doi:10.1016/j.jmb.2008.12.072 J. Mol. Biol. (2009) 387, 451–464

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Predicting Protein Function and Binding Profile via Matching of Local Evolutionary and Geometric Surface Patterns

Yan Yuan Tseng, Joseph Dundas and Jie Liang*

Department of Bioengineering, University of Illinois at Chicago, 851 S. Morgan Street, Room 218, SEO, MC-063, Chicago, IL 60607-7052, USA

Received 16 September 2008; received in revised form 19 December 2008; accepted 23 December 2008
Available online 6 January 2009

Inferring protein functions from structures is a challenging task, as a large number of orphan protein structures from structural genomics project are now solved without their biochemical functions characterized. For proteins binding to similar substrates or ligands and carrying out similar functions, their binding surfaces are under similar physicochemical constraints, and hence the sets of allowed and forbidden residue substitutions are similar. However, it is difficult to isolate such selection pressure due to protein function from selection pressure due to protein folding, and evolutionary relationship reflected by global sequence and structure similarities between proteins is often unreliable for inferring protein function. We have developed a method, called pevoSOAR (pocket-based evolutionary search of amino acid residues), for predicting protein functions by solving the problem of uncovering amino acids residue substitution pattern due to protein function and separating it from amino acids substitution pattern due to protein folding. We incorporate evolutionary information specific to an individual binding region and match local surfaces on a large scale with millions of precomputed protein surfaces to identify those with similar functions. Our pevoSOAR method also generates a probablistic model called the computed binding a profile that characterizes protein-binding activities that may involve multiple substrates or ligands. We show that our method can be used to predict enzyme functions with accuracy. Our method can also assess enzyme binding specificity and promiscuity. In an objective large-scale test of 100 enzyme families with thousands of structures, our predictions are found to be sensitive and specific: At the stringent specificity level of 99.98%, we can correctly predict enzyme functions for 80.55% of the proteins. The overall area under the receiver operating characteristic curve measuring the performance of our prediction is 0.955, close to the perfect value of 1.00. The best Matthews coefficient is 86.6%. Our method also works well in predicting the biochemical functions of orphan proteins from structural genomics projects.

Keywords: protein function prediction; local binding surfaces; binding profile; Bayesian Markov chain Monte Carlo; substitution rates

Edited by B. Honig

**Corresponding author.* E-mail address: jliang@uic.edu.
Present address: Y. Y. Tseng, Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA.

Abbreviations used: CASTp, Computed Atlas of Surface Topography of Proteins; ROC, receiver operating characteristic curve; MCC, Matthews correlation coefficient; EC number, Enzyme Commission number; pevoSOAR, pocket-based evolutionary search of amino acid residues; GO, Gene Ontology; cRMSD, coordinate root mean square distance; oRMSD, orientational root mean square distance.

Introduction

Predicting the molecular functions of a protein and fully characterizing its biochemical roles is an important task. An effective and widely used computational method is to identify evolutionary relationships between a protein of known function and the protein in question through sequence alignment. However, the reliability of this approach deteriorates rapidly when the level of sequence identity between the two proteins becomes lower

than 60–70%.[1,2] In addition, this method cannot provide location information on where functionally important regions and what the key residues are. More sophisticated sequence-based methods employ position-specific scoring matrices, hidden Markov models, and subfamily-specific scoring methods for function predictions.[3-5]

It is well known that very remote evolutionary relationships can be recognized through analysis of protein fold structures.[6-9] However, knowledge of the three-dimensional (3D) fold structure does not necessarily translate into knowledge of protein functions. It is also well known that proteins of the same fold can have different biochemical functions, and proteins of different fold can have similar functions.[10-14] Further challenges come from structural genomics projects,[15] where many proteins have their structures solved first without the knowledge of their biochemical functions. To derive functional information from protein structures, a recent study showed that by integrating information of fold, sequence, motif, and functional linkages, protein functions can be accurately inferred.[16] Success in inferring functions of difficult proteins has also been achieved from analyzing the distance relationship in the protein structure space map.[17]

Because protein carries out its biological roles by interacting with other molecules, binding surfaces on protein structures play important roles in determining protein functions. As functional annotation cannot be transferred reliably based on global sequence or structure similarity,[1,18,19] a promising approach is to examine local spatial regions where binding occurs and to identify similar local spatial patterns on other proteins whose functions are known.[10,20-31] This approach allows the detection of remote functional relationships for proteins in which the global similarity has evolved beyond recognition.

An example of this approach is the pvSOAR method of comparing local surfaces[31] computed using geometric algorithm.[28,32,33] It is based on the analysis of unfilled empty spaces in proteins. There are three types of empty spaces in proteins where binding interactions may occur (see Fig. 1). *Voids* are unfilled spaces inside the protein that are fully enclosed. *Pockets* on protein surfaces are caverns that open to the outside of the protein through mouths that are small relative to cavern dimensions but big enough that a solvent ball has access to the outside of the molecule. The mouth of a pocket is narrower than at least one cross section of the interior of the pocket. *Depressions* are concave regions on protein surfaces that have no constriction at the mouth.[34] Pockets and voids can be computed from protein structures with the alpha shape method, with residues forming the wall delineated and volume size measured.[28,33,35-37] In the pvSOAR method, wall residues of a pocket or a void are concatenated regardless of the separation between residues in the primary sequence into a sequence fragment. The similarity between two

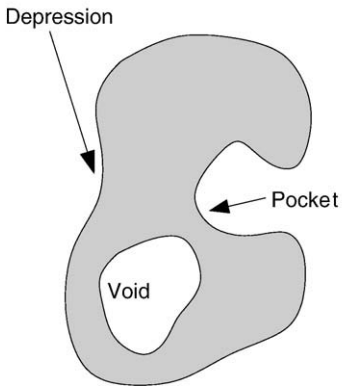

Fig. 1. Pockets and voids in proteins. There are three types of concave regions on protein surfaces: fully enclosed *voids* with no outlet, *pockets* accessible from the outside but with constriction at mouths, and shallow *depressions*. We use the general term *surface pockets* to include both pockets and voids.

surface pockets or voids is first evaluated by assessing the sequence similarity between the two sequence fragments of these surface pockets, then by further assessing spatial and orientational similarity. A number of novel functional relationships between proteins of different families and folds were uncovered by this method.[31]

To scale up this method and to search rapidly through a database of a large number of protein surface pockets, success hinges upon the use of a scoring matrix for assessing similarity between matched local pocket sequence fragments. However, existing scoring matrices such as BLOSUM, PAM, and JTT[38] are not effective for this purpose, because they do not take into account the evolutionary history of the individual protein of interests. These canned matrices have implicit parameters whose values were precomputed, while the information of the particular protein of interest has limited or no influence. In addition, the counting methods behind the derivation of some of these matrices suffer from underestimation of substitutions in certain branches of a phylogeny.[39] Furthermore, these matrices are derived based on the assumption that the whole proteins or domains experience similar selection pressure and therefore have the same substitution rates. This is unrealistic, as residues in different environments may experience different selection pressures.[40] For example, conserved residues on the binding site are under very different selection pressure than are conserved residues in the folding core.[41]

In this study, we further improve the method of function prediction by incorporating evolutionary information specific to an individual binding surface pocket. By estimating substitution rates of the residues located on a surface pocket, we derive customized scoring matrices for assessing surface similarity for predicting and characterizing complex biochemical functions. Our approach, called pevoSOAR (for pocket-based evolutionary[27] search of

amino acid residues), can effectively separate selection pressure due to the need of binding and function from that due to the need of folding and stability. A novel development of our method is a probabilistic model called *computed binding profile*, which summarizes the results of surface similarity comparison. This profile can suggest substrates and help to clarify potentially complex binding activities of a protein as well as possible cross-reactivities. It can be used to predict protein functions with improved sensitivity and specificity.

Our article is organized as follows: Using the example of acetylcholinesterase, we first illustrate how our method works. We then discuss the probabilistic model for constructing the computed binding profile of a protein. This is followed by a discussion of a large-scale test of protein function prediction for 100 protein families. Next, we describe results of the challenging task of predicting the functions of orphan protein structures obtained from structural genomics. We conclude with remarks on the general applicability of our method.

Results

Function prediction by detection of similar binding surfaces

For proteins binding similar substrates and catalyzing similar chemical reactions, the surfaces where such activities occur experience similar physical and chemical constraints. Often these surfaces have similar shapes and physicochemical properties. Due to these constraints, the sets of allowed and forbidden residue substitutions also share some similarity. Our assumption is that such similarity can be detected

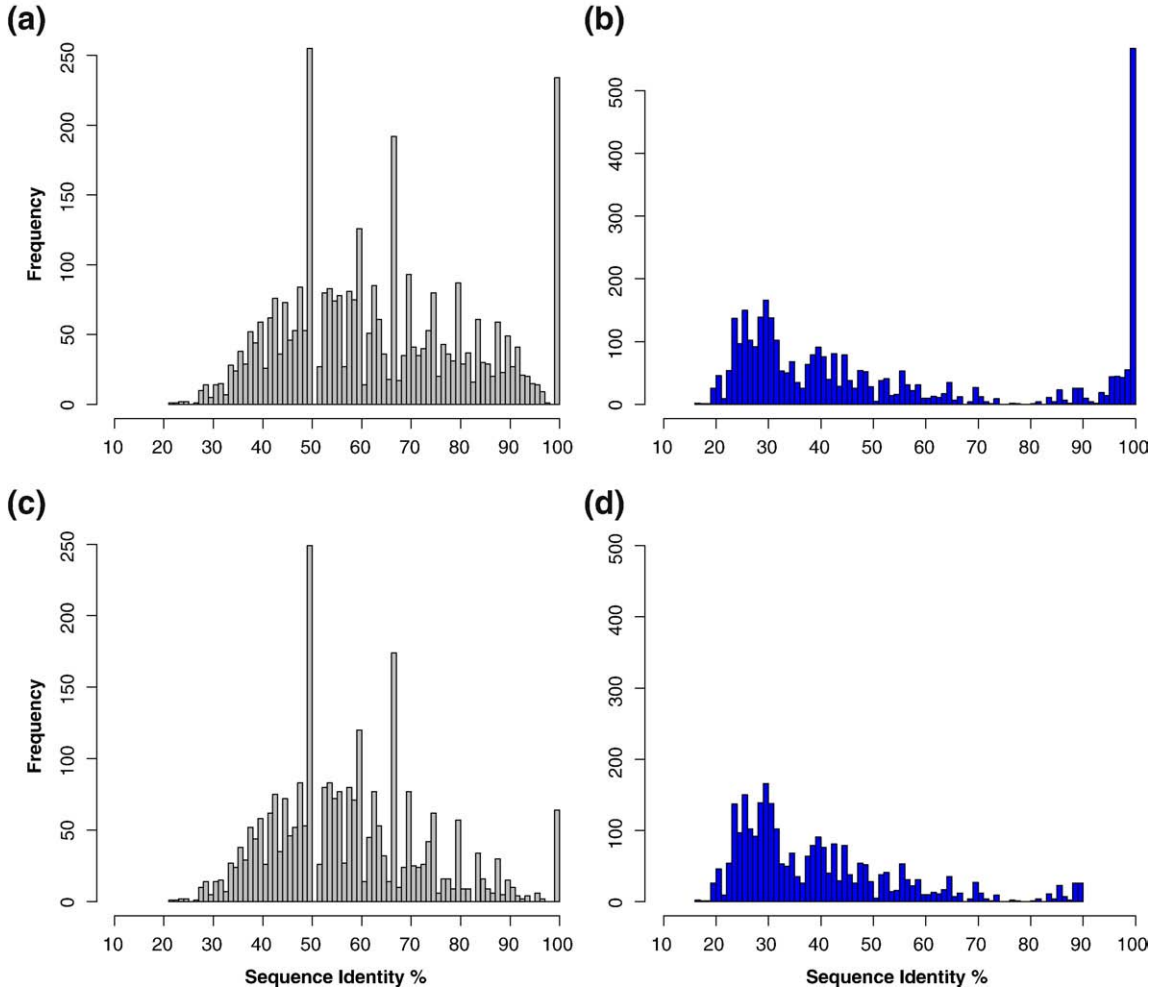

Fig. 2. Distribution of sequence identity values of binding surface pockets sequence fragments and full sequences between members of a protein family for 100 protein families (2196 structures). Distributions of identities of fragment of residues on binding surface walls between members of the same protein family (a) before and (c) after removal of sequences with overall backbone identity >90%. Note that there are still many instances where the level of identity between pocket fragments is >90%. The median sequence identity is 60.5% and 55.6% for (a) and (c), respectively. Distributions of identities of full sequences between members of the same protein family (b) before and (d) after removal of those with overall >90% sequence identity. The median sequence identity is 39.2% and 34.2%, respectively. Overall, binding surfaces are far more conserved than the full sequences.

with the use of a sensitive computational method. We first describe how binding pockets are similar to each other in general. We then discuss how our method works by assessing similarity, using the example of acetylcholinesterase and deformylase.

Sequence fragments of binding pocket and sequence of backbone

It is informative to assess in general the similarity between two binding pockets of similar function. We have collected 2196 protein structures belonging to 100 protein families, each with its own enzyme classification label[42] and Gene Ontology (GO) descriptive terms.[43] Figure 2a shows the distribution of sequence identity of pairs of sequence fragments of the residues located on the surfaces of binding pocket. Here, each pair comes from members of the same protein family. This distribution is characterized by a median of 60.5% for sequence identity. The overall distribution can be regarded as unimodal. Figure 2b shows the distribution of the overall backbone sequence identities of proteins from the same family for this group of 2196 protein structures. Its median sequence identity is 39.2%, and the smallest sequence identity is 16.4%. This distribution is clearly bimodal. After protein pairs with >90% full sequence identity are removed from the data, the distribution of pocket sequence fragments has a median of 55.6% sequence identity (Fig. 2c), and the distribution of the full sequences has a median identity of 34.2% (Fig. 2d). Overall, pocket sequence fragments have about 20% higher identity than that of full sequences.

From these two distributions, it is clear that binding pockets in general have much higher conservation than that of the full sequence. If we use the simple approach of transferring functional annotation between proteins if they share sequence identity greater than threshold values, and even if we go aggressively beyond the recommended threshold of 60–70%,[1,2] we would have failed at the 50% threshold to identify the functions of 1394 out of the 2196 proteins, representing a failure rate of 63.4%.

It seems members of the same protein family often can be clustered into two groups based on backbone sequence identity. Members of one group are closely related with each other and have relatively short evolutionary distance. Members of another group have diverged further and are more remotely related. The mixture of these two groups gives rise to the observed bimodal distribution. However, by the criterion of similarity among binding surfaces as measured by the identity of pairs of sequence fragments, all members of a given enzyme family appear to follow a unimodal distribution, suggesting their functional roles are closely related.

Illustration: predicting functions of acetylcholinesterase

We use acetylcholinesterase to illustrate our method. Acetylcholinesterase [Enzyme Commission number (EC) 3.1.1.7] is found in the synapse between nerve cells and muscle cells. It breaks down acetylcholine molecules into acetic acid and choline upon stimuli. Using a template structure [Protein Data Bank (PDB) code 1ea5], we aim to identify other structures that are acetylcholinesterase with the same EC number at the level of all four digits and to locate the surface regions that are involved in enzyme activities. EC numbers represent a progressively finer classification of an enzyme, with the first digit describing the basic reaction, and the last digit often describing the specific functional group that is cleaved during reaction.

We first exhaustively compute all pockets on the template structure.[28,37] Based on annotation contained in the PDB file, a pocket containing 32 residues (CASTp ID 79, molecular volume of 986.3 Å^3)[37] is identified as the functional pocket (Fig. 3a), which contains the Ser and His residues of the active site triad.

To construct an evolutionary model of this functional pocket, we have collected a set of 17 sequences homologous to 1ea5[44] and built a phylogenetic tree (Fig. 3d).[45] The residue substitution rates on this binding surface are estimated and scoring matrices for similarity assessment are then calculated (see Methods and Designs and Ref. 41). Using the pvSOAR search method with these scoring matrices to search the CASTp database of computed surface pockets for all PDB structures (>30,000, with >2 million surface pockets), and declaring that two proteins are of the same function when, in addition, the RMSD value of their binding pocket residues are at a significant p value of 10^{-4} (see Methods and Designs), we found 70 PDB structures to have functional surfaces similar to that of the query template 1ea5 and hence are predicted as acetylcholinesterase. Indeed, all of them have the same EC 3.1.1.7 label as that of 1ea5. The query protein and an example of matched protein surface are shown in Fig. 3a and b, respectively. There are 71 PDB entries with enzyme class label EC 3.1.1.7 in the Enzyme Structures Database†.[42] Our method successfully identified 70 of them.

Illustration: predicting functions of deformylase

Another approach other than using the EC numbers in describing protein function is to use the hierarchical terms developed by the GO consortium, where the biological role of a protein is described in terms of biochemical functions, cellular components, and biological processes. Following the same strategy as that of acetylcholinesterase, we use a structure (PDB 1lm6) of deformylase from *Streptococcus pneumoniae* as a template and search for other protein structures with similar binding surfaces.

We evaluate the results using the three GO terms associated with the query protein structure. 1lm6

† www.ebi.ac.uk/thornton-srv, for structures of enzymes contained in the ENZYME data bank.

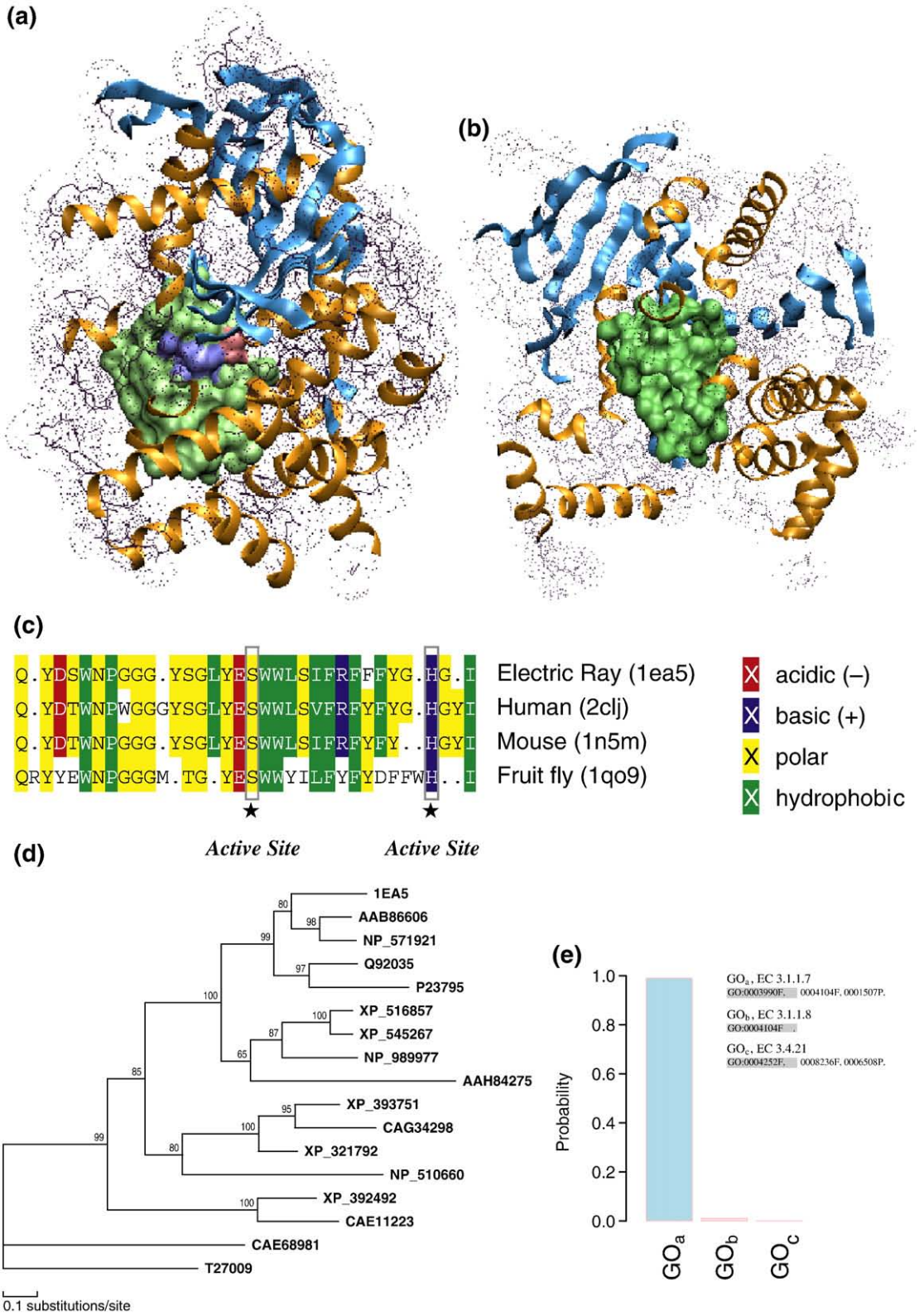

Fig. 3. Function prediction and the computed binding profile of acetylcholinesterase. (a) The functional pocket (CASTp ID 79) on a structure of acetylcholinesterase (1ea5, EC 3.1.1.7). It contains 32 residues and has a molecular volume of 986.3 Å^3. Two residues from the catalytic triad are shown: Ser200 (red) and His440 (blue). (b) A matched binding surface on a human protein structure (2clj, CASTp ID 96), with 34 residues and a molecular volume of 981 Å^3. (c) The multiple alignment of several orthologous sequence fragments of residues located in the binding pockets. The two triad residues Ser200 and His440 are conserved. (d) The phylogenetic tree consisting of 17 sequences of acetylcholinesterase is used for estimating substitution rates of residues at the binding pocket. (e) The structure 1ea5 is predicted to be an acetylcholinesterase, as indicated by the computed binding profile (GO_a EC 3.1.1.7, $\pi_1 \approx 0.99$).

has two GO terms for biochemical functions (0042586, iron binding; 0005506, peptide deformalase activity) and one GO term for biological process (0006412, translation). With scoring matrices derived from the substitution rates of residues located on the binding pocket and a significance p-value threshold of RMSD values at 10^{-4}, a total of 94 protein chains are found to have functional surfaces similar to that of 1lm6. Among these, 50 chains (53%) share all three GO terms as that of 1lm6, and 40 (43%) have no GO annotations. The remaining 4 (4%) are found to have GO terms different from that of 1lm6, and therefore can be considered as incorrect predictions (false positive). Overall, the prediction accuracy among proteins with known GO annotation is 93%. If we make the speculative but reasonably simple assumption that the rest of the 40 chains with unknown GO descriptive terms are sampled from the same distribution as that of the 54 chains with known GO terms, it is expected that the functions of 38 or so will be predicted correctly, and only 3 would be false positives.

Some of the predictions would have eluded sequence alignment methods. Among the 50 chains correctly predicted to have functions similar to that of 1lm6, 12 chains from 10 PDB structures have sequence identities >60% with the query protein, and these would have been predicted by a sequence alignment method following the recommendation from Refs. 1 and 2. However, the remaining 38 chains have sequence identities <60% (24 of which are <50%), and their functions would be difficult to predict by using the sequence alignment method. Overall, among the 94 chains where predictions are made, 32 have sequence identities >60% with the query protein, and 62 have sequence identities <60% (30 of which are <50%).

Large-scale enzyme function prediction

To assess the overall applicability of our method, we have carried out a large-scale study of protein function prediction using enzymes. Enzymes are among the best-characterized proteins in the PDB, and are an important class of proteins. Among >30,000 PDB structures (version 2006/12), there are 13,877 protein structures that are annotated as enzymes and have EC labels. In many cases, there is no information about where the active region is located on the structure and what the important residues are.

We obtain a database of computed protein surfaces on all PDB structures by selecting from the CASTp database only surface pockets that contain eight or more residues.[37] A total of 770,466 local surface pockets are collected from 1260 enzyme families. We then randomly select 100 enzyme families, each represented by a different EC label, with the criterion that there are ≥ 10 structures in each enzyme family. Altogether there are 2196 structures in these 100 protein families. For each protein family, we take the structure with the best resolution and R-factor and define the surface pocket containing key residues as annotated either in the PDB records or in the feature tables of SwissProt as the query template of the functional surfaces of this enzyme family. Using the Bayesian Monte Carlo estimator, we then derive a substitution rate matrix for this canonical template surface.[41]

Using customized similarity matrices derived from the estimated rate matrix, we then take each of the 100 template surfaces in turn and query exhaustively against all 770,466 surfaces in the database. For each matched surface from the 770,466 surfaces, if its coordinate root mean square distance (cRMSD) to the query canonical template surface is smaller than the threshold at the significance level of a cutoff p value, we declare a hit is found. This threshold is obtained as in Ref. 31. We then repeat this process for all 100 surface templates of the protein families. After collecting the list of hits for each of the 100 protein families, we identify the correctly predicted protein structure by comparing the EC labels of the hit structure and the template structure. The prediction is correct if all four digits of the two EC labels are identical. The results are summarized in the receiver operating characteristic (ROC) curve shown in Fig. 4. This is obtained by calculating the overall sensitivity and specificity of predictions of all 100 protein families at different significance p values by cRMSD. That is, they are calculated based on the number of true positives and false negatives (for sensitivity) and the number of true negatives and false positives (for specificity) found from searches for each template of the 100 protein families against the whole set of 770,466 local surfaces from 1260 enzyme families. Here an exact match of all four digits of the EC numbers is required for true positives. At the significance level of $p = 10^{-3}$, the specificity of predictions of the functions of all 2196 structures from the 100 protein family is 99.98% at all four digits of the EC labels, and the sensitivity is 80.55%. The Mathews coefficient, another measure evaluating classification quality,[46] is 82.09% at this p value. The best Mathews coefficient is 86.6% at the p value of 10^{-1}. The overall area under the ROC curve is 0.955, close to the perfect value of 1.0.

Similar to what we find from the set of 2196 protein structures, there are 1394 instances of proteins with overall backbone sequence identity less than 50%. As noted here, the sequence identity is measured between a query protein and its hit, 1058 and 608 of which have sequence identity below 40% and 30%, respectively. This indicates that the task of accurately predicting the functions of these 100 protein families is challenging, as 63.4% of them have below 50% sequence identity.

Predicting binding activities and profiling protein functions

The *computed binding profile* is a probabilistic model that can be used to identify substrates and to predict enzyme specificity. It is derived from querying results of searching a template surface against a

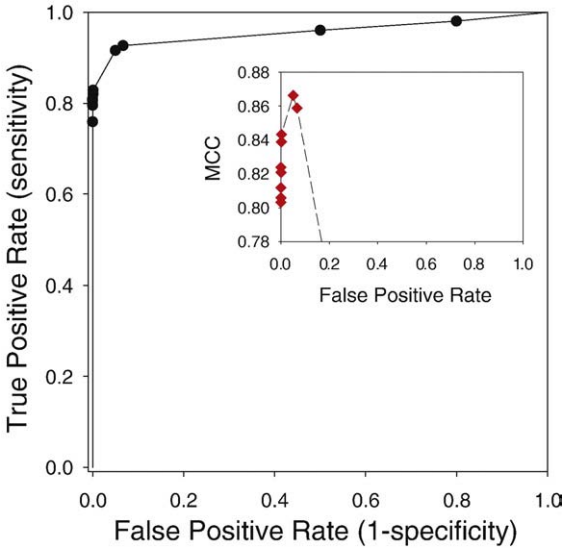

Fig. 4. Results of a large-scale test of protein function prediction for 100 protein families. For a declared hit of matched surface, if it comes from a protein structure with the same EC number (up to the fourth digit) as that of the query protein, the prediction is regarded as correct. Results are summarized in the ROC curve, where the *x*-axis represents the false-positive rate at different statistical significance *p* value of cRMSD measurement. Here the false-positive rate is 1−specificity, namely, 1−TN/(TN+FP), where TN is the number of true negatives and FP is the number of false positives. The *y*-axis represents the true positive rate or sensitivity, defined as TP/(TP+FN), where FN is the number of false negatives. An overall performance measure is the area under the ROC curve, which is 95.5%. At the confidence level of cRMSD $p=10^{-3}$, the average specificity of predictions of the functions of all 2196 proteins in these 100 protein families is 99.98%, and the average sensitivity is 80.55%. The Matthews coefficient[46] is plotted in the inset.

large library of protein surfaces. When EC numbers are used, the binding profile contains a varying number of EC labels, each with an associated probability π_i value for the *i*th label. This is interpreted as the likelihood of binding the same substrate as enzymes of that EC label. We can infer that the biochemical functions of certain enzymes are likely to be highly specific, namely, they act on most likely only one type of substrates and therefore may have very specific biochemical reactions. The computed binding profile of such enzymes contains only one EC label with a high probability π_i value.

As an example, flavoenzyme (structure 1trb) from *Escherichia coli* belongs to a subclass of oxidoreductase. The computed binding profile of flavoenzyme indicates that this protein is a thioredoxin disulfide reductase (EC 1.8.1.9) at a high specificity, with a π_1 value of ≈ 1.00 (Fig. 5a).

Our method can also identify enzymes that catalyze multiple substrates and hence can predict possible cross-reactivities. Cyclodextrin glycosyltransferase degrades starch to cyclodextrins [circular (1,4)-linked glucoses] through cyclization of 1,4-alpha-D-glucan.[47] This enzyme is also closely related to alpha-amylases and can act on glycogen, related polysaccharides, and oligosaccharides. The predicted binding profile suggests that the functional surface on the structure of 1d3c (CASTp id 78) from *Bacillus circulans* acts like a cyclodextrin glycosyltransferase (EC 2.4.1.19, the correct label) with probability $\pi_1 \approx 0.77$ (Fig. 5a). It also correctly indicates that this enzyme may bind and, hence, catalyze like an alpha-amylase to a lesser extent (EC 3.2.1.1, with a probability $\pi_2 \approx 0.22$).

Predicting the biochemical function of orphan protein structures: challenging examples from structural genomics

Orphan protein structures obtained from structural genomics have unknown biochemical functions. It is challenging to predict their functions. Several recent studies addressed this issue and reported success in the computational prediction of functions of orphan proteins.[24,25]

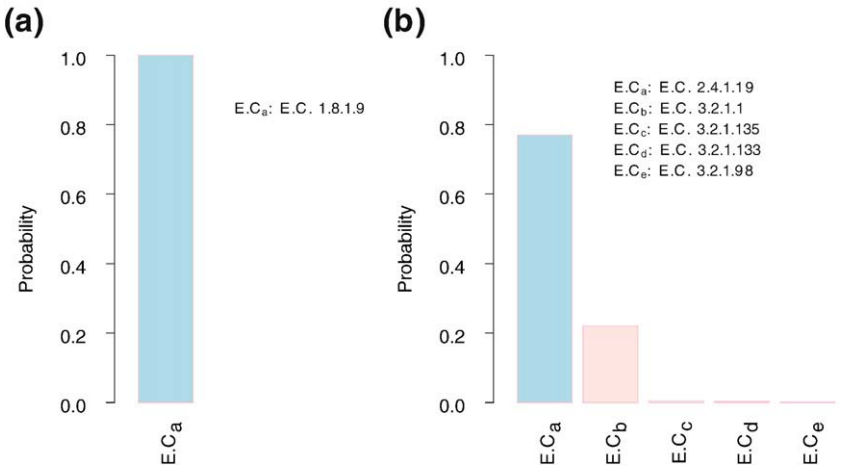

Fig. 5. Assessing enzyme specificity and promiscuity from computed binding profiles. (a) Flavoenzyme (1trb) from *E. coli* and its computed binding profile. This protein belongs to a subclass of oxidoreductases and possesses the activity of thioredoxin disulfide reductase. The computed binding profile gives the correct EC label (EC 1.8.1.9). It also suggests that this enzyme is highly specific ($\pi_1 \approx 1.00$). (b) The computed binding profile of cyclodextrin glycosyltransferase using the template 1d3c. It indicates that this enzyme is promiscuous and has cross-reactivities. It has the enzyme activity of cyclodextrin glycosyltransferase ($EC_a = 2.4.1.19$) at $\pi_a \approx 0.77$, and may also bind and hence catalyze like an alpha-amylase ($EC_b = 3.2.1.1$) at $\pi_b \approx 0.22$. The computed binding profile also suggests trace amounts of other related biochemical activities (EC 3.2.1.135, 3.2.1.133, and 3.2.1.98).

BioH. The conformation of the BioH protein from *E. coli* has unknown biological functions, but is conjectured to be involved in biotin biosynthesis.[48] It is a challenging task to infer the functional roles of BioH, because all structural homologs have ≤20% sequence identity, and some sequence homologs with between 30% and 90% sequence identity are hypothetical proteins. Using a phylogenetic tree of 28 related sequences (Fig. 6), we estimated the substitution rates of residues on the predicted binding pocket (the union of pockets with CASTp ID 28, 35, and 40 containing 35 residues and a molecular volume of 500.2 Å^3), which contains the suspected triad residues (Fig. 6). Since orphan protein structures such as BioH have no related known structures, we use the orientational root mean square distance (oRMSD) measure developed in Ref. 31 instead of the cRMSD measure for shape similarity.

The computed binding profile suggests that BioH is most likely related to a carboxylic ester hydrolase (EC 3.1.1.–), and more specifically, it may react as an acetylcholinesterase (EC 3.1.1.7, PDB 1w76, $\pi_1 \approx 1.0$). BioH was tested independently for 12 different enzyme activities with EC numbers different from our predictions, but the highest activity was found to be that of an carboxylic esterase (EC 3.1.1.1), which has the same first three digits as our prediction (EC 3.1.1.–).[48] Work by Sanishvili *et al.* also reported prediction results of the functional roles of BioH, where BioH was predicted to possess lipase, protease, or esterase activities, with additional structural features suggesting possible roles as acyltransferases and thioesterases.[48]

An orphan protein from *V. cholerae*. The structure of a hypothetical protein (Fig. 7a) from *Vibrio cholerae* (PDB ID 1u9d) is solved by Binkowski *et al.* (unpublished results) at the Midwest Center of Structural Genomics of Argonne National Laboratory. None of the sequence-based methods (e.g., BLAST and Pfam), structural alignment methods (e.g., CE, DALI, and 3DPSSM), structural classification systems (e.g., SCOP and CATH), and the GO database provide any information about the functional roles of this protein. All of the significant hits obtained by these comparison methods are hypothetical proteins with unknown biological functions. It is very challenging to predict the functions of this protein.

Using a method based on properties of shape and chemical texture of protein surfaces, we first identified the putative functional pocket, which is located in the homodimer interface.[49] This pocket is used as a template to search for similar surfaces in the database. Our results (Fig. 7c) show that 1u9d is likely to be related to phosphotransferase (with the EC label starting with 2.7.), at a probability of $\pi_1 \approx 0.95$. Because the oRMSD measure is less specific than the cRMSD measure, we conservatively estimate that 1u9d has a similar function up to two EC digits as that of the hit protein. The other hit of 1u9d is a choline kinase ($\pi_2 = 0.02$), which is also a member of the phosophotranferases. In addition, 1u9d may also have trace of activities as carbon–carbon lyases (EC 4.1.–.–). In summary, our computed binding profile suggests a limited number of biochemical assays, which can be carried out to further determine the functional profile of 1u9d.

Discussion

In this study, we have significantly improved the pvSOAR method for predicting protein functions by incorporating evolutionary information specific to individual binding surfaces. This can be illustrated by the example of alpha-amylase from *Bacillus subtilis* (1bag, CASTp ID 60). Using an updated database, we correctly identified 131 structures as alpha amylase with our current method, pevoSOAR, while the original pvSOAR method correctly predicted 116 structures. The additional 15 structures predicted by pevoSOAR are more challenging. They are more distantly related to the query protein, as their pairwise backbone sequence identities with 1bag are all less than 25%, with only one exception at 27%. In addition, our method can predict the profile of protein-binding activities, which may involve multiple substrates or ligands. Our method can be used to predict protein functions, to identify potential substrates, and to assess binding specificity.

Comparison with other methods

Although sequence-based methods such as PSI-BLAST will often find many proteins homologous to a query protein, they require significant overall sequence identity (>60–70%) for confident prediction of protein functions,[1,2] without the benefit of identifying the regions or residues that are functionally important. Our approach takes advantage of structural information and can directly identify functionally important local surface regions and can confidently predict functions of proteins with low levels of sequence identity. For example, several structures that we found using the acetylcholinesterase template 1ea5 have low levels of sequence identity with the query template but high levels of local surface sequence identity (e.g., 1qo9, 38% full-length and 60% functional surface identities with 1ea5 in Fig. 3).

Our pevoSOAR method shares some similarities to several recent works. The method of Ref. 24 is most similar to ours in that it uses manually constructed as well as automatically generated local 3D templates to assess the similarity in local structure for inferring protein functions). Although an exact direct comparison is difficult, as the underlying data set and the methodology are different, these two studies each involve about 100 different protein families. There are important differences in the criteria of prediction evaluation. In our study, the assignment of enzyme functions needs to be identical at all four digit levels of the EC labels, whereas the study of Ref. 24 is about prediction of the correct CATH domain labels. Although not perfect, EC numbers are directly related to biochemical reactions, whereas the same classification label of CATH domain does not

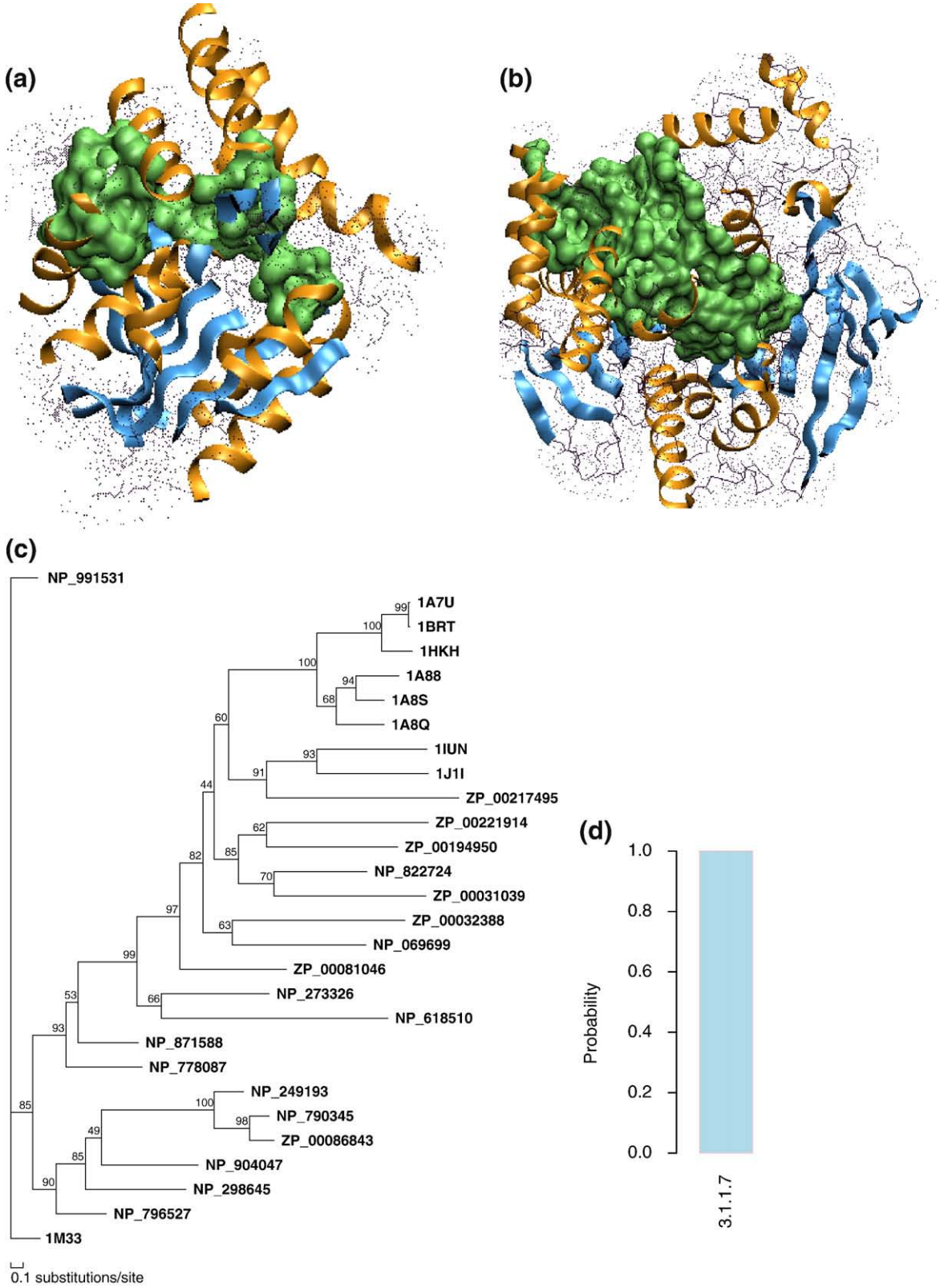

Fig. 6. Predicting functions of protein BioH obtained from structural genomics. (a) The structure of BioH (1m33) with the putative binding pocket shown. The catalytic residues (Ser82, Asp207, and His235) are located in the candidate binding pocket. (b) A similar functional surface detected from carboxylic ester hydrolases (1w76, CASTp ID 128, EC 3.1.1.7), with full sequence identify of only $\leq$20%. (c) The phylogenetic tree of 28 sequences related to BioH. Some are hypothetical proteins. (d) The computed binding profile of BioH.

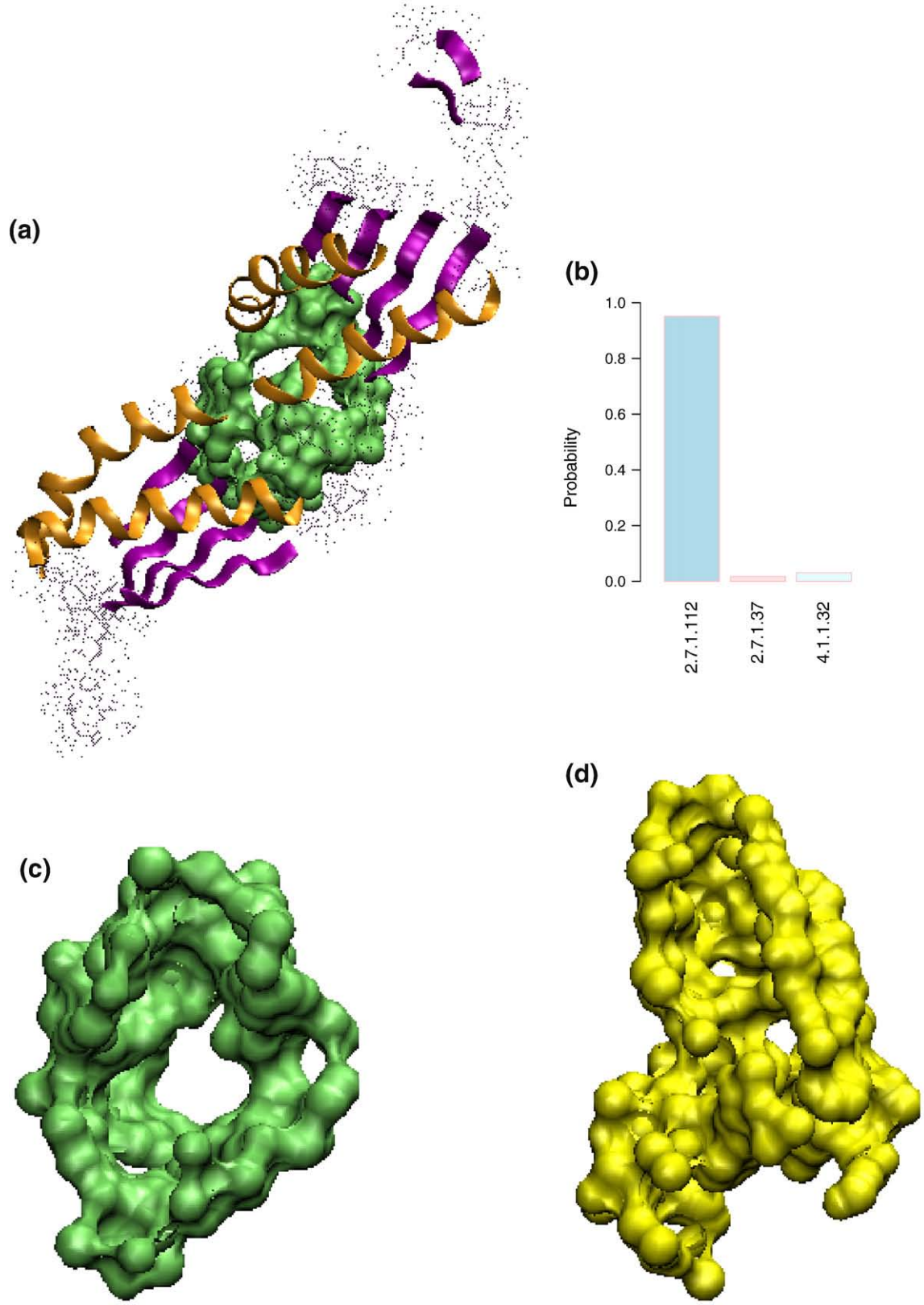

Fig. 7. Predicted functions of an orphan protein from *V. cholerae*, whose structure (PDB 1u9d) was obtained from structural genomics project. (a) The candidate binding pocket on 1u9d (CASTp ID 25) is located on the interface of the homodimer. (b) 1u9d is predicted to have a functional surface acting like a phosphotransferase (EC 2.7.–.–, $\pi_1 \approx 0.95$). (c) The functional surface on the query structure 1u9d. (d) A similar functional surface pocket (containing 62 residues with a molecular volume of 2252.63 Å^3) is found on human protein tyrosine kinase (2hck, CASTp ID 132, with an EC label of 2.7.1.112). The identity of residues on the functional pockets between these proteins is ≈48%, which is much higher than that of the full backbone sequences (≤15%).

necessarily guarantee the same protein function.[50] For example, aldehyde reductase (1ads, EC 1.1.1.21, CATH fold 3.20.20.100) has very similar fold structure with phosphotriesterase (1dpm, EC 3.1.8.1, CATH fold 3.20.20.140), yet their functions are quite different. On the other hand, aspartate aminotransferase (1yaa, EC 2.6.1.1) has similar function with D-amino acid aminotransferase (3daa, EC 2.6.1.21), but they belong to different folds (CATH 3.90.1150.10 and 3.40.640.10; CATH 3.30.470.10 and 3.20.10.10, respectively). It is well known that proteins of the same SCOP fold and CATH domain may have acquired different functions during evolution.[51-54]

With this difference in evaluation criteria, our result compares favorably with that of Ref. 24, as the measure of area under the ROC curve in Fig. 4 is 95.5%, compared to 82% in Fig. 4 of Ref. 24. We therefore conclude that our method can provide accurate information about enzymatic functions with high accuracy.

Challenges in assessing local similarity

Although the idea of inferring protein functions by assessing similarity of local spatial patterns is appealing,[55] there are significant challenges. First, it is difficult to identify the relevant small number of residues that are most informative of the function of a protein. Second, because the number of selected residues is small, it is difficult to extract evolutionary information, as the pattern of conservation is more difficult to detect from a smaller amount of data.

The Catalytic Site Atlas project provides a solution to the problem of identifying key residues by painstakingly constructing a library of 3D templates of key residues important for enzyme functions. These residues are selected manually from the literature and structural analysis.[56] It provides an important resource for studying enzyme function.

A difference between our method and those based on manually constructed 3D functional templates is that our method is fully automated. Because surface pockets are computed automatically, there is no need to manually construct 3D templates. The only requirement for our method is the knowledge that a specific computed surface pocket contains functionally important residues. The identification of such pockets can be obtained from information in annotation or can be the outcome of a functional site prediction method.[49]

Our method also differs from several other methods based on automatically generated 3D templates. The size of the surfaces in our method can be small or large, depending on the geometry of the binding pocket, whereas methods based on 3D template often are limited with the number of residues that can be included (e.g., a few residues).[24]

For uncovering the evolutionary pattern from a relatively small number of residues, we have shown that the Bayesian Monte Carlo method we developed works well. By explicitly constructing a phylogenetic tree, by using a continuous-time Markov process to describe the evolutionary process, and by using a Bayesian framework and a Markov chain Monte Carlo estimator,[41] we showed that evolutionary information specifically relevant to binding surface residues and unaltered by other constraints such as folding stability can be obtained. We believe this approach is generally applicable for problems of assessing evolutionary patterns of small regions. It also allows estimation of selection pressure due to protein function that is unaltered by selection pressure due to protein folding.

The role of hypothetical protein sequences

A limitation of our method is that we require knowledge of the structure of a protein whose function is to be predicted. However, once the structure of one protein is known, sequences with unknown structures and unknown functions (e.g., hypothetical proteins obtained from genome sequencing projects) that can be aligned to the sequence of the known structure become an important source of information about the evolution of protein functional surface. After the surface sequence fragment of binding site residues is extracted from geometric computation, our method does not require the availability of any other protein structures. Sequences that are used to construct the substitution rate matrix can be all of unknown structures, or unknown functions. As an example, several sequences contained in the phylogenetic tree in Fig. 6 are hypothetical proteins with unknown structures and unknown functions (e.g., NP_871588), but they provide critical evolutionary information for predicting protein function (Fig. 6).

Characterizing complex protein functions

In the large-scale study, we used the EC label of the highest probability as the predicted enzyme function. Although enzymes often are characterized well by the EC labels, there are several reasons why additional characterizations are important. First, protein structures may have mislabeled EC numbers, e.g., a domain is assigned the EC number of a different domain simply because they belong to the same peptide chain. Second, for many proteins, EC labels do not provide accurate information on the biochemical reactions: an enzyme may be able to react with multiple substrates. Such complex activities cannot be easily characterized. Third, knowledge of the EC label *per se* does not imply knowledge of the location of the active site or binding surface, nor the identities of the key residues. The computed binding profile generated by our method provides a more realistic picture of protein activities than just a single label of functions, as shown in the examples of phosphoglycerate mutase, which is very specific, and the example of cyclodextrin glycosyltransferase, which has broader cross-reactivities. The matched surface helps to locate residues important for binding and for function.

General applicability for enzyme characterization

Here we estimate the number of enzyme proteins in the PDB in which our method can provide useful functional information. Among ca 30,000 structures in the PDB (version 2006/12), we found that there are 13,877 protein structures annotated as enzymes with EC numbers assigned. We then select surface pockets on the enzymes in the CASTp database that contain residues annotated as functionally important either in a PDB record or in the feature table of SwissProt. Altogether we found 3275 enzyme structures whose surface pockets contain annotated functional residues.

For estimation, we use BLOSUM50 as a crude scoring matrix that does not reflect accurately the bias of residue composition in functional surfaces. This canned matrix does not account for specific evolutionary history of individual protein, or individual local surface. After clustering the 3275 enzyme structures in the PDB by EC labels, we obtain 343 clusters. We then selected the representative structure in each cluster by the criteria of best resolution and *R*-factor. Using the surface-matching method but with the canned matrix to query each of the 343 representative proteins against the surfaces contained in the CASTp database with >30,000 proteins, we are able to identify a total number of ≈11,000 protein structures as hits, namely, proteins satisfying the stringent confidence criterion of p value <0.001 for coordinate RMSD for aligned surfaces. The study with 100 protein families reported above shows that matched enzyme surfaces at this p-value threshold gives few false-positive predictions.

Based on preliminary studies of alpha amylase and other enzymes reported in Ref. 41, the number of proteins with related functions that can be established with confidence will be increased conservatively by a factor of about 3.0–3.4 when the evolutionary history of the functional surface is analyzed and the binding-surface-specific rate matrix is used. A rough estimation is that our method will characterize the functional surfaces of about 9800–11,000 structures among the 13,877 known enzyme structures, i.e., for over 70–80% of the PDB structures known to be enzymes. After removing mislabeled, incorrectly assigned, and low-quality enzyme structures, it is likely the percentage of enzyme structures whose functions our method will help to characterize will further increase. The binding surfaces of these proteins will also be identified. This represents a significant portion of all known enzyme structures.

Our pevoSOAR method is based on comparing the similarity of protein surfaces. It builds upon three techniques: First, we use geometric algorithms to quantify accurately protein local surfaces;[28,57] second, we use a Bayesian Monte Carlo estimator to characterize the evolutionary history specifically for a local surface;[41] third, we compare surfaces by assessing evolutionary similarity of residues on local surfaces,[41] in shape, and in orientation.[31] In principle, our method of function characterization by matching protein surfaces is general and can be applied to protein functions other than enzyme activities such as protein–protein interactions. In this case, a prerequisite is the ability to generate a library of surface patches that represent the interfaces of protein–protein interaction accurately.

Methods and Designs

Estimating substitution rates

The success in rapid detection of functionally related protein surfaces through the alignment of sequence fragment of binding surface residues[58] depends on the use of a scoring matrix that determines the similarity between residues. The instantaneous rate matrices of amino acid residue substitution is the basis for developing such scoring matrices. We use a reversible continuous-time Markov process as our evolutionary model.[39,59-61] Details of Bayesian estimator based on the technique of Markov chain Monte Carlo, including the construction of the phylogenetic tree, are described in Ref. 41.

Scoring matrices of similarity for surfaces at different evolutionary time intervals

To derive the scoring matrix for assessing functional similarity between two surfaces and for database search, we calculate the residue similarity scores $b_{ij}(t)$ between residues i and j at evolutionary time t.[62] From the rate matrix, we use the Altschul model to calculate similarity score $b_{ij}(t)$:[62]

$$b_{ij}(t) = \frac{1}{\lambda}\log\frac{p_{ij}(t)}{\pi_j} = \frac{1}{\lambda}\log\frac{m_{ij}(t)}{\pi_i\pi_j},$$

where $m_{ij}(t)$ is the joint probability of observing both residue types i and j at the two sequences separated by time t, and λ is a scalar. Here, $p_{ij}(t)$ can be computed from the instantaneous rate matrix.[41]

Matching local surfaces

Because *a priori* we do not know how far a particular candidate protein is separated in evolutionary time from the query template protein, we calculate a series of 300 scoring matrices, each characterizing the residue substitution pattern at a different time separation, ranging from 1 to 300 time units. Here, 1 time unit represents the time required for 1 substitution per 100 residues.[63] We use the Smith–Waterman algorithm as implemented in the Ssearch program using each of the 300 scoring matrices to align sequence fragments of candidate binding surfaces against the database of sequence fragments of protein surface pockets derived from the CASTp database.[37]

In addition, surfaces matched by sequence fragment similarity are subject to further shape analysis. We compare surfaces by either the coordinate RMSD values or the orientational oRMSD value we developed in Ref. 31 when specified. Those that can be superimposed to the residues of the query surface at a statistically significant level (e.g., p value <0.001 by coordinate RMSD measure) are declared as hits.[31,41] The p value for cRMSD and oRMSD is estimated through extensive randomization simulations as described in Ref. 31.

Probabilistic model for profiling protein-binding activities

We introduce a probabilistic model, the *computed binding profile*, for characterizing specific binding activities and for inferring protein functions. We use each of the 300 scoring matrices representing time intervals from 1 unit to 300 units to search the surface database in turn. Assuming each time interval is equally likely, the probability of a query protein belonging to the ith EC label is calculated as:

$$\pi_i = \frac{\sum_t \mathrm{EC}_i(t)}{\sum_t N(t)}, \quad (1)$$

where $\mathrm{EC}_i(t)$ is the number of PDB hits belonging to the ith EC label using matrix of time distance t and $N(t)$ is the total number of PDB hits with a known EC number using matrix of time distance t. When a protein has a number of different hits with different EC labels with associated probability values, this set of EC labels and the corresponding π_i values provide a computed binding profile that help to characterize the potentially complex binding activities of a protein.

Acknowledgements

We thank Dr. Andrew Binkowski for previous implementation of pvSOAR, and Drs. Andrew Binkowski and Andzrej Joachimiak for suggesting BioH and 1u9d for this study. This work is supported by grants from the NSF (DBI-0646035 and DMS-0800257), NIH (GM079804-01A1 and GM081682), and ONR (N000140310329).

References

1. Rost, B. (2002). Enzyme function less conserved than anticipated. *J. Mol. Biol.* **318**, 595–608.
2. Tian, W. & Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* **333**, 863–882.
3. Hannenhalli, S. & Russell, R. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **303**, 61–76.
4. Jensen, L., Gupta, R., Staerfeldt, H. & Brunak, S. (2003). Prediction of human protein function according to gene ontology categories. *Bioinformatics*, **19**, 635–642.
5. Sjolander, K. (2004). Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**, 170–179.
6. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a Structural Classification of Proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
7. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
8. Thornton, J. M. (2001). From genome to function. *Science*, **292**, 2095–2097.
9. Zarembinski, T., Hung, L., Dieckmann, H., Kim, K., Yokota, H., Kim, R. & Kim, S. H. (1998). Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc. Natl Acad. Sci. USA*, **95**, 15189–15193.
10. Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **279**, 1211–1227.
11. Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J., Skolnick, J. & Godzik, A. (1999). From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions. *Protein Sci.* **8**, 1104–1115.
12. Copley, S., Novak, W. & Babbitt, P. (2004). Divergence of function in the thioredoxin fold suprafamily: evidence for evolution of peroxiredoxins from a thioredoxin-like ancestor. *Biochemistry*, **43**, 13981–13995.
13. Wang, K. & Samudrala, R. (2005). FSSA: a novel method for identifying functional signatures from structural alignments. *Bioinformatics*, **21**, 2969–2977.
14. Polacco, B. & Babbitt, P. (2006). Automated discovery of 3D motifs for protein function annotation. *Bioinformatics*, **22**, 723–730.
15. Chandonia, J. & Brenner, S. (2006). The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.
16. Pal, D. & Eisenberg, D. (2005). Inference of protein function from protein structure. *Structure*, **13**, 121–130.
17. Hou, J., Jun, S., Zhang, C. & Kim, S. (2005). Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc. Natl Acad. Sci. USA*, **102**, 3651–3656.
18. Wilson, C., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233–249.
19. Todd, A., Orengo, C. & Thornton, J. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
20. Fischer, D., Norel, R., Wolfson, H. & Nussinov, R. (1993). Surface motifs by a computer vision technique: searches, detection, and implications for protein–ligand recognition. *Proteins*, **16**, 278–292.
21. Norel, R., Fischer, D., Wolfson, H. J. & Nussinov, R. (1994). Molecular surface recognition by computer vision-based technique. *Protein Eng.* **7**, 39–46.
22. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **236**, 412–420.
23. Glaser, F., Pupko, T., Paz, I., Bell, R., Shental, D., Martz, E. & Ben-Tal, N. (2003). Consurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
24. Laskowski, R., Watson, J. & Thornton, J. (2005). Protein function prediction using local 3D templates. *J. Mol. Biol.* **351**, 614–626.
25. Pazos, F. & Sternberg, M. (2004). Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl Acad. Sci. USA*, **101**, 14754–14759.
26. Ferre, F., Ausiello, G., Zanzoni, A. & Citterich, M. (2005). Functional annotation by identification of local surface similarities: a novel tool for structural genomics. *BMC Bioinformatics*, **6**, 194.
27. Gold, N. & Jackson, R. (2006). Fold independent structural comparisons of protein–ligand binding sites for exploring functional relationships. *J. Mol. Biol.* **355**, 1112–1124.
28. 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**, 1884–1897.

29. Laskowski, R. A. (1995). Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graphics*, **13**, 323–330.
30. Laskowski, R. A., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438–2452.
31. Binkowski, T. A., Adamian, L. & Liang, J. (2003). Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.* **332**, 505–526.
32. Edelsbrunner, H. & Mücke, E. (1994). Three-dimensional alpha shapes. *ACM Trans. Graphics*, **13**, 43–72.
33. Edelsbrunner, H., Facello, M. & Liang, J. (1998). On the definition and the construction of pockets in macromolecules. *Discrete Appl. Math.* **88**, 83–102.
34. Liang, J. & Dill, K. A. (2001). Are proteins well-packed? *Biophys. J.* **81**, 751–766.
35. Edelsbrunner, H. (1995). The union of balls and its dual shape. *Discrete Comput. Geom.* **13**, 415–440.
36. Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V. & Subramaniam, S. (1998). Analytical shape computing of macromolecules II: Identification and computation of inaccessible cavities inside proteins. *Proteins*, **33**, 18–29.
37. Binkowski, T. A., Naghibzadeh, S. & Liang, J. (2003). CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res.* **31**, 3352–3355.
38. Jones, D. T., Taylar, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *CABIOS*, **8**, 275–282.
39. Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699.
40. Tourasse, N. & Li, W. (2000). Selective constraints, amino acid composition, and the rate of protein evolution. *Mol. Biol. Evol.* **17**, 656–664.
41. Tseng, Y. & Liang, J. (2006). Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach. *Mol. Biol. Evol.* **23**, 421–436.
42. Bairoch, A. (1993). The ENZYME data bank. *Nucleic Acids Res.* **21**, 3155–3156.
43. Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J. *et al.* (2004). The Gene Ontology Annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* **32** (Database issue), D262–D266.
44. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
45. Adachi, J. & Hasegawa, M. (1996). A computer program package for molecular phylogenetics ver 2.3. Computer Science Monographs, 28:1–150. Institute of Statistical Mathematics, Tokyo, Japan.
46. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
47. Uitdehaag, J., Kalk, K., van, D., Dijkhuizen, L. & Dijkstra, B. (1999). The cyclization mechanism of cyclodextrin glycosyltransferase (CGTase) as revealed by a gamma-cyclodextrin–CGTase complex at 1.8-Å resolution. *J. Biol. Chem.* **274**, 34868–34876.
48. Sanishvili, R., Yahunin, A. F., Laskowski, R. A., Skarina, T., Evdokimova, E., Doherty-Kirby, A. *et al.* (2003). Integrating structure, bioinformatics, and enzymology to discover function: BioH, a new carboxylesterase from *Escherichia coli*. *J. Biol. Chem.* **278**, 26039–26045.
49. Tseng, Y. & Liang, J. (2007). Predicting enzyme functional surfaces and locating key residues automatically from structures. *Ann. Biomed. Eng.* **35**, 1037–1042.
50. Meng, E., Polacco, B. & Babbitt, P. (2004). Superfamily active site templates. *Proteins*, **55**, 962–976.
51. Wistow, G., Mulders, J. & de, J. (1987). The enzyme lactate dehydrogenase as a structural protein in avian and crocodilian lenses. *Nature*, **326**, 622–624.
52. Acharya, K., Ren, J., Stuart, D., Phillips, D. & Fenna, R. (1991). Crystal structure of human alpha-lactalbumin at 1.7 Å resolution. *J. Mol. Biol.* **221**, 571–581.
53. Orengo, C., Todd, A. & Thornton, J. (1999). From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374–382.
54. Jeffery, C. (2004). Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. *Curr. Opin. Struct. Biol.* **14**, 663–668.
55. Najmanovich, R., Hassani, A., Morris, R., Dombrovsky, L., Pan, P., Vedadi, M. *et al.* (2007). Analysis of binding site similarity, small-molecule similarity and experimental binding profiles in the human cytosolic sulfotransferase family. *Bioinformatics*, **23**, e104–e109.
56. Porter, C., Bartlett, G. & Thornton, J. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **32**, D129–D133.
57. Edeslbrunner, H., Facello, M. & Liang, J. (1998). On the definition and the construction of pockets in macromolecules. *Discrete Appl. Math.* **88**, 18–29.
58. Binkowski, T., Freeman, P. & Liang, J. (2004). pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic Acid Res.* **32**, W555–W558.
59. Yang, Z., Nielsen, R. & Hasegawa, M. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**, 1600–1611.
60. Felsenstein, J. & Churchill, G. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**, 93–104.
61. Siepel, A. & Haussler, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21**, 468–488.
62. Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
63. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). Atlas of Protein Sequence and Structure, vol. 5, pp. 345–352, Nat Biomed Research Foundation, Washington, DC.