Appeared in: Conf. Systems Bioinformatics (CSB), 2006

CAVITY-AWARE MOTIFS REDUCE FALSE POSITIVES IN PROTEIN FUNCTION PREDICTION

Brian Y. Chen[a⋆], Drew H. Bryant[b⋆], Viacheslav Y. Fofanov[c], David M. Kristensen[d], Amanda E. Cruess[a], Marek Kimmel[c], Olivier Lichtarge[d,e], Lydia E. Kavraki[a,b,e†]

[a] *Department of Computer Science,*
[b] *Department of Bioengineering,*
[c] *Department of Statistics,*
Rice University
Houston, TX 77005, USA
[⋆] *= Equal Contribution*

[d] *Program in Structural Computational Biology and Molecular Biophysics,*
[e] *Department of Molecular and Human Genetics,*
Baylor College of Medicine
Houston, TX 77030, USA
[†] *Corresponding Author: kavraki@rice.edu*

Determining the function of proteins is a problem with immense practical impact on the identification of inhibition targets and the causes of side effects. Unfortunately, experimental determination of protein function is expensive and time consuming. For this reason, algorithms for computational function prediction have been developed to focus and accelerate this effort. These algorithms are comparison techniques which identify *matches* of geometric and chemical similarity between *motifs*, representing known functional sites, and substructures of functionally uncharacterized proteins (*targets*). Matches of statistically significant geometric and chemical similarity can identify targets with active sites cognate to the matching motif. Unfortunately, statistically significant matches can include false positive matches to functionally unrelated proteins. We target this problem by presenting Cavity Aware Match Augmentation (CAMA), a technique which uses *C–spheres* to represent active clefts which must remain vacant for ligand binding. CAMA rejects matches to targets without similar binding volumes. On 18 sample motifs, we observed that introducing C–spheres eliminated 80% of false positive matches and maintained 87% of true positive matches found with identical motifs lacking C–spheres. Analyzing a range of C–sphere positions and sizes, we observed that some *high-impact* C–spheres eliminate more false positive matches than others. High-impact C–spheres can be detected with a geometric analysis we call Cavity Scaling, permitting us to refine our initial cavity-aware motifs to contain only high-impact C–spheres. In the absence of expert knowledge, Cavity Scaling can guide the design of cavity-aware motifs to eliminate many false positive matches.

1. INTRODUCTION

Exhaustive knowledge of the biological *function* of a large number of proteins would have a broad impact on the identification of drug targets and the reduction of potential side effects. Unfortunately, the experimental determination of protein function is an expensive and time consuming process. In an effort to guide and accelerate the experimental process, computational techniques have been developed to predict protein function by identifying distinct similarities to known proteins. Algorithms like Geometric Hashing[34], JESS[14], pvSOAR[33] and Match Augmentation (MA)[5], search functionally uncharacterized protein structures (*targets*), for substructures with geometric and chemical similarity (*matches*), to known active sites (*motifs*). Finding a match with statistically significant geometric and chemical similarity can imply that the target has an active site similar to the motif, suggesting functional homology[1, 14, 33, 5].

One fundamental subproblem of protein function prediction is the design of *effective motifs*. Ideally, effective motifs have geometric and chemical characteristics which have matches to functionally homologous targets (*sensitive* motifs), and do not have matches to functionally unrelated targets (*specific* motifs). In practice, however, many matches are identified within functionally unrelated targets. For this reason, statistical models[1, 14, 33, 5] can establish a threshold of similarity necessary to imply functional homology. Predictions from any non-trivial statistical model will inevitably contain some false positive matches which identify statistically significant geometric similarity to functionally unrelated proteins. In the context of actual function predictions, where expensive resources could be deployed to verify computational predictions, false positive matches must be avoided to minimize wasted resources, while preserving as many true positive matches to functional homologs as possible. This paper proposes a method that reduces false positive matches while preserving most true positive matches, by adding biological information that rejects matches to functionally unrelated targets.

It is hypothesized that ligand binding proteins often contain active clefts or cavities which create chemical microenvironments essential for biological function. In several instances, large surface concavities have been associated with protein function[30, 13]. Inspired by seminal work in the modeling and search for protein cavities[30, 8, 33], we seek to use cavities to eliminate false positive matches. If the matching atoms of the target truly form a cognate active site with similar function, the matching atoms of the target should surround an empty cavity with similar shape.

This paper presents Cavity-Aware Match Augmentation (CAMA), an adaptation of Match Augmentation[5], which searches for motifs built from *motif points*, while requiring specific geometric volumes, represented with sets of *C–spheres*, to remain empty. On 18 *cavity-aware* motifs derived from ligand binding proteins, we compared the number of false positive matches found relative to identical motifs without C–spheres. Cavity-aware motifs eliminated a large proportion of false positive matches that were identified with point-based motifs, while preserving most true positive matches. We also compared the relative effect of many C–sphere positions and sizes to the number of false positive matches eliminated. This led us to observe trends indicating that certain *high-impact* C–spheres contribute more to the elimination of false positive matches than others. We exploited these trends to produce *Cavity Scaling*, a technique for identifying high-impact C–spheres *a priori*. Cavity Scaling allowed us to refine our existing motifs to contain only high-impact C-spheres, guiding the design of cavity-aware motifs that eliminate many false positive matches.

2. RELATED WORK

Motif Types The search for effective motifs has led to many different geometric representations of protein active sites, including *point-based* motifs and *cavity-based* motifs. Point-based motifs represent active sites as sets of *motif points* in three dimensions, labeled with varying chemical and biological definitions. Depending on how motif points are defined, they have different labels associated with them and these labels need to be taken into account with varying comparison algorithms. Motif points have been used to represent evolutionarily significant amino acids[5], "pseudo-centers" representing protein-ligand interactions[17, ?], atoms in catalytic sites[2, 14], points on the Connolly surface[21] with labels representing electrostatic potentials[15], and even atoms in flexible motifs[18].

Clefts and cavities, on the surface or within protein structures, have many different volumetric representations. These cavity-based representations include spheres[12, 6, 26, 20], alpha-shapes[9, 8, 33, 32], and grid-based techniques[28].

Geometric Comparison Algorithms Many algorithms exist for identifying matches between motifs and targets. These methods differ fundamentally in that they are optimized for comparing different types of motifs. There are algorithms for comparing graph-based motifs[27], algorithms for finding catalytic sites[14], and the seminal Geometric Hashing framework[10] which can search for many types of motifs, including motifs based on atom position[22], points on Connolly face centers[16], catalytic triads[2], and flexible protein models[18]. The comparison algorithm we use in this work is based on Match Augmentation[5], because of its availability and compatibility with our selected motif type.

Statistical Models of Geometric Similarity Finding a match with MA indicates only that substructural geometric and chemical similarity exists between the motif and a substructure of the target, not that the motif and the target have functionally similar active sites. We measure geometric similarity with *LRMSD*, root mean square distance (RMSD) between matching points in 3D, aligned with least RMSD. In order to use matches to imply functional similarity, it is essential to understand the degree of similarity, in LRMSD, sufficient to imply functional similarity. However, a simple LRMSD threshold is insufficient to indicate functional similarity between any motif and a matching target. Some motifs match functional homologs at lower values of LRMSD than other motif-target pairs, and LRMSD itself is affected by the number of matching points[5]. Fortunately, earlier work has demonstrated that motif-specific LRMSD thresholds can be produced with statistical models of functional similarity[5]. Many important statistical models have been designed, including parametric[1, 14], empirical[33], and nonparametric[5] statistical models.

Geometric comparison algorithms operate on

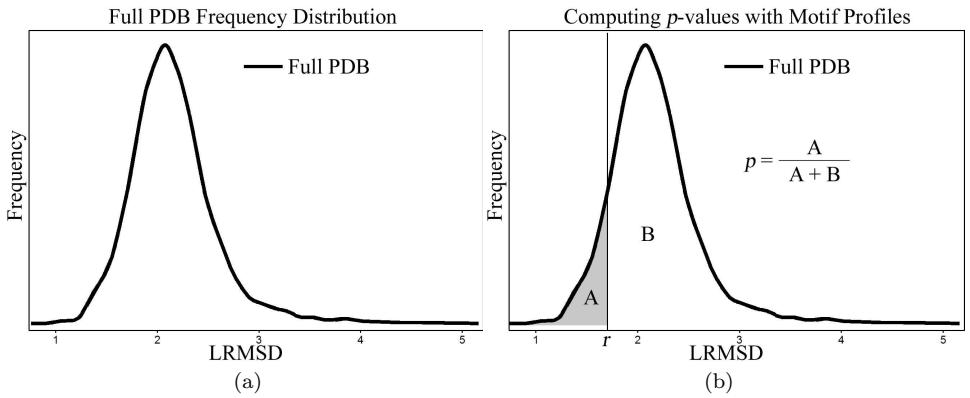

Fig. 1. A frequency distribution of matches between a motif and all functionally unrelated proteins in the PDB (a). Comparing the area under the curve to the left of some LRMSD r, relative to the entire area under the curve (b).

the assumption that substructural and chemical similarity implies functional similarity. Our statistical model can be used to identify the degree of similarity sufficient to follow this implication. Given a match m with LRMSD r between motif S and target T, exactly one of two hypotheses must hold:

H_0**:** S and T are structurally dissimilar
H_A**:** S and T are structurally similar

Our statistical model tests these hypotheses by computing a *motif profile*. Motif profiles are frequency distributions (see Figure 1a) of match LRMSDs between S and the entire Protein Data Bank (PDB)[11], which is essentially a large set of functionally unrelated proteins. A motif profile is basically a histogram (see example plotted in Figure 1a), where the vertical axis indicates the number of matches at each specific LRMSD, indicated by the horizontal axis. Motif profiles provide very complete information about matches typical of H_0. If we suspect that a match m has LRMSD r indicative of functional similarity, we can use the motif profile to determine the probability p of observing another match m' with smaller LRMSD by computing the volume under the curve to the left of r, relative to the entire volume (see Figure 1b). The probability p, referred to as the p-value, is the measure of statistical significance. With a standard of statistical significance α, if $p < \alpha$, then we say that the probability of observing a match m' with LRMSD $r' < r$ is so low that we reject the null hypothesis (H_0) in favor of the alternative hypothesis (H_A). We call m statistically significant.

In the context of controlled experiments, where we know when matches identify functional homologs and when they do not, there are four possibilities: True positives (TP), False positives (FP), True negatives (TN), and False negatives (FN). A match is a TP, if it identifies a functional homolog, and if the match is statistically significant. A match is a FP, if the match identifies a functionally unrelated protein, and is statistically significant. A match is a TN if it is not statistically significant and matches a functionally unrelated protein. A match is a FN if it identifies a functional homolog, but is not statistically significant.

In practice, our statistical model occasionally identifies false positive matches. Designing motifs which generate fewer FP matches is an essential aspect of motif design, especially when we consider the possibility that expensive experimental resources could be wasted in an attempt to verify predicted functions. In the next section, we discuss a method for designing motifs which strongly reduces false positives.

3. METHODS

Cavity-Aware Motifs The cavity-aware motifs used in this work are an integration of a point-based motif and a cavity-based motif. Cavity-aware motifs contain motif points taken from atom coordinates labeled with evolutionary data[23, 24, 5, 7]. A motif S contains a set of $|S|$ motif points $\{s_1, \ldots, s_{|S|}\}$ in three dimensions, whose coordinates are taken from backbone and side-chain atoms. Each *motif point* s_i in the motif has an associated *rank*, a measure of the functional significance of the motif point.

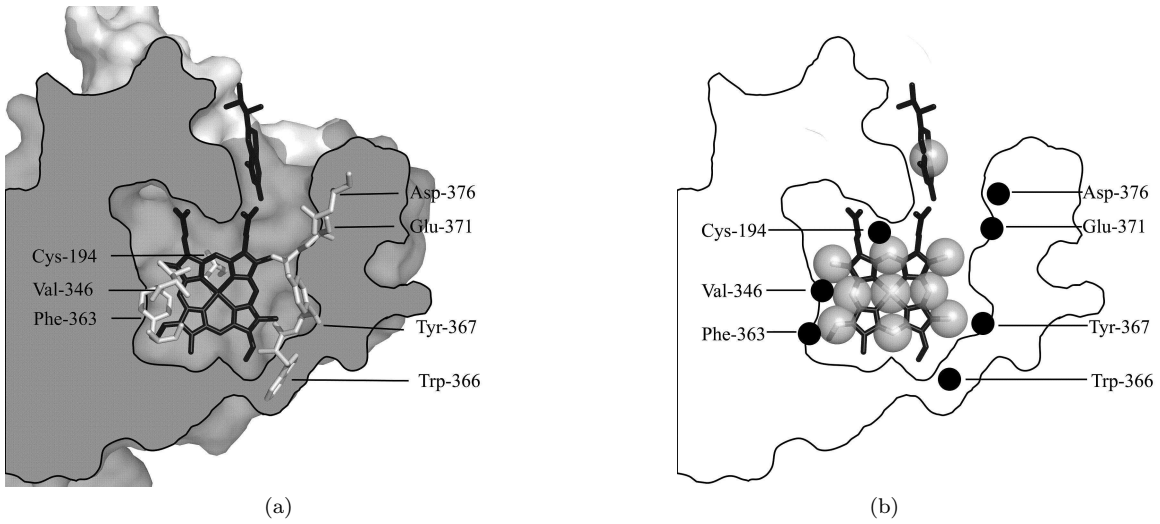

Fig. 2. A diagram of a cavity-aware motif. Beginning with functionally relevant amino acids and bound ligand coordinates (a), cavity-aware motif points are positioned at alpha carbon coordinates (black dots, (b)), and C–spheres are positioned at ligand atom coordinates (transparent spheres, (b)).

Each s_i also has a set of alternate amino acid *labels* $l(s_i) \subset \{GLY, ALA, ...\}$, which represent residues to which this amino acid has mutated during evolution. Labels permit our motifs to simultaneously represent many homologous active sites with slight mutations, not just a single active site. In this paper, we obtain labels and ranks using the Evolutionary Trace[23, 24].

Cavity-aware motifs also contain a set of C–spheres $C = \{c_1, c_2, \ldots c_k\}$ with radii $r(c_1), r(c_2), \ldots r(c_k)$, which are rigidly associated with the motif points. $\forall c_i, 1 < i < k$, a maximum radius, $r_{max}(c_i)$, is defined to be the largest radius (rounded to the nearest integer) such that c_i contains no atoms from the protein which gave rise to the motif. C–spheres are a loose approximation of solvent exposed volumes essential for ligand binding. C–spheres can have arbitrary radii, and can be centered at arbitrary positions. While this work targets the functional prediction of active sites that bind small ligands, this representation could be used to represent protein-protein interfaces and other generalized interaction zones.

C–sphere positions in this work were selected based on the coordinates of atoms in bound ligands. For example, in Figure 2, we modeled the heme-dependent enzyme nitric oxide synthase, which catalyzes the synthesis of nitric oxide (NO) from an L-arginine substrate. This multi-step reaction takes place in a deep cleft and involves zinc, tetrahydrobiopterin, and hydride-donating (NADPH or H_2O_2) cofactors[4, 31]. Using PDB structure 1dww, we centered C–spheres at several atom coordinates on the heme, in order to fill the heme-binding cavity, and placed one C–sphere to represent tetrahydrobiopterin, which was further outside from the main cavity, as shown in Figure 2.

In our experimentation, a small number (usually 10) of C–spheres were manually placed for each motif. In some cases, not all atoms of the ligand were used, such as in heme in Figure 2, but selections were made to approximate the shape of the ligand binding cavity based on the atom coordinates available. C–spheres could have been made to fit better by moving the C–sphere centers, but we used atom coordinates to standardize our experimentation. Future work will explore the generalized positioning of C–spheres.

Matching Criteria Cavity Aware Match Augmentation compares a cavity-aware motif S to a target T, a protein structure encoded as $|T|$ *target points* referred to as $T = \{t_1, \ldots t_{|T|}\}$, where each t_i is taken from atom coordinates, and labeled $l(t_i)$ for the amino acid t_i belongs to. A match m is a bijection correlating all motif points in S to a subset of T of the form $m = \{(s_{a_1}, t_{b_1}), (s_{a_2}, t_{b_2}) \ldots (s_{a_{|S|}}, t_{b_{|S|}})\}$. Referring to Euclidean distance between points a and b as $||a - b||$, an acceptable match requires:

Criterion 1 $\forall i$, s_{a_i} and t_{b_i} are label compatible: $l(t_{b_i}) \in l(s_{a_i})$.

Criterion 2 $\forall i, ||A(s_{a_i}) - t_{b_i}|| < \epsilon$, our

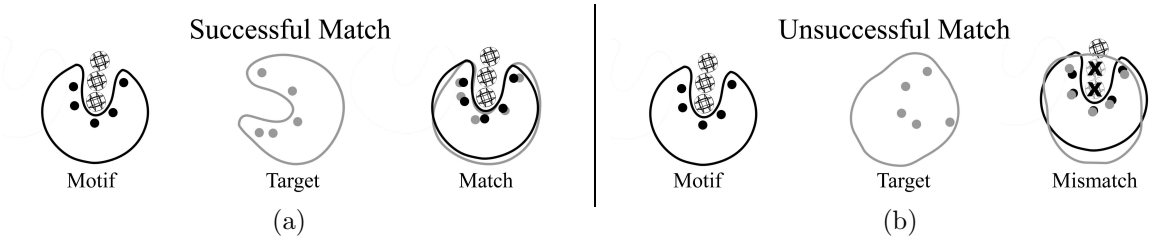

Fig. 3. Two cases of cavity-aware matching. Every time a match is generated by CAMA, an alignment of the motif points is generated to the matching points of the target. This specifies the precise positions of the C–spheres in the motif relative to the target. CAMA accepts matches to targets where no C–spheres contain any target atoms (a), and rejects matches where any target atom is within one or more C–spheres (b).

threshold for geometric similarity.

Criterion 3 $\forall t_i \forall c_j \; ||t_i - A(c_j)|| > r(c_j)$

where motif S is in LRMSD alignment with a subset of target T, via rigid transformation A. Criterion 1 assures that we have motif and target amino acids that are identical or vary with respect to important evolutionary divergences. Criterion 2 assures that when in LRMSD alignment, all motif points are within ϵ of correlated target points. Finally Criterion 3 assures that no target point falls within a C–sphere, when the motif is in LRMSD alignment with the matching target points. CAMA outputs the match with smallest LRMSD among all matches that fulfill these criteria. Partial matches correlating subsets of S to T are rejected.

Matching Algorithm CAMA is a three staged hierarchical matching algorithm which identifies correlations for motif points in order of rank. The first stage, *Seed Matching* is a hashing technique which exploits pairwise distances between motif points to rapidly identify correlations between the three highest ranking motif points, and triplets of target points. These triplets are passed to the second stage, *Augmentation*, which expands seed matches to full correlations of all motif points. The final stage, *Cavity Filtering*, identifies the aligned position of the C–spheres in each full correlation, and checks to see if any target points fall within a C–sphere. The correlation with the smallest LRMSD that has no target points within any C–sphere is returned as the resulting match. Seed Matching and Augmentation are documented in earlier work[5], but we summarize them below for completeness.

Seed Matching Seed Matching identifies all sets of 3 target points $T' = \{t_A, t_B, t_C\}$ which fulfill our matching criteria with the highest ranked 3 motif points, $S' = \{s_1, s_2, s_3\}$. In this stage, we represent the target as a geometric graph with colored edges. There are exactly three unordered pairs of points in S', and we name them red, blue and green. In the target, if any pair of target points t_i,t_j fulfills our first two criteria with either red, blue or green, we draw a corresponding red blue or green edge between t_i,t_j in the target. Once we have processed all pairs of target points, we find all three-colored triangles in T. These are the Seed Matches, a set of three-point correlations to S' that we sort by LRMSD and pass to Augmentation.

Augmentation Augmentation is an application of depth first search that begins with the list of seed matches. Assuming that there are more than four motif points, we must find correspondences for the unmatched motif points within the target. Interpret the list of seed matches as a stack of partially complete matches. Pop off the first match, and considering the LRMSD alignment of this match, plot the position P of the next unmatched motif point s_i relative to the aligned orientation of the motif. In the spherical region V around P, identify all target points t_i, compatible with s_i, inside V. Now compute the LRMSD alignment of all correlated points, include the new correlation (s_i, t_i). If the new alignment satisfies our first two criteria and there are no more unmatched motif points, put this match into a heap which maintains the match with smallest LRMSD. If there are more unmatched motif points, put this partial match back onto the stack. Continue to test correlations in this manner, until V contains no more target points that satisfy our criteria. Then, return to the stack, and begin again by popping off the first match on the stack, repeating this process until the stack is empty.

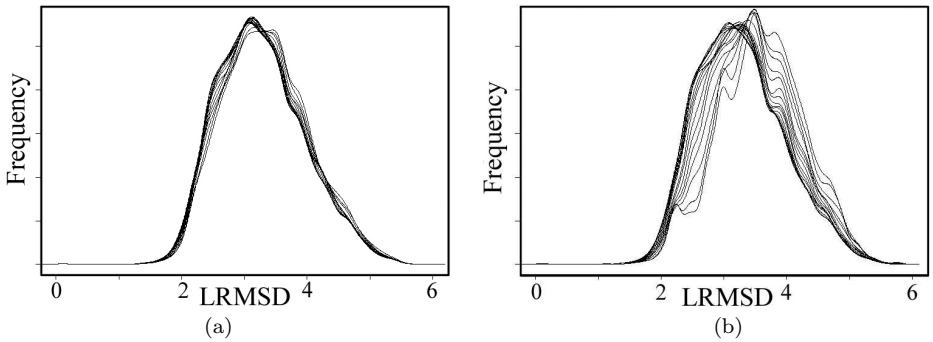

Fig. 4. 20 motif profiles for a low-impact C–sphere (a), and a high-impact C–sphere (b), as radius increases.

Cavity Filtering Augmentation results in a heap of completed matches. Beginning from the match with lowest LRMSD, for each match, retrieve the alignment of the motif onto the target. Using this alignment, we plot the positions of the C–spheres in rigid alignment with the motif. Then, for each C–sphere, we check if a target point exists within the C–sphere. If any target point is found within any C–sphere, the match is discarded, and we continue to the match with next-lowest LRMSD. This is diagrammed in Figure 3b. If we identify a match with no target points in any C–spheres, as in Figure 3a, we return this match as the output.

Discussion Standard MA would accept the match with lowest LRMSD, regardless of the C–spheres. Cavity Filtering rejects matches in order of ascending LRMSD, starting with the match with lowest LRMSD, causing CAMA to potentially increase the LRMSD of matches found, in comparison to MA. When C–sphere radii are all zero, CAMA and MA are therefore identical.

Cavity-Aware Statistical Significance We evaluate p-values for matches to a cavity-aware motif S in the same manner as for point-based motifs. We first generate a point-based version S' of S, and use S' to compute a motif profile. Then, given a match m of S with LRMSD r, we compute the p-value of r relative to this motif profile. p-values for cavity-aware motifs are computed relative to point-based motif profiles because the purpose of a cavity-aware motif is to eliminate matches which would have been statistically significant relative to the point-based motif. Since matches with cavity-aware motifs have equal or greater LRMSDs than matches with identical point-based motifs, matches found with cavity-aware motifs have equal or higher p-values.

Cavity-aware motifs are not perfect; due to variations in active site structure, some functional homologs have atoms which occupy C–spheres. In our experimentation, we measured both the number of FP matches eliminated, as well as the number of TP matches lost by adding C–spheres, and demonstrate that the number of TP matches lost is small in comparison to the number of FP matches eliminated.

High-Impact C–spheres In our experimentation, we observed that some *high-impact* C–spheres eliminated more FP matches than other C–spheres. Identifying high-impact C–spheres is essential, because a cavity-aware motif without high-impact C–spheres would not eliminate many more FP matches than an identical point-based motif. More importantly, a computational technique for identifying high-impact C–spheres could simplify the design of cavity-aware motifs by ensuring that only high-impact C–spheres are used.

We have observed that motif profiles derived from cavity-aware motifs that include high-impact C–spheres have a tendency of shifting towards higher LRMSDs as C–sphere radius increases. In Figure 4a, we demonstrate motif profiles computed with a motif that has exactly one C–sphere. Each motif profile corresponds to identical motif points with a C–sphere at an identical position, where the only difference is that radius changes evenly between zero and the C–sphere's maximum size. As size increases, the motif profile changes very little. This is a low-impact C–sphere. In comparison, in Figure 4b, for the same motif points and a C–sphere in a different position, as radius changes uniformly between zero and the

Motifs Used in Experimentation

PDB id	Amino Acids Used	Ligands Used	#C	Range
16pk*	R39,P45,G376,G399,K202	$C_{15}H_{22}N_5O_{12}F_4P_3$	10	4-6
1ady*	E81,T83,R112,E130,Y264,R311	$C_{16}H_{21}N_8O_8P$	10	4-6
1ani*	D51,D101,S102,R166,H331,H412	Zn^{2+}, O_4P^{3-}	10	2-6
1ayl	L249,S250,G251,G253,K254,T255	ATP, $C_2O_4^{2-}$	10	4-8
1b7y*	W149,H178,S180,E206,Q218,F258,F260	$C_{19}H_{25}N_6O_7P$, Mg^{2+}	10	4-8
1czf	D180,D201,D202,A205,G228,S229,R256,K258,Y291	$C_8H_{15}NO_6$, Zn^{2+}	10	2-8
1did*	F25,H53,D56,F93,W136,K182,	Mn^{2+}, $C_6H_{13}NO_4$	10	2-6
1dww*	C194, V346, F363, W366, Y367, E371, D376,	Heme, NHA	10	4-10
1ggm*	E188,R311,E239,E341,E359,S361	$C_{12}H_{17}N_6O_8P$	10	4-10
1ja7	S36,C76,W108,Q57,I58,W63,	$C_8H_{15}NO_6$	10	4-8
1jg1	E97,G99,G101,D160,L179,G183,	$C_{14}H_{20}N_6O_5S$	10	6-8
1kp3	R106,F139,E202,L286,R288,Y331	ATP	10	6-8
1kpg	D17,G72,G74,W75,G76,F200	$C_5H_{11}NO_2Se$	10	6-6
1lbf	E51,S56,P57,F89,G91,F112,E159,N180,S211,G233	$C_{12}H_{18}NO_9P$	10	4-6
1ucn	K12,P13,G92,R105,N115,H118	O_4P^{3-}, Ca^{2+}, ADP	8	4-8
2ahj	P53,L120,Y127,V190,D193,I196	Fe^{3+}, NO, $C_4H_8O_2$, Zn_{2+}	10	4-10
7mht	P80,C81,S85,E119,R163,R165	$C_{14}H_{20}N_6O_5S$	10	4-8
8tln*	M120,E143,L144,Y157,H231	C_2H_6OS, Ca^{2+}, Zn_{2+}	9	2-8

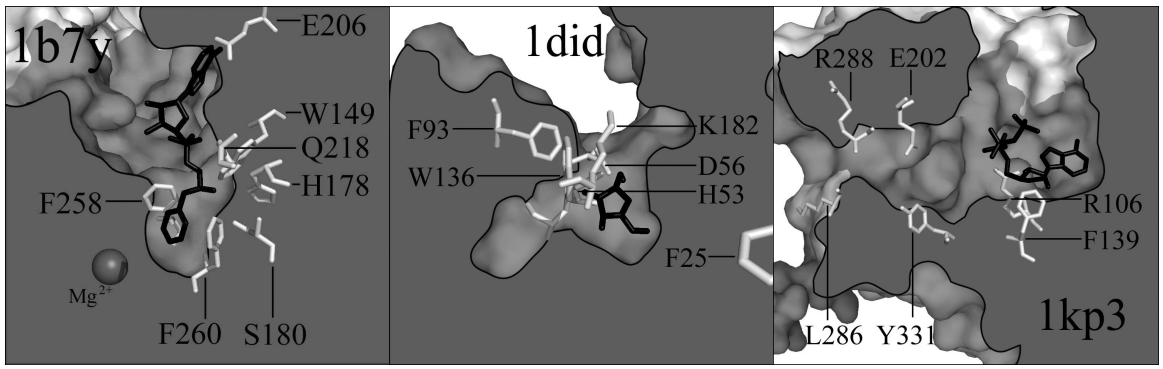

Fig. 5. Motifs used, with example diagrams below. Starred (*) motifs use functionally documented amino acids. The column marked "#C" denotes the number of C–spheres in each motif. "Range" denotes the range of C–sphere maximum diameters (in Å) for the motif. Experimental details can be found at: **http://www.cs.rice.edu/∼brianyc/papers/CSB2006**

C–sphere's maximum size, many more matches shift towards higher LRMSDs, as mentioned in Section 3. This is a high-impact C–sphere.

We have designed a technique which uses this effect to identify high-impact C–spheres, called Cavity Scaling. Cavity Scaling takes as input a single C–sphere, and a set of motif points. Using this cavity-aware motif, we generate a spectrum of cavity-aware test motifs which differ only in the radius of the single C–sphere. The C–sphere radius in each test motif ranges from zero to the maximum size of the input C–sphere. We then compute a motif profile for each test motif, and compare the motif profile medians. If the motif profile medians change significantly as C–sphere radius increases, then we consider the input C–sphere a high-impact C–sphere. The process of Cavity Scaling is then repeated for each C–sphere that has been defined, individually.

Cavity Scaling permitted us to refine C–sphere selections in cavity-aware motifs. As we will show later, refined cavity-aware motifs eliminate most FP matches and maintain TP matches in comparison to manually defined cavity-aware motifs. More importantly, even though this work tests C–spheres centered on ligand atom coordinates, Cavity Scaling is independent of C–sphere centers, making it a general test for high-impact C–spheres. In the future, this could be applied at a larger scale to explore more general representations of cavity-aware motifs, and provide feedback about C–sphere placements in motif design.

4. EXPERIMENTAL RESULTS

Motifs The motifs used in this work begin as 18 point-based motifs designed to represent a range of unrelated active sites in unmutated protein structures with biologically occurring bound ligands. These are documented in Figure 5. Earlier work has produced examples of motifs designed with evolutionarily significant amino acids[5, 7] and amino acids with documented function[29], so these princi-

ples were followed in the design of our point-based motifs. Amino acids for use in 10 of the motifs were selected by evolutionary significance, and are taken directly from earlier work[7], and the remaining 8 motifs were identified by functionally active amino acids documented in the literature (marked $\star$ in Figure 5).

For example, in the case of nitric oxide synthase, we selected active site residues which bind NHA and heme. Cys-194 is axially coordinated to heme. Glu-371 and Trp-366 form hydrogen bonds with the guanidinium group of NHA while Tyr-367 and a protonated Asp-376 form hydrogen bonds to the carboxylate group of NHA[3]. We also selected Val-346 and Phe-363, which create a small hydrophobic cavity within the larger heme-binding cavity, allowing dioxide (O_2) to bind end-on to heme without steric interference[4]. C-sphere positions and sizes were defined in Section 3.

The selection of motif points strongly influences motif sensitivity and specificity. However, in this work, we seek to demonstrate that adding C–spheres can improve point-based motifs. For this reason, we take the selection of motif points and the number of TP and FP matches found, for each point-based motif, as given.

Functional Homologs In order to count TP and FN matches, it is essential to fix a benchmark set of functional homologs. We use the functional classification of the Enzyme Commission[25] (EC), which identifies distinct families of functional homologs for each motif used. Proteins with PDB structures in these families form the set of functional homologs we search for. Structure fragments and mutants were removed to ensure accuracy.

Unrelated Proteins In order to measure FP and TN matches, it is essential to fix the set of functionally unrelated protein structures. The set we use is, initially, a snapshot of the PDB from Sept 1, 2005. For each motif, the set of functional homologs is removed, producing a homolog-free variation of the PDB specific for each motif. Furthermore, the PDB was processed to reduce sequential and structure redundancy. In structures with multiple chains describing the same protein, only one copy of each redundant chain was used, and all mutants and protein fragments were removed. This produced 13599 protein structures. The set of structures used was not strictly filtered for sequential nonredundancy because eliminating one member of any pair with too much sequence identity involves making arbitrary choices. Eliminating fragments and mutated structures, which seem to be the largest source of sequential redundancy, was the most reproducible and well defined policy.

Implementation Specifics CAMA was implemented in C/C++. Large scale comparison of many potential C–sphere radii was accomplished with a distributed version of CAMA, which used the Message Passing Interface[19] (MPI) protocol for interprocess communication. Code was prototyped on a 16-node Athlon 1900MP cluster and the Rice TeraCluster, a cluster of 272 800Mhz Intel Itanium2 processors. Final production runs ran on Ada, a 28 chassis Cray XD1 with 672 2.2Ghz AMD Opteron cores.

4.1. C-Spheres Eliminate False Positives, Preserve True Positives

We first demonstrate that C–spheres affect the elimination of FP matches and the retention of TP matches. We compared the number of TP and FP matches found with 18 point-based motifs to cavity-aware versions of the same motifs. For completeness, we show how 20 increments of varying C–sphere radii affect the number of TP and FP matches found.

Our data begins as 18 motifs $\{S_1, S_2, ...S_{18}\}$. For each motif S_i, we generated 20 C–sphere size variations called $\{S_{i_0}, S_{i_1}, \ldots, S_{i_{19}}\}$. If S_i has C–spheres $\{c_1, c_2, \ldots c_k\}$, with individual maximum sizes $r_{max}(c_1), r_{max}(c_2), \ldots r_{max}(c_k)$, then the variation $S_{i_j} \in \{S_{i_0}, S_{i_1}, \ldots, S_{i_{19}}\}$ S_{i_j} has C–spheres of radii $(\frac{j}{19} r_{max}(c_1)), (\frac{j}{19} r_{max}(c_2)), \ldots (\frac{j}{19} r_{max}(c_k))$. For example, $S_{i_{19}}$ has C–spheres of radii $r_{max}(c_1), r_{max}(c_2), \ldots r_{max}(c_i)$, and S_{i_0} would have only C–spheres of radii 0, making S_{i_0} equivalent to a point-based motif.

Since matches to $S_{i_1}, S_{i_2}, \ldots, S_{i_{19}}$ have p-values greater than or equal to S_{i_0}, because they have C–spheres with non-zero radii, the number of FP and TP matches identified among $S_{i_1}, S_{i_2}, \ldots, S_{i_{19}}$ is less than or equal to that of S_{i_0}. The number of homologs matched by each point-based motif, S_{i_0}, is listed in the left of Figure 6. The number of TP and FP matches eliminated is calculated relative to the number matched by the point-based motif, and thus all S_{i_0} have 100% of TP and FP matches, as in the leftmost point of the graph in Figure 6. Second

Matches to Homologs by Point-based Motifs

Motif	# Homologs	# TPs	# FPs
16pk	20	14	216
1ady	22	20	200
1ani	75	75	205
1ayl	8	8	170
1b7y	9	0	170
1czf	14	14	117
1did	149	149	80
1dww	192	181	76
1ggm	7	5	195
1ja7	1008	448	57
1jg1	13	13	196
1kp3	35	35	162
1kpg	13	11	151
1lbf	11	11	50
1ucn	153	133	162
2ahj	23	6	186
7mht	10	9	160
8tln	59	56	187

Average Percentage of TP and FP Matches

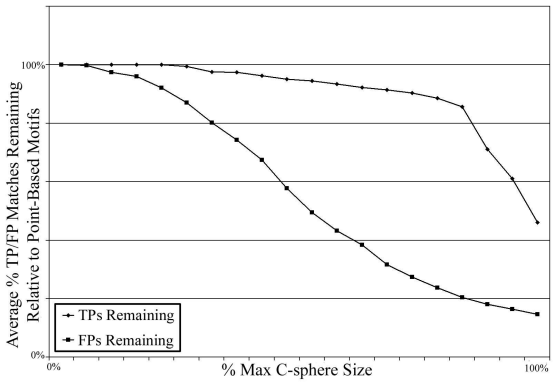

Fig. 6. Average effect of cavity-aware motifs on TP and FP matches, over all motifs. The horizontal axis charts C–sphere radius, where the radius of all C–spheres scales simultaneously from zero to individual maximum size (see Section 4.1). The vertical axis charts the average percentage, per motif, of TP and FP matches remaining, relative to their respective point-based motifs. The number of TP and FP matches for each point-based motif is shown at left. FP matches are dramatically reduced while most TP matches are preserved. Before TP matches begin to fall off, cavity-aware motifs eliminate 80% of FP matches while maintaining 87% of TP matches.

from the left, we plot the percentage of TP and FP matches retained among S_{i_1}, relative to S_{i_0}, for all i, and then average these percentages over all S_{i_1}. Continuing from left to right, we compute the average percentage of TP and FP matches, over all S_{i_2}, then all S_{i_3}, etc., again relative to S_{i_0}.

Observations Demonstrated in Figure 6, as C–sphere radius increases, the number of FP matches are reduced dramatically, while the number of TP matches falls slightly. Also, large percentages of TP matches were maintained as C–sphere radius increased, with few losses, until approximately 80% of maximum size, when the number of true positives began to fall off, for most motifs. This was expected since maximum size was computed only on the primary motif structure, and not on homologs.

One motif, Phenylalanyl-TRNA Synthetase (1b7y), exhibited 0 sensitivity. The point-based version of 1b7y matched no functional homologs, so no cavity-aware motifs based on 1b7y matched any functional homologs either. For this reason, the percentage of TP matches eliminated by cavity-aware variations of 1b7y is undefined, and therefore no TP and FP data (for consistency) is included in the averages plotted in Figure 6. Cavity-aware variations on 1b7y still rejected more FPs as C–sphere radius increased. Point-based motifs from 1ja7 and 2ahj exhibited low sensitivity, identifying less than 20% of the total number of true positives. Having a very flexible active site, cavity-aware variations of 16pk were significantly less sensitive than its point-based counterparts. Overall, cavity-aware motifs eliminate many FP matches, while preserving most TP matches.

4.2. Analysis of Individual C–spheres

Some C–spheres may have a greater impact on FP match elimination than other C–spheres. We performed Cavity Scaling on each C–sphere in each of our 18 motifs, identifying which C–spheres were high–impact. 1ayl, used in Figure 7 is an excellent example, having several high- and low-impact C–spheres. All motifs had related behavior: Some motifs had many high-impact C–spheres, and others (1czf, 16pk, 8tln) had none, but significant increases in motif profile medians remained correlated to the elimination of FP matches in all examples.

Observations Motif profiles of some single-C-sphere motifs, computed over increasing radii, shift significantly in the median towards higher LRMSDs. These single-C-sphere motifs eliminate more FP matches as radii increase. Alternatively, motif profile medians of other single-C-sphere motifs that do not eliminate many FP matches also do not shift towards higher LRMSDs as radii increase. This is apparent in Figure 7, where we detail this effect for single C-sphere motifs based on 1ayl. In the inset graphs, identical copies of the 1ayl motif that contain only C-spheres 4 or 6 undergo significant changes in motif profile medians, towards higher LRMSDs, as ra-

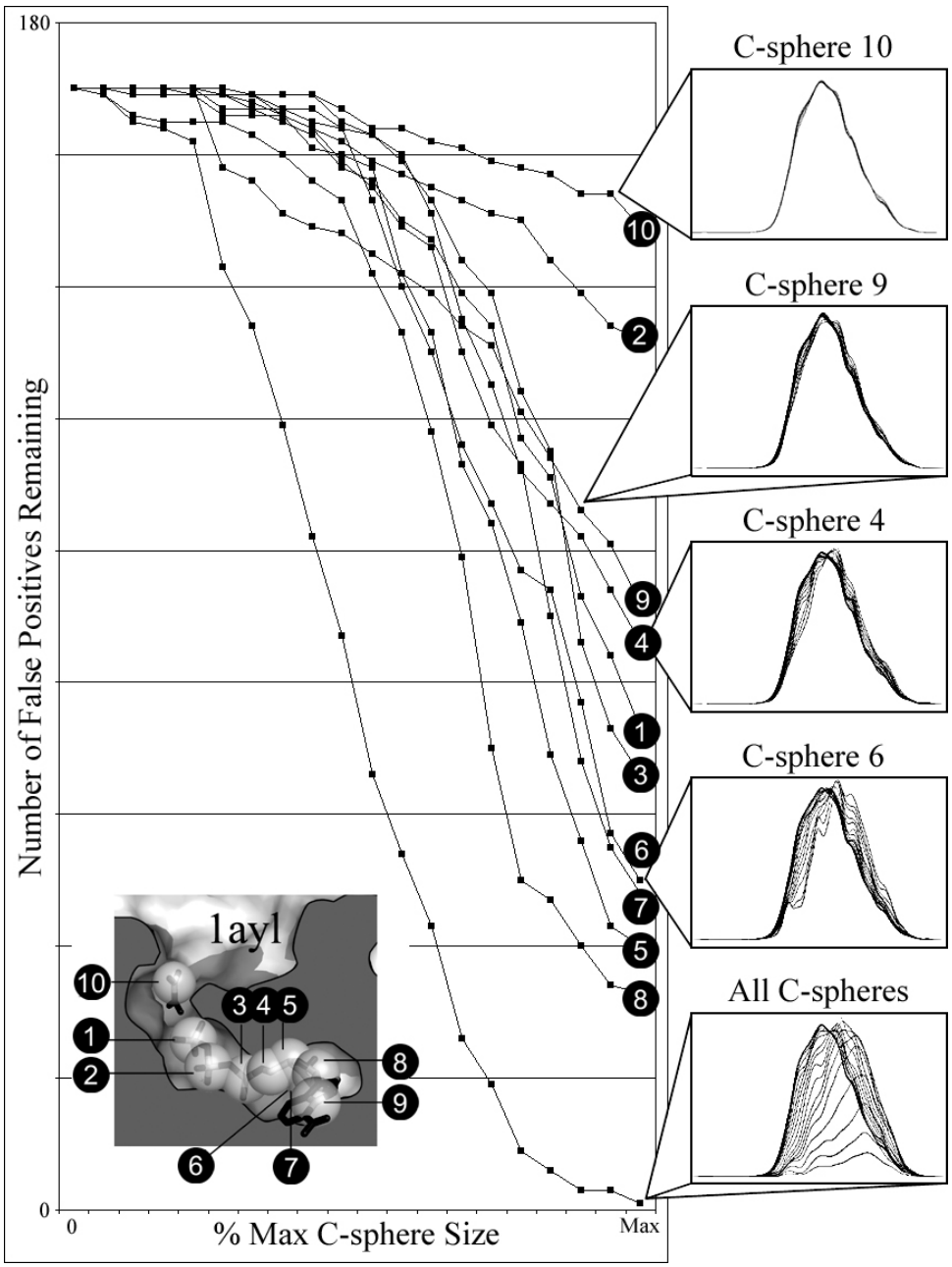

Fig. 7. Effect of Individual C–spheres on Motif Specificity. As C–sphere size uniformly increases, as described in Section 4.1 (horizontal axis), some high-impact C–spheres, such as 4 and 6, eliminate more FP matches (vertical axis) than others, such as 10 and 9. Line plots show the number of remaining FP matches for a specific single-C-sphere motif, and for a motif containing all C-spheres. C-sphere positions relative to cavity shape are illustrated in the inset graphic. High-impact C–spheres, such as C–sphere 6, generate motif profiles whose medians shift towards higher LRMSDs as C–sphere radius increases. Other C–spheres, which do not eliminate as many FP matches, such as C–sphere 10, do not affect motif profiles as much. Cavity Scaling identifies C–spheres which eliminate more FP matches.

dius increases. Simultaneously, as seen in the main graph, these single-C-sphere motifs, containing only C-sphere 4 or 6, rapidly eliminate FP matches. 1ayl motif copies with only C-spheres 9 or 10 experience insignificant changes in motif profile medians, and also eliminate FP matches more slowly, as radius increases. C-sphere positions relative to active site geometry are provided in the inset graphic in Figure 7. No correlation between high-impact C–spheres and cavity topography was apparent, emphasizing

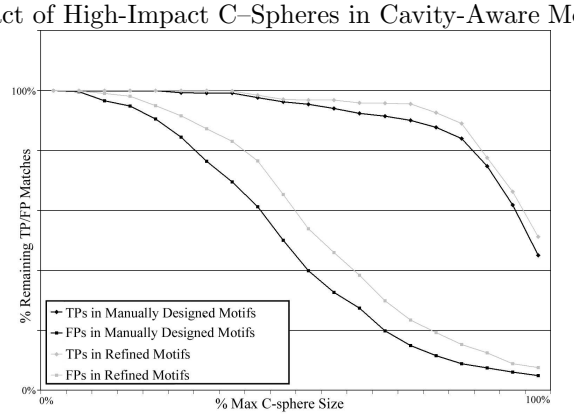

Fig. 8. TP/FP matches preserved when using automatically refined cavity-aware motifs. Axes here are identical those of Figure 6. Automatically refined motifs (gray) reject a large majority of FP matches, retaining slightly more than manually designed (black) motifs. Automatically refined motifs also preserve slightly more TP matches than manually designed motifs.

the difficulty of designing motifs with high-impact cavities.

Motifs with only one C–sphere eliminate very few TP matches, but careful inspection indicates that individual cavities cause different TP matches to be rejected. This effect accumulates into the slow loss of TP matches observed in section 4.1.

4.3. Automatically Refined Cavity-aware Motifs

In an experimental function prediction setting, rules and automated techniques for defining sensitive and specific motifs are important for high throughput function predictions. Having shown in the previous section that Cavity Scaling can identify high-impact C-spheres, we use Cavity Scaling to generate motifs containing only high-impact C–spheres, and demonstrate that they are reasonably effective.

Experiment: We applied Cavity Scaling on every C–sphere in every motif, which identified a set of high-impact C–spheres for all motifs except 1czf, 16pk and 8tln. We repeated the experiment described in Section 4.1 for the remaining motifs, using only high-impact C–spheres. We refer to these as automatically refined motifs. We compared our results to manually designed motifs used in Section 4.1, which contained all C–spheres.

Observations: Like the axes of Figure 4.1, Figure 8 plots percent of maximum size (horizontal axis) versus the average percent of remaining TP and FP matches (vertical axis). Automatically refined cavity-aware motifs reject a large majority of FP matches, retaining a few more than manually designed motifs. This is expected, because low-impact cavities still eliminate some FP matches, which are not eliminated in automatically refined motifs. Automatically refined motifs retained more TP matches on average than manually designed motifs, for the same reasons.

5. CONCLUSIONS

In order to design more sensitive and specific motifs, we have integrated atom geometry and active cavity volumes into cavity-aware motifs. On 18 nonhomologous motifs, cavity-aware motifs eliminated most false positive matches while preserving most true positive matches. We also observed that some high-impact C–spheres have a greater influence on the number of true positive and false positive matches eliminated, and that high-impact C–spheres can be identified with Cavity Scaling. Cavity Scaling refines the selection of C–spheres in cavity-aware motifs, ensuring that motifs used in practice will contain high-impact C–spheres.

Cavity Scaling is particularly relevant for cavity-aware motif design because it operates independently of C–sphere centers. C–spheres centered on general spatial locations could be filtered with Cavity Scaling for high-impact C–spheres, providing a general approach to C–sphere placement, independent of bound ligands. Cavity Scaling does not entirely answer the problem of designing cavity-aware motifs, because it does not provide quantitative reasons for selecting specific sphere sizes, but from our experience with

this data set, C–spheres at approximately 80-85% of maximum size seem best.

ACKNOWLEDGEMENTS

This work is supported by a grant from the National Science Foundation NSF DBI-0318415. Additional support is gratefully acknowledged from training fellowships of the W.M. Keck Center (NLM Grant No. 5T15LM07093) to B.C. and D.K.; 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. A.C. has been partially supported by a CRA-W Fellowship.

References

1. Stark A., Sunyaev S., and Russell RB. A model for statistical significance of local similarities in structure. *J. Mol. Biol.*, 326:1307–1316, 2003.
2. Wallace A.C., Laskowski R.A., and Thornton J.M. Derivation of 3D coordinate templates for searching structural databases. *Prot. Sci.*, 5:1001–13, 1996.
3. Crane B.R., Arvai A.S., Ghosh D.K., Wu C., Getzoff E.D., Stuehr D.J., and Tainer J.A. Structure of nitric oxide synthase oxygenase dimer with pterin and substrate. *Science*, 279:2121–2126, 1998.
4. Crane B.R., Arvai A.S., Ghosh S., Getzoff E.D., Stuehr D.J., and Tainer J.A. Structures of the n^{ω}-hydroxy-l-arginine complex of inducible nitric oxide synthase oxygenase dimer with active and inactive pterins. *Biochemistry*, 39:4608–4621, 2000.
5. Chen B.Y., Fofanov V.Y., Kristensen D.M., Kimmel M., Lichtarge O., and Kavraki L.E. Algorithms for structural comparison and statistical analysis of 3d protein motifs. *Proceedings of Pacific Symposium on Biocomputing 2005*, pages 334–45, 2005.
6. Levitt D.G. and Banaszak L.J. Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of Molecular Graphics*, 10(4):229–34, 1992.
7. Kristensen D.M., Chen B.Y., Fofanov V.Y., Ward R.M., Lisewski A.M., Kimmel M., Kavraki L.E., and Lichtarge O. Recurrent use of evolutionary importance for functional annotation of proteins based on local structural similarity. *Protein Science*, in press, 2006.
8. Liang J. Edelsbrunner H., Facello M. On the definition and the construction of pockets in macromolecules. *Discrete Applied Mathematics*, 88:83–102, 1998.
9. Edelsbrunner H. and Mucke E.P. Three-dimensional alpha shapes. *ACM Trans. Graphics*, 13:43–72, 1994.
10. Wolfson H.J. and Rigoutsos I. Geometric hashing: An overview. *IEEE Comp. Sci. Eng.*, 4(4):10–21, 1997.
11. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., and Bourne P.E. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.
12. Kuntz I.D., Blaney J.M., Oatley S.J., Langridge R., and Ferrin T.E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.
13. Liang J., Edelsbrunner H., and Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Science*, 7:1884–1897, 1998.
14. Barker J.A. and Thornton J.M. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinf.*, 19(13):1644–1649, 2003.
15. Kinoshita K. and Nakamura H. Identification of protein biochemical functions by similarity search using the molecular surface database ef-site. *Protein Science*, 12:15891595, 2003.
16. Rosen M., Lin S.L., Wolfson H., and Nussinov R. Molecular shape comparisons in searches for active sites and functional similarity. *Prot. Eng.*, 11(4):263–277, 1998.
17. Shatsky M., Shulman-Peleg A., Nussinov R., and Wolfson H.J. Recognition of binding patterns common to a set of protein structures. *Proceedings of RECOMB 2005*, pages 440–55, 2005.
18. Shatsky M., Nussinov R., and Wolfson H.J. Flexprot: Alignment of flexible protein structures without a predefinition of hinge regions. *Journal of Computational Biology*, 11(1):83–106, 2004.
19. Snir M. and Gropp W. *MPI: The Complete Reference (2nd Edition)*. The MIT Press, 1998.
20. Williams M.A., Goodfellow J.M., and Thornton J.M. Buried waters and internal cavities in monomeric proteins. *Protein Science*, 3:1224–35, 1994.
21. Connolly M.L. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221:709–713, 1983.
22. Bachar O., Fischer D., Nussinov R., and Wolfson H. A computer vision based technique for 3-d sequence independent structural comparison of proteins. *Prot. Eng.*, 6(3):279–288, 1993.
23. Lichtarge O., Bourne H.R., and Cohen F.E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257(2):342–358, 1996.
24. Lichtarge O., Yamamoto K.R., and Cohen F.E. Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J.Mol.Biol.*, 274:325–7, 1997.
25. International Union of Biochemistry. Nomenclature Committee. *Enzyme Nomenclature*. Academic Press: San Diego, California, 1992.

26. Smart O.S., Goodfellow J.M., and Wallace B.A. The pore dimensions of gramacidin a. *Biophysics Journal*, 65:2455–2460, 1993.
27. Artymiuk P.J., Poirrette A.R., Grindley H.M., Rice D.W., and Willett P. 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, 1994.
28. Laskowski R.A. Surfnet: A program for a program for visualizing molecular surfaces, cavities, and intramolecular interactions. *Journal Molecular Graphics*, 13:321–330, 1995.
29. Laskowski R.A., Watson J.D., and Thornton J.M. Protein function prediction using local 3d templates. *Journal of Molecular Biology*, 351:614–626, 2005.
30. Laskowski R.A., Luscombe N.M., Swindells M.B., and Thornton J.M. Protein clefts in molecular recognition and function. *Protein Science*, 5:2438–2452, 1996.
31. Adak S., Wang Q., and Stuehr D.J. Arginine conversion to nitroxide by tetrahydrobiopterin-free neuronal nitric-oxide synthase. *J. Biol. Chem.*, 275:33554–33561, 2000.
32. Binkowski T.A., Joachimiak A., and Liang J. Protein surface analysis for function annotation in high-througput structural genomics pipeline. *Protein Science*, 14:2972–2981, 2005.
33. Binkowski T.A., Adamian L., and Liang J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.*, 332:505–526, 2003.
34. Lamdan Y. and Wolfson H.J. Geometric hashing: A general and efficient model based recognition scheme. *Proc. IEEE Conf. Comp. Vis.*, pages 238–249, 1988.