Composite Motifs Integrating Multiple Protein Structures Increase Sensitivity for Function Prediction

Brian Y. Chen[1], Drew H. Bryant[2], Amanda E. Cruess[1], Joseph H. Bylund[3], Viacheslav Y. Fofanov[4], Marek Kimmel[4], Olivier Lichtarge[5], Lydia E. Kavraki[1,2]

Abstract. The study of disease often hinges on the biological *function* of proteins, but determining protein function is a difficult experimental process. To minimize duplicated effort, algorithms for *function prediction* seek characteristics indicative of possible protein function. One approach is to identify substructural *matches* of geometric and chemical similarity between *motifs* representing known active sites and *target* protein structures with unknown function. In earlier work, statistically significant matches of certain *effective motifs* have identified functionally related active sites. Effective motifs must be carefully designed to maintain similarity to functionally related sites (*sensitivity*) and avoid incidental similarities to functionally unrelated protein geometry (*specificity*).

Existing techniques design motifs using the geometry of a single protein structure. Poor selection of this structure can limit motif effectiveness if the selected functional site lacks similarity to functionally related sites. To address this problem, this paper presents *composite motifs*, which combine structures of functionally related active sites to potentially increase sensitivity. Our experimentation compares the effectiveness of composite motifs with *simple motifs* designed from single protein structures. On six distinct families of functionally related proteins, leave-one-out testing showed that composite motifs had sensitivity comparable to the most sensitive of all simple motifs and specificity comparable to the average simple motif.

On our data set, we observed that composite motifs simultaneously capture variations in active site conformation, diminish the problem of selecting motif structures, and enable the fusion of protein structures from diverse data sources.

[1] Department of Computer Science, Rice University
[2] Department of Bioengineering, Rice University
[3] Department Ecology and Evolutionary Biology, Rice University
[4] Department of Statistics, Rice University
[5] Department of Molecular and Human Genetics, Baylor College of Medicine

1 Introduction

Developing an improved understanding of biological systems, the molecular basis of disease, and the design of novel and effective drugs are important efforts which could be enhanced with a broader understanding of the biological *function* of proteins. However, elucidating protein function is an expensive and time consuming experimental process, depending on the insight of experienced investigators and expensive laboratory equipment. To support and accelerate this cause, computational techniques for protein *function prediction* have been developed to gather evidence suggesting hypothetical functions of *target* proteins.

This paper focusses on one family of function prediction techniques that we call *motif matching* algorithms, such as Match Augmentation (MA) [11], Jess [2], PINTS [44], and pvSOAR [3], among many others. The evidence gathered by motif matching algorithms are instances of geometric and chemical similarity, *matches*, between *motif* structures, representing sites of known biological function, and substructures of *target* proteins, for which functional information is unavailable. In the past, matches with statistically significant geometric and chemical similarity have identified targets with sites functionally similar to the motif [44, 2, 3, 11], suggesting that matches may provide meaningful evidence of similar function.

One major challenge confronting the motif matching strategy is the fact that motifs are imperfect templates for geometric and chemical comparison. While generally they are designed to represent a known active site, the geometric form and chemical composition of active site characteristics can drastically affect the number of matching functionally related targets (motif *sensitivity*), as well as the number of unintended matches to unrelated sites (motif *specificity*). *Effective* motifs, which are both sensitive and specific, are critical for a successful application of motif matching. For this reason, motif refinement towards heightened sensitivity and specificity is a critical open problem. This paper contributes one practical method for motif refinement.

Motif refinement strategies in earlier work [10, 9, 41, 42] implement analyses which ultimately select geometric components for motifs from only one protein

Published in the Computational Systems Bioinformatics Conference (CSB), 2007, pp. 343–355.

structure. We refer to these motifs as *Simple Motifs*. In response, this paper asks if *Composite Motifs*, which combine the geometry of several active site structures, could better capture the natural variability inherent in functionally related active sites. We also asked if the design of motifs based on multiple protein structures could escape the potentially negative effects of using simple motifs.

This paper proposes two specific types of composite motifs, *averaged motifs* and *centered motifs*, which are constructed from a multiple structural alignment of related active sites. Beginning with a data set of 6 distinct families of functionally related proteins, we conducted a series of leave-one-out experiments to test the sensitivity and specificity of averaged and centered motifs. In comparison to all possible simple motifs from the same family, averaged and centered motifs performed with high sensitivity and average specificity, while simple motifs exhibited wildly varying sensitivity and specificity, demonstrating that composite motifs diminish the need to select individual motifs. Furthermore, the high sensitivity of averaged motifs also demonstrates that composite motifs can better capture geometric variations within a family of related sites.

This paper does not argue that composite motifs are a solution to the difficult problem of motif design. Rather, we propose that composite motifs are one method for achieving effective motifs which could compliment existing strategies for motif refinement, such as MULTIBIND [41, 42], Geometric Sieving [10], Cavity Scaling [9], and Surfnet-Consurf [25].

Composite motifs contribute to the study of motif refinement with three unique strengths: First, composite motifs capture variations in active site conformations, which are not apparent in any individual protein structure. Improved representation of active site conformations can enhance motif effectiveness. Second, composite motifs eliminate the problem of selecting an individual protein structure, eliminating the risk of selecting ineffective simple motifs. Finally, composite motifs provide a novel opportunity for the integration of protein structures from novel sources. Since the effectiveness of the motif is based on the geometry of a potentially large set of protein structures, alternative sources of protein structure data, such as snapshots from molecular dynamics simulations and NMR data, could be incorporated into the design of composite motifs. Composite motifs are a first step towards the synthesis of multiple protein structures for improved function prediction.

2 Related Work

The application of motif matching to protein function prediction is affected by at least three distinct subproblems:

1. selecting a functional site representation
2. designing a matching algorithm
3. filtering biologically irrelevant matches

This paper describes composite motifs, which contribute to the first subproblem. However, a complete demonstration of the effectiveness of composite motifs, in the context of function prediction, also requires solutions to the other two subproblems. This section explains existing approaches to all three subproblems in relation to our contributions.

2.1 Related Work in Motif Design

The design of effective motifs is a two stage problem requiring a computational representation of protein structure, or motif *type*, and the choice of specific active site elements to include, the motif *design*.

Motif types in earlier work can be loosely classified into two classes: *point-based* motifs, and *volume-based* motifs. Point-based motifs have used points in space to represent alpha carbon atoms [11, 45], sidechain atoms [1, 44], points [28] on the connolly surface [13], and chemical binding patterns [41, 42]. These *motif points* can be labeled with atomic and residue identity [11, 45, 2, 44], electrostatic potential [28], and evolutionary significance and variation [11], among many other chemical and biological properties. Labeling motif points allows additional chemical and biological knowledge to be mapped to an otherwise purely geometric comparison process, increasing the relevance of the motif type.

Volume-based motifs use spheres [9, 30, 33, 43, 46], grids [33] and other geometric representations, such as alpha shapes [3, 5, 14, 15], to represent active clefts and cavities in protein structures. Rather than directly representing atomic structure, volumetric motifs represent volumes that can be functionally significant, such as ligand or cofactor binding sites. While volume-based motifs are not always labeled, some techniques which apply volume-based motifs also integrate sequence analysis and point-based comparison with volumetric comparisons.

Once the motif type is chosen, given a specific active site to represent, a specific motif design must be established for the active site. For point-based motifs, this can involve the selection of the atoms thought to be most closely involved with the function of the protein. In the past, functionally documented amino acids from the literature [10], databases of catalytic

sites [2], and evolutionarily significant amino acids [11] have been used to design point-based motifs. Volumetric motifs have been designed by identifying statistically significant cavities and indentations on protein surfaces [6].

Given the active site to be represented, recent results suggest that a selection of amino acids can then be refined for geometric and chemical comparison. For example, identifying geometrically conserved binding patterns common among several functionally related active sites [41, 42] could yield additional matches to functionally related proteins. Motifs can be refined to be geometrically unique, recurring rarely among functionally unrelated proteins [10]. Finally, point-based motifs can be augmented with volumetric data and eliminate matches lacking functionally significant cavities [9].

Volumetric motifs have been refined by identifying indentations on the protein surface that are distant from evolutionarily significant amino acids [25]. In addition, high-impact volumes within a surface clefts, which seem to be essential for functionally related matches, can be automatically identified to refine cavity-aware motifs [9].

This paper provides a unique approach to the refinement of point-based motifs. While other motif refinement techniques focus on the selection of amino acids [41, 42, 10] or integrate additional data [10, 29], this paper improves on existing motif designs by incorporating the geometry of other protein structures containing similar active sites. In our experimentation, we asked if this approach would yield motifs that more closely resemble the population of structures with functionally related active sites. The possibility of integrating multiple protein structures yields the first technique, to our knowledge, where motifs can contain geometric information not taken directly from a single protein structure.

Our approach is most related to techniques designed to represent a range of protein structures, such as hinge-bending point-based motifs [40], and motifs representing conserved binding patterns [41, 42]. Hinge-bending motifs can represent multiple protein structures, but only capture structures implied by the range of hinge motions, which can differ from the population of proteins containing similar functional sites. In comparison, the composite motifs studied in this work are built explicitly from populations of protein structures with similar functional sites. Motifs representing conserved binding patterns represent the largest common set of motif points between a set of functionally similar active sites, but the largest common set of motif points may not include functionally significant motif points with geometric variations in active site conformations. In contrast, our techniques for generating composite motifs, described in Section 3, can represent a consensus structure among these variations.

2.2 Earlier Motif Matching Algorithms

Motif matching algorithms are designed for compatibility and efficiency with a specific motif type. In addition to full structure alignment methods such as DALI [26], which could be applied to the motif matching problem, motif matching algorithms for point-based motifs include Geometric Hashing [31, 47], JESS [2], PINTS [39], and Match Augmentation [11, 9] (MA). One unique advantage of composite motifs is that composite motifs are point-based motifs that are assembled in a novel manner but remain compatible with existing point-based motif matching algorithms.

Motif matching algorithms are also designed for compatibility with volume-based motifs, such as pv-SOAR [4, 5]. Other function prediction and analysis techniques based on volume-based motifs analyze a single protein structure in an effort to identify characteristics consistent with an active site: SCREEN [36] identifies cavities which are likely to be drug binding sites, SURFNET [32] and SURFNET-Consurf [25] seek to identify catalytic sites. CASTp [6] analyzes cavities on the protein surface and identifies those probable of biological activity.

2.3 Statistical Models for Motif Matching

Having found a set of matches using a motif matching algorithm, the final subproblem for function prediction via motif matching is to eliminate matches which are unlikely to have any biological relevance. In several approaches to motif matching, statistical models have been developed which model the degree of geometric and chemical similarity observed in matches with functionally related proteins. In comparison to a baseline degree of similarity observed in matches at random, matches to functionally related proteins exhibit statistically significant geometric and chemical similarity. The statistical models employed by PINTS [44], JESS [2], and MA [11], have been shown to be capable of identifying functionally related active sites.

Statistical models can be used to assign p-values to a given match. The p-value estimates the probability of observing another target, selected at random, with greater geometric and chemical similarity than the target identified with the given match. Thus, a match is statistically significant if the p-value falls below a given significance threshold α.

2.4 The MASH pipeline

In earlier work [8], we developed the MASH software pipeline, which contains a matching algorithm and

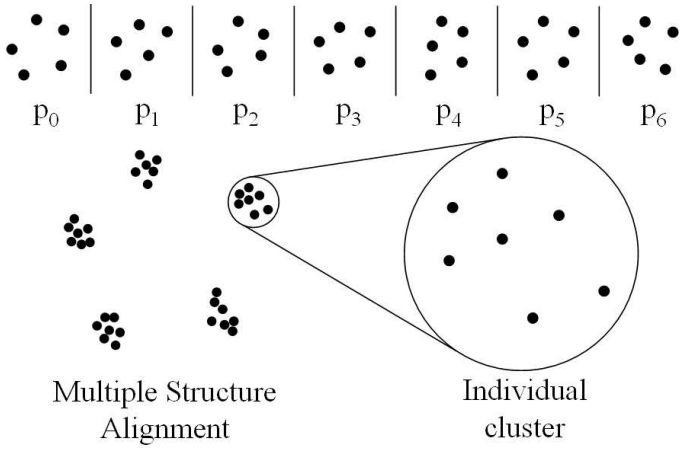

Fig. 1. Composite motif construction begins with the multiple structure alignment of the individual motifs p_0, p_1, etc, yielding clusters of correlated points in the ultimate alignment. We describe this iterative alignment process in Section 3.2.

a statistical model for identifying matches to point-based motifs. Because of its availability and compatibility with composite motifs, we use MASH to benchmark the effectiveness of composite motifs in our experimentation.

As input, MASH takes a simple or composite motif, a target protein structure, and a reference set of protein structures. Using MA [11], MASH computes a match m between the motif and the target as well as a match between the motif and all members of the reference set. Then, applying our statistical model [11], MASH uses these matches to assign a p-value to m. The output of MASH is the match m, and the p-value of m. If $p < \alpha$, then we say that the match m is statistically significant, and a positive prediction of functional similarity. Otherwise, we say that m is statistically insignificant, and a negative prediction of functional similarity.

In our experimentation, we use MASH for experimentation on composite motifs and running control experiments on simple motifs.

3 Generating Composite Motifs

In our experimentation, we asked if composite motifs represent geometric variations in functionally related active sites better than simple motifs. For this reason, we detail both simple and composite motifs here.

3.1 Simple Motifs

Derived originally from a single protein structure P_0, a simple motif p_0 is composed of l points in space $p_{(0,0)}, p_{(0,1)}, \ldots, p_{(0,l)}$, where the coordinates for each $p_{(0,i)}$ are derived from an atom in P_0.

Each *motif point* $p_{(0,i)}$ is also labeled with biological and chemical information. Initially, each motif point is identified with its atom type and amino acid type within P_0. Each motif point also bears a *ranking* $r(p_{(0,i)})$ which is associated with the functional importance of the motif point. The matching algorithm used in this paper, MA [11] is capable of prioritizing its search for motifs in order of functional importance. Finally, each motif point also contains a list of associated amino acids $l(p_{(0,i)})$, called *alternate labels*, which represent acceptable substitutions in matching target amino acids. This permits our motifs to represent amino acids substitutions in major evolutionary divergences [11, 34, 35] or variations between distinct but chemically related amino acids.

3.2 Composite Motifs

Composite motifs are point-based motifs whose motif points are positioned by the geometric consensus of related active site structures. This paper presents averaged and centered motifs which are two examples of composite motifs designed from related active sites.

In the design of composite motifs, we begin with a set of k protein structures P_0, P_1, ..., P_k, where each P_i is contains a functionally related active site, which is defined as an *individual motif* $p_i = \{p_{(i,0)}, p_{(i,1)}, \ldots, p_{(i,n)}\}$ with exactly n motif points. Given that these motifs are functionally related, we list the motif points in p_0, p_1, ..., p_k in such an order that for any i, $0 \leq i \leq n$, the motif points $p_{(0,i)}, p_{(1,i)}, \ldots, p_{(k,i)}$ are functionally identical. Furthermore, for any i, $0 \leq i \leq n$, the motif points $p_{(0,i)}, p_{(1,i)}, \ldots, p_{(k,i)}$ are assigned the same ranking and the same alternate labels.

Using a method from [48], we first compute a multiple structural alignment of the individual mo-

tifs. This is accomplished by first computing a least RMSD (LRMSD) alignment[1] of each p_i to an arbitrarily selected p_j. In each alignment between one p_i and p_j, $p_{(i,0)}$ is correlated to $p_{(j,0)}$, $p_{(i,1)}$ is correlated to $p_{(j,1)}$, etc, resulting in a cluster containing all $p_{(i,0)}$, a cluster containing all $p_{(i,1)}$, and so on. We compute a centroid for each cluster, and refer to each centroid as $c_0, c_1, \ldots, c_l$. In the next iteration, we align each p_i to this set of centroids, instead of the arbitrarily selected individual motif, and recompute the centroids for the new multiple structural alignment. Repeated iterations converge rapidly to a single multiple structural alignment [48], with centroids $C_0, C_1, \ldots, C_l$.

Once the multiple structural alignment is complete, we use the newly aligned formation of structures to finalize averaged and centered motifs.

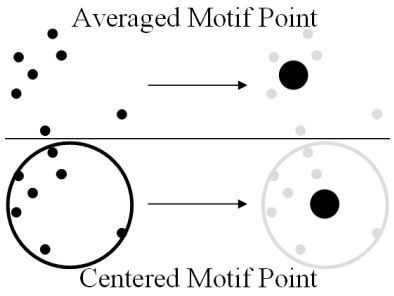

Fig. 2. The multiple structure alignment of the individual motifs generates clusters of correlated motif points, demonstrated on the left side of this figure. As demonstrated above, averaged motif points are positioned at the centroid of the cluster. Centered motifs, demonstrated below, compute the smallest containing sphere around the correlated motif points, and use the center of the sphere for the composite motif point.

Averaged Motifs Averaged motifs use C_0, C_1, …, C_l as the coordinates of their motif points. This is demonstrated in Figure 2. Once we have the coordinates of the averaged motif points, the labels, ranking, and alternate labels, being identical in each of p_0, p_1, …, p_k, are applied respectively to each of C_0, C_1, …, C_l, completing an averaged motif.

Centered Motifs Centered motifs are initially generated with the same iterative multiple structural alignment. However, once the alignment is complete, the smallest sphere containing each cluster of correlated motif points is computed, and the center of the sphere is used for each composite motif point. We demonstrate this in Figure 2. Again, the labels, ranking, and alternate labels are mapped to each of these points.

Advantages of Composite Motifs We designed composite motifs to represent variations in active site structures, to reduce the need to select motif structures, and to promote the fusion of protein structures from varying data sources. Towards the first goal, averaged and centered motifs select points in space to represent the variation exhibited by each motif point. This straightforward approach is strongly applicable to the natural variability of protein structures, under the assumption that geometric identity implies functional similarity.

Generating a single composite motif that represents a set of related sites also reduces the problem of selecting a single protein structure to represent the entire set. In our experimentation, we will test the degree to which composite motifs can identify functionally related proteins, in comparison to simple motifs based on individual related sites. One concern we had was that some sites might be overrepresented in the family of protein structures, thereby affecting motif points in averaged motifs. Since structural overrepresentation is inevitable, due to the fact that structures are unavailable for all proteins, we designed centered motifs, to use the geometric position of the overall cluster (the smallest surrounding sphere) for motif points.

Composite motifs have the distinctive characteristic that protein structure data from many sources could be fused in a single representation. As the availability of protein structures and functional annotations accelerates, composite motifs could provide a useful method for applying additional knowledge towards function prediction. In particular, because hundreds of protein structures can be integrated into composite motifs, additional sources of data, such as snapshots from molecular dynamics simulations and models from structure prediction techniques, could be integrated to counterbalance experimental biases inherent in existing structures and further expand the set of structural variations represented by composite motifs.

4 Experimentation

In controlled experimentation, this section compares the effectiveness of simple motifs against averaged and centered motifs. First, we identified 6 families in the Enzyme Commission (EC) classification which contained many distinct protein structures with functionally related active sites. Treating these classifications as a gold standard for functional similarity, we

[1] An LRMSD alignment of two sets of points A and B rotates and translates A to the position where root mean squared deviation (RMSD) between A and B is minimized

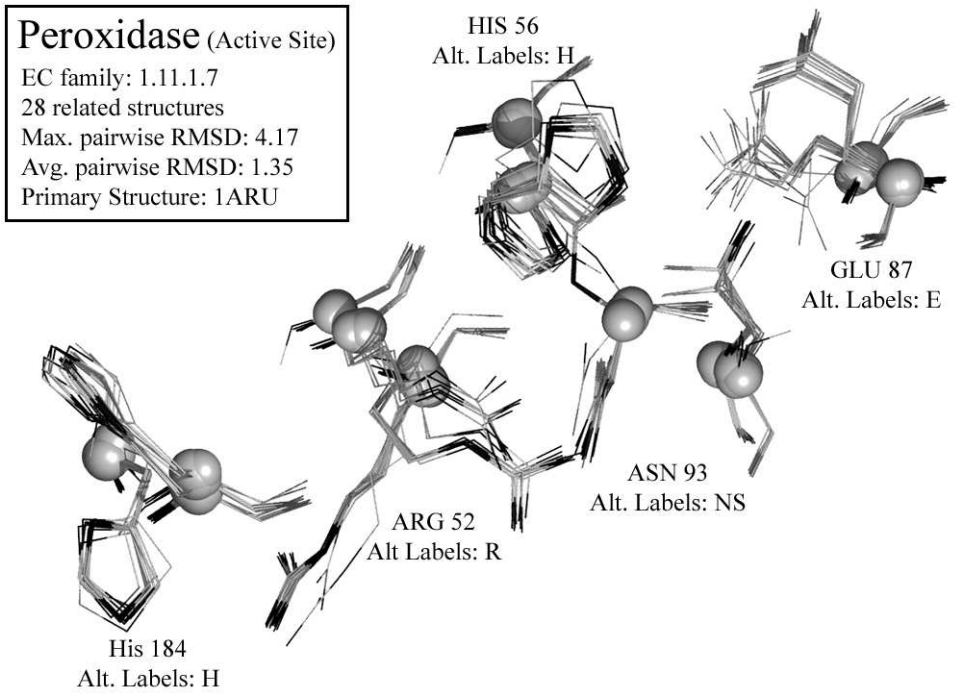

Fig. 3. Multiple structural alignment of Peroxidase active sites in EC family 1.11.1.7. The substructures aligned in this image demonstrate the distinct geometric variability of related sites in each EC family. Structural differences between sites in each structure are apparent in both sidechain conformations as well as alpha carbon (spheres, in this image) positions. Some families, such as 1a3h, were distinctly more variable, while others, such as 1did, exhibited less variability.

used each family to generate averaged and centered motifs on a leave-one-out basis. Finally, we tested the effectiveness of these averaged and centered motifs to identify statistically significant matches with the left out structure, in comparison to simple motifs.

4.1 Protein Families

In this work, we identified six families of proteins within the Enzyme Classification (EC) specified by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology [27], which, although imperfect, is standard and useful for our purposes. In each family, we required one *primary structure*, with functional amino acids documented in the literature, as well as at least 10 other non-mutant protein structures (although EC families with more structures were preferred), all with resolution below 3Å.

The next six paragraphs describe the functionally documented amino acids from each primary structure. For simplicity, in our experimentation, we will refer to each EC family (bolded, below) using the PDB code (bolded, below) of its primary structure.

1a3h/3.2.1.4 Bacillus agaradherans endoglucanase is a cellulase and belongs to EC family 3.2.1.4. Five points were selected for this motif, including tryptophan 262, which exists in an orientation that allows it to interact with substrate, tryptophan 178, which is an invariant residue in the subfamily 5-2 enzymes that is part of the aglycon binding sites, and histidine 206, which may play an important role in catalysis, perhaps as part of substrate binding [19]. Glutamic acid 139 and 228 were also included, being the catalytic acid/base and the enzymatic nucleophile, respectively [19].

1aru/1.11.1.7 Peroxidase from the fungus Arthromyces ramosus is a heme protein belonging to EC family 1.11.1.7. Five points were selected for this motif, including histidine 184, which binds the heme iron [23], and the distal arginine (Arg-52 in this structure [21]), which has been proposed to play a role in substrate binding and stabilization of the product of the first step of the enzyme reaction [24]. Also included was histidine 56, which is suggested to be responsible for proton translocation in the hydrogen peroxide substrate and has been shown to undergo conformational change in complexes with both cyanide and triiodide [21]. Asparagine 93 and glutamic acid 87 form a hydrogen bond network with histidine 56 [21].

	1asy	1did	1k55	1rx7	1a3h	1aru
Min. Å	0.072773	0.000272	0.018086	0.007937	0.000383	0.00021
Max. Å	3.034972	0.820726	7.134243	5.299205	5.754516	4.169486
Avg. Å	1.947437	0.251243	3.790644	1.514413	2.429289	1.346931
# of Structs.	14	93	181	132	119	28

Fig. 4. A summary of the variations in geometric similarity between all pairs of simple motifs used in experimentation, as well as the number of structures in each family. Families denoted by the PDB code of their primary structure.

1asy/6.1.1.12 Aspartyl-tRNA synthetase is a dimeric aminoacyl tRNA synthetase responsible for the translation of genetic information and belongs to EC family 6.1.1.12. Eight points were selected for this motif. Serine 329 is part of a loop that interacts with the discriminator base G73 and the first base pair of the stem of the tRNA molecule, serine 423 and lysine 428 are the endpoints of a segment that interacts with the phosphate groups of A72 and G73, and lysine 293 is the only residue making direct contact with a tRNA molecule bound to the other monomer [17]. Arginine 325 and 531 are involved in binding the ATP substrate, bonded to the α-phosphate and γ-phosphate, respectively [16], while aspartic acid 342 plays a role in binding the amino groups of the aspartic acid substrate [18]. Proline 273 has been confirmed to be essential in the dimerization [20], and enzymatic activity has been shown to decrease markedly when this residue is substituted [16].

1did/5.3.1.5 D-xylose isomerase, belonging to EC family 5.3.1.5, converts xylose to xylulose, such as in the conversion of glucose to fructose. Six points were selected for this motif. It has been proposed that aspartic acid 56 polarizes and activates histidine 53, which acts as a base to catalyze ring opening, and that lysine 182 aides in isomerization, while tryptophan 136 and phenylalanine 93 and 25 from a completely hydrophobic environment in which the hydride shift occurs [12].

1k55/3.5.2.6 Class D β-Lactamase, a member of EC family 3.5.2.6, is responsible for the hydrolysis of β-lactam antibiotics, and as a result, it is one of the causes of bacterial resistance to this group of antibiotics [22]. Eight points were selected for this motif. Serines 67 and 115 and lysine 205 are among the residues active in catalysis, while phenylalanines 69 and 120, valine 117, tryptophan 154, and leucine 155 create a hydrophobic pocket within the active site [22].

1rx7/1.5.1.3 Dihydrofolate reductase, belonging to EC family 1.5.1.3 and required for normal metabolism in prokaryotic and eukaryotic cells, is an enzyme that catalyzes the NADPH-dependent reduction of 7,8-dihydrofolate to 5,6,7,8-tetrahydrofolate [38]. Seven points were chosen for this motif. Histidine 45 creates an ionic interaction with the pyrophosphate moiety of the NADP+ coenzyme and makes a bifurcated hydrogen bond with two oxygens of the ADP group [7]. Glycine 96 also makes such a hydrogen bond with two oxygens of the ADP 5'-phosphate [7]. Aspartic acid is the single polar residue in the folate binding cleft and participates in the catalyzing reduction of 7,8-dihydrofolate in two ways: by indirect protonation of N5 and by the precise positioning of the dihydropteridine ring through H-bonding [7]. Phenylalanine 31 forms a rigid ceiling to the pteridine binding site, which appears to be important for catalysis [7]. Isoleucine 50 is among the residues that create a hydrophobic pocket surrounding the folate tail [7]. Finally glycine 15 is part of group of amino acids that function as a lid that controls that entry and exit of ligands into the enzyme, and tryptophan 22 is involved in the slow, rate-limiting release of product [38].

4.2 Motifs used in Experimentation

Simple Motifs From every structure in every family, we created one simple motif as a control set for our experimentation.

Creating a simple motif for the primary structure in each family was accomplished by running the Evolutionary Trace (ET) [34, 35] to identify alternate labels and a ranking of evolutionary significance (see Section 3.1) for all functionally documented amino acids. The geometric positions of the alpha carbons in functionally documented amino acids, coupled with the alternate labels and ranking provided by ET, complete a *primary motif* for each family.

Creating a simple motif for all non-primary structures in each family is substantially more difficult, because functional documentation was not available for many non-primary structures. For this reason, we applied MA [11, 10] to search for the primary motif in the other structures of each protein family, identifying a set of *similar sites*. In each structure, we use the most geometrically similar site as the simple motif.

The lack of functional documentation in many of the non-primary structures of each family leaves few alternative methods for discovering similar sites, but regardless of which site is used, MA is no substi-

tute for functional documentation. Existing alternative methods, such as sequence comparison and other structure comparison algorithms, do not provide any improved guarantees to identify cognate active sites. A similar approach for identifying related sites was implemented in the Catalytic Site Atlas [37], which uses sequence analysis to relate functionally documented amino acids to similar amino acids in proteins of related function. Sequence analysis does not guarantee functional similarity, but significantly widens the range of similar active sites.

In order to minimize any bias introduced by MA, we used very broad geometric thresholds when searching for similar sites. We used MA to consider all similar sites which had matching alpha carbons as distant as 10Å in the LRMSD alignment, while searching for the site with smallest LRMSD. Geometric thresholds used by MA do not appear to have significantly biased the set of simple motifs. As documented in Figure 4, between the simple motifs of each family, we measured the degree of pairwise geometric similarity, and observed notable geometric variations in all families except 1did.

In our experimentation, a statistically significant match between a simple motif and a structure in the same family is called a true positive (TP) match, and a statistically significant match to a structure outside the family is a false positive (FP) match. A statistically insignificant match to a structure inside the family is a false negative (FN), and a statistically insignificant match to a structure outside the family is called a true negative (TN).

Composite Motifs For each family of k simple motifs, we also created k averaged and k centered motifs in a leave-one-out manner. This is accomplished by identifying the $k-1$ simple motifs that are not left out, and using them as individual motifs in the construction of an averaged or a centered motif, as described in Section 3.2.

Assembling simple motifs creates a test set where each composite motif can be tested against the left out structure. For each leave-one-out motif generated, if the left-out member of the protein family has a statistically significant match, then we call this match a TP. If the left out structure is not statistically significant we call the match a FN. FP and TN matches are counted in the same way as simple motifs.

4.3 Experimental Protocol

For every simple and composite motif, we computed matches between the motif and every member of the associated protein family. We also computed matches between the motif and 5000 randomly sampled structures from the PDB, to represent a set of functionally unrelated proteins. We then assessed the statistical significance of each match computed, and counted the number of TPs, FPs, TNs, and FNs for all motifs.

Given greater computing time, the set of randomly sampled PDB structures could be expanded further. However, in earlier work [11, 10] we observed that sampling 5% (5000 is more than 5%) of the PDB can reasonably represent the geometric composition of the proteins in the PDB. For this reason, sampling 5000 functionally unrelated proteins was deemed sufficient to simulate the number of FP matches observed in general conditions. Overall, approximately 4054 distributed CPU hours were spent gathering these matches.

4.4 Implementation Specifics

This work uses a snapshot of the PDB database from 09.14.2006. Structures with multiple chains were divided into separate structures, producing 93582 structures. While separating chains might block the identification of matches to active sites that span multiple chains, re-integration of separate chains might yield errors which lead to chemically impossible protein structures. None of the motifs used in this experimentation span separate chains.

Composite motifs were computed using C/C++ code developed on an Athlon XP 2600+, with 1Gb of ram, running Debian Linux. Computing averaged and centered motifs, described in Section 3, takes approximately 10-15 seconds on this machine. P-values and matches computed using distributed MASH, documented in [10], was run on Ada, a 28 chassis Cray XD1 with 672 2.2Ghz AMD Opteron cores.

4.5 Averaged and Centered Motifs are Sensitive and Specific

We compared the sensitivity and specificity of averaged and centered motifs to the sensitivity of every possible simple motif in each protein family.

Observed sensitivity is plotted Figure 5. The horizontal axis represents each family of EC proteins, denoted by their primary structure. The vertical axis represents sensitivity: the proportion of TP matches observed relative to the number of proteins in the protein family. The black brackets, each having three hash marks, signify the minimum, mean, and maximum number of TP matches identified by simple motifs in the EC class. Every simple motif in the family corresponding to `1did` matched all members of the family. The dark grey line represents the number of TP matches identified by centered motifs, and the light grey line represents the number of TP matches identified by averaged motifs. Averaged motifs were among the most sensitive of all individual matches.

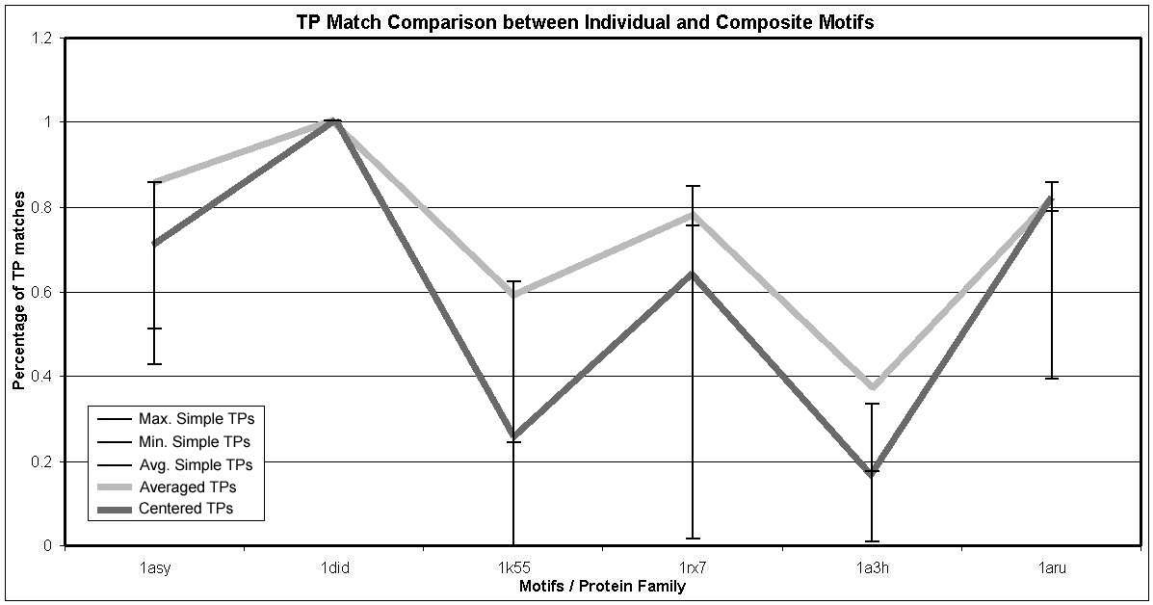

Fig. 5. A comparison of TP matches found by composite motifs, relative to TP matches found by simple motifs from the same family. On the vertical axis, we normalize the total proportion of TP matches for each family; a value at 1.0 demonstrates that the motif identified statistically significant matches to all structures in its EC family. On the horizontal axis, we chart the protein families studied in this work. The vertical black bars indicate the maximum, minimum, and average number of TP matches identified by single-structure motifs from each EC family. It is apparent, with the exception of 1did, that single-structure motifs can fall within a wide range of sensitivity. The dark and light grey lines signify the number of TP matches identified by centered and averaged motifs, respectively. Composite motifs, especially averaged motifs, are significantly more sensitive than most simple motifs on almost all protein families studied.

One family of protein structures, 1did, demonstrated very low structural variability. This is consistant with the observation from Figure 4 that simple motifs in 1did expressed little geometric variability. As a result, composite motifs generated from this family performed perfectly also.

Among individual motifs, sensitivity fluctuates significantly. For example, in the family of 1rx7, some individual motifs identify matches with only 2 out of the 136 remaining members of the family, while other individual motifs identify as many as 112. In the family of 1aru, some individual motifs identify matches with only 11 out of the 27 remaining members of the family, while others identify as many as 24. The choice of individual structures for motif design significantly risks the sensitivity and specificity of the motif created. In comparison, the sensitivity of averaged motifs was consistantly greater than the mean sensitivity of individual motifs, which was similar to the sensitivity of centered motifs as well. With the exception of averaged motifs for 1a3h, composite motifs in general did not outperform all individual motifs. This demonstrates that composite motifs largely avoid the problem of selecting individual motifs, and that averaged motifs can achieve very high sensitivity.

We measured specificity in Figure 6. The horizontal axis again corresponds to each family of EC proteins, and the vertical axis corresponds to the number of FP matches, from the random sample 5000 PDB proteins, observed for each motif. We report the number of FPs observed, instead of specificity, because there are so many more unrelated proteins than functionally related proteins, that specificity is almost always close to 99%. Reporting the number of FPs makes the results easier to observe. The black brackets correspond to the highest, lowest, and mean number of FP matches to each individual M_i. The dark grey and the light grey lines correspond to the number of FP matches to centered and averaged motifs, respectively. The mean number of FP matches observed with simple motifs was very similar to the number of FP matches observed with centered and averaged motifs.

The number of FPs observed can fluctuate significantly among individual motifs. In 1a3h, some individual motifs identify 123 FP matches, whereas others identify only 41. In other families, specificity did

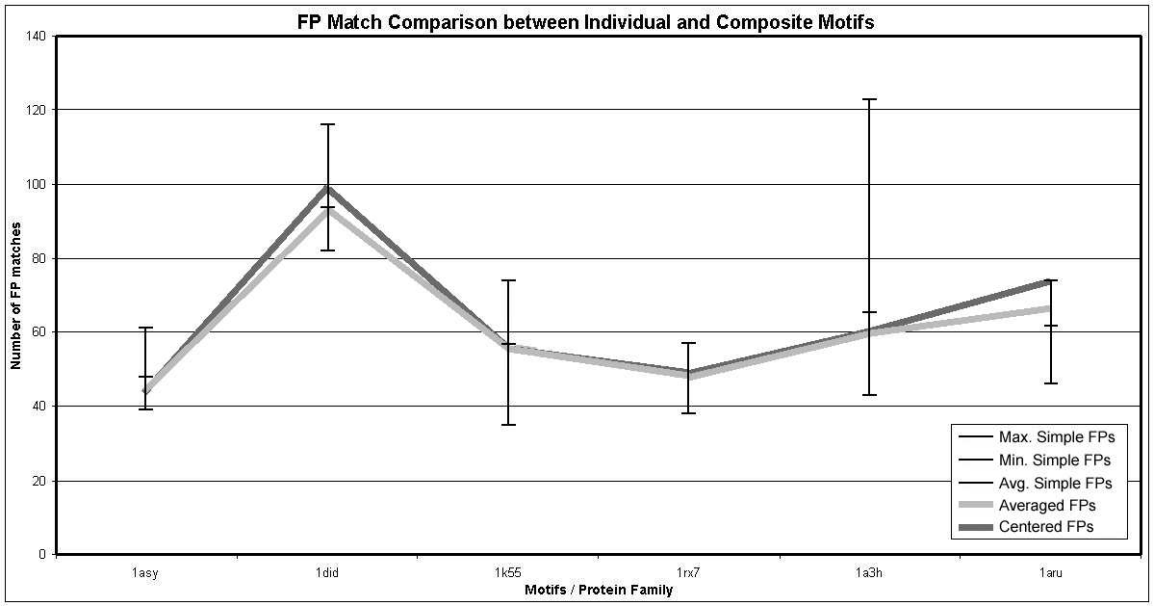

Fig. 6. A comparison of FP matches found by composite motifs, relative to FP matches found by simple motifs from the same family. On the vertical axis, we plot the number of FP matches observed. On the horizontal axis, we chart the protein families studied in this work. The vertical black bars again indicate the maximum, minimum, and average number of FP matches identified by single-structure motifs from each EC family. The dark and light grey lines signify the number of FP matches identified by centered and averaged motifs, respectively. With one exception, composite motifs tend to identify an average number of FP matches, in comparison to single-structure motifs, demonstrating that composite motifs are not an additional source of prediction error.

not fluctuate as much, such as in 1rx7, where individual motifs identified between 38 and 57 FP matches. In comparison, averaged and centered motifs almost always identified an average number of FP matches. Composite motifs appear to avoid high false positive rates which can occur with individual motifs, again reducing the problem of selecting individual protein structures.

5 Conclusions

We have described composite motifs, a unique approach to motif refinement. Overall, composite motifs seem to achieve sensitivity among the most sensitive individual motifs, while maintaining average specificity and eliminating the problem of accidentally selecting an ineffective simple motif.

On 6 families of functionally related proteins, our experimentation demonstrates, on a small scale, that composite motifs can capture variations in active site conformations. We observed that averaged motifs performed with sensitivity comparable to the most sensitive simple motifs, and that centered motifs performed with sensitivity typical of the average simple motif. While increasing sensitivity, averaged and centered motifs tended to identify FP matches typical of the average simple motif.

We also observed that simple motifs had sensitivity and specificity falling in a very wide range. Selecting any individual structure for the design of a motif risks the selection of insensitive or nonspecific simple motifs. In our experimentation, we observed that composite motifs may diminish this problem, because no selection needs to be made, and because they performed with high sensitivity and average specificity.

As the availability of protein structures and functional annotations accelerates, we feel that composite motifs will become increasingly applicable for effective annotation of protein structures and for the integration of additional types of structural information from diverse data sources.

Acknowledgements
This work is supported in part by a grant from the National Science Foundation NSF DBI-0547695 and NSF DBI-0318415 through a subcontract from the Baylor College of Medicine. Additional support is gratefully acknowledged from training fellowships from the W.M. Keck Center for Interdisciplinary Training (NLM Grant No. 5T15LM07093) to B.C.; from March of Dimes Grant FY03-93 to O.L.; from a Sloan Fellowship to L.K; and from a VIGRE Training in Bioinformatics Grant from NSF DMS 0240058 to V.F. Experiments were run on equipment funded by NSF EIA-0216467 and NSF CNS-0523908. Large production runs were done on equipment

supported by NSF CNS-042119, Rice University, and partnership with AMD and Cray. D.B. has been partially supported by the W.M. Keck Undergraduate Research Training Program and by the Brown School of Engineering at Rice University.

References

1. O. Bachar, D. Fischer, R. Nussinov, and H. Wolfson. A computer vision based technique for 3-d sequence independent structural comparison of proteins. *Prot. Eng.*, 6(3):279–288, 1993.
2. J.A. Barker and J.M. Thornton. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinf.*, 19(13):1644–1649, 2003.
3. T.A. Binkowski, L. Adamian, and J. Liang. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.*, 332:505–526, 2003.
4. T.A. Binkowski, P. Freeman, and J. Liang. pvsoar: Detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucl. Acid. Res.*, 32:W555–8, 2004.
5. T.A. Binkowski, A. Joachimiak, and J. Liang. Protein surface analysis for function annotation in high-througput structural genomics pipeline. *Protein Science*, 14:2972–2981, 2005.
6. T.A. Binkowski, S. Naghibzadeh, and J. Liang. Castp: Computed atlas of surface topography of proteins. *Nucl. Acid. Res.*, 31(13):3352–55, 2003.
7. C. Bystroff, S.J. Oatley, and J. Kraut. Crystal structures of *Escherichia coli* dihydrofolate reductase: the nadp$^+$ holoenzyme and the folate-nadp$^+$ ternary complex. substrate binding and a model for the trasition state. *Biochemistry*, 29:3263–3277, 1990.
8. B.Y. Chen. *Geometry-based Methods for Protein Function Prediction*. PhD thesis, Rice University, September 2006.
9. B.Y. Chen, D.H. Bryant, V.Y. Fofanov, D.M. Kristensen, A.E. Cruess, M. Kimmel, O. Lichtarge, and L.E. Kavraki. Cavity-aware motifs reduce false positives in protein function prediction. *Proceedings of the 2006 IEEE Computational Systems Bioinformatics Conference (CSB 2006)*, pages 311–23, August 2006.
10. B.Y. Chen, V.Y. Fofanov, Bryant D.H., B.D. Dodson, D.M. Kristensen, A.M. Lisewski, M. Kimmel, O. Lichtarge, and L.E. Kavraki. Geometric sieving: Automated distributed optimization of 3D motifs for protein function prediction. *Proceedings of The Tenth Annual International Conference on Computational Molecular Biology (RECOMB 2006)*, pages 500–15, April 2006.
11. B.Y. Chen, V.Y. Fofanov, D.M. Kristensen, M. Kimmel, O. Lichtarge, and L.E. Kavraki. Algorithms for structural comparison and statistical analysis of 3D protein motifs. *Proceedings of Pacific Symposium on Biocomputing 2005*, pages 334–45, 2005.
12. C.A. Collyer and D.M. Blow. Observations of reaction intermediates and the mechanism of aldose-ketose interconversion by d-xylose isomerase. *Proc. Natl. Acad. Sci.*, 87:1362–1366, 1990.
13. M.L. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221:709–713, 1983.
14. H. Edelsbrunner, M. Facello, and J. Liang. On the definition and the construction of pockets in macromolecules. *Discrete Applied Mathematics*, 88:83–102, 1998.
15. H. Edelsbrunner and E.P. Mucke. Three-dimensional alpha shapes. *ACM Trans. Graphics*, 13:43–72, 1994.
16. Cavarelli J. et. al. Yeast aspartyl-trna synthetase: a structural view of the aminoacylation reaction. *Biochimie*, 75:1117–1123, 1993.
17. Cavarelli J. et. al. Yeast trna^{Asp} recognition by its cognate class ii aminoacyl-trna synthetase. *Nature*, 362:181–184, March 1993.
18. Cavarelli J. et. al. The active site of yeast aspartyl-trna synthetase: structural and functional aspects of the aminoacylation reaction. *EMBO J.*, 13(2):327–337, January 1994.
19. Davies G.J. et. al. Structure of the *Bacillus agaradherans* family 5 endoglucanase at 1.6 åand its cellobiose complex at 2.0 åresolution. *Biochemistry*, 37:1926–1932, 1998.
20. Eriani G. et. al. Role of dimerization in yeast aspartyl-trna synthetase and importance of the class ii invariant proline. *Proc. Natl. Acad. Sci. USA*, 90(22):10816–10820, November 1993.
21. Fukuyama K. et. al. Crystal structures of cyanide- and triiodide-bound forms of *Arthromyces ramosus* peroxidase at different ph values: perturbations of active site residues and their implication in enzyme catalysis. *J. Biol. Chem.*, 270(37):21884–21892, September 1995.
22. Golemi D. et. al. The first structural and mechanistic insights for class d β-lactamases: evidence for a novel catalytic process for turnover of β-lactam antibiotics. *J. Am. Chem. Soc.*, 122:6132–6133, 2000.
23. Kunishima N. et. al. Crystal structure of the fungal peroxidase from arthromyces ramosus at 1.9 a resolution. structural comparison with the lignin and cytochrome c peroxidases. *J. Mol. Biol.*, 235(1):331–344, January 1994.
24. Vitello L.B. et. al. Effect of arginine-48 replacement on the reaction between cytochrome c peroxidase and hydrogen peroxide. *Biochemistry*, 32(37):9807–9818, September 1993.
25. F. Glaser, R.J. Morris, R.J. Najmanovich, R.A. Laskowski, and J.M. Thornton. A method for localizing ligand binding pockets in protein structures. *Proteins*, 62(2):479–88, 2006.
26. L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233:123–138, 1990.
27. Nomenclature Committee. International Union of Biochemistry. *Enzyme Nomenclature*. Academic Press: San Diego, California, 1992.

28. K. Kinoshita and H. Nakamura. Identification of protein biochemical functions by similarity search using the molecular surface database ef-site. *Protein Science*, 12:15891595, 2003.
29. D.M. Kristensen, B.Y. Chen, V.Y. Fofanov, R.M. Ward, A.M. Lisewski, M. Kimmel, L.E. Kavraki, and O. Lichtarge. Recurrent use of evolutionary importance for functional annotation of proteins based on local structural similarity. *Protein Science*, 15(6):1530–6, Jun 2006.
30. I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, and T.E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.
31. Y. Lamdan and H.J. Wolfson. Geometric hashing: A general and efficient model based recognition scheme. *Proc. IEEE Conf. Comp. Vis.*, pages 238–249, Dec 1988.
32. R.A. Laskowski. SURFNET: A program for a program for visualizing molecular surfaces, cavities, and intramolecular interactions. *Journal Molecular Graphics*, 13:321–330, 1995.
33. D.G. Levitt and L.J Banaszak. Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of Molecular Graphics*, 10(4):229–34, Dec 1992.
34. O. Lichtarge, H.R. Bourne, and F.E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257(2):342–358, 1996.
35. I. Mihalek, I. Res, and O. Lichtarge. A family of evolution-entropy hybrid methods for ranking of protein residues by importance. *J. Mol. Biol.*, 336(5):1265–82, 2004.
36. M. Nayal and B. Honig. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, 63(4):892–906, 2006.
37. C.T. Porter, G.J. Bartlett, and J.M. Thornton. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32:D129–D133, 2004.
38. V.M. Reyes, M.R. Sawaya, K.A. Brown, and J. Kraut. Isomorphous crystal structures of *Escherichia coli* dihydrofolate reductase complexed with folate, 5-deazafolate, and 5,10-dideazatetrahydrofolate: mechanistic implications. *Biochemistry*, 34:2710–2723, 1995.
39. R.B. Russell. Detection of protein three-dimensional side chain patterns. new examples of convergent evolution. *J. Mol. Biol.*, 279:1211–27, 1998.
40. M. Shatsky, R. Nussinov, and H.J. Wolfson. Flexprot: Alignment of flexible protein structures without a predefinition of hinge regions. *Journal of Computational Biology*, 11(1):83–106, 2004.
41. M. Shatsky, A. Shulman-Peleg, R. Nussinov, and H.J. Wolfson. Recognition of binding patterns common to a set of protein structures. *Proceedings of RECOMB 2005*, pages 440–55, 2005.
42. M. Shatsky, A. Shulman-Peleg, R. Nussinov, and H.J. Wolfson. The multiple common point set problem and its application to molecule binding pattern detection. *J. Comp. Biol.*, 13(2):407–28, 2006.
43. O.S. Smart, J.M. Goodfellow, and B.A. Wallace. The pore dimensions of gramacidin a. *Biophysics Journal*, 65:2455–2460, 1993.
44. A. Stark, S. Sunyaev, and R.B. Russell. A model for statistical significance of local similarities in structure. *J. Mol. Biol.*, 326:1307–1316, 2003.
45. G. Verbitsky, R. Nussinov, and H.J. Wolfson. Structural comparison allowing hinge bending. *Prot: Struct. Funct. Genet.*, 34(2):232–254, 1999.
46. M.A. Williams, J.M. Goodfellow, and J.M. Thornton. Buried waters and internal cavities in monomeric proteins. *Protein Science*, 3:1224–35, 1994.
47. H.J. Wolfson and I. Rigoutsos. Geometric hashing: An overview. *IEEE Comp. Sci. Eng.*, 4(4):10–21, Oct 1997.
48. Wang X. and Snoeyink J. Multiple structure alignment by optimal rmsd implies that the average structure is a consensus. *Proceedings of Computational Systems Bioinformatics 2006 (CSB2006)*, 2006.