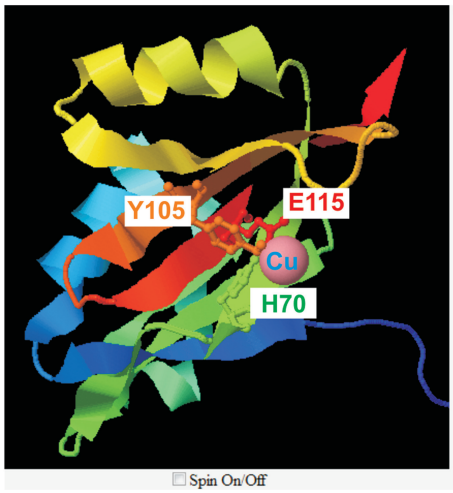
The last table contains information on protein–ligand binding location in the query structure. Top 10 predictions are presented with the information on the template protein, the template ligand and the query residues which are likely to be involved in binding interactions. These predictions and interactions are visualized using the Jmol applet, where the ligand atoms are shown as spheres and binding site residues in query are highlighted using ball and stick (Figure 3).

PERFORMANCE OF WEB SERVER

The COFACTOR algorithm has been extensively trained and tested on large-scale benchmarks. In a recent study (12), COFACTOR was tested on 501 proteins, which harbor 582 natural and drug-like ligand molecules. Starting from the low-resolution structural models generated by I-TASSER (24), the method successfully identifies ligand-binding pocket locations for 65% of apo receptors with an average distance error 2 Å. The average precision of binding-residue assignments is 46 and 137% higher than that by FINDSITE (4) and ConCavity (25), which were designed to identify protein–ligand binding sites.

In the recent community-wide CASP9 experiment where all predictions were made before the experimental results were released (2), COFACTOR achieved a binding-site prediction precision 72% and Matthews correlation coefficient 0.69 for the 31 blind test proteins, which was significantly higher than all other participating methods. As CASP9 assessors concluded, among all 33 participant groups 'Two groups (FN096, Zhang; FN339, I-TASSER_FUNCTION) performed better than the rest, while the following 10 prediction groups performed comparably well' (2).

To examine the ability of this approach to predict two other unambiguously defined concepts of functions: EC numbers (10) and GO (11) terms, especially with new settings taken by the COFACTOR server, we tested the server approach on a large benchmark set of 450 non-homologous proteins collected from PDB. As experimental controls, we select those commonly used approaches that are based on sequence–profile alignment (26), profile–profile alignment (27) and HMM–HMM alignment (28). In all experiments, close homologs of query proteins were intentionally removed from the template libraries using a sequence identity cut-off 30%, before the predictions were made. Supplementary Figure S1 summarizes the performance of COFACTOR to identify the correct function and the improvement achieved in function prediction. For instance, if we consider the identity of first three digits of EC number as a criteria to evaluate the correctness of prediction, functional annotations were transferred correctly from the top hit of COFACTOR in 156/318 enzymatic test proteins,



Template proteins with similar binding site:

	Rank	Cscore[LB]	PDB Hit	TM-score	RMSD[a]	IDEN[a]	Cov.	BS-score	Lig. Name	Download Complex	Predicted binding site residues in the model
○	1	0.28	3r4qA	0.785	2.35	0.270	0.925	1.21	CO	Download	70,105,115
○	2	0.17	1jifB	0.716	2.57	0.174	0.908	0.90	BLM	Download	67,68,69,72,105,107,111,112,113,115
○	3	0.16	1lqpA	0.753	2.31	0.209	0.917	0.85	FCN	Download	71,97,105,115
○	4	0.15	1qinA	0.756	2.55	0.133	0.942	0.86	GIP	Download	70,93,103,105,115,120
○	5	0.14	1ewjA	0.701	2.32	0.143	0.867	0.98	BLM	Download	65,67,68,69,70,74,99,103,107,111,113,115,118,11
○	6	0.10	1froA	0.751	2.62	0.133	0.942	0.83	GSB	Download	73,75,98,101,117,118
○	7	0.04	2p7kA	0.732	2.53	0.171	0.925	1.39	CIT	Download	9,11,47,56
○	8	0.04	1xrkA	0.722	2.30	0.168	0.892	1.10	BLM	Download	34,35,45,56,58,59,60
○	9	0.03	2wl3A	0.700	2.86	0.098	0.933	0.83	GOL	Download	79,83,92,95
○	10	0.01	3ed4B	0.296	5.55	0.067	0.625	0.87	GOL	Download	79,85,91,104

Figure 3. An excerpt of the result page showing ligand-binding site analysis for a Glyoxalase family protein from *Bacillus anthracis* (PDB ID: 2qqz). The server identifies high global and local similarity to Lactoylglutathione lyase of *Agrobacterium tumefaciens*, suggesting that the query also has a similar metal-ion binding site, which is required for catalysis in *Glyoxalase* I enzymes. The protein–ligand interactions are visualized using the Jmol applet.

which is approximately 27, 9 and 12% higher than the results obtained using the top hit by PSI-BLAST (26), MUSTER (27) and HHsearch (28), respectively. Similarly after removing close homologs from the template library, for the 337 test proteins, GO terms are annotated correctly ($F\text{sim} > 0.5$, see Supplementary Material) by COFACTOR for 49 and 64% proteins using the top one and the best in top five template proteins, respectively (Supplementary Table S1). Using the top one (best in top 5) template proteins, PSI-BLAST, MUSTER and HHsearch can predict GO terms correctly for 38% (49%), 44% (60%) and 41% (56%), respectively.

Here, we should note that the PSI-BLAST, MUSTER and HHsearch methods start only from query sequences, which are therefore much faster than the entire pipeline of sequence-to-structure-to-function in COFACTOR since the latter starts from the structural models predicted by I-TASSER (although the procedure of structure search by COFACTOR itself takes only less than 1 h in general). Nevertheless, these data demonstrate encouraging results that the use of protein structure information can help to obtain significant gains in the function annotations.

CONCLUSIONS

We have developed the COFACTOR server for automated structure-based functional annotation. One of the major advantages of the COFACTOR algorithm is the combination of the global and local structural comparisons. Although the global structural similarity is important for functional inference, we have witnessed a number of examples in both the benchmark and the CASP experiments, where COFACTOR successfully identified the correct functional homologs, which have different global folds but with similar binding sites, using the local structural comparisons (12).

Meanwhile, since COFACTOR scoring function includes the global structure similarity, it is more robust to the local structural variations in the target structural models than other methods, such as ConCavity (25), which rely only on the local pocket comparisons. This allows for the COFACTOR server to identify correct function homologs even using low-resolution structure models, which is of practical importance and usefulness, given the fact that most protein sequences lack experimental structure and only low-resolution structure can be generated by computational protein structure predictions (12).

Nevertheless, since the COFACTOR is essentially a template-based comparative method, no function predictions can be correctly generated if there is no homologous template protein present in the function libraries. It is therefore critical for the COFACTOR to have complete and updated template libraries. Currently, we have had the structure and ligand-binding libraries updated every week, since the information is collected directly from the PDB library (18). However, the data of GO and EC classifications are collected from other secondary resources (19,29,30), the updates of which are therefore not as regular and rely on the update of these resources. All the libraries are freely downloadable at http://zhanglab.ccmb.med.umich.edu/COFACTOR/library.html. Finally, the current algorithm is designed for single chain proteins. If multiple chains are submitted, the first chain in the PDB file is used by server automatically. We plan to extend the algorithm for multiple chain proteins and add the feature to the server in near future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figure 1 and Supplementary References [31–33].

FUNDING

Funding for open access charge: National Science Foundation Career Award [DBI 0746198]; National Institute of General Medical Sciences [GM083107, GM084222].

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H.M. and Westbrook,J.D. (2004) The impact of structural genomics on the protein data bank. *Am. J. Pharmacogenom.*, **4**, 247–252.
2. Schmidt,T., Haas,J., Cassarino,T.G. and Schwede,T. (2011) Assessment of ligand-binding residue predictions in CASP9. *Proteins*, **79(Suppl. 10)**, 126–136.
3. Zhang,Y. (2009) Protein structure prediction: when is it useful? *Curr. Opin. Struct. Biol.*, **19**, 145–155.
4. Brylinski,M. and Skolnick,J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl Acad. Sci. USA*, **105**, 129–134.
5. Oh,M., Joo,K. and Lee,J. (2009) Protein-binding site prediction based on three-dimensional protein modeling. *Proteins*, **77(Suppl. 9)**, 152–156.
6. Nagano,N., Orengo,C.A. and Thornton,J.M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.*, **321**, 741–765.
7. Watson,J.D., Laskowski,R.A. and Thornton,J.M. (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.*, **15**, 275–284.
8. Gherardini,P.F. and Helmer-Citterich,M. (2008) Structure-based function prediction: approaches and applications. *Brief Funct. Genomic Proteomic.*, **7**, 291–302.
9. Arakaki,A.K., Zhang,Y. and Skolnick,J. (2004) Large scale assesment of the utility of low resolution protein structures for biochemical function assignment. *Bioinformatics*, **20**, 1087–1096.
10. Barrett,A.J. (1997) Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *Eur. J. Biochem.*, **250**, 1–6.
11. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
12. Roy,A. and Zhang,Y. (2012) Recognizing Protein-Ligand Binding Sites by Global Structural Alignment and Local Geometry Refinement. *Structure*, doi:10.1016/j.str.2012.03.009, in press.
13. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
14. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
15. Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.
16. Xu,J. and Zhang,Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**, 889–895.
17. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
18. 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.
19. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.
20. Yu,C., Zavaljevski,N., Desai,V., Johnson,S., Stevens,F.J. and Reifman,J. (2008) The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation. *BMC Bioinformatics*, **9**, 52.
21. Krivov,G.G., Shapovalov,M.V. and Dunbrack,R.L. Jr (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**, 778–795.
22. Herraez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**, 255–261.
23. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/.
24. 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.
25. Capra,J.A., Laskowski,R.A., Thornton,J.M., Singh,M. and Funkhouser,T.A. (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol*, **5**, e1000585.
26. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
27. Wu,S. and Zhang,Y. (2008) MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*, **72**, 547–556.
28. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
29. Velankar,S., McNeil,P., Mittard-Runte,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.
30. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009: an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
31. Pesquita,C., Faria,D., Falcao,A.O., Lord,P. and Couto,F.M. (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, **5**, e1000443.
32. Schlicker,A., Domingues,F.S., Rahnenfuhrer,J. and Lengauer,T. (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, **7**, 302.
33. Wang,J.Z., Du,Z., Payattakool,R., Yu,P.S. and Chen,C.F. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.