BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 6 2006, pages 723–730
doi:10.1093/bioinformatics/btk038

Structural bioinformatics

Automated discovery of 3D motifs for protein function annotation

Benjamin J. Polacco and Patricia C. Babbitt*

Department of Biopharmaceutical Sciences, University of California, San Francisco, CA 94143-2250, USA

Received on October 5, 2005; revised on December 17, 2005; accepted on January 3, 2006
Advance Access publication January 12, 2006
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Function inference from structure is facilitated by the use of patterns of residues (3D motifs), normally identified by expert knowledge, that correlate with function. As an alternative to often limited expert knowledge, we use machine-learning techniques to identify patterns of 3–10 residues that maximize function prediction. This approach allows us to test the assumption that residues that provide function are the most informative for predicting function.

Results: We apply our method, GASPS, to the haloacid dehalogenase, enolase, amidohydrolase and crotonase superfamilies and to the serine proteases. The motifs found by GASPS are as good at function prediction as 3D motifs based on expert knowledge. The GASPS motifs with the greatest ability to predict protein function consist mainly of known functional residues. However, several residues with no known functional role are equally predictive. For four groups, we show that the predictive power of our 3D motifs is comparable with or better than approaches that use the entire fold (Combinatorial-Extension) or sequence profiles (PSI-BLAST).

Availability: Source code is freely available for academic use by contacting the authors.

Contact: babbitt@cgl.ucsf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The increasing availability of structural data for proteins of unknown function creates a demand for *in silico* methods to infer the function of these proteins using structural information (Teichmann *et al*., 2001). But while comparison of overall structures can extend homology detection to evolutionary distances where sequence similarity is undetectable (Chothia and Lesk, 1986), fold comparison often does not identify similarities among functionally significant residues or atoms involved in a protein function's mechanism. Together, the coordinates of these residues or atoms can define a 3D motif. There are many available motif-matching methods that can be used to identify a protein with a matching motif and thus a similar function and mechanism (e.g. Artymiuk *et al*., 1994; Fetrow and Skolnick, 1998; Barker and Thornton, 2003). Such methods offer useful complements to fold-based homology comparisons, especially in cases where homologs have diverged in function.

In earlier studies, 3D motifs have typically been chosen based on expert knowledge of functionally important residues in enzyme-active sites such as the catalytic triad of the serine proteases. These motifs have been successful at identifying specific enzymatic activities (Torrance *et al*., 2005), binding relationships (Artymiuk *et al*., 1994) and superfamily membership (Meng *et al*., 2004). However, in the absence of a large data source of functional information, accumulation of motifs is slow. The catalytic site atlas (CSA) is a new effort to create a comprehensive database of functional information gleaned from the literature (Porter *et al*., 2004). It currently provides 147 non-redundant active site motifs for enzymes (Torrance *et al*., 2005). Similarly, Arakaki *et al*. (2004) presented an automated method that used the functional information in feature records of the Swiss-Prot database to construct 3D motifs for 162 different enzymes. Even this method is limited by the shortage of functional information in Swiss-Prot. There are numerous other examples of computational approaches to predict functionally important residues (e.g. Zvelebil and Sternberg, 1988; Elcock, 2001; Wangikar *et al*., 2003), but these may not be accurate enough to translate to useful motifs (see Discussion).

An alternative is the use of automated 3D motif detection methods. These have shown some success, though none has mapped motifs to specific protein functions with the design goal of characterizing novel proteins with high accuracy. PINTS detects repeated patterns of sidechains between pairwise comparisons of diverse structures, and has generated a large set of repeated motifs (Russell, 1998). A similar data-mining approach that compares all patterns across an entire library of structures finds the catalytic triads of proteases along with metal binding sites, salt bridges and similar structural features (Oldfield, 2002). Although such general structural features do not provide much specific functional information, they dominate the databases of motifs generated by these types of methods.

We present here a new approach for automated 3D motif generation named Genetic Algorithm Search for Patterns in Structures (GASPS). GASPS was developed with two basic design goals. First, for any specified group of proteins, GASPS should find the motif most useful for identifying the group. Second, GASPS should rely as little as possible on the knowledge about what is likely a predictive or functionally important residue. We validate the effectiveness of GASPS on four highly divergent groups of enzymes: the

*To whom correspondence should be addressed.

convergent serine proteases (SP), the amidohydrolase superfamily (AHS), the enolase superfamily (ES) and the haloacid dehalogenase superfamily (HADS). These motifs verify that many, but not all of the previously known functionally important residues are the best predictive residues (along with additional unexpected residues). We describe the crotonase superfamily (CS) as an example of a group that is not well suited for characterization by 3D motifs as they are typically defined.

2 METHODS

2.1 Motif representation and matching

As an initial test of principle, we adopted the motif model and matching algorithm of SPASM (Kleywegt, 1999), although GASPS can be adapted for use with other motif matching algorithms as well. A motif is a small set of residues (<10 for this study) taken from a single chain, here called the query chain. For each position, SPASM requires a matching residue to be of the identical type with no substitutions. Alternatively, a unique set of residues at each position may be specified that can be substituted with no penalty, though in the course of our study we were unable to use this feature effectively (see Supplementary Materials). SPASM models each residue with just two points, backbone Cα and the sidechain geometrical center. SPASM computes a superposition root-mean-squared deviation (RMSD) for each match it finds within user-defined thresholds of RMSD, sidechain distance deviations (SCD) and Cα distance deviations (CαD). For this study, thresholds were set to RMSD = 3.2 Å, SCD = 3.8 Å and CαD = 5.0 Å. SPASM allows the use of several additional constraints that were not used for this study. Only the match with the best RMSD is considered from each structure.

2.2 GASPS

GASPS generates motifs by selecting residues from a single query chain. Here, functional sites and motifs that span more than one chain are not directly addressed. These motifs are scored for their ability to accurately discriminate the positive from the negative sets. There are four main components to a GASPS run: query processing, initial guesses, scoring and refined guesses.

Query processing To limit the search space, only the 100 most conserved residues in the query chain are considered for inclusion in a motif. Conservation is calculated from a multiple sequence alignment by weighting sequences to reduce the effects of redundancy, considering conservative substitutions based on a substitution matrix and to penalize gaps (Valdar, 2002). All multiple sequence alignments were generated by a two-iteration PSI-BLAST (Altschul *et al*., 1997) search against nrdb90 (Holm and Sander, 1997) built in February, 2004.

Initial guesses A total of 50 candidate motifs are initially chosen spread equally across the linear sequence of the query chain to provide coverage of all regions. For each random guess, a first residue is selected from the query chain and then four other residues are randomly chosen such that each alpha carbon is within 12 Å of the first alpha carbon.

Scoring function Candidate motifs are scored for their ability to discriminate between the positive and negative proteins based on the best RMSD matches from a SPASM search. The query structure, which is always a perfect match to the motif, is excluded from the positive set. The scoring function is primarily the normalized area under a receiver-operator characteristic (ROC) plot to five false positives (a false positive rate of ~0.001). If the sorted RMSD scores for structures in the negative set are (f_1, f_2, $f_3, \ldots, f_n$), then this area, called R, can be computed explicitly as

$$R = \frac{1}{5}\sum_{i=1}^{5}\frac{T(f_i)}{T_{\max}},$$

where $T(f)$ is the number of true positives with a better RMSD match than a given false positive and $T_{\max}$ is the size of the positive set. R ranges from 0 to 1. Because R is based on discrete counts, different motifs will frequently have identical R scores. To avoid ties, we include an additional term in the GASPS scoring function. This term, S, is the normalized difference in median RMSD between the true positives and false positives, only considering those that score better than the fifth false positive (f_5). This can be explicitly defined as

$$S = \frac{\text{median}(f_{1-5}) - \text{median}(t_{1-m})}{\text{median}(f_{1-5})},$$

where t_{1-m} is the set of RMSDs from the true positive matches that are better (less) than the fifth false positive (f_5). When no true positives are hit ($R = 0$), S is set to zero. The overall GASPS score (G) is then the sum of S and R weighted to emphasize the ROC score, and is composed as

$$G = 1.0R + 0.1S$$

Refined guesses The 16 highest scoring motifs of any round are included in the next round and are used as parents for constructing 36 novel motifs via one random process: deletion, insertion, mutation or recombination. The only restriction on the new motifs is that they contain at least 3 residues and at most 10. Deletions and insertions generate a new motif by removing or adding a residue to a single parent motif. A mutation is a combination of a deletion and an insertion. A recombination is a random subset of the combination of two parent motifs. The top-scoring motif after 50 rounds of refinement is considered the final winner. Most GASPS runs in this study took between 12 and 18 h on a single 2.667 GHz Intel Xeon processor. Most of this time was spent completing the SPASM searches against the negative set, which time scales directly with the number of proteins in the negative set.

2.3 Structure library

Most analyses in this study used a set of structures selected from the Protein Data Bank (PDB) (Berman *et al*., 2000) to represent all families in The Structural Classification of Proteins (SCOP) version 1.65 (Murzin *et al*., 1995). The selection algorithm treats each SCOP family individually and has three main goals: (1) mutant removal based on text matching PDB fields, (2) sequence redundancy filter to 40% identity and (3) favoring the highest quality structures based on resolution. No distinction is made between *apo* and *holo* structures. The entire corresponding PDB chains for each of the SCOP domains are included, regardless of similarities at other domains. In SCOP version 1.65, this selection results in 5440 unique domains on 4243 unique chains.

2.4 Positive and negative sets

We chose five well-characterized positive groups so that all members within each group share a similar function, and this shared function is dependent on known functional residues (Table 1). Definitions for the four superfamilies were taken from the Structure-Function Linkage Database (SFLD) (Pegg *et al*., 2005). However, the SFLD as yet contains only a few superfamilies, so to mimic a more typical usage of GASPS on less than perfect classifications, for all five groups of proteins studied here, a positive set of structures was selected based on SCOP superfamily and family classifications (given in Table 1). Each positive set is a subset of the structure library with the modification that all chains within a PDB structure file are included. Sequence identities between all pairs of homologous chains used as query chains range from 14 to 40%. The negative set is the entire structure library, excluding all chains that contain at least one domain meeting the criteria for inclusion in the positive set.

2.5 Cross-validation

Complete rounds of leave-one-out cross-validation were performed for several query structures in each group. For the smaller groups, each structure

Table 1. Functionally similar protein groups

Group	SCOP groups	N[a]	Known functional residues
Amidohydrolase superfamily	c.1.9	16	(1a4m) H15 H17 H214 H238 D295
Enolase superfamily	c.1.11	7	(2mnr) K160 D191 E219 D244 K268
Crotonase superfamily	c.14.1.3	7	(1mj3) backbone A98 G141
Haloacid dehalogenase superfamily	c.108.1	12	(1fez) D12 T126 R160 D186 D190
Serine proteases	b.47.1.1, b.47.1.2, b.47.1.3, c.41.1.1	38	(2hlc) H57 D102 S195

[a]Number of non-redundant structures in positive set.

in the positive set was used once as a query structure. For the larger groups, AHS and SP, a randomly selected subset of the structures was used. For each query structure, all possible positive training sets were produced by excluding one other (non-query) positive structure. The corresponding positive test sets each contained just the excluded structure. Similarly, the negative set was equally divided to produce as many negative test sets as positive test sets. The corresponding negative training sets are simply the entire negative set excluding a negative test set. Using ES as an example, this procedure required 42 runs of GASPS (7 query structures multiplied by 6 left-out positive structures). The reported sensitivity on the test sets is the portion of GASPS runs where the final GASPS motif from each training run was able to discriminate the left-out positive member from the left-out negative test set at an RMSD threshold equal to the RMSD of the fifth-best false positive match on the training sets. Those runs for which the final trained GASPS motif did not score significantly on the training set were excluded.

2.6 PSI-BLAST and Combinatorial-Extension libraries

For BLAST (Altschul *et al*., 1997) and PSI-BLAST comparisons with GASPS the libraries were the set of unique chains from the same PDB files used in the positive and negative sets for GASPS (described above). For the Combinatorial-Extension (CE) algorithm (Shindyalov and Bourne, 1998), to avoid computing all-by-all pairwise comparisons, the negative sets were reduced to the probable high-scoring members for each positive group. For most groups, this meant limiting the negative set to those chains with the same SCOP fold as a catalytic domain in the positive set. However, according to SCOP, HADS is the only superfamily of the HAD-like fold, so its negative set for CE was chosen based on CATH (Orengo *et al*., 1997) instead. For SP, there were an insufficient number of same-fold structures that were not SP to provide negative sets for both the subtilisins and trypsins. An additional SCOP fold (b.43: Reductase/isomerase/elongation factor common domain) was included in the library that commonly scored highly against SP folds according to the CE internet database (http://cl.sdsc.edu/ce.html).

3 RESULTS

3.1 Validation of GASPS

3.1.1 Significance of optimized GASPS scores To determine whether any GASPS motif likely represents more than a chance co-occurrence of residues, we computed significance cutoffs from empirical distributions of GASPS motifs due to chance alone. To ensure that any motif was due to chance, artificial positive groups were generated by randomly selecting structures from the structure library, each with a different fold. Based on these distributions, provided in Supplementary Materials, we can set GASPS score thresholds for moderate significance ($P < 0.01$): for groups of ~5 structures motifs must score >0.55 and in larger groups of 10 or more structures they must score >0.4.

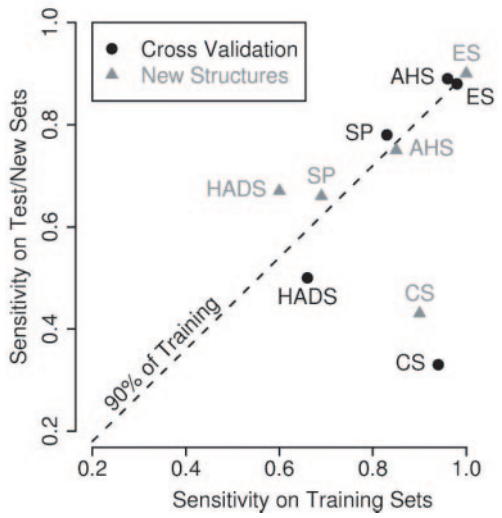

Fig. 1. Generality of GASPS motifs based on sensitivity from two experiments: cross-validation and detection of newer structures. Black filled circles show average sensitivities of motifs from leave-one-out runs on the cross-validation training (*x*-axis) and test (*y*-axis) sets. Gray triangles show sensitivities of motifs generated on the full training sets (all motifs in Fig. 3) when used to detect structures in the full training set (*x*-axis) compared with novel structures solved after the training set was established (*y*-axis).

3.1.2 Cross-validation studies To estimate the performance of GASPS on new proteins, leave-one-out cross-validation studies were completed on each of the groups in Table 1. RMSD thresholds were chosen for each top GASPS motif to give a false positive rate of ~0.0013 (5 false positives) on the training set. With the exception of CS, sensitivity is high and there is a close correspondence between the training and test sets (Fig. 1). Sensitivity on the test sets for most cases is ~90% of that on the training cases. The false positive rate (and its complement, sensitivity) shows an even tighter correspondence with an average rate of 0.0014 on the test cases. The fact that CS is one of the smallest groups and also that it lacks highly conserved sidechains in the active site, as described below, likely contribute to the poor performance of GASPS for this superfamily.

3.1.3 Detection of new structures Across all groups, we identified 12 new structures in the PDB that were not yet classified by

SCOP (as of version 1.65, December 2003), by a combination of searches based on literature, annotation and sequence similarity, along with communications with collaborators. These 12 proteins all share <40% sequence identity with each other or with any protein in the original training set. Motifs generated on the full training set, one for each query structure (shown in Fig. 3), were tested for their ability to match the appropriate new structures within the RMSD thresholds determined on the full training set. For these 12 structures, the group-based average rate of matches is 68% compared with 81% on the structures included in the full training set. If CS is excluded, the group-based average rate of matches is 75%, compared with 79% on the training set (Fig. 1). This is an average across all motifs in each group including those with insignificant scores and very poor match rates. The expected match rate for any given motif appears linked to its original GASPS score. Excluding CS, no top-scoring motifs in any one group missed any of the new structures, and only 1 of 9 insignificantly scoring motifs matched any new structures. No new structures, CS included, failed to match any motif in their group.

3.1.4 Comparisons with other 3D motif methods A key benefit of GASPS is that it requires no knowledge of functionally important residues. However, even on groups where functionally important residues are known, GASPS is still useful if it is able to select a more sensitive motif. We constructed motifs built from the functionally important residues (see Table 1) for all possible query structures and compared their sensitivity with GASPS motifs. For all groups except SP, the GASPS motifs have higher sensitivity than simply using these functional residues (Fig. 2, 'FUN').

Of other available techniques, the closest to GASPS in principle is DRESPAT (Wangikar *et al*., 2003) that detects similar patterns of residues within a group of structures. We used DRESPAT with previously published parameters and a pattern size of four residues to generate patterns for the groups in our dataset. The resulting top ranked patterns identify some functionally important residues for all groups in this study except for CS (data not shown). However, they fail to identify superfamily members with similar specificity and sensitivities to those of GASPS motifs (Fig. 2). It may be possible to adjust the parameters and desired pattern size to improve the performance on a case-by-case basis, but the DRESPAT technique is not designed to automate or aid such a procedure.

Two other 3D motif libraries have recently been used to identify functional or homologous relationships, the motifs used by PINTS and the CSA. As these libraries were not specifically constructed to identify members of the groups in this study, it is impossible to run a parallel experiment for a direct comparison with the techniques shown in Figure 2. PINTS has been used to confirm superfamily membership and binding relationships of structural genomics proteins by finding matches to motifs derived from proximity to ligands or SITE annotations in PDB records (Stark *et al*., 2004). We tested this same technique (made available at http://www.russell.embl.de/pints/) by asking whether the structures used in our study matched with high-specificity those motifs that came from other non-redundant (<40% sequence identity) group members. The measured sensitivities of GASPS motifs (Figs 1 and 2) greatly outperform PINTS for all five groups at similar or even much lower rates of specificity. To be generous, PINTS could adequately serve its purpose if for any query structure it only detects a single true positive motif and ranks it highest among all matches. Even using this much less stringent definition of sensitivity for PINTS than used for GASPS, only for SP does PINTS score better than our reported sensitivities for GASPS motifs.

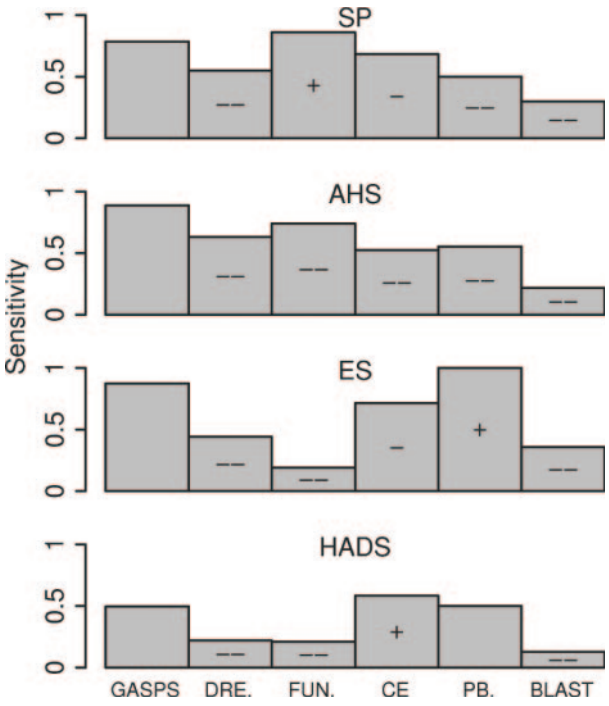

Fig. 2. Sensitivity of GASPS motifs compared with other techniques. Sensitivity shown for GASPS is measured by cross validation (Fig. 1). For all other techniques, the sensitivity is measured at the threshold of the fifth false positive. Other techniques are DRESPAT (DRE.) motifs, motifs built from functional residues (FUN.), CE, PSI-BLAST (PB.) and BLAST. Plus signs (+) indicate significantly better sensitivity than GASPS within the protein group, and minus signs indicate significantly worse performance at $P < 0.05$. Double signs (++ or −−) indicate a greater degree of statistical significance ($P < 0.0001$). CS results are not shown because no 3D motif methods were able to characterize it effectively.

The motifs derived from functional knowledge in the CSA are available for searching by the program Jess (Barker *et al*., 2003) at their website (http://www.ebi.ac.uk/thornton-srv/databases/CSA/). We used each of the structures in our positive sets to search the CSA with Jess and scored true positives according to whether the motif originated from the same group (defined in Table 1) as the original query. Maintaining similar specificity as required for GASPS, we should require that Jess, with only 147 motifs, identify true positives with greater *E*-value than any false positive. Only structures from SP reliably matched any true positives. Outside SP, only three structures (one each from AHS, ES and HADS) matched any CSA motif, but all three of these motifs came from structures that shared >40% sequence identity with the query. Relaxing the specificity to five false positives only allowed two other HADS structures to match the haloacid dehalogenase motif. Even though several of the false positive matches had *E*-values that suggested significance ($\sim 10^{-4}$), none correctly predicted the function or group membership of the original query. These high quality motifs in the CSA are useful for detecting certain specific functions, but they cannot adequately detect the diverse functions or distant relationships covered by the superfamilies studied here.

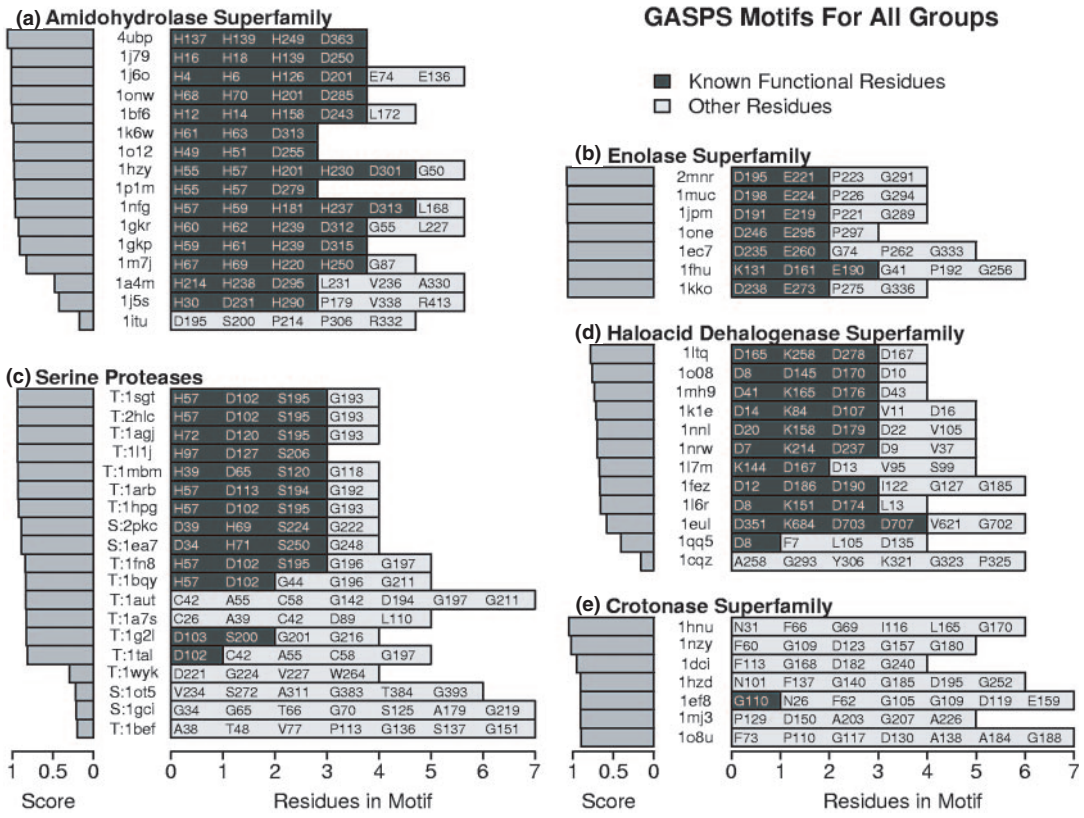

Fig. 3. Scores and functional significance of GASPS motifs. The results of a single GASPS run are presented for each named query structure. Residues in the motif that correspond to previously identified functional residues or known active-site motif residues are darkly shaded. All other residues are lightly shaded regardless of subsequent determination of their functional significance. For SP, query structures are labeled '*T*:' to denote trypsin-like folds or '*S*:' for subtilisin-like folds.

3.1.5 Prediction ability compared with whole-chain tools We compared the sensitivity of GASPS motifs (as estimated by cross validation) with other tools that use the whole protein chain including the sequence-based tools BLAST and PSI-BLAST (Altschul *et al*., 1997) and the fold comparison tool CE (Shindyalov and Bourne, 1998). All members of all positive groups were used as queries for each of the methods, and these were searched against the appropriate library as described in Methods. All sensitivities were measured by counting the fraction of positives that scored better than the fifth best-scoring negative for each query (Fig. 2). No single method is better than all other methods for all of the groups in this study. CS is easily grouped by most methods with the exception of GASPS motifs. The fold comparison tool CE performs well on groups with unique folds such as HADS. AHS and ES, on the other hand, share the common $(\beta/\alpha)_8$ fold with many other superfamilies, which may help explain why CE performs worse than GASPS in these cases. PSI-BLAST performs better than GASPS only for the least divergent superfamilies considered, ES and CS, where PSI-BLAST performs perfectly. With the exception of CS, GASPS motifs outperform BLAST on all groups.

3.2 Detection of key functional residues

An advantage of our method is that the selection of the residues in a motif is unbiased towards any preconceived notions of functionally important residues except indirectly via our exclusion of the least conserved residues. This allows us to ask if there is a relationship between the residues that discriminate proteins of related function and the residues that we know from experimental studies provide function. Table 1 shows the residues that are known to be directly involved in shared function or used in previous functional motif studies. These are used as our gold standard of key functional residues. Every positive structure was used once as a query structure except for SP from which only a diverse subset of structures was chosen. The best motifs from each of these runs are presented in detail in Figure 3. There is a clear trend for the proportion of functional residues in a motif to increase as the motif score rises.

As a stochastic search method, GASPS can be expected to produce different motifs in identically configured runs, and we expect that several of the lower scoring motifs presented in Figure 3 are not the best motif a given query can provide. The results of repeated runs for several configurations are presented in Supplementary Materials. Clearly, multiple GASPS runs per group are necessary to ensure that an optimal motif is found for any group. For some single query structures, however, repeated runs suggest there is no combination of residues that provide a useful motif. Meanwhile, the optimal motifs for the majority of other query structures are highly similar. Taken together, these results suggest that to generate a set of the most useful and inclusive motifs for any group of proteins, limited resources are better spent on running GASPS on

many different query structures than on running GASPS multiple times on the same structure.

3.2.1 Amidohydrolase superfamily AHS is a functionally diverse superfamily composed of homologs with a $(\beta/\alpha)_8$ (TIM) barrel fold that share a conserved mechanistic step mediated by a conserved set of active site residues (Holm and Sander, 1997; Gerlt and Raushel, 2003). All known members of the superfamily are metal-dependent and require either one or two divalent metal ions. Five conserved metal ligands comprising four histidines and an aspartic acid have been identified as functionally important in all groups within the superfamily. Only one GASPS run on this superfamily resulted in a motif with an insignificant score and no overlap with any of these metal ligands (Fig. 3a). The remaining runs all resulted in motifs that contained at least three of the five conserved ligands. The other residues in the significant motifs are all distant from the metal ligands and thus probably not directly involved in the enzyme's active site.

3.2.2 Enolase superfamily Like AHS, ES is made up of homologs with a C-terminal $(\beta/\alpha)_8$ barrel fold plus an N-terminal domain representing a unique fold. All functionally diverse members share a common mechanistic step (Babbitt *et al*., 1996; Gerlt *et al*. 2005). Past studies have carefully documented conserved elements responsible for the shared aspects of mechanism, and motifs based on this functional information have been generated with success (Meng *et al*., 2004). In this study, the conserved residues considered to play a functional role consist of three divalent metal ligands and two basic residues. All motifs resulting from GASPS runs contained at least the same two metal ligands, and one run contained one of the basic residues (Fig. 3b). The remaining metal ligand and both basic residues are known to have variable residue types across members of the superfamily, possibly explaining why GASPS has trouble locating them. A highly conserved residue among the GASPS motifs that has not been identified as functionally important is a proline that is two positions downstream from the second metal ligand. Here called the 'downstream proline', it appears in all ES motifs.

3.2.3 Haloacid dehalogenase superfamily HADS comprises enzymes with diverse functions, yet all members share a common mechanistic step associated with hydrolytic nucleophilic substitution via a conserved aspartate and a few other residues (Allen and Dunaway-Mariano, 2004). The HADS fold is unique according to SCOP, though CATH divides it into two domains: a common Rossman fold domain and a domain unique to the superfamily. Our laboratory has previously developed motifs in a manual process based on expert knowledge (Meng *et al*., 2004), and the residues in these motifs are here considered the conserved functional residues. While the catalytic roles may be conserved at each of these positions, all but the obligate aspartate are substituted in diverse members of the superfamily, as apparently required to accommodate differences in their specific mechanisms and overall functions. Despite these substitutions, most GASPS runs still contain three of the five functional residues (Fig. 3d). The nucleophilic aspartate appears in all significant motifs where possible. (The 1l7m structure contains two alternate conformations listed for this aspartate, D11, which precluded it from inclusion in a motif.) Nearly as frequent as the nucleophilic aspartate is another aspartate two positions downstream that has been implicated by others in binding and protonation of the substrate (Allen and Dunaway-Mariano, 2004).

3.2.4 Serine protease families SP are a polyphyletic group consisting mainly of two non-homologous families: the subtilisins and trypsins. They are grouped together by virtue of their common functions and use of a structurally similar catalytic triad in their active sites that appear to be the result of convergent evolution (Dodson and Wlodawer, 1998). Slightly more than half of the motifs and the highest scoring (10 of 19) included the entire triad (Fig. 3c). Most triad-containing motifs included only one additional residue: a glycine involved in formation of the conserved oxyanion hole (e.g. 2hlc G193) in trypsins. Though this glycine matches a conserved glycine in the subtilisins, the NH group in the subtilisins points away from the active site cavity. Of the nine remaining motifs, four had insignificant scores, three included partial catalytic triads and one was built from a heparin binding protein (1a7s) that, despite its homology to the trypsins, does not contain the catalytic triad or perform protease activity. Many significant runs seemed to be distracted by a disulfide bridge and neighboring alanine near the active site (C42–C58, A55, Fig. 3c and Supplementary Materials), which are well conserved among the trypsins but not the subtilisins.

3.2.5 Crotonase superfamily Members of CS display great catalytic diversity, yet all are unified by a common structure-based stabilization of an enolate anion intermediate of acyl-CoA substrates (Holden *et al*., 2001). Unlike the other groups given in Table 1, however, this shared chemistry is not performed by sidechains but by two structurally conserved NH groups of the peptide backbone that function as part of an oxyanion hole. The sidechains of these residues are not strictly conserved across the superfamily nor are there any other sidechains known or predicted to act in catalysis that are conserved across the entire superfamily. CS therefore provides a test of GASPS and sidechain-based motifs on a group that may not contain a structural motif focused on sidechains. As expected, an insignificant number of residues in the motifs (1 of 33, for all motifs) is involved in the formation of the characteristic oxyanion hole (Fig. 3e). The common residues in the motifs that do discriminate this superfamily seem unlikely to play a direct role in the enzyme's function, based on their distance from the active site. Examples include a conserved phenylalanine (1hnu F66) that is buried but lines an interior cavity and an aspartate (1hnu D135) involved in a conserved salt bridge.

4 DISCUSSION

4.1 Using GASPS for function identification

The performance of GASPS-generated motifs is comparable with that of 3D motifs generated based on expert knowledge of functional sites in other proteins (Artymiuk *et al*., 1994; Wallace *et al*., 1997; Fetrow and Skolnick, 1998; Kleywegt, 1999; Torrance *et al*., 2005). Furthermore, GASPS motifs improve the coverage of protein functions offered by publicly available sources of 3D motifs (Stark and Russell, 2003; Porter *et al*., 2004). Searching with protein fragments in three-dimensional motifs developed by GASPS was also found to be comparable or better than commonly used methods of annotation transfer that use an entire protein chain such as PSI-BLAST or CE. Unlike these methods that use an entire protein or domain, GASPS is able to focus on the features of protein structure

most likely to tell us the most about protein function. GASPS therefore provides a method of generating motifs useful for function or superfamily prediction in an automated fashion with no prior knowledge of mechanistic details. Such motifs can be used to verify similarity of active sites in proteins in which only similarity of fold has been previously identified. For example, GASPS motifs could be used for distinguishing functional differences among families of $(\beta/\alpha)_8$ proteins.

GASPS requires only a prior grouping of related proteins, so GASPS is limited only to groups with sufficient available structures. We cannot say for certain how many structures are required, but it appears to depend on the variability among the available structures. In the current study, all structures shared <40% sequence identity, and GASPS still was able to find general motifs for groups with as few as seven structures. While only 14% of superfamilies and 6% of families in the structure library used here have this many structures, these superfamilies and families make up the majority of protein structures (60 and 32%, respectively). Theoretically, it would seem possible for two highly diverged structures to share only their unique functional motif. However, for most proteins, even of different folds, it appears that sharing similar residues in 3D space occurs frequently enough by chance alone to require more than two structures to produce a trusted motif (Wangikar *et al*., 2003).

SPASM (and therefore GASPS, as used in this study) considers only a single point for each Cα and sidechain. With most catalysis carried out by sidechains (Bartlett *et al*., 2002), we believe the inclusion of the sidechains allows for better characterization of functional sites than if only the backbone placement were considered. Motifs could be represented with more precision by using the location of chemical groups, or even individual functional atoms. However, given the variability in placements of functional atoms in crystallographic structures, (DePristo *et al*., 2004; Torrance *et al*., 2005), approximating the entire sidechain by a single rigid point may be more appropriate.

4.2 Location of functional information

GASPS makes no assumptions about the location of functional information except that it can be resolved to individual residues and that it will be relatively well conserved. The observed correspondence between information useful for classification and functionally significant residues is a result of the choice of positive sets based on shared chemical activities used in this study. The use of GASPS on sets based on other shared characteristics, such as homology, binding partners or cofactors, may identify the residues most attributable to those shared characteristics.

It should be noted that the motifs generated by GASPS may not be the only, or even the most informative structural features. GASPS is expected to miss informative structural features if the features are either inconsistent between members of the group, such as the substituted residues in HADS, or if the features are not based on individual sidechains such as backbone interactions or helix dipoles. The CS results provide a case in point.

In addition to the previously identified functionally important positions, other positions occur with high frequency among the motifs for these groups. These positions may, for example, merely provide a simple geometric positioning constraint for the other motif residues that aid specificity. However, based on their conservation in 3D space, these positions are likely to play an important role for the protein, especially when located in the active site. For example, when the conserved 'downstream proline' in the ES is mutated to alanine in the muconate lactonizing enzyme from *Pseudomonas putida* (equivalent to structure 1muc) it results in an insoluble protein, (R. Nagatani and P. Babbitt, personal communication,) suggesting that this proline may be important for folding or stability of the soluble globular protein.

Based on its ability to identify at least a subset of the functionally important residues, GASPS appears similar to the fully automated DRESPAT, which successfully locates functionally important residues by identifying shared structural patterns in a set of functionally similar protein structures (Wangikar *et al*., 2003). The main differences between GASPS and DRESPAT are that GASPS compares patterns with a negative set and chooses patterns based on their predictive ability. Wangikar *et al*., (2003) suggest that DRESPAT patterns may represent useful 3D motifs. However, in the course of this study we found that when DRESPAT patterns were converted to motifs for use by the search tool SPASM, they were not as accurate as those motifs generated by GASPS.

4.3 Inference of function for diverse groups

Four of the five groups in this study have been described as 'mechanistically diverse superfamilies' (Gerlt and Babbitt, 2001) consisting of divergent enzymes that perform many different overall biochemical functions, but utilize a common mechanistic step such as a partial reaction. Any motifs that identify proteins to these groups will therefore identify the shared mechanistic step but not the overall biochemical function. By mapping a specific mechanistic step to specific structural elements, we are using a finer-resolution view of protein function than overall biochemical function, but one that is more appropriate for such diverse groups (Babbitt, 2003).

4.4 Future applications

If applied to an exhaustive functional classification of proteins, GASPS has the potential to generate an unbiased set of 3D motifs that can aid in function prediction for novel proteins. In addition to aiding protein classification, the collection of 3D motifs can represent hypotheses about determinants of function shared among related proteins. In this regard, the high-scoring motifs can serve as starting points for studies attempting to link function to structure, especially in a superfamily context. Additionally such a study would systematically investigate the utility of 3D motifs at identification of functions other than catalysis, such as ligand binding.

For groups with few experimental structures available, especially those coming from structural genomics initiatives, GASPS would have insufficient structures without the use of predicted structures, e.g. generated by homology modeling. Past work has specifically demonstrated the effectiveness of predicted structures for matching previously determined functional motifs (Arakaki *et al*., 2004). It remains unclear whether predicted structures can be used reliably for generating motifs. Work is in progress in our laboratory to investigate this issue.

ACKNOWLEDGEMENTS

We thank Elaine Meng for her useful discussions and careful reading of an earlier version of the manuscript. This work was supported by

NSF grant DBI-0234768. Funding to pay the Open Access publication charges was provided by NSF grant DBI-0234768 (to P.C.B.).

Conflict of Interest: none declared.

REFERENCES

Allen,K.N. and Dunaway-Mariano,D. (2004) Phosphoryl group transfer: evolution of a catalytic scaffold. *Trends Biochem. Sci.*, **29**, 495–503.

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.

Arakaki,A.K. *et al.* (2004) Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics*, **20**, 1087–1096.

Artymiuk,P.J. *et al.* (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.*, **243**, 327–344.

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

Babbitt,P.C. *et al.* (1996) The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry*, **35**, 16489–16501.

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.

Bartlett,G.J. *et al.* (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *Embo J.*, **5**, 823–826.

DePristo,M.A. *et al.* (2004) Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure*, **12**, 831–838.

Dodson,G. and Wlodawer,A. (1998) Catalytic triads and their relatives. *Trends Biochem. Sci.*, **23**, 347–352.

Elcock,A.H. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.*, **312**, 885–896.

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.

Gerlt,J.A. and Babbitt,P.C. (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.*, **70**, 209–246.

Gerlt,J.A. *et al.* (2005) Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. *Arch. Biochem. Biophys.*, **433**, 59–70.

Gerlt,J.A. and Raushel,F.M. (2003) Evolution of function in (beta/alpha)8-barrel enzymes. *Curr. Opin. Chem. Biol.*, **7**, 252–264.

Holden,H.M. *et al.* (2001) The crotonase superfamily: divergently related enzymes that catalyze different reactions involving acyl coenzyme a thioesters. *Acc. Chem. Res.*, **34**, 145–157.

Holm,L. and Sander,C. (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins*, **28**, 72–82.

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

Meng,E.C. *et al.* (2004) Superfamily active site templates. *Proteins*, **55**, 962–976.

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.

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

Orengo,C.A. *et al.* (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Pegg,S.C. *et al.* (2005) Representing structure-function relationships in mechanistically diverse enzyme superfamilies. *Pac. Symp. Biocomput.*, 358–369.

Porter,C.T. *et al.* (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.

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

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.

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*, **12**, 1405–1412.

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

Torrance,J.W. *et al.* (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J. Mol. Biol.*, **347**, 565–581.

Valdar,W.S. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.

Wallace,A.C. *et al.* (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.

Wangikar,P.P. *et al.* (2003) Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J. Mol. Biol.*, **326**, 955–978.

Zvelebil,M.J. and Sternberg,M.J. (1988) Analysis and prediction of the location of catalytic residues in enzymes. *Protein Eng.*, **2**, 127–138.