BIOINFORMATICS

Vol. 20 no. 7 2004, pages 1087–1096
DOI: 10.1093/bioinformatics/bth044

Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment

*Adrian K. Arakaki, Yang Zhang and Jeffrey Skolnick**

Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington St, Buffalo, NY 14203–1199, USA

Received on August 19, 2003; revised on November 5, 2003; accepted on November 26, 2003
Advance Access publication February 5, 2004

ABSTRACT

Motivation: Several protein function prediction methods employ structural features captured in three-dimensional (3D) descriptors of biologically relevant sites. These methods are successful when applied to high-resolution structures, but their detection ability in lower resolution predicted structures has only been tested for a few cases.

Results: A method that automatically generates a library of 3D functional descriptors for the structure-based prediction of enzyme active sites (automated functional templates, 593 in total for 162 different enzymes), based on functional and structural information automatically extracted from public databases, has been developed and evaluated using decoy structures. The applicability to predicted structures was investigated by analyzing decoys of varying quality, derived from enzyme native structures. For 35% of decoy structures, our method identifies the active site in models having 3–4 Å coordinate root mean square deviation from the native structure, a quality that is reachable using state of the art protein structure prediction algorithms.

Availability: See http://www.bioinformatics.buffalo.edu/resources/aft/

Contact: skolnick@buffalo.edu

INTRODUCTION

One basic step in the process of learning how a living organism works is the determination of the biochemical functions of all the proteins encoded by its genome (Alberts *et al*., 1994). Although the biochemical function of a protein does not necessarily determine its biological role in the cell, it does provide an excellent starting point for further investigations. This kind of large-scale functional assignment requires the application of predictive methods, since the exhaustive experimental characterization of complete proteomes is not currently feasible.

The most commonly used approaches to the prediction of protein function are based on the transfer of annotation from homologous sequences (Bork *et al*., 1994; Koonin *et al*., 1996). However, as was extensively analyzed by Todd *et al*. (2001), the functional promiscuity exhibited by many protein families limits the success of annotation transfer between homologs. If uncertain functional assignments are used as an origin for successive predictions, the homology-based annotation strategy may lead to a propagation of errors (Galperin and Koonin, 1998). Quality evaluations of functional assignments have shown that this type of homology-based inference is a common source of error in genome annotation (Devos and Valencia, 2001; Iliopoulos *et al*., 2003). Different groups studied the sequence identity threshold that allows the transfer of function as defined by the Enzyme Commission (EC) numbers (Barrett, 1997), but the specific value for this threshold is still the subject of controversy (Devos and Valencia, 2000; Rost, 2002; Tian and Skolnick, 2003; Todd *et al*., 2001; Wilson *et al*., 2000). By comparing domain pairs at varying levels of sequence similarity extracted from SCOP (Murzin *et al*., 1995), Wilson *et al*. (2000) established that over 90% of the pairs conserved the first three EC components at 30–40% sequence identity. The same result was found by Devos and Valencia (2000), by analyzing pairs of aligned enzyme sequences from the FSSP database (Holm and Sander, 1994). Similarly, by sequence comparison of members of CATH superfamilies (Orengo *et al*., 1997), Todd *et al*. (2001) concluded that above 30% sequence identity, the first three EC components may be predicted with an accuracy of at least 90%, but below that threshold, structural data are essential to determine the function. Recently, Rost (2002) argued that these groups overestimated the extent of conservation of enzyme function because their datasets were either too biased, or too small. By performing an analysis over data whose bias was reduced by clustering of similar sequences, he found that even at 50% sequence identity, the transfer of three or four EC components is $<30\%$ accurate (Rost, 2002). However, when Tian and Skolnick (2003) reduced the bias by classifying the enzymes not only according to the level of sequence similarity, but also considering functional similarity, they observed that 40% sequence identity is enough to transfer the first three EC

*To whom correspondence should be addressed.

components with an accuracy of at least 90%. Thus, most of analyses agreed that 40% sequence identity allows the transfer of the first three EC components with an accuracy of at least 90%. However, to increase this level of accuracy, the development of other methods is required.

Because structure is more conserved than sequence, protein function prediction should, in principle, benefit by the addition of structural information (Skolnick *et al*., 2000). However, as mentioned before, divergent and convergent evolution have resulted in a non-unique relationship between function and fold, that is, fold type by itself is not enough for correct function prediction (Hegyi and Gerstein, 1999). Several protein function prediction methods take advantage of more detailed structural features, captured in libraries of three-dimensional (3D) descriptors of biologically relevant sites (Fetrow and Skolnick, 1998; Hamelryck, 2003; Jonassen *et al*., 1999; Kleywegt, 1999; Liang *et al*., 2003; Oldfield, 2002; Pennec and Ayache, 1998; Russell, 1998; Wallace *et al*., 1997; Zhao *et al*., 2001). These methods based on structural templates are successful when applied to high-resolution structures, but their ability to detect functional sites in lower resolution predicted structures has only been tested for a few specific descriptors (Fetrow and Skolnick, 1998; Wei *et al*., 1999). Therefore, it is unclear how applicable is this approach to the large-scale function prediction involving a complete library of structural templates and medium-to-low resolution protein models. Given recent improvements in the performance of protein structure prediction algorithms (Bonneau *et al*., 2002; Skolnick *et al*., 2003; Zhang *et al*., 2003), a template-based method for protein function prediction that does not require high-resolution structures would be very advantageous and of immediate practical value.

Here, we present a method for large-scale protein function prediction based on structural descriptors, termed 'automated functional templates' (AFTs) that focuses on the recognition of enzyme active sites in predicted protein structures. Conceptually similar to fuzzy functional forms (FFFs) (Fetrow and Skolnick, 1998), the AFTs are defined by pairwise distances between atoms and pseudo-atoms related to functionally important residues. However, as opposed to the FFFs, the AFTs are automatically generated; therefore, they are readily implemented on a large scale. By relying on functional and structural information automatically extracted from the Swiss-Prot and Protein DataBank (PDB), the method generates 593 descriptors for 162 different enzymes. The algorithm first evaluates the specificity of each AFT on a representative set of PDB high-resolution structures, and then sets confidence intervals for the identification of putative active sites in functionally uncharacterized structures. To investigate its applicability to predicted structures, we perform an all-against-all scanning of the 593 AFTs over a set of 593 000 decoy structures exhibiting up to a 10 Å Cα coordinate root mean square deviation (crmsd) from the native structure. The decoys are derived from 593 native structures of enzymes whose active sites are true positives for each of the AFTs in the functional descriptor library. To illustrate the possible annotation scenarios for the application of the AFT approach, we present two examples of function prediction for *Escherichia coli* proteins. The result of this analysis provides for the first time a comprehensive view of the expected success rate of a function prediction method based on structural templates, when applied to the identification of functional sites in predicted protein structures.

SYSTEMS AND METHODS

Structural and sequence databases

Structural information is obtained from the April 15, 2003 PDB (http://www.rcsb.org/pdb/). To reduce database redundancy, a list containing the largest sequence-unique subset of PDB chains is retrieved from the April 3, 2003 evaluation of automatic (EVA) protein structure prediction website (http://cubic.bioc.columbia.edu/eva/res/weeks.html). The resulting representative subset of the PDB contains 2965 chains, where no pair has more than 33% identical residues over more than 100 aligned residues.

Swiss-Prot Release 41.2 and TrEMBL Release 23.4 of March 28, 2003 (http://us.expasy.org/sprot/) are the sources of all the sequence and functional information. Although there are many other sequence databases available, Swiss-Prot is chosen because of its high level of annotation and minimal redundancy.

The standard system for enzyme classification is based on the EC numbers, which consist of four components: the first three components (class, subclass and sub-subclass) define the reaction catalyzed by the enzyme, and the fourth component is a unique identifier or serial number that can represent the type of chemical bond, molecular mechanism or substrate specificity (Barrett, 1997). An EC number/Swiss-Prot cross reference table is compiled by parsing the annotated EC numbers in the DEscription (DE) lines of all Swiss-Prot entries. EC numbers with missing components (e.g. EC 3.4.11.-) are ignored.

We created a PDB/Swiss-Prot cross reference table based on the April 5, 2003 database offered by the IMB Jena Image Library of Biological Macromolecules (http://www.imb-jena.de/ImgLibPDB/pages/SWP/index.php). This database not only contains more cross references than either PDB or Swiss-Prot, but also corrects some obvious errors found in the DataBase REFerence (DBREF) section of many PDB files. For example, although the PDB entry 1JAN incorrectly refers to the Swiss-Prot entry COG8_HUMAN, the IMB Jena database provides the right cross reference, i.e. MM08_HUMAN. The IMB Jena database only links Swiss-Prot entries to PDB id codes, and not to individual protein chains. Therefore, for each Swiss-Prot—PDB cross reference, the chain identifier information is obtained by detecting the best alignment

between the Swiss-Prot sequence and the sequences of every chain present in the corresponding PDB file.

Representation of an AFT

An automated functional template or AFT consists of the spatial arrangement of N functional building blocks ($3 \leq N \leq 5$). A functional building block is composed of the Cα atom of a functional residue, the two Cα atoms of its N- and C-terminal adjacent residues, and (for non-glycine residues) one pseudo-atom corresponding to its side-chain centroid (SC). An AFT is defined by the amino acid types of 3–5 functional residues, and the set of pairwise distances between the atoms and pseudo-atoms that constitute the functional building blocks. The range of the number of functional building blocks forming an AFT, as well as the reduced description of the amino acids are chosen after optimization for sensitivity (true positive rate), specificity (true negative rate) and calculation speed.

Outline of the AFT building process

The AFTs are based on the 3D arrangement of residues that are important for defining the molecular function of a given enzyme. A summary of the procedure for building an AFT is shown in Figure 1. The algorithm consists of three basic steps: (i) retrieval of functionally important substructures from all PDB structures associated with a specific EC number; (ii) generation of tentative distance-based templates describing the active site and (iii) a specificity assessment of the AFTs. It should be stressed that the whole procedure is fully automated. The details are provided here.

Retrieval of functionally important substructures from PDB

First, in order to data mine for functional residues, we collect all the Swiss-Prot entries that refer to a given EC number. Entries linked to more than one EC number are ignored (e.g. viral polyproteins), because in the annotation provided by Swiss-Prot there are no links between functional residues and particular EC numbers. We extract information about the residues that are potentially important for the enzymatic activity from the feature table (FT) lines of the Swiss-Prot entries. Each FT line contains a key name that describes a region of interest, as well as two numbers indicating the position of the region in the sequence. We collect functional residue information from FT lines containing the following key names: (a) ACT_SITE, indicating amino acids involved in the enzyme activity; (b) METAL, referring to metal binding sites; (c) BINDING, referring to binding sites for any prosthetic group or co-enzyme and (d) SITE, indicating any other interesting site in the sequence.

For each functional residue, we store: (a) the accession number of the originating Swiss-Prot entry; (b) the sequence position; (c) the type of amino acid; (d) the FT key name; and, after mapping the residues in all the cross referenced PDB chains; (e) the coordinates necessary for the AFT generation. We exclude all PDB structures for which either of the following applies: (a) the coordinates of at least three functional residues are unavailable; (b) coordinates of at least one residue annotated as an ACT_SITE are unavailable; (c) any functional residue is mutated; (d) the structure is a theoretical model and (e) the experimental technique used to solve the structure is NMR and a minimized or averaged model is not provided. Many heteromultimeric enzymes have catalytic residues in different polypeptides, with active sites located at the subunit interface (Bartlett *et al*., 2002). These cases are beyond the scope of this study; therefore, we rule out PDB structures exhibiting functional residues distributed across two or more chains.

After this first step is applied to all existent EC numbers, we obtain the coordinates of 15 561 functional building blocks

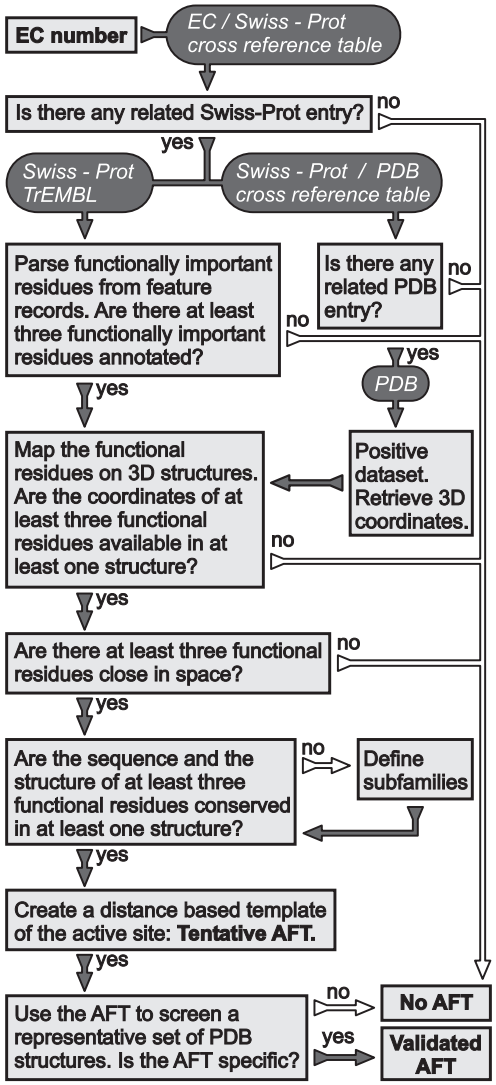

Fig. 1. Overview of the procedure for building an AFT.

from 3506 PDB chains. These protein structures are linked to a total of 395 Swiss-Prot sequences representing 214 different enzymes.

Generation of a tentative AFT

The second step of the algorithm consists of processing the gathered functional information to build provisional AFTs. Since the goal is to select amino acids involved in a common active site, we require all the AFT functional residues to be within interacting distance. The residues are defined as potentially interacting if their SC pseudo-atoms are within 17 Å. After applying this filter, and depending upon the EC number analyzed, we end up with a family ranging from 1 to 207 functional substructures. Each functional substructure belongs to a different PDB chain and is composed of 3–12 functional building blocks.

When more than five functional residues are available, we select a subset of five, according to the following preference rules designed to increase the biological significance of the functional descriptor: (a) maximize the number of functional substructures containing the five residue subset; (b) maximize the number of residues annotated under the key name ACT_SITE; (c) minimize the average distance between residues and (d) minimize the pairwise distance root mean square deviation (drmsd) between substructures.

If the amino acid types and the structure of at least three functional residues are not conserved through all functional substructures (using a drmsd threshold of 0.3 Å), we define subfamilies by complete linkage clustering so that both criteria are satisfied. Then, we create a tentative template for the family, or one for each subfamily if clustering was necessary. The template is defined by the amino acid types of the 3–5 functional residues, plus the set of averaged distances between every possible pair of atoms or pseudo-atoms that define the functional building blocks.

We opt for a template based on averaged distances instead of one based on averaged coordinates, to be able to recognize mirror images of known functional sites. This is important since there are at least two reported cases of isoenzymes displaying mirror image active sites: carbonic anhydrase (EC 4.2.1.1) (Kimber and Pai, 2000) and methionine sulfoxide reductase (EC 1.8.4.5) (Lowther *et al*., 2002).

The averaged distances are calculated hierarchically, to avoid bias toward overrepresented structures. First, we average the pairwise interatomic distances from functional substructures belonging to different chains of the same PDB entry, then those from representatives of different PDB entries linked to the same Swiss-Prot sequence, and finally those from representatives of different Swiss-Prot sequences. We define the 'positive dataset' as the collection of all the substructures utilized to build the AFT. The size of the positive dataset ranges from 1 to 51 (for the AFT associated with EC 1.11.1.5, cytochrome *c* peroxidase) with an average value of 3.8 substructures per AFT.

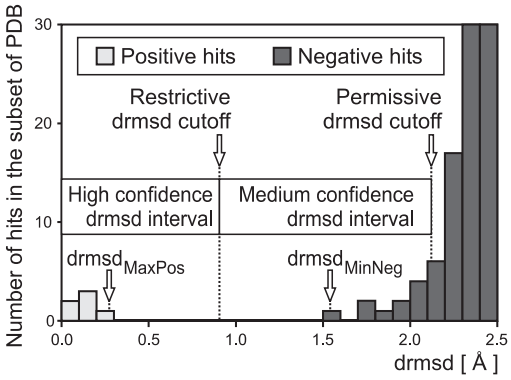

Fig. 2. Specificity parameters of an AFT.

After this second step is applied to all existent EC numbers, we obtain 605 provisional AFTs associated with 167 enzymes.

Validation of an AFT

In the final step of the procedure, we evaluate the ability of an AFT to specifically detect the enzyme active site for which the AFT was designed, and we establish criteria to recognize significant predictions. First, we prepare a 'negative dataset' from the representative subset of PDB by removing all the positive structures, i.e. PDB chains from which the positive dataset was extracted. We also remove the structures of enzymes whose EC numbers share the first three components with the EC number under analysis.

We scan the negative structures for sets of residues that satisfy the amino acid type requirements of the tentative AFT, and we calculate the drmsd from the AFT of every detected hit. For a given AFT, the number of hits in the negative dataset ranges from 5×10^5 to 5×10^{10}, depending upon the number of residues and the sequence complexity of the functional descriptor. As shown in Figure 2, we define $\text{drmsd}_{\text{Neg}}$ as the drmsd from the AFT of a hit in the negative dataset, and $\text{drmsd}_{\text{MinNeg}}$ as the minimum $\text{drmsd}_{\text{Neg}}$ detected in the negative dataset. Similarly, $\text{drmsd}_{\text{Pos}}$ is defined as the drmsd from the AFT of a substructure in the positive dataset, and $\text{drmsd}_{\text{MaxPos}}$ as the maximum $\text{drmsd}_{\text{Pos}}$ detected in the positive dataset (Fig. 2). Our criterion to promote a tentative AFT to a validated AFT is the existence of a gap of at least 0.2 Å between the distributions of drmsd for the negative and the positive hits, i.e. we require that $\text{drmsd}_{\text{MinNeg}} - \text{drmsd}_{\text{MaxPos}} > 0.2$ Å.

We propose two drmsd cutoffs for establishing the significance of a match to an AFT: (i) a restrictive cutoff calculated as the average of $\text{drmsd}_{\text{MinNeg}}$ and $\text{drmsd}_{\text{MaxPos}}$, and (ii) a permissive cutoff defined so that the expected number of false positive matches is less than 0.005 per negative structure. After testing different schemas to evaluate the significance of a hit to decoy structures (see below), we found that the current concept of permissive cutoff provides an adequate balance

between sensitivity and specificity for the application of the AFTs to low resolution structures. Based on the restrictive and the permissive cutoffs, we define high confidence and medium confidence intervals of drmsd (Fig. 2).

After the last step of the algorithm is applied to all existent EC numbers, we obtain 593 validated AFTs associated with 162 enzymes.

Generation of decoy structures

In order to generate the decoys structures, we first select one representative positive protein for each of the 593 AFTs. Then, we use these representative structures as starting points for parallel hyperbolic Monte Carlo simulations of CABS model (Zhang *et al*., 2002). Only Cα atoms and SC pseudo-atoms are modeled. For each protein, we randomly select 1000 decoys, equally distributed in 10 bins according to their crmsd to the native structure, i.e. $[i-1$ Å$, i$ Å$]$, $i = 1, 2, \ldots, 10$. To avoid structural redundancy, we add a filter in the decoy selection procedure so that the crmsd of any pair of structural decoys in the ith bin is $> i/2$ Å. For the purpose of native-like decoy generation, contact information from native structures is partially incorporated as loose tertiary restraints into the CABS force field. Therefore, the low (high) crmsd decoys are mainly from low (high) temperature replicas.

During the generation of the decoys, we have an energy term in our force field, which puts a propensity to the predicted secondary structure. The purpose of this energy term is to help the decoy conformations keep protein-like local structures. However, if other interactions, such as short-range correlation, distant restraints, or steric clashes, do not favor the predicted secondary structures, the local conformations can still be different from the predicted or native secondary structures. For the side-chain construction, two kinds of rotamers have been used, corresponding to residues with extended or helical secondary structure. Depending on the local backbone structures, the appropriate side-chain packing is produced under the guide of the CABS potential. Therefore, although on average the protein-like secondary structure and the side-chain packing are well preserved in the decoy conformations, variation is still allowed because of the local structure flexibility in CABS model. As the quality of the models deteriorate, the protein-like secondary structure and the side-chain packing become worse. The global and local structural features of these decoys closely resemble those of the predicted protein structures generated by the TOUCHSTONE II (Zhang *et al*., 2003) and TASSER (Zhang and Skolnick, 2003) *ab initio* algorithms.

RESULTS AND DISCUSSION

Overview of the obtained AFTs

The different stages of AFT generation are described in Systems and methods section, and outlined in the flowchart of Figure 1. The application of our procedure to all known enzymes results in 605 AFTs associated with 167 different enzymes. After the sensitivity and specificity assessment step, the number is reduced to 593 AFTs for 162 enzymes, whose EC numbers are shown in Table 1. The fractions of enzymes listed in Table 1 belonging to each EC class are similar to those from Release 30 of the Enzyme nomenclature database, of March 2003 (http://us.expasy.org/enzyme/), with some overrepresentation of class 3.

As shown in Table 2, the validated AFTs cover 24% of the known sub-subclasses of enzymes, but only 4% of the total number of existent enzymes as defined by the four components of the EC number. However, since our procedure is strictly structure-dependent, it is also useful to refer our results to the number of enzymes with at least one solved structure. In that case, the percentage increases to 19%. In Systems and methods section, we explain all the reasons that

Table 1. List of the 162 enzymes defined by the 593 AFTs

Class 1	**Class 3**				**Class 4**
1.1.1.23	3.1.1.3	3.1.1.7	3.1.1.45	3.1.1.47	4.1.1.50
1.1.1.27	3.1.1.61	3.1.1.74			4.1.2.13
1.1.1.28	3.1.2.22				4.1.3.7
1.1.1.29	3.1.3.1	3.1.3.2	3.1.3.16		4.2.1.1
1.1.1.37	3.1.11.2				4.2.1.11
1.1.1.95	3.1.21.1	3.1.21.4			4.2.1.24
1.1.3.15	3.1.26.4				4.2.99.18
1.1.3.38	3.1.27.1	3.1.27.3	3.1.27.4	3.1.27.5	4.6.1.1
1.3.99.1	3.1.31.1				**Class 5**
1.4.3.6	3.2.1.1	3.2.1.3	3.2.1.4	3.2.1.10	5.1.2.2
1.7.2.2	3.2.1.17	3.2.1.18	3.2.1.60	3.2.1.68	5.3.1.5
1.11.1.5	3.2.1.91	3.2.1.135			5.3.3.1
1.11.1.6	3.3.2.3	3.3.2.6			5.4.2.1
1.11.1.7	3.4.11.5	3.4.11.10	3.4.11.18		5.5.1.1
1.11.1.10	3.4.16.5	3.4.16.6			**Class 6**
1.11.1.11	3.4.17.1	3.4.17.2	3.4.17.8	3.4.17.15	6.3.2.4
1.14.99.1	3.4.17.18				6.3.5.2
1.15.1.1	3.4.19.3	3.4.19.12			
1.17.4.1	3.4.21.1	3.4.21.4	3.4.21.5	3.4.21.6	
Class 2					
2.1.1.6	3.4.21.7	3.4.21.9	3.4.21.10	3.4.21.12	
2.1.4.1	3.4.21.20	3.4.21.21	3.4.21.22	3.4.21.26	
2.1.4.2	3.4.21.32	3.4.21.35	3.4.21.36	3.4.21.37	
2.3.1.41	3.4.21.39	3.4.21.42	3.4.21.46	3.4.21.47	
2.3.2.13	3.4.21.50	3.4.21.59	3.4.21.62	3.4.21.64	
2.4.1.19	3.4.21.66	3.4.21.68	3.4.21.69	3.4.21.71	
2.4.2.36	3.4.21.73	3.4.21.79	3.4.21.80	3.4.21.81	
2.7.1.11	3.4.21.82	3.4.21.87	3.4.21.88	3.4.21.97	
2.7.1.37	3.4.22.1	3.4.22.2	3.4.22.14	3.4.22.16	
2.7.1.40	3.4.22.17	3.4.22.25	3.4.22.30	3.4.22.38	
2.7.1.50	3.4.22.39	3.4.22.40	3.4.22.43		
2.7.2.8	3.4.24.7	3.4.24.11	3.4.24.16	3.4.24.17	
2.7.7.7	3.4.24.21	3.4.24.23	3.4.24.26	3.4.24.27	
2.7.7.19	3.4.24.28	3.4.24.29	3.4.24.34	3.4.24.40	
2.7.9.1	3.4.24.42	3.4.24.65	3.4.24.69		
2.8.1.1	3.5.1.1	3.5.1.38	3.5.1.52		
	3.5.2.6				
	3.5.4.4	3.5.4.5			
	3.8.1.5				

Table 2. Fraction of known enzymes covered by the AFTs

Components in the EC number	Known enzymes[a]	Enzymes with known structure[b]	Enzymes with tentative AFTs[c]	Enzymes with validated AFTs[d]
n_1. - . - . -	6	6 (100%)[e]	6 (100%)[e]	6 (100%)[e]
$n_1.n_2$. - . -	59	52 (88%)	27 (46%)	27 (46%)
$n_1.n_2.n_3$. -	211	160 (76%)	54 (26%)	54 (24%)
$n_1.n_2.n_3.n_4$	3780	871 (23%)	167 (4%)	162 (4%)

[a]Number of known enzymes at the corresponding level of nomenclature definition.
[b]Number of enzymes with at least one available PDB structure.
[c]Number of enzymes described by at least one of 605 non-validated AFTs.
[d]Number of enzymes described by at least one of 593 validated AFTs.
[e]Percentage of the number of known enzymes.

account for the failure of our approach in producing AFTs for the remaining structurally characterized enzymes. The main causes are: (a) insufficient functional annotation in Swiss-Prot and (b) insufficient structural information in the PDB.

Significant cutoffs for active site detection

The application of the AFTs to identify putative active sites requires a criterion to assess the significance of a hit. Thus, we compute a restrictive and a permissive drmsd cutoff for each AFT, as detailed in Systems and methods section. Figure 3 shows the cumulative relative frequency of the cutoffs for the 593 validated AFTs. The restrictive (permissive) cutoffs range from 0.12 to 1.73 Å (0.53 to 3.67 Å), with a median value of 0.76 Å (1.88 Å). Hits whose drmsd from the AFT are below the restrictive or the permissive cutoffs are considered of high or medium confidence, respectively (Fig. 2). The choice of a specific drmsd cutoff is a trade-off between sensitivity and specificity. The negative and the positive datasets include only experimentally determined structures, and by definition, the scanning of these structures using restrictive cutoffs achieves sensitivity and specificity values of 100% (see Systems and methods section). It is expected that the scanning of high-resolution structures of uncharacterized enzymes using restrictive cutoffs would results in similar values of sensitivity and specificity, provided that an AFT for the active site is available. Such restrictive cutoffs cannot, in general, be applied to low-resolution structures due to their higher chance of having a distorted active site. Thus, the use of the permissive cutoffs, which sacrifice specificity for sensitivity, allows the method to be applied to the detection of active sites in low-resolution structures.

Several AFTs have low-restrictive cutoffs due to one of the main limitations of the EC classification system: the description of the enzyme function is based on overall reaction, even though the structure–function correlation is higher at the level of partial chemical reactions (Babbitt, 2003). For example, our library includes an AFT for 1,4-α-D-glucan glucanohydrolase (EC 3.2.1.1) that has a good permissive cutoff (2.3 Å), but a poor restrictive cutoff (0.2 Å) as a consequence

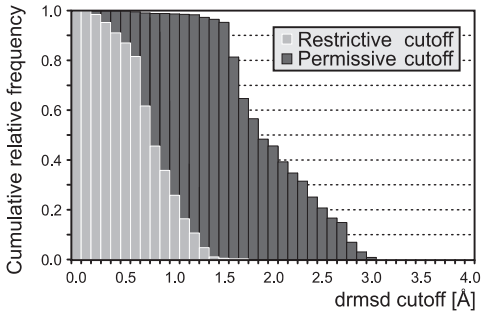

Fig. 3. Distribution of the drmsd cutoffs for active site identification.

of $drmsd_{MinNeg}$ (0.3 Å) being too close to $drmsd_{MaxPos}$ (0.1 Å). The negative hit that defines the $drmsd_{MinNeg}$ belongs to cyclomaltodextrin glucanotransferase (EC 2.4.1.19), an enzyme that degrades starch to cyclodextrins by formation of a 1,4-α-D-glucosidic bond. Thus, although even the first components of the EC numbers are different in these enzymes, they catalyze related partial reactions, which is reflected in the structural similarity exhibited by their active sites.

As mentioned in Systems and methods section, for calculation of the drmsd cutoffs, we remove from the negative set the structures of enzymes whose EC numbers share the first three components with the EC number under analysis. This is done because enzymes that belong to the same EC sub-subclass exhibit highly related active sites. For example, Wallace *et al*. (1996) developed a single functional descriptor able to specifically recognize the active site shared by as many as 13 different serine proteases (EC 3.4.21.-). Therefore, although each AFT is associated with a specific four-component EC number, the enzymes of the same sub-subclass were not considered as false positive hits during the validation process, implying a specificity of the AFT at the first three EC components level.

Detection of active sites in decoy structures

To estimate the applicability of our method to predicted protein structures with different levels of resolution, we scan the

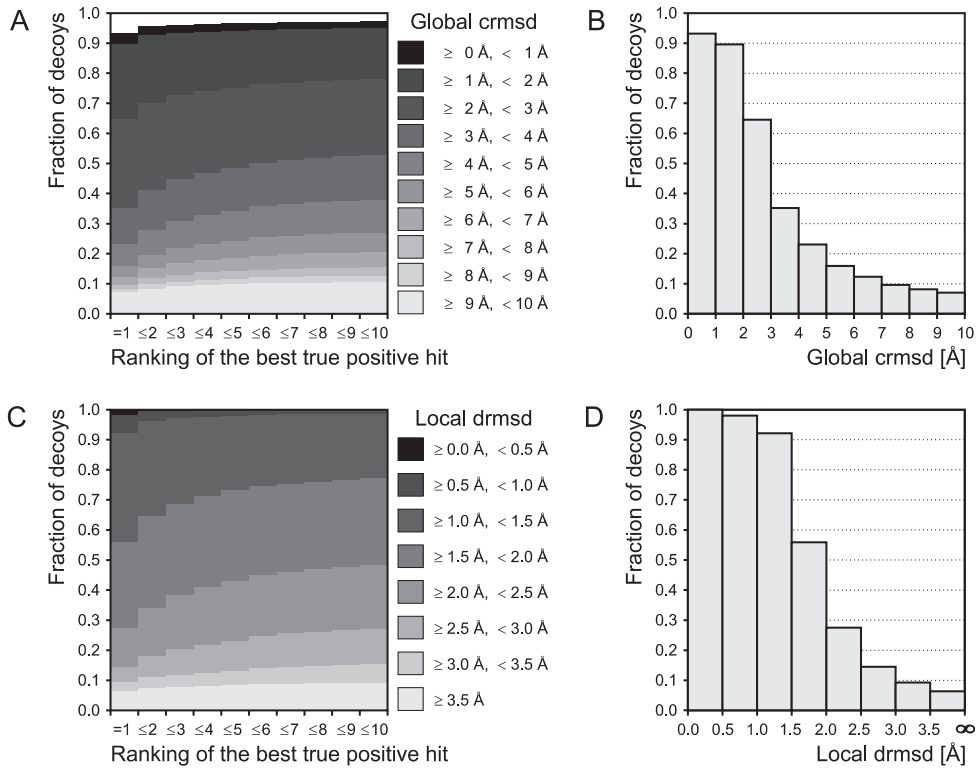

Fig. 4. Application of AFTs to decoy structures. The 593 000 decoys are distributed in bins according to: (**A** and **B**) the global Cα crmsd from the native structure, or (**C** and **D**) the local drmsd from the active site substructure. (A and C) Fraction of decoys correctly annotated versus ranking of the best true positive hit at different rmsd intervals. (B and D) Fraction of decoys correctly annotated (top 1 hit) versus rmsd interval. The recognition by an AFT matching the first three components of the true EC number is considered a true positive hit.

complete AFT library against decoy structures, using permissive drmsd cutoffs. For each AFT, we first select one representative structure from the set of positive structures utilized for its construction, and then generate 1000 decoys as described in Systems and methods section. After an all-against-all scanning of the 593 descriptors over the 593 000 decoys, we rank the medium confidence hits for each decoy in terms of increasing drmsd from the AFTs. In other words, for each decoy we generate a list of possible functions, ranked by the degree of structural similarity between the putative active site and the corresponding AFT. Then, we calculate the overall fraction of decoys whose function at the level of the first three EC components is correctly assigned in the top 1–10 ranking positions. We present the result of this analysis for two different partitions of the 593 000 decoys: (i) according to their global crmsd from the native structures (Fig. 4A and B), and (ii) according to the local drmsd from the substructure recognized by the corresponding AFT in the native structure (Fig. 4C and D). The same analysis using restrictive cutoffs or different definitions of the permissive cutoff shows a severe drop in the fraction of decoys correctly assigned (data not shown), confirming that permissive cutoffs are more appropriate for biochemical function prediction in low-resolution structures. The global crmsd reflects an averaged distortion for the entire structure, but a substructure belonging to a set of models of a given global crmsd, still spans a considerable range of local distortions. In particular, functional substructures such as active sites tend to be better defined than randomly selected substructures. Thus, even though the permissive cutoffs range from 0.53 to 3.67 Å drmsd, they can partially detect distorted functional substructures in models whose global Cα crmsd from the native structure are in the interval of 3–4 Å. However, the lower the permissive cutoff of the AFT, the higher the quality of the model should be in order to generate significant results.

The graphs presented in Figure 4 allow us to estimate the expected success rate of our method when it is applied to predicted structures. For instance, for models in the range of 3–4 Å Cα crmsd from the native structure, one can expect a correct functional assignment in 35% of the cases (Fig. 4A and B); if the top 5 hits are considered, that success rate increases to 48% (Fig. 4A). As a result of local similarities shared by a subset of the active sites recognized by our AFT library, the top hit detects at least one true functional residue in 76% of the decoys belonging to the abovementioned range of crmsd, independent of the correctness of the functional assignment. Similarly, for models with local distortions in the range of 1.5–2.0 Å drmsd from the active site

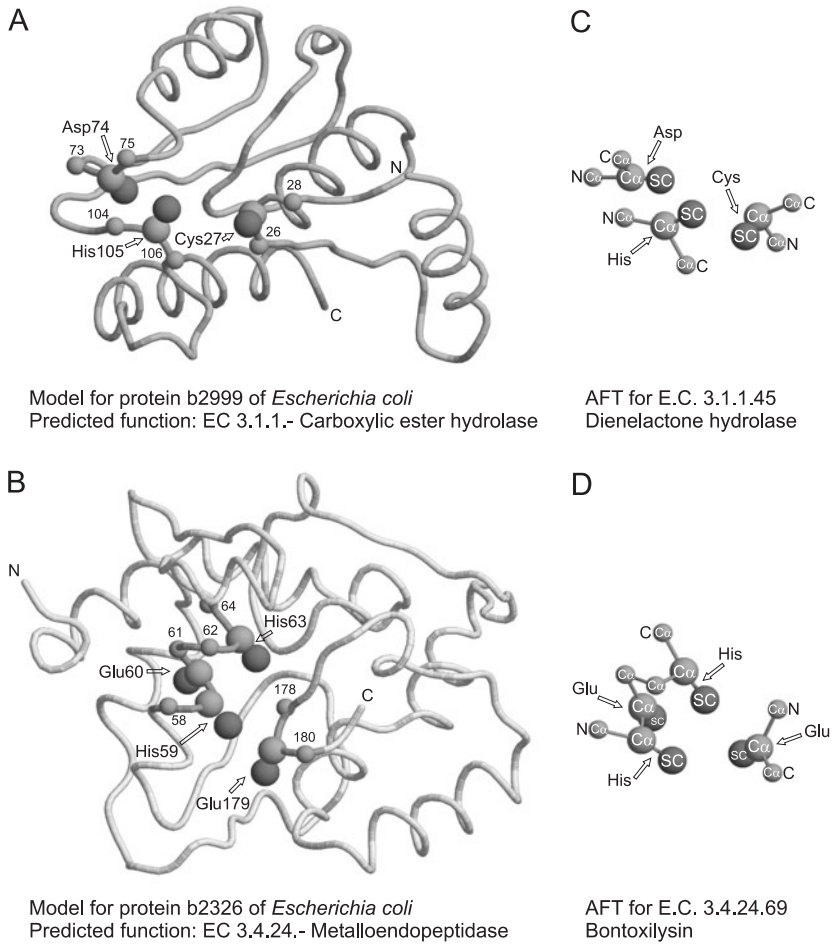

Fig. 5. Application of AFTs to predicted structures. Models for *E.coli* proteins b2999 (**A**) and b2326 (**B**). The spheres represent the location of the hits by an AFT for EC 3.1.1.45 (**C**) in protein b2999 and an AFT for EC 3.4.24.69 (**D**) in protein b2326. Light gray spheres correspond to Cα atoms and dark gray spheres denote the SC pseudo-atoms.

(typically corresponding to 2.5–3.4 Å crmsd), success rates of 56% (Fig. 4C and D) and 73% (Fig. 4C) can be expected for the top 1 and top 5 hits, respectively.

Detection of active sites in predicted structures

In order to illustrate the scenarios of function annotation in which our approach can be useful, we present two examples of the application of AFTs to predicted structures of *E.coli* proteins. Both models (Fig. 5A and B) were generated using TASSER, a new hierarchical approach to protein structure prediction that consists of the template identification by threading, followed by the assembly of tertiary structures via rearranging continuous template fragments under the guide of an optimized Cα and side chain based potential (Zhang and Skolnick, 2003). Foldability estimators calculated by TASSER indicate that the model for b2999 (Fig. 5A) is of very high quality, and the model for b2326 (Fig. 5B) is of low quality, consistent with its absence of strong threading templates. In order to retrieve the most up-to-date annotation for these test cases, we consulted two different databases: (i) GenProtEC (Riley, 1998) of September 17, 2003 that includes information from multiple sources including sequence similarity to orthologs as defined by Darwin (Gonnet *et al*., 2000), identification of groups with similar protein sequence within *E.coli* that are clustered by transitive relationships, and updated literature references; and (ii) Swiss-Prot Release 42.1 and TrEMBL Release 25.1 of October 24, 2003.

The first example corresponds to protein b2999 or ORF_F136, the 136 amino acid product of the *E.coli* gene *YGHX*. The functional assignment for b2999 in GenProtEC is ‘putative enzyme with alpha/beta-hydrolase-like domain’; and TrEMBL (entry Q46849) shows no information about function, but displays cross-references to: (i) Gene Ontology (Ashburner *et al*., 2000) term GO:0016787, which corresponds to ‘hydrolase activity (EC 3.-.-.-) inferred from electronic annotation’ and (ii) PFAM (Bateman *et al*., 2002) PF01738, dienelactone hydrolase family (EC 3.1.1.45). A Blast search of b2999 against EXProt (Ursing *et al*., 2002),

a database of proteins with experimentally verified function, shows that the best hit is in fact a dienelactone hydrolase (Swiss-Prot O67988). However, the identity between both sequences over the region aligned by BLAST is 29% (37/127 residues), a value below the 40% sequence identity threshold required to transfer the first three EC components to an uncharacterized sequence (Tian and Skolnick, 2003). When we scan our model for b2999 (Fig. 5A) with the AFT library, we find a hit (with a 0.81 Å drmsd) below the permissive cutoff (1.86 Å drmsd) for the AFT associated with EC 3.1.1.45 (Fig. 5C). Therefore, our predicted function for this target is carboxylic ester hydrolase (EC 3.1.1.-). This test case exemplifies one of the situations in which the AFT approach can be advantageous, i.e. when the level of sequence identity is sufficient to detect an evolutionary relationship and build a good model, but not enough for a confident homology-based functional annotation. In this case, the AFT method generates a high confidence prediction, by verifying the compatibility between a known active site and the spatial arrangement of the putative active site residues in the model.

The second example corresponds to protein b2326 or yfcM, the 182 amino acid product of the *E.coli* gene *YFCM*. The functional assignment for b2326 in GenProtEC is 'unknown'; and neither Swiss-Prot (entry P76938) nor the databases cross-referenced by Swiss-Prot shows any information about function. A Blast search of b2326 against EXProt does not produce any significant hit. The scanning of the model for b2326 (Fig. 5B) with the AFT library produces a hit (with a 1.50 Å drmsd) below the permissive cutoff (2.22 Å drmsd) for the AFT associated with EC 3.4.24.69 (Fig. 5D). Thus, our functional annotation for this target is metalloendopeptidase (EC 3.4.24.-). This test case represents a different scenario for the application of the AFT method; i.e. when absolutely no functional information can be transferred by homology, and no structural relatives can be easily recognized. In this case, the confidence of the functional prediction is limited by the estimated quality of the model. Still, the prediction can be useful as a hint for experimentalists interested in a specific target.

The main advantage of the AFT approach is that, as opposite to the annotation transfer based on homology that depends on global parameters such as sequence identity, the AFT focuses on the 3D arrangement and amino acid type of a few putative active site residues. Thus, our method is potentially able to detect new relationships between active sites and fold types. As pointed out by Todd *et al*. (2001) we are likely to observe even more extensive and unexpected variations in function within many superfamilies in the future. On the other hand, the most important limitations of the AFT strategy are: (i) its dependency of the quality of the protein model, which cannot always be assessed and (ii) the reduced number of different EC numbers for which experimental structures are available, which limits the number of AFTs and consequently the number of predictions in a large-scale analysis.

CONCLUSION

We have described a method that automatically generates a library of functional 3D descriptors (automated functional templates or AFTs) for the structure-based prediction of enzyme active sites. We have shown that 593 AFTs specifically recognize, at the level of the first three components of the EC number, the experimentally determined structures currently available for 162 different enzymes. We have performed a large-scale test to measure the scope of the AFT approach on low-resolution structures. Previously, based on the analysis of calcium-binding sites, Wei *et al*. (1999) questioned the utility of predicted protein structures for identification of functional active sites. However, in 35% of the cases, our method can correctly identify the biochemical function of enzymes whose models are in the range of 3–4 Å crmsd from the native structure, a quality that is reachable using state of the art *ab initio* algorithms for protein structure prediction (Zhang *et al*., 2003; Zhang and Skolnick, 2003). For instance, 46% of the sequences belonging to a benchmark set composed by 1489 non-homologous proteins with 41–200 residues, the TASSER algorithm can generate models whose global Cα atom crmsd is below 4 Å (Zhang and Skolnick, 2003). Our next step will be to apply the AFT method on a large-scale to predicted structures of proteins with uncharacterized biochemical function. By greatly reducing the number of functions to be tested, the predictions generated by our approach would significantly accelerate experimental screening for the determination of enzyme activity.

ACKNOWLEDGEMENTS

This research was supported in part by National Institutes of Health Grant GM-48835. A.K.A. also acknowledges partial support from the Pew Latin American Fellows Program in Biomedical Sciences.

REFERENCES

Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J.D. (1994) *Molecular Biology of the Cell*. Garland Pub., New York, pp. xliii, 1294 [1267].

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.

Babbitt,P.C. (2003) Definitions of enzyme function for the structural genomics era. *Curr. Opin. Chem. Biol*., **7**, 230–237.

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.

Bartlett,G.J., Porter,C.T., Borkakoti,N. and Thornton,J.M. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol*., **324**, 105–121.

Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and

Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res*., **30**, 276–280.

Bonneau,R., Strauss,C.E., Rohl,C.A., Chivian,D., Bradley,P., Malmstrom,L., Robertson,T. and Baker,D. (2002) De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol*., **322**, 65–78.

Bork,P., Ouzounis,C. and Sander,C. (1994) From Genome Sequences to Protein Function. *Curr. Opin. Struct. Biol*., **4**, 393–403.

Devos,D. and Valencia,A. (2000) Practical limits of function prediction. *Proteins*, **41**, 98–107.

Devos,D. and Valencia,A. (2001) Intrinsic errors in genome annotation. *Trends Genet*., **17**, 429–431.

Fetrow,J.S. and Skolnick,J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol*., **281**, 949–968.

Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol*., **1**, 55–67.

Gonnet,G.H., Hallett,M.T., Korostensky,C. and Bernardin,L. (2000) Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics*, **16**, 101–103.

Hamelryck,T. (2003) Efficient identification of side-chain patterns using a multidimensional index tree. *Proteins*, **51**, 96–108.

Hegyi,H. and Gerstein,M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol*., **288**, 147–164.

Holm,L. and Sander,C. (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res*., **22**, 3600–3609.

Iliopoulos,I., Tsoka,S., Andrade,M.A., Enright,A.J., Carroll,M., Poullet,P., Promponas,V., Liakopoulos,T., Palaios,G., Pasquier,C. *et al*. (2003) Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics*, **19**, 717–726.

Jonassen,I., Eidhammer,I. and Taylor,W.R. (1999) Discovery of local packing motifs in protein structures. *Proteins*, **34**, 206–219.

Kimber,M.S. and Pai,E.F. (2000) The active site architecture of *Pisum sativum* beta-carbonic anhydrase is a mirror image of that of alpha-carbonic anhydrases. *EMBO J*., **19**, 1407–1418.

Kleywegt,G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol*., **285**, 1887–1897.

Koonin,E.V., Tatusov,R.L. and Rudd,K.E. (1996) Protein sequence comparison at genome scale. *Methods Enzymol*., **266**, 295–322.

Liang,M.P., Brutlag,D.L. and Altman,R.B. (2003) Automated construction of structural motifs for predicting functional sites on protein structures. *Pac. Symp. Biocomput*., 204–215.

Lowther,W.T., Weissbach,H., Etienne,F., Brot,N. and Matthews,B.W. (2002) The mirrored methionine sulfoxide reductases of Neisseria gonorrhoeae pilB. *Nat. Struct. Biol*., **9**, 348–352.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.

Oldfield,T.J. (2002) Data mining the protein data bank: residue interactions. *Proteins*, **49**, 510–528.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Pennec,X. and Ayache,N. (1998) A geometric algorithm to find small but highly similar 3D substructures in proteins. *Bioinformatics*, **14**, 516–522.

Riley,M. (1998) Genes and proteins of *Escherichia coli* K-12. *Nucleic Acids Res*., **26**, 54.

Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol*., **318**, 595–608.

Russell,R.B. (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol*., **279**, 1211–1227.

Skolnick,J., Fetrow,J.S. and Kolinski,A. (2000) Structural genomics and its importance for gene function analysis. *Nat. Biotechnol*., **18**, 283–287.

Skolnick,J., Zhang,Y., Arakaki,A.K., Kolinski,A., Boniecki,M., Szilagyi,A. and Kihara,D. (2003) TOUCHSTONE: a unified approach to protein structure prediction. *Proteins*, **53**, 469–479.

Tian,W. and Skolnick,J. (2003) How well is enzyme function conserved as function of pairwise sequence identity? *J. Mol. Biol*., **333**, 863–882.

Todd,A.E., Orengo,C.A. and Thornton,J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol*., **307**, 1113–1143.

Ursing,B.M., van Enckevort,F.H., Leunissen,J.A. and Siezen,R.J. (2002) EXProt: a database for proteins with an experimentally verified function. *Nucleic Acids Res*., **30**, 50–51.

Wallace,A.C., Borkakoti,N. and Thornton,J.M. (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci*., **6**, 2308–2323.

Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1996) Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci*., **5**, 1001–1013.

Wei,L., Huang,E.S. and Altman,R.B. (1999) Are predicted structures good enough to preserve functional sites? *Struct. Fold Des*., **7**, 643–650.

Wilson,C.A., Kreychman,J. and 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.

Zhang,Y., Kihara,D. and Skolnick,J. (2002) Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins*, **48**, 192–201.

Zhang,Y., Kolinski,A. and Skolnick,J. (2003) TOUCHSTONE II: a new approach to *ab initio* protein structure prediction. *Biophys. J*., **85**, 1145–1164.

Zhang,Y. and Skolnick,J. (2003) Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl Acad. Sci., USA*, submitted.

Zhao,S., Morris,G.M., Olson,A.J. and Goodsell,D.S. (2001) Recognition templates for predicting adenylate-binding sites in proteins. *J. Mol. Biol*., **314**, 1245–1255.