BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 13 2005, pages 2969–2977
doi:10.1093/bioinformatics/bti471

Structural bioinformatics

FSSA: a novel method for identifying functional signatures from structural alignments

Kai Wang and Ram Samudrala*

Computational Genomics Group, Department of Microbiology, University of Washington, Seattle, WA 98195, USA

Received on January 20, 2005; revised and accepted on April 25, 2005
Advance Access publication April 28, 2005

ABSTRACT

Motivation: It is commonly believed that sequence determines structure, which in turn determines function. However, the presence of many proteins with the same structural fold but different functions suggests that global structure and function do not always correlate well.

Results: We propose a method for accurate functional annotation, based on identification of functional signatures from structural alignments (FSSA) using the Structural Classification of Proteins (SCOP) database. The FSSA method is superior at function discrimination and classification compared with several methods that directly inherit functional annotation information from homology inference, such as Smith–Waterman, PSI-BLAST, hidden Markov models and structure comparison methods, for a large number of structural fold families. Our results indicate that the contributions of amino acid residue types and positions to structure and function are largely separable for proteins in multi-functional fold families.

Availability: The FSSA software is available at http://software.compbio.washington.edu/fssa

Contact: ram@compbio.washington.edu

Supplementary information: http://data.compbio.washington.edu/fssa/bioinformatics_supplement

INTRODUCTION

The success of structural genomics initiatives requires the development and application of tools for structure analysis, prediction and annotation (Goldsmith-Fischman and Honig, 2003). Once the structures are determined experimentally, one of the biggest challenges is to infer their biological and physiological functions. Several methods have been used widely to infer functional knowledge from structural information, when sequence data alone are not enough to infer function confidently (Thornton *et al*., 2000; Teichmann *et al*., 2001). For a given structure, comparison of structural folds (Taylor and Orengo, 1989; Shindyalov and Bourne, 1998; Holm and Sander, 1999; Ortiz *et al*., 2002) or sequential structural motifs (Jonassen *et al*., 1999, 2002; Kasuya and Thornton, 1999; Jones *et al*., 2003) with other proteins with known function may give insights about its function. In addition, the essence of biochemical function can be captured from structural motifs, independent of the overall fold (Kobayashi and Go, 1997; Kleywegt, 1999; Barker and Thornton, 2003; Jambon *et al*., 2003; Stark and Russell, 2003; Pazos and Sternberg, 2004). With the development of novel algorithms (Russell and Barton, 1992; Yang and Honig, 2000; Guda *et al*., 2001; Leibowitz *et al*., 2001b; Dror *et al*., 2003), multiple structural alignments may also be used to infer function, with better discrimination power than pairwise comparison methods (Leibowitz *et al*., 2001a). However, in the absence of significant sequence and structure similarities, other prediction methods must be used: for example, the size of clefts on the surface of a protein may be used to predict enzyme function (Laskowski *et al*., 1996), while protein surface patches may be used to analyze protein–protein interactions (Jones and Thornton, 1997).

Although it is commonly believed that structure determines biological function, protein global structure and function do not always correlate well with each other, since only a limited number of structural folds are expected to be found in nature (Orengo *et al*., 1999). Given the large number of functions exerted by cellular proteins, this suggests that some diverse and distinct functions must be derived from the same structural folds (Anantharaman *et al*., 2003). Todd *et al*. (1999, 2001, 2002) have shown examples of a variety of biochemical functions that are performed by proteins with the same structural fold, or even by members of a single homologous family. The TIM barrel proteins, which have eight alpha/beta motifs folded into a barrel structure, are the most frequently observed folds in nature (Branden, 1991), and are probably the most famous example of a multi-functional fold family (Nagano *et al*., 2002). The Structural Classification of Proteins (SCOP) scheme (Murzin *et al*., 1995) is a widely used classification method that classifies protein structures into hierarchical levels of class, fold, superfamily and family to embody structural and evolutionary relationships. Proteins within the same SCOP superfamily suggest common evolutionary origin, and there are 26 superfamilies within the TIM barrel fold. Some other famous and well-studied multi-functional fold families include proteins with the immunoglobulin fold, the RRM-like fold, the HUP fold and the Rossman fold.

The fact that multi-functional fold families exist in nature suggests that the contribution of amino acid residue types and positions to protein structure and function may be largely separable. The analysis of local structure profiles within a fold family, in the context of protein function, may thus provide insights into the functional role of specific amino acid residue types and positions, where local structure is defined as a distinct spatial organization composed of a few amino acid residues. Studies have been reported on such structure–function relationships among a group of structurally similar proteins: Matsuo and Bryant (1999) presented a concept called homologous core structures (HCS), which is defined as the subset of C_α coordinates whose spatial locations are conserved across structure–structure alignments with previously identified homologues. They showed that discrimination between homologues and analogues, on the basis of HCS overlap, is clearly superior to discrimination by local root mean

*To whom correspondence should be addressed.

square (RMS) superposition residual, the percentage of identical residues, or structure–structure alignment length as a fraction of domain length. Russell *et al*. (1998) presented a method to assess the significance of binding site similarities within superimposed protein three-dimensional structures, and applied it to all similar structures in the Protein Data Bank (PDB). The supersites were defined as structural locations on groups of analogous proteins (i.e. superfolds) showing a statistically significant tendency to bind substrates, despite little evidence of a common ancestor for the proteins considered. The analysis of these supersites may, thus, provide a guide for predicting function from structure.

These studies in total suggest that we can retrieve functional information by analyzing subtle structural differences in proteins sharing the same fold. The method we propose here focuses on the analysis of distribution of local structure profiles in a group of proteins with the same structural fold, where 'local structure profile' refers to a combination of local structure and local sequence similarity. The similarities of local structures are usually indicative of functional conservation, and have been used in the discrimination of SCOP superfamilies (Hou *et al*., 2003). In addition, it has been reported that active-site structural similarity, rather than overall structural similarity, can better describe the functional profile (Fetrow and Skolnick, 1998; Cammer *et al*., 2003), and some structure-based functional descriptors have been used for function classification (Di Gennaro *et al*., 2001; Stark and Russell, 2003; Pazos and Sternberg, 2004). In addition to local structure similarity, local sequence similarity can also be indicative of functional importance, and it forms the basis of motif-based methods to search for functionally important residues (Henikoff *et al*., 2000; Attwood *et al*., 2003; Hulo *et al*., 2004). Since those proteins in a multi-functional fold family may be classified into distinct functional categories, we hypothesize that functionally important residues tend to adopt the same local structure profiles in the same category, but have diverse local structure profiles across different categories. On the other hand, structurally important residues may adopt local conformations that are largely independent of function. Therefore, we can estimate the probability of a residue being functionally important, based on its local structure profile conservation in the same functional category, relative to conservation in different functional categories. Based on this hypothesis, we have developed a method called functional signature from structural alignments (FSSA), to estimate the log odds of a residue being functionally important, relative to its structural importance. For every protein, the collection of log odds scores for all its residues comprises its 'functional signature'. The functional signature may be used to predict function for a new query structure that is known to adopt a certain structural fold. We evaluated the performance of the FSSA method in function discrimination experiments and function classification experiments using datasets from the SCOP database. The FSSA method has displayed good performance overall in these experiments, and it can be used to supplement other function prediction methods based on global sequence and structure comparison.

METHODS

Data source

The domain structures and corresponding sequences used in our analysis were downloaded from the ASTRAL database (Chandonia *et al*., 2004) version 1.67. We pre-processed each of the structures and renumbered the residues to make them consecutive. A few structures with large missing segments (consecutive C_α atoms more than 10 Å away) were not used in our study, since structure alignment programs cannot reliably align them. Structures, with and without ligands bound, were treated in the same manner owing to the small percentage of unliganded structures available, even though ligand binding is likely to have an effect on local structure.

Construction of functional signatures

In our study, we define proteins within the same SCOP superfamily as homologues, while those belonging to different superfamilies but possessing the same fold as structural analogues. Suppose we have N domain structures with the same structural fold, and they are classified into several functional categories. For each structure $S_i (1 \leq i \leq N)$ with length $L_i (1 \leq i \leq N)$ we perform a global structure alignment with every other structure, using the MAMMOTH structure comparison program (Ortiz *et al*., 2002). MAMMOTH is a fast and accurate program that performs sequence-independent structure alignments using C_α backbone coordinates. MAMMOTH uses the URMS Distance (Kedem *et al*., 1999) between two heptapeptides to define whether or not two segments have similar local structure and annotates them by '*' in the alignment outputs. Gaps in the alignments are treated as non-matches, and they generally only account for a small percentage of all non-matched residues. For each amino acid residue $R_{ij} (1 \leq i \leq N, 1 \leq j \leq L_i)$ in the structure S_i, we count the frequencies of similarity of local structure profiles in structures in the same functional category and in different functional categories. Similar local structure profile refers to both similar local structures (as judged by the annotation in the MAMMOTH output) and similar amino acid residue types (residue pairs where the BLOSUM50 matrix score ≥ 0). We then calculate the likelihood ratio (LR) and log-likelihood ratio (LLR) score for R_{ij}, as represented by the logarithms of the ratio of the two frequencies:

$$\mathrm{LR}_{ijm} = \frac{\mathrm{counts_{hm}}/\mathrm{counts_{h}}}{\mathrm{counts_{am}}/\mathrm{counts_{a}}},$$

$$\mathrm{LR}_{ijn} = \frac{(\mathrm{counts_{h}} - \mathrm{counts_{hm}})/\mathrm{counts_{h}}}{(\mathrm{counts_{a}} - \mathrm{counts_{am}})/\mathrm{counts_{a}}},$$

$$\mathrm{LLR}_{ijm} = \log(\mathrm{LR}_{ijm}),$$

$$\mathrm{LLR}_{ijn} = \log(\mathrm{LR}_{ijn}),$$

where LLR_{ijm} and LLR_{ijn} represent the log-likelihood ratio of finding matched local structure profiles and not finding matched local structure profiles in homologous proteins for residue R_{ij} in structure S_i, respectively. $\mathrm{counts_h}$ and $\mathrm{counts_a}$ represent the number of homologous and structurally analogous proteins, respectively. $\mathrm{counts_{hm}}$ and $\mathrm{counts_{am}}$ represent the number of homologous proteins with matching local structure profiles and the number of structurally analogous proteins with matching local structure profiles, respectively. Pseudocounts are used when $\mathrm{counts_{hm}}$ or $\mathrm{counts_{am}}$ are equal to zero. The collection of LLR_{ijm} for all residues in a structure S_i represents the functional signature for this structure.

Calculation of posterior odds for a query structure

All structures with a known functional signature are used as reference structures to classify a query structure, with the same fold, into a particular functional category. After performing structure alignment between a reference structure $S_i (1 \leq i \leq N)$ and the query structure, the collection of residues with matching local structure in S_i is L_M and $L_M \subseteq (1, 2, \ldots, L_i)$. According to Bayes' rule, the log posterior odds that the query structure belongs to the same functional category as the reference structure S_i, can be expressed as:

$$\log(\mathrm{odds(posterior)}) = \log(\mathrm{odds(prior)}) + \sum_{j \in L_M} \mathrm{LLR}_{ijm} + \sum_{j \notin L_M} \mathrm{LLR}_{ijn}.$$

For a query with an unknown function, the log prior odds can be treated as a constant for a given functional category. Usually the use of Bayes' rule requires data independence assumption, which means that the likelihood ratios for different residues are independent. Given the fact that usually

only a few residues are functionally important and well conserved in a given structure, this assumption is relaxed here. When we have the posterior log odds score for every reference structure, we assign the function of the query structure into the functional category that has the highest average log odds scores.

Function discrimination experiments

The purpose of these experiments was to test whether an algorithm can confidently discriminate homologues from structural analogues. We evaluated the performance by the receiver operator characteristic (ROC) area under curve scores. ROC is a widely used means to evaluate the discrimination ability of binary classification methods, when the test results are continuous measures. ROC curves display the relationship between sensitivity (true positive rate) and 1−specificity (false positive rate) across all possible threshold values that define the positivity of a condition (in our case, whether a domain structure belongs to a particular SCOP superfamily). The area under the ROC curve ranges from 0 to 1 with a higher score indicating better discriminatory power. Protein domains within each family were used as positive testing samples, and domains outside the family but within the same superfamily, were used as positive training samples. Negative samples were all domains outside the superfamily but within the same structural fold family, and were randomly split into training and testing sets in the same ratio as the positive samples. This yielded 37 SCOP families containing at least 5 family members (positive testing set), at least 5 superfamily members outside of the family (positive training set) and at least 10 members outside the superfamily, but within the same fold (negative training and testing sets). The ROC scores were calculated for the positive and negative testing samples for different discrimination methods as described below.

We compared the performance of several function discrimination methods. For the Smith–Waterman sequence alignment method, we used the programs search in the FASTA program suite version 3.4 (Pearson and Lipman, 1988). We searched a given query sequence against every sequence in a training set, using default parameters, and kept the lowest E-value found for this sequence. For a group of query sequences containing both positive and negative samples, we calculated the ROC score based on their E-values. For the PSI-BLAST method, we used the program blastpgp in the NCBI-BLAST program suite version 2.2.6 (Altschul *et al*., 1997). We searched a given query sequence against all sequences in a training set, using three iterations and all other default parameters; we then used the lowest E-value for this sequence for the calculation of ROC score. For the hidden Markov model (HMM) method, we used the program HMMER version 2.3.1 (Eddy, 1998). Although pre-generated HMMs are available from the Pfam database, these models contain information on our testing set; we, therefore, constructed a multiple sequence alignment using CLUSTAL W version 1.83 (Thompson *et al*., 1994) for each superfamily, and then built a HMM using the resulting multiple alignment. For a given query sequence, we aligned it with the HMM and used the E-value for the calculation of ROC score. All the E-values were used here to measure relative similarity without stringent statistical meaning, since all the database sequences were similar to the query and they violated the 'sequence unrelatedness' assumption to calculate accurate E-values. For the global and local structure comparison methods, we used the programs MAMMOTH (Ortiz *et al*., 2002) and CE (Shindyalov and Bourne, 1998), respectively. We searched every query structure against a training set, and used the highest Z-score for the calculation of the ROC score. For the FSSA method, we trained a model using a training set, calculated the log odds score for every query sequence and used the calculated score for ROC evaluation.

Function classification experiments

Our goal here was to test how well a method can assign a query sequence with a known structural fold into a functional category, as defined by SCOP superfamily. We used those SCOP folds that were represented in the function discrimination experiments above. To investigate how performance changes with respect to homology among testing and training sequences, we used four different datasets retrieved from the ASTRAL database, representing proteins whose pairwise sequence identities were <=10, 20, 30 and 95%, respectively. For each fold in each dataset, those superfamilies with less than eight sequences were combined into a single 'OTHER' category, and those folds containing only one superfamily (excluding the 'OTHER' category) were not used. For each fold in each dataset, we then divided the corresponding sequences into four parts of similar sizes, ensuring that each functional category has approximately the same frequency in each part. In each of the four-fold cross-validation experiments, 75% of the sequences were used as a database and 25% of the sequences were used for queries. For the Smith–Waterman and PSI-BLAST methods, we searched each query sequence against the database and assigned the query into the same functional category with the database sequence having the lowest E-value. For the HMM method, we built and calibrated a model using the Clustal W and hmmbuild programs for each functional category using sequences in the database, and then used the hmmpfam program to assign each query into the functional category based on the lowest E-value. For structure comparisons, we used either the MAMMOTH or the CE program to search each query against the database structures, and assigned the query into the same functional category as the database structure with the highest Z-score. For the FSSA method, we assigned the query into the functional category that had the highest average posterior log odds score.

RESULTS

Construction of functional signatures

We constructed functional signatures for protein domains in the ASTRAL database (Chandonia *et al*., 2004) whose pairwise sequence identities are $\leq$30%, using the FSSA method. Figure 1 shows examples of functional signatures for proteins in the metallo-dependent hydrolase (SCOP superfamily identifier: c.1.9) and aldolase (SCOP superfamily identifier: c.1.10) superfamilies. Both superfamilies belong to the TIM barrel structural fold and contain similar numbers of proteins. The functional signature consists of a score for each residue in the protein domain, indicating the log odds of finding similar local structure profile in homologues from structural analogues. These signatures are somewhat similar to the idea of Homologous Core Structures (HCS) (Matsuo and Bryant, 1999), in that higher scores correspond to functionally more important residues. However, the construction of HCS uses whole-structure segments that can be aligned, while the construction of FSSA uses only individual residues with similar local structure profiles. In addition, the construction of HCS uses only structure information, while the construction of FSSA uses both structure and sequence information. Furthermore, HCS uses pairwise alignments between homologous proteins, whereas FSSA uses pairwise alignments between both homologous and structural analogues, thus enhancing signal for functionally important residues.

Figure 1 indicates that most domains in the metallo-dependent hydrolase superfamily have similar functional signatures, with the C-terminal portion of the protein having relatively higher log odds scores compared with the rest of the protein. A visual examination of the domain structures reveals that this region corresponds to an additional α-helix in the C-terminal end of the barrel. The helix functions as a 'cap' to the barrel, and is one of the criteria used to classify this SCOP superfamily. In comparison, the distributions of log odds scores for protein domains in the aldolase superfamily are more heterogeneous. For some protein domains (e.g. SCOP identifiers d1o5ka_ and d1f74a_), the highest log odds scores tend to accumulate around the C-terminal end of the sequence. But for other protein domains (e.g. SCOP identifiers d1pe1a_ and d1of8a_),

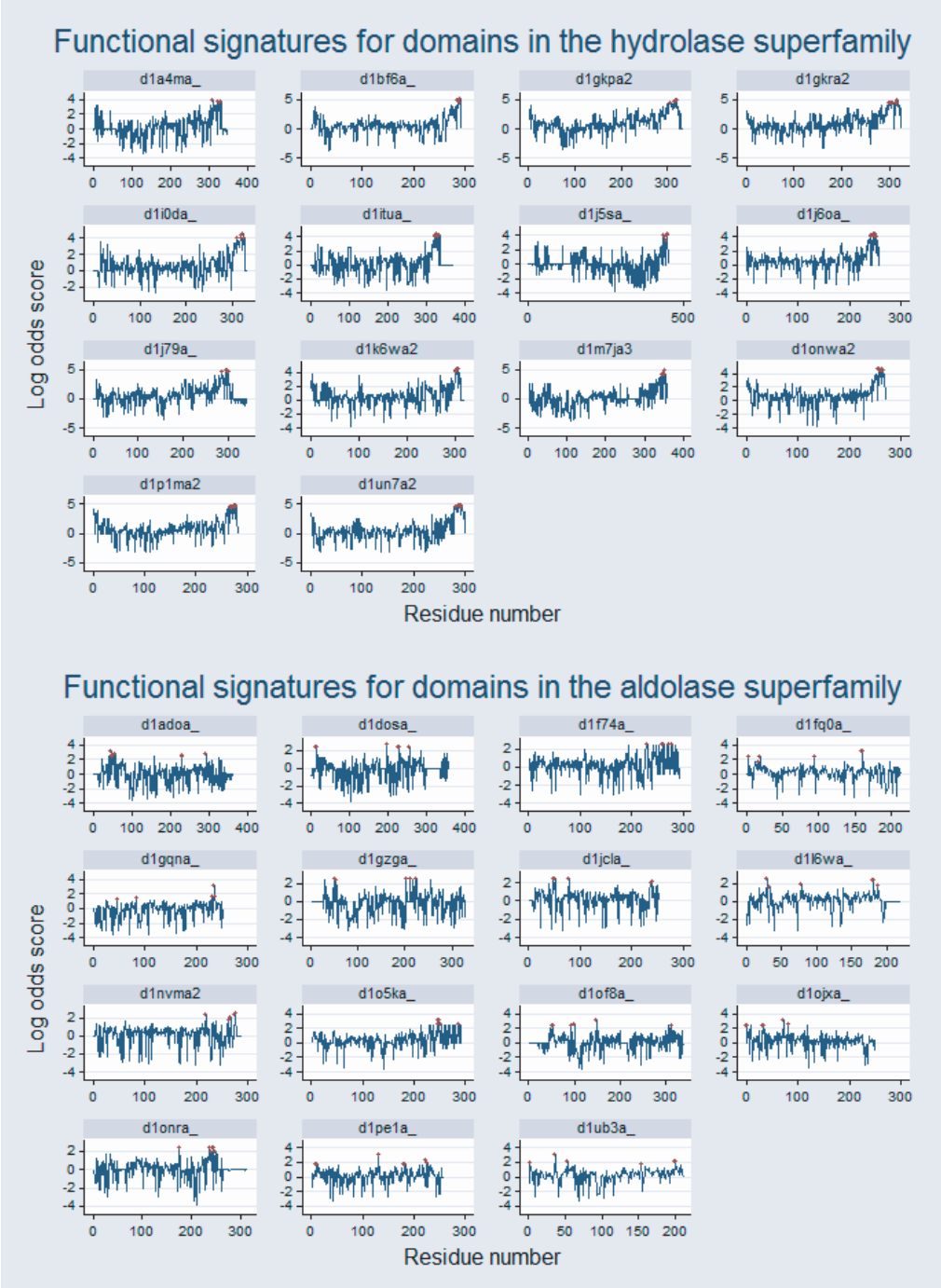

Fig. 1. Comparison of the functional signatures of protein domains within the metallo-dependent hydrolases (SCOP superfamily identifier: c.1.9) and the aldolase (SCOP superfamily identifier: c.1.10) superfamilies. The functional signature for each protein domain is represented by plotting the log odds score versus residue number. For each signature the five residues with the highest log odds scores are highlighted by red dot symbols. In general, domains in the hydrolase superfamily have similar signatures whereas domains in the aldolase superfamily have heterogeneous signatures.

the distribution of the highest log odds scores are scattered all over the sequence. The similarity of functional signatures for protein domains in a particular superfamily may thus dictate whether the FSSA method works well for that superfamily in function prediction applications.

Function discrimination experiments

The value of a function prediction method depends on whether it can successfully discriminate between homologues and structural analogues. We define proteins within the same SCOP superfamily as homologues, and proteins within the same SCOP fold but different

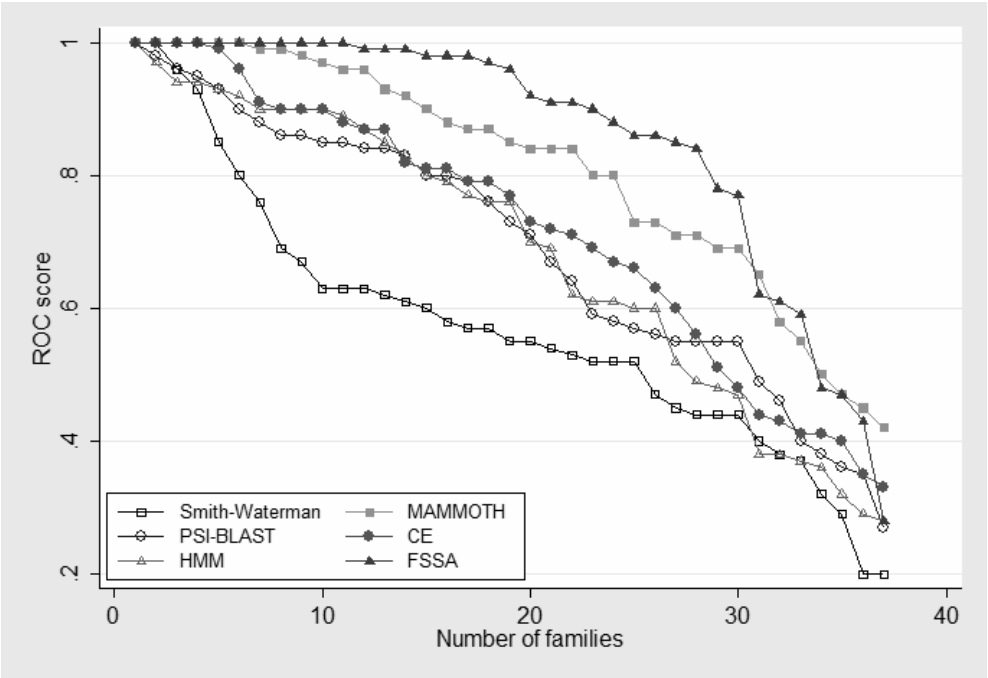

Fig. 2. Relative performance of six function discrimination methods on 37 datasets from the SCOP database that has been filtered by 30% pairwise sequence identity. For each function discrimination method, the number of SCOP families is plotted against the minimum ROC score achieved by that method. The FSSA method has the best performance in discriminating homologues from structural analogues.

superfamilies as structural analogues. We compared the performance of the FSSA method in distinguishing homologues and structural analogues to several other function discrimination methods, including Smith–Waterman, PSI-BLAST, HMMs and two structure comparison methods, MAMMOTH and CE (see Methods section). Of the two structure comparison programs we used, MAMMOTH performs global structure alignments, while CE performs local structure alignments. We used 37 SCOP families in our experiments, with the data preparation techniques aimed at minimizing sequence identity between training and testing sets (see Methods section). We measured the performance of each method by the ROC area under the curve score for these families, and compared the distribution of these ROC scores for different methods (Fig. 2 and Supplementary Table 1). Overall, the FSSA method has the best performance, with the highest ROC score for 24/37 families. In addition, the FSSA method also has the highest average ROC scores (0.86), among the six function discrimination methods. Since the calculation of ROC score for each family involves different number of sequences, the average score is not a valid means to compare function discrimination methods, though it provides some insight into the general performance of different methods. Also, some folds (such as the immunoglobulin and the TIM barrel folds) are enriched in these datasets, so they may not be representative of protein fold space in general.

Function classification experiments

Although the FSSA method performs well in terms of function discrimination, such a test is not adequate to demonstrate its value in function prediction applications. First, the FSSA method uses a 'negative training set', that contains sequences that share the same structural fold but belong to different superfamilies relative to the sequences that we are considering. Since other methods cannot incorporate information from negative samples, the better performance evaluated by ROC may be due only to the inclusion of this additional information. Further, although these function discrimination tests are commonly used to evaluate function prediction methods (Ben-Hur and Brutlag, 2003; Hou *et al*., 2003, 2005; Liao and Noble, 2003; Saigo *et al*., 2004), they have little practical use. Many function discrimination methods, such as those employing logistic regression or support vector machine techniques, are binary classifiers in nature, and are very difficult, if not impossible, to use for multi-category classification problems. In reality, when we can confidently assign a given sequence to a structural fold, we want to clearly identify the functional category that this sequence belongs to, as opposed to a binary answer of whether or not it belongs to a particular functional category. Therefore, a more rigorous and useful test for the performance of function prediction methods would be to see if proteins could be accurately assigned into functional categories, such as those defined by SCOP superfamilies. Figure 3 shows an example, where all five proteins have the same TIM barrel structural fold, but their catalytic sites and catalytic residues are quite different from each other [PDB identifiers 2dor (Rowland *et al*., 1998), 1qpr (Sharma *et al*., 1998), 1a4m (Wang and Quiocho, 1998), 1jcl (Heine *et al*., 2001) and 1n55 (Kursula and Wierenga, 2003), respectively]. These five proteins belong to different SCOP superfamilies and different EC primary classes. This example suggests that local structural differences, instead of overall structural folds, determine the function uniquely among proteins in multi-functional fold families. Given a collection of query structures, such as those determined by structural genomics projects, the goal of function classification experiments is

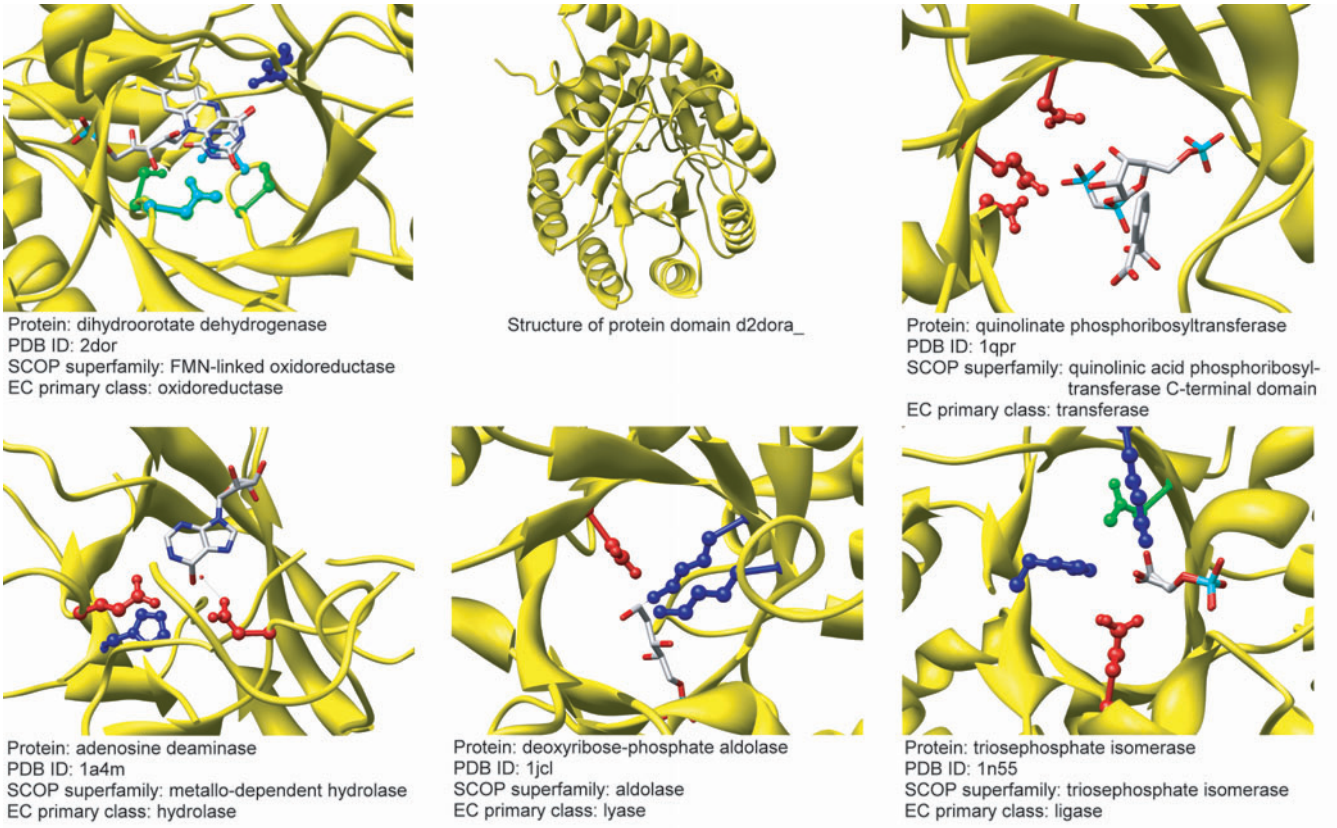

Fig. 3. Catalytic residues inside the barrel structure for five TIM barrel proteins (PDB identifier: 2dor, 1qpr, 1a4m, 1jcl and 1n55, respectively). The side chains for catalytic residues are shown by stick and ball representations and colored as red (acidic residue), blue (basic residue) and green (polor residue). The substrates or substrate analogs are shown by stick representations and are colored by elements. The structure of protein domain d2dora_ is also shown as an example of the overall TIM barrel structural fold. The pairwise C_α RMSDs between these folds range from 2.4 to 4.2 Å, with an average of 3.4 Å. These five proteins have quite different organizations of catalytic residues and biochemical activities, despite the similarity of their overall structural folds.

to identify the particular functional categories (SCOP superfamilies) these query structures belong to.

We performed function classification experiments on several SCOP folds derived from the function discrimination experiments (see Methods section). To investigate the correlation between performance and homology among testing and training sequences, we used four different datasets retrieved from the ASTRAL database, representing proteins whose pairwise sequence identities are <=10, 20, 30 and 95%, respectively. For all sequence identity levels, these structural folds in our datasets contain all-alpha, all-beta, alpha/beta, alpha+beta and small proteins, and are good representatives of the fold space. We used four-fold cross-validation experiments to test the function classification accuracy for the six methods: Smith–Waterman, PSI-BLAST, HMM, MAMMOTH, CE and FSSA (Fig. 4 and Supplementary Table 2). Overall, the FSSA method has the best function classification performance, when pairwise sequence identity in the datasets is $\leq$30%, though the differences are subtle between all methods utilizing structural information. Sequence-homology based function classification methods perform relatively poorly at low sequence identity levels. The poor performance of the HMM method is not unexpected, since the multiple alignment quality is low when sequence identity is low. We expect that HMMs constructed from manual alignments will have better performance. Despite the overall best performance of the FSSA method, we also notice that it does not work well for some folds, such as the OB-fold (SCOP identifier: b.40) and the adenine nucleotide alpha hydrolase-like fold (SCOP identifier: c.26), due to a heterogeneity in the functional signatures (see below). Our results, therefore, highlight the importance of using multiple methods to provide evidence for function. Because the performance of the FSSA method is relatively consistent for particular folds at different sequence identity levels (Supplementary Table 2), we may use the above results as *a priori* information to judge when to use FSSA to complement homology-based function prediction methods for new query sequences.

We further examined the prediction accuracies of the FSSA method on the TIM barrel structural fold family (SCOP fold identifier: c.1), which is one of the largest structural fold families. For the datasets with pairwise sequence identity $\leq$30%, the FSSA method correctly predicts the function for 11/14 (79%) proteins for the hydrolase superfamily, but only 1/15 (7%) for the aldolase superfamily. As we have shown in Figure 1, the members in the hydrolase superfamily have similar functional signatures, whereas the signatures in the aldolase superfamily are more heterogeneous. Therefore, the similarity of signatures within a superfamily may dictate if the FSSA will work well for that superfamily. This suggests that functional signatures from heterogeneous superfamilies should be interpreted

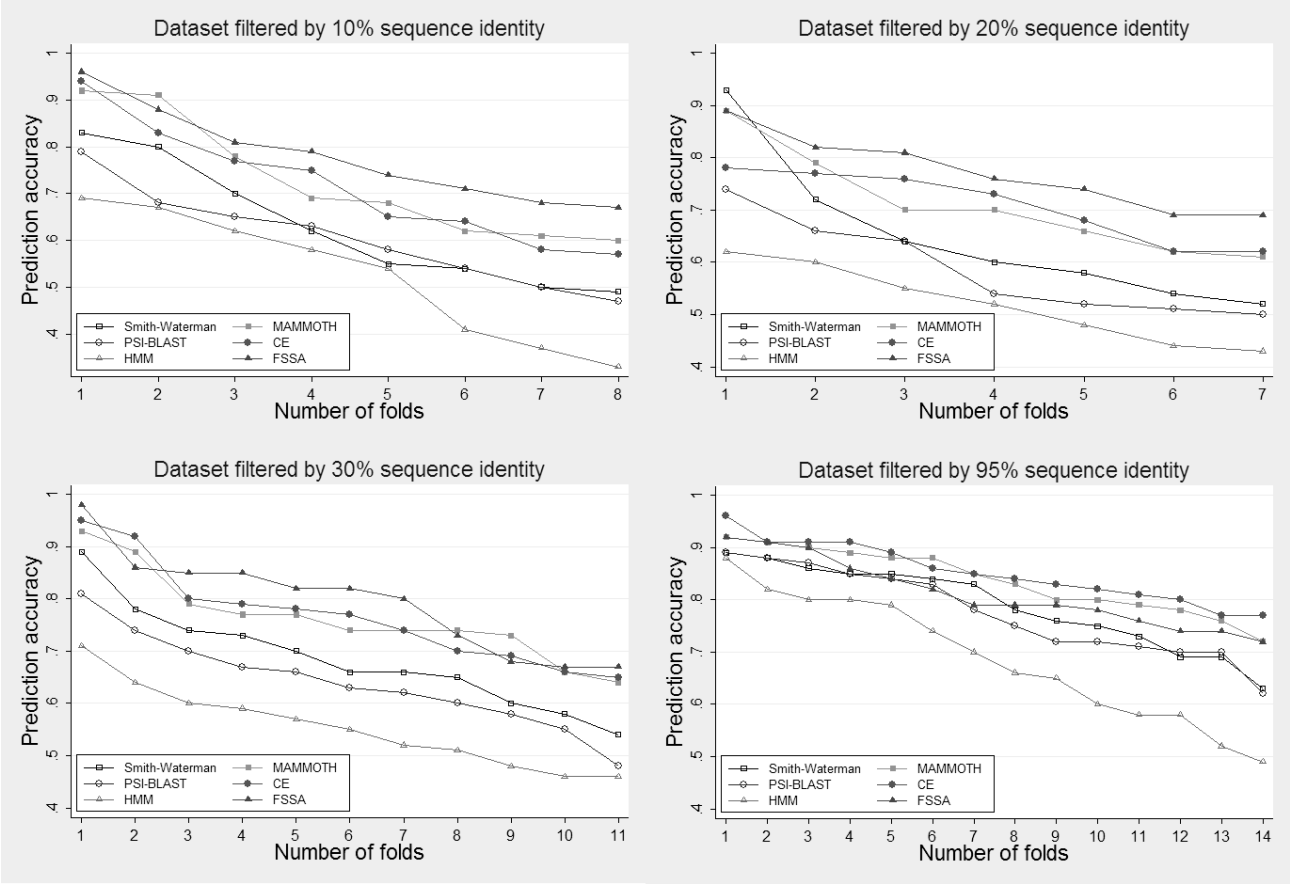

Fig. 4. Relative performance of six function classification methods on datasets from the SCOP database that has been filtered by 10, 20, 30 and 95% pairwise sequence identity, respectively. For each function classification method, the number of SCOP folds is plotted against the minimum prediction accuracy achieved by that method. The FSSA method has the overall best performance in function classification when sequence identity is $\leq$30%.

with more caution, since they may contain considerable amounts of noise.

Effects of excluding sequence information from the FSSA method

The FSSA method uses information on both local structure similarity (from the MAMMOTH program output) and local sequence similarity (from the BLOSUM50 substitution matrix). To deconvolute these contributions, we performed additional experiments on the FSSA method, using only local structure information. The functional signatures generated by the two forms of FSSA were generally quite similar to each other. Likewise, for the function discrimination and function classification experiments, the performance of the two forms of FSSA correlated very well with each other (see Supplementary Table 1 and Supplementary Table 2). When comparing the structure-only FSSA method to the other five homology-based methods, we found that it still has the best performance, with the highest ROC scores for 22/37 families, as well as the highest average prediction accuracies when sequence identity is $\leq$30%. However, the FSSA method, using both structure and sequence information, has slightly better performance than the structure-only FSSA, achieving higher or equal ROC values for 30/37 families in the function discrimination experiments and higher average prediction accuracies at all sequence identity levels in the function classification experiments. This suggests that both structure and sequence information contribute to the better performance of the FSSA method, and further improvements can be made by more sophisticated utilization of the sequence and structure information.

In summary, the FSSA is a novel method that explicitly estimates the relative contribution to function and structure for every residue in a protein sequence. The generated log odds scores may be used to interpret functional importance of individual residue types and positions, as well as to classify protein structures into functional categories. Together with other homology-based function prediction methods, the FSSA method will be valuable in function annotation applications for structural genomics projects.

DISCUSSION

Structural genomics projects are producing large amounts of new structures, prior to any functional knowledge of the target proteins (Goldsmith-Fischman and Honig, 2003). In addition, genome sequencing projects are producing a wealth of sequence data, many of which are homologous to a protein with known structure. However, determining the biological and physiological functions of a protein, even with a known structure, is still an open problem. Usually,

genome annotators may assign the function of a protein to be the same as the protein with the most similar sequence or structure. However, global sequence- or structure-based function classification methods usually do not have enough accuracy for experimental validation of the predictions. Therefore, novel prediction methods, such as the FSSA method presented here, are necessary to be developed to give accurate function prediction when the sequence identity levels between the query and the database are relatively low.

Our results presented in this paper indicate that at least for proteins in multi-functional fold families, the contribution of amino acid residue types and positions to structure and function are largely separable. Thus, we can construct functional signatures for proteins with known structures, and use the signatures to interpret the structural and functional importance of individual amino acid residues. Once these particular residues are identified, site-directed mutagenesis experiments can be performed for further functional characterization of these proteins. In addition, the FSSA method may be used in protein design applications to help modify existing functions or produce novel ones.

Fold similarities often require additional investigation of key residues before functions can be confidently inferred, and many algorithms have been developed to achieve this goal (see references in the Introduction section). For structural genomics targets with unknown function, comparing functional sites, instead of the overall structural fold, can reveal more clues about the biological activity of a protein (Stark *et al*., 2004). Compared with other functional site identification algorithms, our approach has some marked differences: the functional signature is a collection of log odds scores that are continuously distributed along the whole sequence, rather than a small collection of catalytic residues. Also, instead of trying to capture a common pattern from a group of homologous proteins, the FSSA method maintains a separate signature for each individual protein, thus allowing more sensitive functional analysis. Because of these differences, functional signatures should not be interpreted to be catalytic sites. When we examine the catalytic sites in Figure 3, none of them are positions with the highest log odds scores. For example, as shown in Figure 1, the highest log odds score of the 1a4m protein is accumulated in the C-terminal region, which does not contain the catalytic sites. However, the local structures in the C-terminal region capture the characteristics of the hydrolase superfamily, and can be used to classify function accurately, which is what the FSSA method demonstrates. We envision that other methods, aimed at finding discrete structural motifs or distributions of catalytic sites, can be used to validate whether the functional sites identified by the FSSA method are biologically relevant, and the combination may result in enhanced and comprehensive functional information for newly determined structures.

The FSSA method uses pairwise structure alignments. Multiple sequence alignment based methods have been developed extensively for constructing profiles for function prediction (Krogh *et al*., 1994; Bateman *et al*., 1999) and it has been shown that structural information can improve the quality of sequence alignments and can be used to generate better profiles (Al-Lazikani *et al*., 2001). However, these methods aim at identifying remote homologues, or discriminating functionally related proteins from unrelated ones. They may not work well at discriminating homologues from structural analogues, or at classifying homologues into functional subfamilies. To solve these problems, multiple sequence alignment based methods must be adapted to identify key residues for determining functional specificity (Hannenhalli and Russell, 2000), but such algorithms require relatively high-sequence identity to generate accurate multiple sequence alignments. Using structure information may generate better alignments, but having automated and accurate multiple structure alignments for a large number of proteins across different superfamilies is a difficult problem. Reliable multiple structure alignments (generated manually, for example) however, may improve the accuracy of the FSSA method for specific fold families.

The FSSA method uses the MAMMOTH program output as well as the BLOSUM50 matrix to obtain a binary definition of whether or not two residues have a similar local structure profile. To investigate the relative contribution of structure and sequence information on the quality of the signatures, we also tested a modified FSSA method using only structure information. Generally, the FSSA using both structure and sequence information performed better than the one using only structure information, showing that incorporating additional sequence information does improve performance. A more sophisticated definition of local structure profile similarity may further improve the performance of the FSSA method. However, this is a difficult problem, since, unlike sequence alignments, structural alignments may contain slight alignment shifts between adjacent residues. In such cases, different amino acid types can be aligned with each other, resulting in incorrect functional signatures.

Given the fact that the contents of sequence databases are significantly greater than those of structure databases, it would be more desirable if we can directly use sequence information for function classification. We envision that this problem may be solved by using sophisticated sequence-to-structure alignments or using high-quality *de novo* structure predictions (Bradley *et al*., 2003; Skolnick *et al*., 2003; Hung *et al*., 2005). The latter can be used for functional annotation, based on structure comparison as well as FSSA. Extension of the FSSA method such that sequence only information is used will have a greater impact on genome annotation, function prediction and protein design applications.

ACKNOWLEDGEMENTS

We thank the Samudrala group for helpful discussions and comments on the manuscript. We also thank the two anonymous reviewers for their constructive suggestions on our methodology as well as the presentation of the manuscript. This work was supported by a Searle Scholar Award, NSF grant DBI-0217241 and NIH grant GM068152-01 to R.S.

REFERENCES

Al-Lazikani,B. *et al*. (2001) Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. *Proc. Natl Acad. Sci. USA*, **98**, 14796–14801.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Anantharaman,V. *et al*. (2003) Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Curr. Opin. Chem. Biol.*, **7**, 12–20.

Attwood,T.K. *et al*. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.

Barker,J.A. and Thornton,J.M. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, **19**, 1644–1649.

Bateman,A. *et al*. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.*, **27**, 260–262.

Ben-Hur,A. and Brutlag,D. (2003) Remote homology detection: a motif based approach. *Bioinformatics*, **19** (Suppl. 1), i26–i33.

Bradley,P. *et al*. (2003) Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins*, **53** (Suppl. 6), 457–468.

Branden,C.I. (1991) The TIM barrel—the most frequently occuring folding motif in proteins. *Curr. Opin. Struct. Biol.*, **1**, 978–983.

Cammer,S.A. *et al*. (2003) Structure-based active site profiles for genome analysis and functional family subclassification. *J. Mol. Biol.*, **334**, 387–401.

Chandonia,J.M. *et al*. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32** (Database issue), D189-D192.

Di Gennaro,J.A. *et al*. (2001) Enhanced functional annotation of protein sequences via the use of structural descriptors. *J. Struct. Biol.*, **134**, 232–245.

Dror,O. *et al*. (2003) Multiple structural alignment by secondary structures: algorithm and applications. *Protein Sci.*, **12**, 2492–2507.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

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.

Goldsmith-Fischman,S. and Honig,B. (2003) Structural genomics: computational methods for structure analysis. *Protein Sci.*, **12**, 1813–1821.

Guda,C. *et al*. (2001) A new algorithm for the alignment of multiple protein structures using Monte Carlo optimization. *Pac. Symp. Biocomput.*, 275–286.

Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional subtypes from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.

Heine,A. *et al*. (2001) Observation of covalent intermediates in an enzyme mechanism at atomic resolution. [Erratum (2001) *Science*, **294**, 2096.] *Science*, **294**, 369–374.

Henikoff,J.G. *et al*. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.

Holm,L. and Sander,C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*, **27**, 244–247.

Hou,Y. *et al*. (2003) Efficient remote homology detection using local structure. *Bioinformatics*, **19**, 2294–2301.

Hou,J. *et al*. (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.

Hulo,N. *et al*. (2004) Recent improvements to the PROSITE. *Nucleic Acids Res.*, **32** (Database issue), D134–D137.

Hung,L.-H. *et al*. (2005) Protinfo: new algorithms for enhanced protein structure prediction. *Nucleic Acids Res.*, in press.

Jambon,M. *et al*. (2003) A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins*, **52**, 137–145.

Jonassen,I. *et al*. (1999) Discovery of local packing motifs in protein structures. *Proteins*, **34**, 206–219.

Jonassen,I. *et al*. (2002) Structure motif discovery and mining the PDB. *Bioinformatics*, **18**, 362–367.

Jones,S. and Thornton,J.M. (1997) Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121–132.

Jones,S. *et al*. (2003) Using structural motif templates to identify proteins with DNA binding function. *Nucleic Acids Res.*, **31**, 2811–2823.

Kasuya,A. and Thornton,J.M. (1999) Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.*, **286**, 1673–1691.

Kedem,K. *et al*. (1999) Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. *Proteins*, **37**, 554–564.

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

Kobayashi,N. and Go,N. (1997) ATP binding proteins with different folds share a common ATP-binding structural motif. *Nat. Struct. Biol.*, **4**, 6–7.

Krogh,A. *et al*. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Kursula,I. and Wierenga,R.K. (2003) Crystal structure of triosephosphate isomerase complexed with 2-phosphoglycolate at 0.83 Å resolution. *J. Biol. Chem.*, **278**, 9544–9551.

Laskowski,R.A. *et al*. (1996) Protein clefts in molecular recognition and function. *Protein Sci.*, **5**, 2438–2452.

Leibowitz,N. *et al*. (2001a) Automated multiple structure alignment and detection of a common substructural motif. *Proteins*, **43**, 235–245.

Leibowitz,N. *et al*. (2001b) MUSTA—a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins. *J. Comput. Biol.*, **8**, 93–121.

Liao,L. and Noble,W.S. (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.*, **10**, 857–868.

Matsuo,Y. and Bryant,S.H. (1999) Identification of homologous core structures. *Proteins*, **35**, 70–79.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nagano,N. *et al*. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.*, **321**, 741–765.

Orengo,C.A. *et al*. (1999) The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279.

Ortiz,A.R. *et al*. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Pazos,F. and Sternberg,M.J. (2004) Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl Acad. Sci. USA*, **101**, 14754–14759.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Rowland,P. *et al*. (1998) The crystal structure of *Lactococcus lactis* dihydroorotate dehydrogenase A complexed with the enzyme reaction product throws light on its enzymatic function. *Protein Sci.*, **7**, 1269–1279.

Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.

Russell,R.B. *et al*. (1998) Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.*, **282**, 903–918.

Saigo,H. *et al*. (2004) Protein homology detection using string alignment kernels. *Bioinformatics*, **20**, 1682–1689.

Sharma,V. *et al*. (1998) Crystal structure of quinolinic acid phosphoribosyltransferase from *Mycobacterium tuberculosis*: a potential TB drug target. *Structure*, **6**, 1587–1599.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Skolnick,J. *et al*. (2003) TOUCHSTONE: a unified approach to protein structure prediction. *Proteins*, **53** (Suppl. 6), 469–479.

Stark,A. and Russell,R.B. (2003) Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res.*, **31**, 3341–3344.

Stark,A. *et al*. (2004) Finding functional sites in structural genomics proteins. *Structure* (*Camb.*), **12**, 1405–1412.

Taylor,W.R. and Orengo,C.A. (1989) Protein structure alignment. *J. Mol. Biol.*, **208**, 1–22.

Teichmann,S.A. *et al*. (2001) Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.*, **11**, 354–363.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Thornton,J.M. *et al*. (2000) From structure to function: approaches and limitations. *Nat. Struct. Biol.*, **7** (Suppl.), 991–994.

Todd,A.E. *et al*. (1999) Evolution of protein function, from a structural perspective. *Curr. Opin. Chem. Biol.*, **3**, 548–556.

Todd,A.E. *et al*. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, **307**, 1113–1143.

Todd,A.E. *et al*. (2002) Sequence and structural differences between enzyme and nonenzyme homologs. *Structure* (*Camb.*), **10**, 1435–1451.

Wang,Z. and Quiocho,F.A. (1998) Complexes of adenosine deaminase with two potent inhibitors: X-ray structures in four independent molecules at pH of maximum activity. *Biochemistry*, **37**, 8314–8324.

Yang,A.S. and Honig,B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *J. Mol. Biol.*, **301**, 691–711.