Appeared in: J. Bioinformatics and Comp. Biology 5(2a):353–382, 2007

Journal of Bioinformatics and Computational Biology

CAVITY SCALING: AUTOMATED REFINEMENT OF CAVITY-AWARE MOTIFS 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

[d] *Program in Structural Computational Biology and Molecular Biophysics,*
[e] *Department of Molecular and Human Genetics,*
Baylor College of Medicine
Houston, TX 77030, USA

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Algorithms for geometric and chemical comparison of protein substructure can be useful for many applications in protein function prediction. These *motif matching* algorithms identify *matches* of geometric and chemical similarity between well-studied functional sites, *motifs*, and substructures of functionally uncharacterized proteins, *targets*. For the purpose of function prediction, the accuracy of motif matching algorithms can be evaluated with the number of statistically significant matches to functionally related proteins, *true positives* (TPs), and the number of statistically insignificant matches to functionally unrelated proteins, *false positives* (FPs).

Our earlier work developed *cavity-aware* motifs which use *motif points* to signify functionally significant atoms and *C–spheres* to represent functionally significant volumes. We observed that cavity-aware motifs match significantly fewer FPs than matches containing only motif points. We also observed that *high-impact* C–spheres, which significantly contribute to the reduction of FPs, can be isolated automatically with a technique we call *Cavity Scaling*.

This paper extends on our earlier work by demonstrating that C–spheres can be used to accelerate point-based geometric and chemical comparison algorithms, maintaining accuracy while reducing runtime. We also demonstrate that the placement of C–spheres can significantly affect the number of TPs and FPs identified by a cavity-aware motif. While the optimal placement of C–spheres remains a difficult open problem, we compared two logical placement strategies to better understand C–sphere placement.

Keywords: Protein Structure; Protein Function; Refinement of Protein Motifs.

*= Equal Contribution †Corresponding Author: kavraki@rice.edu

1. Introduction

Geometric and chemical comparison of protein substructures can be applied to the problem of protein function prediction by identifying new sites that share geometric and chemical characteristics with well studied active sites. Several *motif matching* algorithms have been designed for this application, including JESS [7], PINTS [67], Geometric Hashing [41,74], pvSOAR [11,13], and Match Augmentation [18]. These algorithms search for *matches* of geometric and chemical similarity between *motifs*, representing known active sites, and the structures of *target* protein. Within this approach, one difficult subproblem is how to choose active site characteristics, for representation in motifs, so that matches to functionally related targets are found.

In the past, motifs have been created to represent many different types of data, in an effort to characterize different aspects of active site structure and chemistry. These include *point-based* motifs [5,7,18,38,55,56,60,63,64,67,68,70,69], which use *motif points* to represent labeled atomic coordinates, and *cavity-based* motifs [11,13,16,27,26,32,40,42,47,66,72], which use volumetric constructs to represent protein cavities and clefts associated with protein function.

Given comparison algorithms for computing matches of each motif type, individual motifs can be evaluated by measuring the number of statistically significant matches to functionally related proteins, *true positives* (TPs), and the number of statistically significant matches to functionally unrelated proteins, *false positives* (FPs).

In earlier work [16], we hypothesized that point-based and cavity-based motifs could be combined for complementary benefits. This hypothesis led us to propose *cavity-aware* motifs, which combine motif points to represent atomic structure and *C–spheres* to represent protein cavities and clefts. We observed that cavity-aware motifs identify nearly as many TP matches as do point-based motifs, while eliminating a large proportion of FP matches. In earlier work [16], also developed *Cavity Scaling*, an algorithm for refining cavity-aware motifs by identifying *high-impact* C–spheres that tend to eliminate more FP matches.

Contributions While cavity-aware motifs have the potential to identify most TP matches while eliminating many FP matches, it remains a difficult open problem to position C–spheres so that this potential is achieved. Beginning with manually selected motif points, this paper expands on our earlier work to compare two C–sphere placement strategies: Ligand-based C–sphere placement and Volume-based C–sphere placement, which does not require ligated protein structures, expanding the applicability of our technique.

This paper also expands on our earlier work by combining aspects of point-based and cavity-based motifs to accelerate the matching computation. In particular, we demonstrate a method for using C–spheres to reduce comparisons in our matching algorithm relative to point-based motifs.

Finally, this paper provides expanded evidence that Cavity Scaling is capable of identifying high-impact C–Spheres. While our earlier work used CS to refine

cavity-aware motifs with ligand-based C–sphere placement, we demonstrate here that CS is equally effective at refining motifs with volume-based C–spheres placement. CS enables the identification of high-impact C–spheres even for unligated protein structures, generalizing the applicability of cavity-aware motifs.

Organization In Section 2 this paper first surveys existing types of motifs, algorithms for motif matching, and statistical models of geometric matching. Section 3 describes cavity-aware motifs in depth, and then explains two strategies for C–sphere placement. Section 4 presents an expanded description of our pipeline for identifying and statistically scoring matches, describing how we use C–spheres to accelerate our matching algorithm. Section 5 presents the CS algorithm in detail, describing how modifications for large-scale computation can yield significantly improved performance. In Section 6, we present our experimental results. Finally, a discussion of our conclusions is given in Section 7.

2. Related Work and Contributions

Most approaches to the problem of motif matching include a specific representation of motifs, an algorithm for identifying matches for those motifs, and a statistical model for measuring the statistical significance of matches found. This section highlights recent innovations in motif matching, and explains our contributions in the context of this work.

2.1. *Active Site Representations*

Every approach to the problem of motif matching must include motifs: a formal representation of active sites that can be used for geometric and chemical comparison. Developing such a representation is very difficult because biological characteristics relevant for active site function may not be well understood, or may be difficult to represent within the formal motif definition. The two most popular active site representations are *point-based* motifs and *cavity-based* motifs.

Point Based Motifs Point-based motifs are composed of motif points in three dimensions that represent atoms taken from protein structures and active sites. Point-based motifs have represented amino acid C-alpha atoms [68,18], sidechain atoms [5,67], atoms in hinge-bending flexible active sites [68], atoms in catalytic sites [7,69], catalytic triads [70], and conserved binding patterns [63,64]. Motif points have also been used to represent points [60,55,56] and electrostatic potentials [38] on Connolly surfaces [22], and pairs of points have been used to represent vectors of sidechain orientation [31].

Motif points can be labeled with atom and residue information, evolutionary significance and mutation data [18] from the Evolutionary Trace [49,53], hydrogen donor/acceptor and hydrophobic/hydrophilic properties [64], and electrostatic potentials [38]. These labels allow additional discriminating information, derived from protein structures or sequences, to create more selective or more general comparisons.

The design of point-based motifs is fundamentally related to the selection of atoms chosen to represent the active site. Choosing certain atoms rather than others may cause the motif to have less geometric and chemical similarity to functionally related active sites. Alternatively, other atom choices may cause the motif to have greater geometric and chemical similarity to functionally unrelated proteins. To address the former problem, MULTIBIND [63,64] can be applied to identify conserved binding patterns among functionally related proteins, so that motifs retain similarity to functionally related active sites. In response to the latter problem, we developed Geometric Sieving [17], which refines motifs to have less similarity to functionally unrelated proteins.

Cavity-based Motifs Active sites and functional regions can also be represented using the shape of the active cleft or cavity. These *volumetric motifs* have been represented with spheres [40,47,66,72,16], alpha-shapes [27,26,11,13], and grid-based techniques [42,47]. The design of volumetric motifs requires an understanding of which regions the motif should occupy and what amino acids should border the motif. Two examples of volumetric motif refinement are SURFNET-Consurf [32], which modifies the boundaries of computationally identified active clefts to avoid regions distant from highly conserved amino acids, and CS [16], which is summarized later in this work.

2.2. *Geometric Comparison Algorithms*

After selecting a formal representation of active sites, it is necessary to develop an algorithm for motif comparison. Many algorithms have been designed to identify matches, but all methods are optimized for distinct active site representations and thus performance and accuracy is difficult to compare.

Point-based Comparison Algorithms The comparison of point-based motifs is dependent on finding point-to-point correlations between points in the target structure (*target points*) and motif points. One excellent example of a point-based comparison algorithm is Geometric Hashing, [41,74], which uses rotationally and translationally invariant representations of points in space to identify substructural similarity. In addition to being used to search for point-based motifs [60,68,5,69,63], Geometric Hashing has also been used to simultaneously align multiple [46,45], even hinge-bent [62], protein structures. Other point-based comparison algorithms test possible point-to-point correlations in a depth-first-search manner, such as the database search algorithm used in PINTS [61], JESS [7], and Match Augmentation [18] (MA).

Volume-based Comparison Algorithms pvSOAR [11,13] compares volumes in protein structure using cavity-based motifs derived from alpha-shapes and protein sequences. Earlier work on volumetric representations features analysis of only a single protein without comparison. Using varying representations of protein surfaces, these studies, using grid-based algorithms SURFNET [42] and SURFNET-ConSURF [32], and alpha-shapes technique CASTp [48], observed that ligand binding sites are often the largest "pocket" on the protein surface.

2.3. *Statistical Models of Geometric Similarity*

Finding a match 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. This observation adds the third subproblem to motif matching: eliminating matches that are inconsistent with functional similarity. In order to use matches to imply functional similarity, it is essential to understand the degree of similarity consist ant with matches to between motifs and functionally related targets.

A simple LRMSD [a] threshold is insufficient to indicate functional similarity between any motif and a matching target. Some motifs match functionally related proteins at lower values of LRMSD than other motif-target pairs, and LRMSD itself is affected by the number of matching points [18]. Fortunately, it has been observed that matches to functionally related proteins tend to have geometric and chemical similarity which is *statistically significant* [7,11,18,67].

A match indicates statistically significant geometric similarity if the motif has unusually high geometric and chemical similarity to a subset of the target, in comparison to a baseline degree of similarity. Statistical significance is measured, for a given match m, by first determining or estimating the geometric and chemical similarity of the motif relative to a reference set of protein structures. Next, the probability p, called the p-value, of observing another match with equal or greater geometric and chemical similarity is measured. If p is sufficiently low, it is said that m indicates an instance of statistically significant geometric and chemical similarity.

The PINTS [67] database begins with a reference set based on a nonredundant subset of SCOP [54]. The low-RMSD tail of the frequency distribution of matches follows the extreme value distribution, with parameters that can be estimated from motif data, such as the number and type of amino acids. Careful calibration of these parameters to existing matches allows PINTS to generate the extreme value distribution for a wide range of motifs *a priori*. Using the extreme value distribution with a given motif and match LRMSD, PINTS can explicitly evaluate p.

JESS [7] uses a reference set based on set of nonredundant multi-domain representatives from the CATH database [58]. Distributions of matches between a motif and this reference set are modeled using a parametric model of mixtures of normal distributions. JESS applies this approach to comparatively evaluate the significance of matches between a library of motifs and a given target structure. The most significant match in the library provides evidence of functional similarity between the given target and the matching motif.

pvSOAR [11,12], assesses the statistical significance of volume matches between two surface pockets. Given an input match, pvSOAR gathers approximately 38 million other pairs of pockets at random. Ordering these pairs based on geometric

[a] We use LRMSD, the root mean square distance (RMSD) between matching points in 3D when aligned with smallest RMSD [36], to measure geometric similarity.

similarity, pvSOAR finds the number of pairs with greater geometric similarity. The fraction of pairs with greater similarity, relative to the total number of pairs, provides the measure of statistical significance.

3. Cavity-Aware Motifs

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[43,48]. Inspired by seminal work in the modeling and search for protein cavities [43,26,11], we developed cavity-aware motifs, presented in earlier work [16], which combine our own point-based motifs, developed earlier [18,17], with C–spheres that represent volumes essential for protein function.

Cavity-aware motifs, as shown in Figure 1, contain motif points taken from atom coordinates labeled with evolutionary data[18,39]. 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 that we use to prioritize the search for matches. Each s_i also has a set of alternate amino acid *labels* $l(s_i) \subset \{GLY, ALA, ...\}$, that allow our motifs to represent homologous active sites with residues corresponding to evolutionary divergences. In this paper, we obtain labels and ranks using the Evolutionary Trace[49,50].

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 such that c_i contains no atoms from the protein structure used to create the motif. C–spheres are a loose approximation of empty volumes essential for protein function, such as cavities for ligand or cofactor 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, the generality of this representation could also be used to represent protein-protein interfaces and other interaction zones.

In Section 4, we describe how we use cavity-aware motifs in a pipeline for motif matching. Our pipeline uses C–spheres that model molecular volumes which are essential for protein function: If a matching target site truly forms an active site with similar function, a similar cleft represented by $C = \{c_1, c_2, \ldots, c_k\}$ should exist in the target. Matches lacking a similar cleft violate this constraint and can be eliminated.

3.1. *Cavity-Aware Motif Design*

Using C–spheres to eliminate matches, in the manner above, makes the position and radius of each $c_i \in C$ critically related to the set of matches identified. In order

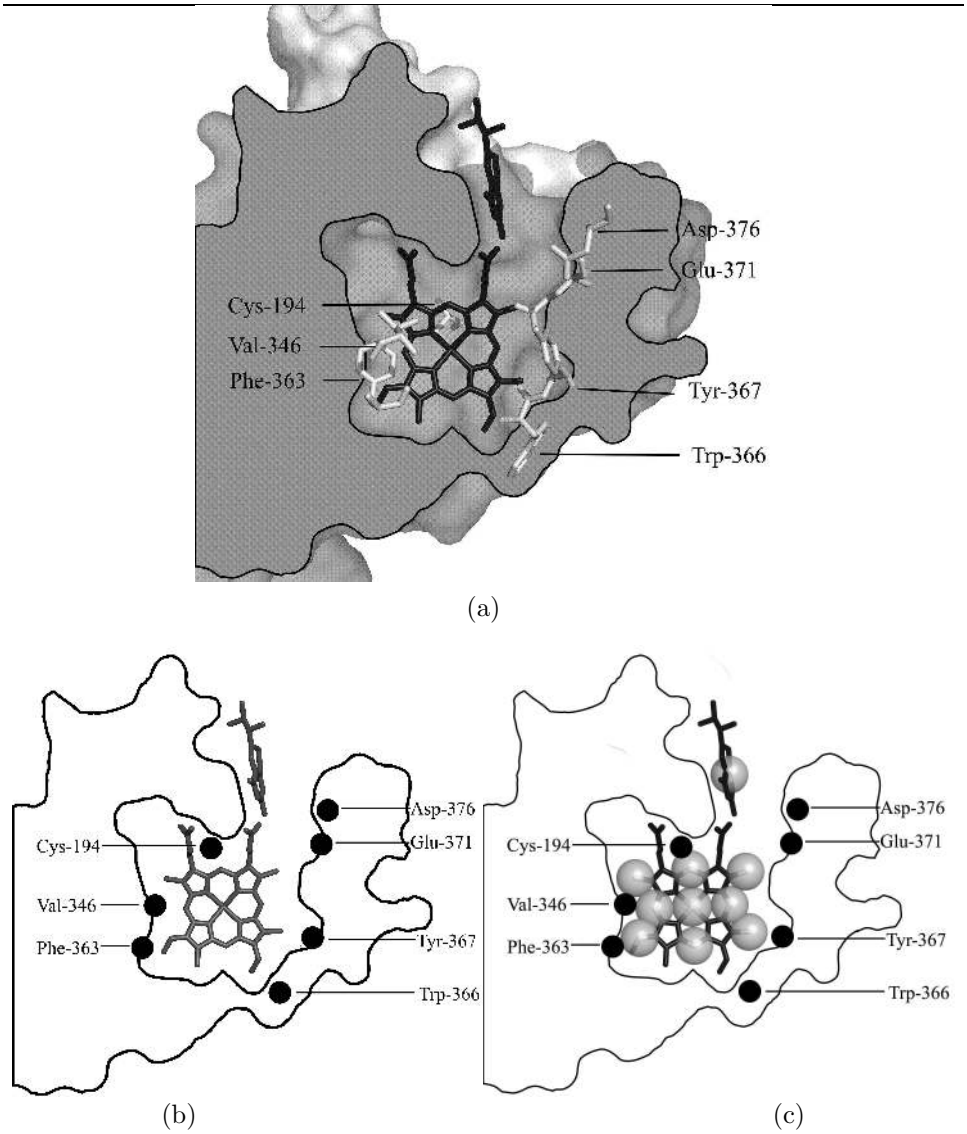

Fig. 1. A diagram of a cavity-aware motif representing the heme-dependent enzyme nitric oxide synthase (pdb id: 1dww), using a ligand-based C–sphere placement strategy. Beginning with functionally relevant amino acids and bound ligand coordinates (a), motif points are positioned at alpha carbon coordinates (black dots, (b)), and C–spheres are positioned at ligand and cofactor atom coordinates (transparent spheres, (c)).

to maximize TPs and FPs, the practical *design* of cavity-aware motifs, realized in the placement and radius of C–spheres, is paramount.

Unfortunately, since C–spheres can occupy any position and any radius, an opti-

mal cavity-aware motif cannot be identified through exhaustive testing. The design of effective cavity-aware motifs is thus a difficult open problem affected by geometric and biological phenomena. While this paper does not offer a solution to this problem, we extend our earlier work by studying two logical strategies for C–sphere placement: Centering C–spheres at atom coordinates of bound ligands, and positioning C–spheres to maximally occupy cleft volume.

3.1.1. *Ligand Based C–spheres*

The first strategy manually places C–spheres in the atom positions of bound ligands. This strategy uses C–spheres to represent volumes that must remain empty for ligand or co-factor binding. While all of the volume within the ligand- or cofactor-binding cleft is not necessarily occupied, this strategy is intended to model volumes essential for protein function.

In general, we did not use every ligand atom, but rather selected atoms so that C–spheres would generally occupy the volume occupied by the ligand or cofactor. No other constraints were used, except that we generally chose 10 or less atom coordinates.

3.1.2. *Volume Based C–spheres*

The second strategy places C–spheres to maximally occupy volumes believed to be related to protein function. We accomplish this in two phases. In the first phase, given the atom coordinates of the protein originating the motif, we use the Qhull library [6] to compute 3D Voronoi cells surrounding each atom in the protein structure.

A Voronoi cell is the distinct region in space, surrounding an atom point a, where any point in the region is closer to a than some other atom point b. In three dimensions, Voronoi cells are always bounded or unbounded polyhedra bordered by Voronoi planes which are equidistant to two points. Voronoi planes intersect in Voronoi lines which form the equidistant boundary between at least 3 atom points. Voronoi lines intersect in Voronoi points which are equidistant to at least four atom points. A C–sphere centered at a Voronoi point, at maximum radius, is the largest sphere in contact with the equidistant atom points.

In the second phase, we eliminate all C–spheres which have a maximum radius below 15 Å. Then, for each member of a set of *indicator points*, we select the C–sphere with largest radius containing the indicator point. All unselected C–spheres are then eliminated.

This filtering phase eliminates C–spheres of large maximum radius, which, in our observations, are mostly outside of the protein. This paper uses ligand atom coordinates as indicator points for functional volumes, but any points, positioned by experts or algorithms for identifying functional volumes, are sufficient for use in the second phase. Thus, the second strategy is not dependant on ligated protein

structures: given a known active cleft or cavity, the placement of indicators in the region is sufficient.

3.2. *Discussion*

There are infinite ways to place C–spheres, but determining which placement strategy is most successful is a difficult open problem. This paper studies two placement strategies as an initial investigation of this difficult problem. Our strategies for C–sphere placement are not techniques for identifying functional volumes, and assume that the functional volume is already known.

The complementary problem of identifying functional volumes has been studied carefully in earlier work [11,14,13,42,43,48,72]. Methods for identifying functional volumes have been applied in the past for the design of cavity-based motifs [12], and have been validated to identify functional volumes [11,42,43]. In the future, that these techniques could be applicable for the design of cavity-aware motifs.

In Section 4, we will demonstrate CS, which refines existing C–sphere placements. CS is not a solution to the C–sphere placement problem, because it is impossible to test all possible placements. Instead, CS is a refinement tool designed to compliment any C–sphere placement strategy by filtering out C–spheres who do not substantially contribute to the elimination of FP matches.

4. A Pipeline for Cavity-Aware Matching

This section summarizes the operation of Cavity Aware Match Augmentation (CAMA) and a statistical model used for identifying statistically significant matches, adding details not found in earlier work. We then explain a new technique for exploiting C–spheres, also not found in earlier work, which substantially accelerates CAMA with small loss of accuracy.

4.1. *Matching Criteria*

CAMA 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 to which t_i belongs. 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 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. This is diagrammed in Figure 2. CAMA outputs the match with smallest LRMSD among all matches that fulfill these criteria. Partial matches correlating subsets of S to T are rejected.

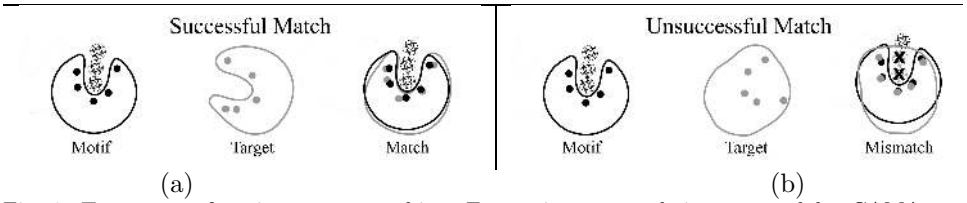

Fig. 2. Two cases of cavity-aware matching. Every time a match is generated by CAMA, an alignment of the motif points is generated to the matching points of the target. This specifies the precise positions of the C–spheres in the motif relative to the target. CAMA accepts matches to targets where no C–spheres contain any target atoms (a), and rejects matches where any target atom is within one or more C–spheres (b).

It is important to note that CAMA seeks the match with lowest LRMSD, but eliminates matches where target points occupy a C–sphere. This causes CAMA to output matches with equal or higher LRMSD, relative to an identical motif lacking C–spheres.

4.2. *Matching Algorithm*

CAMA is a two stage hierarchical matching algorithm, based on MA, 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. As an improvement over our method from earlier work [16], as correlations are being expanded, we insist that C–spheres remain empty. The final output is the correlation with the smallest LRMSD, satisfying all matching criteria.

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, 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.

If there are no more unmatched motif points, we put this match into a heap which maintains the match with smallest LRMSD. If there are more unmatched motif points, we put this partial match back onto the stack. We 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.

C–spheres Prune the Search Space In addition to eliminating matches that do not satisfy our matching constraints, C–Spheres can also eliminate some potential matches being considered by CAMA, increasing algorithmic efficiency. This is because the Augmentation stage is a depth first search which can be represented as a branching search tree. Correlations of motif points and target points represent nodes in this tree, where seed matches represent root nodes. An edge between a parent node and child node represents an instance where the highest ranking unmatched motif point can be aligned with a target point, generating an expanded partial match with an additional correlated pair. Since multiple target points may be available to expand a partial match, the tree can branch from a parent node to several child nodes. This is depicted in the left of Figure 3, while the next unmatched motif points s_3, s_4, and s_5, are shown on the right.

When testing an alignment, if the C–spheres contain a target point, then the children of this node, having correlations with only one additional motif-target pair, will have similar alignments and are likely to have C–spheres which also contain the same target point. Heuristically, we can eliminate the parent node, rather than continue to test additional partial matches. Pruning the tree in this manner reduces the number of comparisons necessary. Empirical testing indicates that this optimization causes CAMA to be approximately 3 times faster than simply testing complete matches (See "Cavity Filtering" in Figure 3).

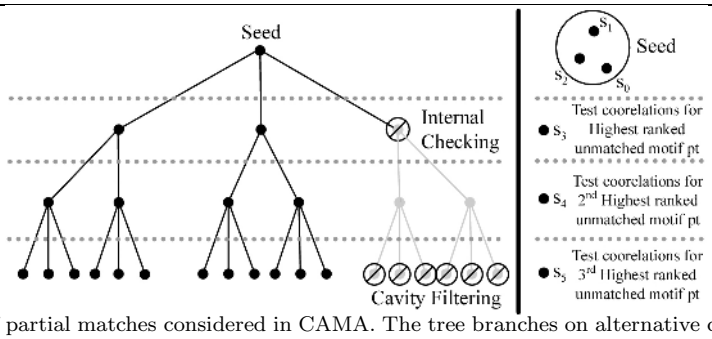

Fig. 3. Tree of partial matches considered in CAMA. The tree branches on alternative correlations between the highest ranked unmatched motif point and an unmatched target point. For example, the three branches from the seed match illustrate that there are three target points that the highest unranked motif point can be correlated with. If optimal alignment of the motif with the correlated target points causes a target point to fall within one or more C–spheres, we can immediately eliminate the match without considering further correlations.

4.3. *Statistical Model*

In earlier work [18], we demonstrated a statistical model for assessing the statistical significance of matches between a point-based motif and target. For a given match m with LRMSD r between motif S and target T, our earlier model assessed the probability p of observing a match with similar LRMSD r', when comparing the same motif and any protein with known structure. First, a match is computed between S and every member of a representative set of proteins, in order to establish a baseline degree of geometric similarity between S and the space of known protein structures. This set of matches is depicted as a frequency distribution, or *motif profile*, in Figure 4a. Figure 4b indicates how p, or the p-value, our measure of statistical significance, is computed. Given a standard of statistical significance α, we say that m is statistically significant if $p < \alpha$.

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.

The purpose of C–spheres is to convert FP matches under point-based motifs into TNs with cavity-aware motifs, by making them statistically insignificant. Given a match m between point-based motif S_p and a functionally unrelated target T, suppose that m is statistically significant, even though T is functionally unrelated to S, making m a FP match. We first compute a motif profile of S_p called P_{S_p}. A cavity-aware version of S_p, called S, contains C–spheres which cause LRMSD to be higher for matches to proteins lacking similar active site clefts, such as T. As we can see in Figure 4, the increase in LRMSDs causes the p-value of the match

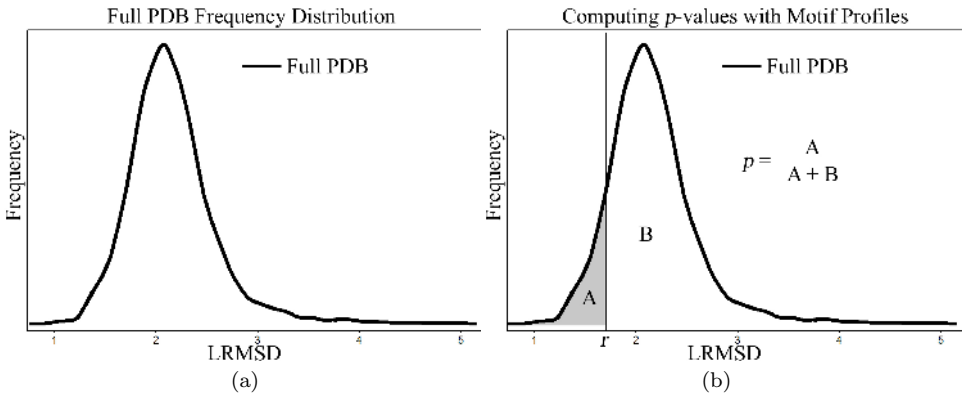

Fig. 4. A frequency distribution of matches between a motif and all functionally unrelated proteins in the PDB (a). Comparing the area under the curve to the left of some LRMSD r, relative to the entire area under the curve (b).

to be higher, possibly over α, in which case the match between S and T becomes statistically insignificant, changing from FP to TN.

Due to variations in active site structure, some functional homologs have atoms which occupy C–spheres, when the match and the motif are optimally superimposed. In our earlier experimentation, which we review in Section 6, we measure both the number of FP matches eliminated, as well as the number of TP matches lost by adding C–spheres. Given effective motifs, the number of TP matches lost is small in comparison to the number of FP matches eliminated.

5. Cavity Scaling

The design of effective motifs is a critical component in the search for similar active sites. As demonstrated earlier [17], the selection of motif points is essential for effective motifs. The position and size of C–spheres in cavity-aware motifs is no exception. In our experiments, we have observed that the selection of C–sphere positions and radii can drastically affect the number of TP and FP matches eliminated, significantly influencing the effectiveness of some cavity-aware motifs.

In order to assist in the design of effective motifs, we have designed CS, a motif refinement algorithm which takes a cavity-aware motif, identifies high-impact C–spheres, and returns a refined cavity-aware motif containing only high-impact C–spheres as output. This section describes how CS identifies high-impact C–spheres.

5.1. *Markers of High-impact C–spheres*

We have observed that motif profiles derived from cavity-aware motifs that include certain C–spheres have a tendency of shifting towards higher LRMSDs as C–sphere radius increases. Figure 5a demonstrates 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 radius. As size increases, the motif profile changes very little. In comparison, in Figure 5b, for the same motif points and a C–sphere in a different position, as radius changes uniformly between zero and the C–sphere's maximum radius, many more matches shift towards higher LRMSDs, as mentioned in Section 4.2.

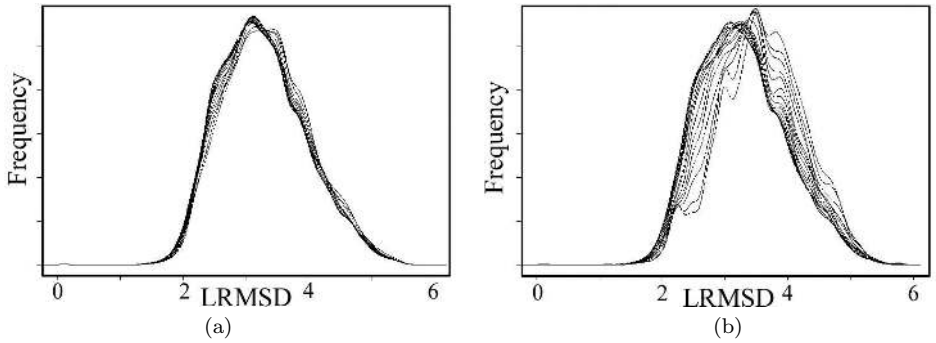

Fig. 5. Motif profiles for a low-impact C–sphere (a), and a high-impact C–sphere (b), as radius increases. For clarity, we provide 20 motif profiles for each C–sphere, showing how much the motif profile changes for a high-impact C–sphere. CS normally inspects only the motif profile with no C–spheres (the profile at the furthest left in both (a) and (b), and the motif profile corresponding to the C–sphere at maximum radius, at the furthest right in both (a) and (b).

As matches shift towards higher LRMSDs, according to our statistical model in 4.3, statistically significant matches become statistically insignificant. This causes FP matches, which make up the dominating majority of matches computed in a motif profile, as mentioned in Section 4.3, to become TN matches. Therefore, C–spheres which cause more substantial shifts towards higher LRMSDs, as radius increases, cause more FP matches to become TN matches, relative to C–spheres which cause less substantial shifts in LRMSD. C–spheres which cause substantial shifts towards higher LRMSDs, therefore, are high-impact C–spheres. This is the primary principle which allows CS to distinguish high-impact C–spheres from low-impact C–spheres.

5.2. *The CS Algorithm*

As diagrammed in Figure 6, CS independently examines motif profiles for each C–sphere of the input, identifying which C–spheres are high-impact. We measure changes in motif profiles by comparing the median LRMSD, in order to distinguish shifts towards higher LRMSDs. While other statistics could be used, our experimentation suggests that motif profile medians are sufficient to identify high-impact C–spheres. Given an input motif S and one of its C–spheres, c_i, CS generates a

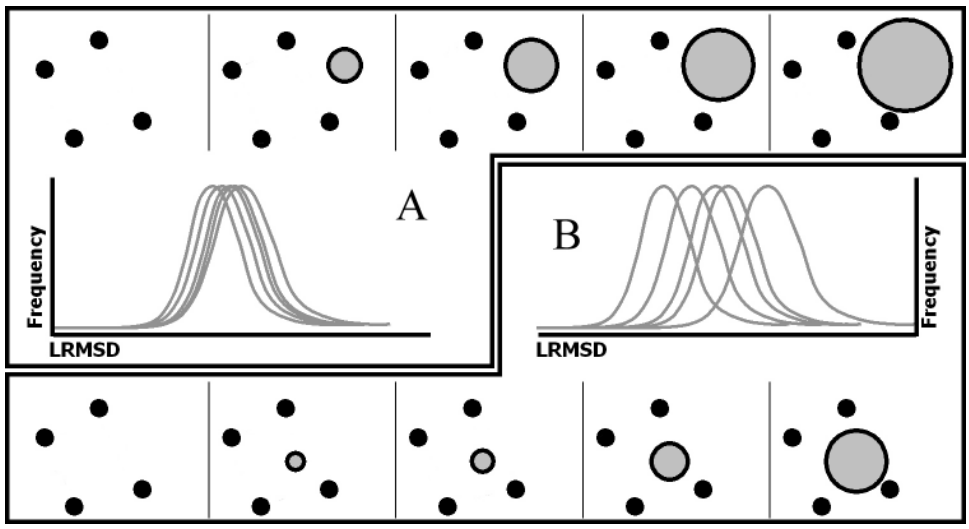

Fig. 6. How CS detects low-impact C–spheres (a) and high-impact C–spheres (b). Motif profiles corresponding to high-impact C–spheres vary significantly in their medians as C–sphere radius increases. Medians for low-impact C–spheres vary little.

variation of S which has no C–spheres, called S_p. Using S_p, CS applies CAMA to compute a motif profile against the PDB, which we call P_{S_p}. We then generate another variation of S, called S_{c_i}, which has only C–sphere c_i at its maximum radius, and compute a motif profile of S_{c_i}, called P_{c_i}, against the PDB. Comparison of the medians of P_{S_p} and P_{c_i}, $med(P_{S_p})$, and $med(P_{c_i})$, respectively, determines if c_i is a high-impact C–sphere. In order to determine if $med(P_{S_p})$, and $med(P_{c_i})$ vary substantially enough to identify c_i as a high-impact C–sphere, we used a simple empirical threshold of .5 LRMSD. An alternative threshold can be computed using confidence thresholds from a method of Efron and Tibshirani [29,28,30].

5.3. *Inherent Parallelism of CS*

The computation of individual matches, for the assembly of many motif profiles, is a computationally expensive but fundamentally parallel task. For this reason, CS could be easily distributed across a cluster of computers. In our own implementation, we have left the computation of individual motif profiles to separate processes, but we have not fully parallelized CS. Operation of many instances of CAMA in parallel leads to an important bandwidth issue which would have to be addressed in a parallel implementation of CS as well: If all CAMA processes are computing matches to a set of targets kept on shared storage, file servers can be easily overwhelmed. We circumvented this problem by storing duplicate targets on local storage, to prevent this bottleneck. The inherent parallelism of CS could be exploited to produce a very powerful too, and in our implementation, the full potential of CS is not yet realized.

CAMA was implemented in C/C++. 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.

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. CS only tests existing C-spheres to determine which are high-impact, and does not address the problem of finding high-impact C–sphere positions from the general set of all possible C–sphere positions. This is a subject of continuing investigation.

6. Experimental Results

In earlier work [16], we demonstrated that cavity-aware motifs, with Ligand Based C–spheres, could eliminate many FPs while preserving most TPs. We also demonstrated that CS could identify high-impact C–spheres, which contribute to the elimination of many FPs while still preserving many FPs, and that motifs refined with CS eliminated almost as many FPs as pre-refinement motifs, while preserving more TPs.

The experiments in this section add a new dimension to our earlier observations by providing a comparison between Ligand Based and Volume Based C–sphere placements. While the optimal placement of C–spheres remains a difficult open problem, these results contribute a comparison of two logical placement strategies. This section also expands our demonstration of CS, providing further evidence that CS is a capable refinement technique for cavity-aware motifs.

6.1. *Input*

Point 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 7. Earlier work has produced examples of motifs designed with evolutionarily significant amino acids[18,39] and amino acids with documented function[44], so these principles 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[39], and the remaining 8 motifs were identified by functionally active amino acids documented in the literature (marked * in Figure 7). Manual selection of evolutionarily significant amino acids and literature search limited the overall number of motifs we considered to 18. The biochemical mechanisms inspiring these motifs are carefully described in Appendix A.

The selection of motif points strongly influences motif sensitivity and specificity. 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. These values are provided in Figure 8.

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

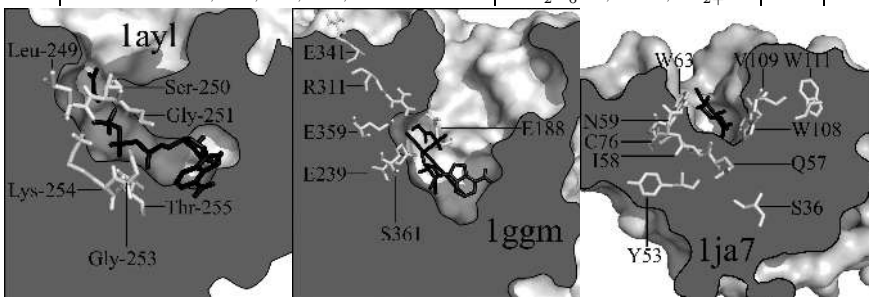

Fig. 7. Motifs used, with example diagrams below. Starred (⋆) motifs use functionally documented amino acids. The column marked "#C" denotes the number of C–spheres in each motif. "R" denotes the range of C–sphere maximum diameters (in Å) for the motif. Functionally documented are described in depth in Appendix A.

C–Spheres C–spheres used in our experimentation were generated using the ligand-based and volume-based strategies described in Section 3. For both C–sphere designs, the maximum radius of any C–sphere was the distance to the nearest atom in the protein structure used to generate the motif. The two strategies for C–sphere placement generated two sets of 18 cavity-aware motifs having with identical motif points and different C–sphere placements.

Functionally Related and Unrelated Proteins 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[57] (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.

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.

6.2. *C-Spheres Eliminate FPs, Preserve TPs*

We compared the number of TP and FP matches found by two different sets of cavity-aware motifs. Both sets had the same motif points, but one set has C–spheres placed using the ligand-based strategy, and the other set had C–spheres placed with the volume-based strategy, as mentioned in Section 3.

We refer to the motifs in each set as $\{S_1, S_2, ...S_{18}\}$. 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 radii $r_{max}(c_1), r_{max}(c_2), \ldots, r_{max}(c_k)$, then the variation $S_{i_j} \in \{S_{i_0}, S_{i_1}, \ldots, S_{i_{19}}\}$ 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 8. 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 8. Second 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 In Figure 8, as C–sphere radius increases, for both C–sphere placement strategies, the number of FP matches are reduced dramatically. The cavity-aware motifs designed using the ligand-based strategy eliminated very few matches until C–sphere radius increased to approximately 80% of maximum radius, whereas motifs designed using the volume-based strategy eliminated TPs more rapidly.

One motif, Phenylalanyl-TRNA Synthetase (1b7y), exhibited 0 sensitivity for both C–sphere placement strategies. 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 8. 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

Point-based Motif Perf.

Motif	#H	TP	FP
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

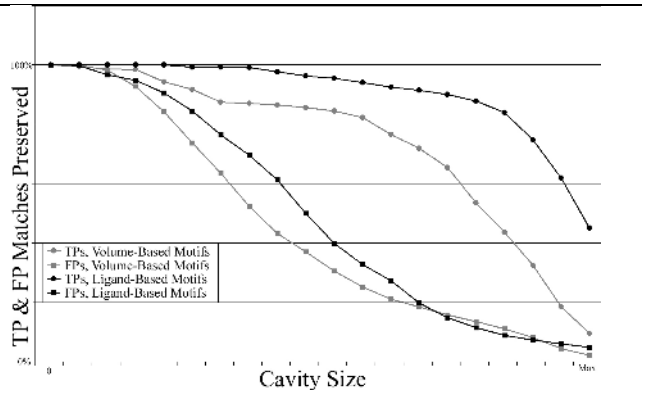

Fig. 8. A comparison of the number of TP and FP matches observed when using cavity-aware motifs, relative to point-based motifs. Since we use C–spheres to eliminate potential matches, our cavity-aware matches only identify the same or less matches than an identical point-based motif. Thus, the 100% line at the top of the graph represents the number of TP and FP matches identified by a point-based motif, normalized accross all 18 motifs. Cavity-aware motifs with C–spheres of radius zero are identical to point-based motifs, and thus all lines in this line graph begin at the upper left corner. As C–sphere radius increases (horizontal axis – see Section 6.2 for a more detailed explanation), lines, corresponding to the number of TP and FP matches observed, monotonically fall from 100%, as matches are eliminated. This graph shows that while fewer TP matches are observed when using cavity-aware motifs, in comparison to an identical point-based motif, FP matches are eliminated in far greater proportions. Using cavity-aware motifs eliminates many FP matches at a small sacrifice of TP matches, in cavity-aware motifs with ligand-based (black) and also with volume-based (grey) C–sphere placement strategies. Cavity-aware motifs with volume-based C–sphere placement strategies tended to eliminate more matches overall in comparison to ligand-based motif designs.

the total number of true positives. Having a 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.

C–spheres designed using the ligand-based strategy seemed to eliminate fewer matches (both TP and FP) than C–spheres designed using the volume-based strategy. In combination with the earlier observation that C–spheres that preserved the most TP matches while eliminating the most FP matches were not the largest C–spheres, but instead around 80% of maximum radius, these observations emphasize the point that positioning and sizing C–spheres to maximize TP matches while minimizing FP matches is a difficult open problem.

6.3. *Analysis of Individual C–spheres*

Some C–spheres have a greater impact on FP match elimination than other C–spheres. We performed CS on each C–sphere in each of our ligand-based and volume-based motifs, identifying which C–spheres were high–impact. 1ayl, with ligand-based

C–sphere placements, used in Figure 9, 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, over increasing radii, shift significantly in the median towards higher LRMSDs and 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 9, which details this effect for single C-sphere motifs based on 1ayl. In the inset graphs, copies of the 1ayl motif containing only C-spheres 4 or 6 undergo significant increases in motif profile medians, as radius increases. In the main graph, 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, eliminating FP matches more slowly as radius increases. C-sphere positions relative to active site geometry are provided in the inset graphic in Figure 9. No correlation between high-impact C–spheres and location within the cavity was apparent, emphasizing again the difficulty of cavity-aware motif design.

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 6.2.

6.4. *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 CS can identify high-impact C-spheres, we use CS to generate ligand-based and volume-based motifs containing only high-impact C–spheres, and demonstrate that they are reasonably effective.

Experiment We applied CS to every C–sphere in both our ligand-based and volume-based motifs. Among the ligand-based versions of each motif, CS identified a set of high-impact C–spheres for all motifs except 1czf, 16pk and 8tln. Among the volume-based versions, CS identified high-impact C–spheres for all motifs except 1ayl, 1czf, 1did, 1lbf, and 2ahj.

We repeated the experiment described in Section 6.2 for the remaining motifs, using only high-impact C–spheres. Motifs without high-impact C–spheres were not included. We refer to these as automatically refined motifs. We compared our results to unrefined the motifs used in Section 6.2.

Observations Like the axes of Figure 6.2, Figure 10 plots percent of maximum radius (horizontal axis) versus the average percent of remaining TP and FP matches (vertical axis). Refined cavity-aware motifs reject a large majority of FP matches,

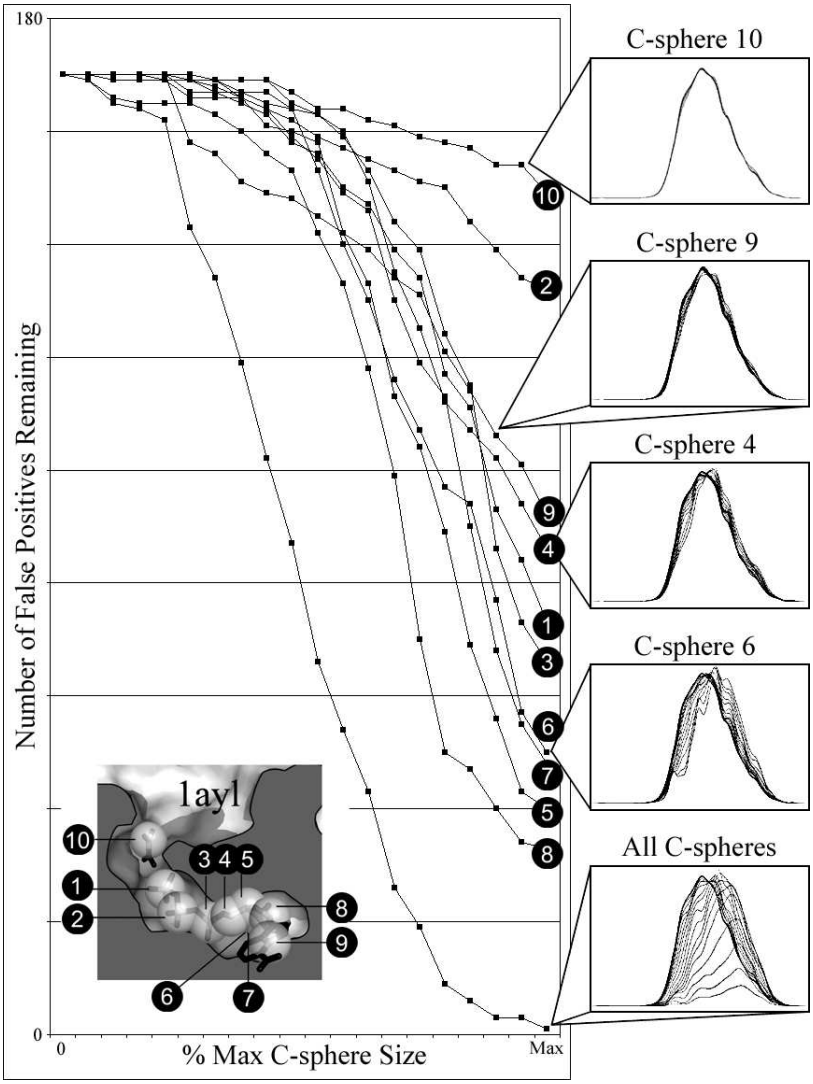

Fig. 9. Effect of Individual C–spheres on Motif Specificity.

As C–sphere size uniformly increases, as described in Section 6.2 (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. CS identifies C–spheres which eliminate more FP matches.

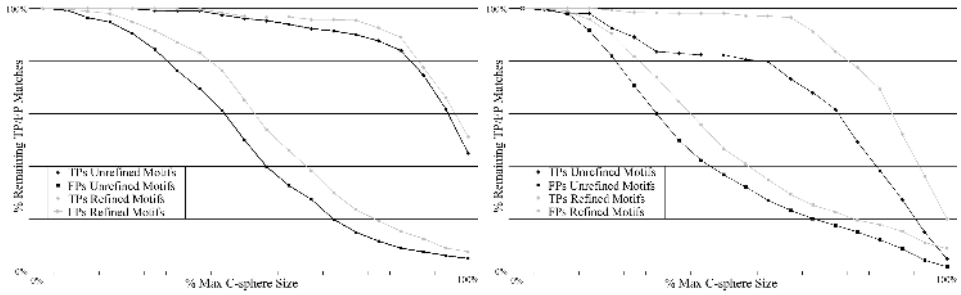

Fig. 10. TP/FP matches preserved when using automatically refined cavity-aware motifs. Axes here are identical those of Figure 8. Cavity-aware motifs with ligand-based (a) and volume-based (b) C–sphere placements, when refined with CS, tended to identify more TP matches, while still eliminating many FP matches, in comparison to unrefined cavity-aware motifs. Motifs refined with CS (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.

retaining a only few more than unrefined motifs. In addition, refined motifs tended to identify additional TP matches.

These results, achieved automatically by applying CS to refine C–spheres based on bound ligands, demonstrate that it is possible to automatically refine configurations of C–spheres to produce improved cavity-aware motifs with little expert knowledge.

7. Conclusions

In earlier work [16], we introduced cavity-aware motifs, and demonstrated that cavity-aware motifs can reduce many FP matches while retaining motifs TP matches. We also demonstrated that CS was capable of refining existing cavity-aware motif designs.

This paper expanded on our earlier work by offering an initial investigation into the design of cavity-aware motifs. Studying ligand-based and volume-based design strategies, we observed that ligand-based designs seemed to eliminate less matches overall, even tho both techniques eliminated many FP matches while preserving TP matches. Alternatively, volume-based motif design strategies can be used to design cavity-aware motifs without ligated protein structures. One additional advantage of cavity-based motifs is that C–spheres can be used to accelerate the matching computation by pruning the search space.

We have also expanded our demonstration of the effectiveness of CS, an algorithm which refines cavity-aware motifs by selecting high-impact C–spheres. CS is particularly relevant to the problem of cavity-aware motif design because it operates independently of expert knowledge. C–spheres centered on general spatial locations, or selected by nonexpert users, could be filtered with CS for high-impact

C–spheres. This approach provides one way to take advantage of existing techniques for identifying functional volumes in protein structures, and also makes the problem of cavity-aware motif design more accessible to nonexpert users. On ligand-based and volume-based motifs, we observed that CS identifies additional true positives while still eliminating many false positive matches.

CS does not entirely answer the problem of refining cavity-aware motifs, because it does not provide quantitative reasons for selecting specific sphere sizes. In the future, by developing a testing apparatus which generates and tests general C–sphere positions, it may be possible to improve the CS process and further assist human motif design.

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. A. Aberg, A. Yarenchuk, M. Yukalo, B. Rasmussen, and S. Cusack. Crystal structure analysis of the activation of histidine by thermus thermophilus histidyl-trna synthetase. *Biochemistry*, 36:3084–3094, 1996.
2. S. Adak, Q. Wang, and D.J. Stueher. Arginine conversion to nitroxide by tetrahydrobiopterin-free neuronal nitric-oxide synthase. *J. Biol. Chem.*, 275:33554–33561, 2000.
3. R.A. Anderson, W.F. Bisron, F.S. Kennedy, and V.L. Vallee. The role of magnesium in escherichia coli alkaline phosphatase. *Proc. Natl. Acad. Sci. USA*, 72:2989–2993, 1975.
4. J.G. Arnez, A.C. Dock-Bregeon, and D. Moras. Glycyl-trna synthetase uses a negatively charged pit for specific recognition and activation of glycine. *J. Mol. Biol.*, 286:1449–1459, 1999.
5. 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.
6. C.B. Barber, D.P. Dobkin, and H.T. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Trans. on Mathematical Software*, 22(4):469–483, 1996.
7. 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.
8. H. Belrhali, A. Yarenchuk, M. Tukalo, C. Berthet-Colominas, B. Rasmussen, P. Bosecke, O. Diat, and S. Cusack. The structural basis for seryl-adenylate and ap4a synthesis by seryl-trna synthetase. *Structure*, 3:341–352, 1995.
9. H. Belrhali, A. Yarenchuk, M. Tukalo, K. Larsen, C. Berthet-Colominas, R. Leberman, B. Beijer, B. Sproat, J. Als-Nielsen, and G. Grubel et al. Crystal structures at

2.5 angstrom resolution of seryl-trna synthetase complexed with two analogs of seryl adenylate. *Science*, 263:1432–1436, 1994.

10. B.E. Bernstein, P.A. Michels, and W.G. Hol. Synergistic effects of substrate induced conformational changes in phosphoglycerate kinase activation. *Nature*, 385:275–278, 1997.
11. 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.
12. 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.
13. 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.
14. T.A. Binkowski, S. Naghibzadeh, and J. Liang. Castp: Computed atlas of surface topography of proteins. *Nucl. Acid. Res.*, 31(13):3352–55, 2003.
15. J. Cavarelli, G. Eriani, B. Rees, M. Ruff, M. Boeglin, A. Mitschler, F. Martin, J. Gangloff, J.C. Thierry, and D. Moras. The active site of yeast aspartyl-trna synthetase: structural and functional aspects of the aminoacylation reaction. *EMBO Journal*, 13:327–337, 1994.
16. 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.
17. 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.
18. 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.
19. J.E. Coleman. Structure and mechanism of alkaline phosphatase. *Annu. Rev. Biophys. Biomol.*, 21:441–483, 1992.
20. 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.
21. C.A. Collyer, K. Henrick, and D.M. Blow. Mechanism for aldose-ketose interconversion by d-xylose isomerase involving ring opening followed by a 1,2-hydride shift. *J. Mol. Biol.*, 212(1):211–235, 1990.
22. M.L. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221:709–713, 1983.
23. B.R. Crane, A.S. Arvai, D.K. Ghosh, C. Wu, E.D. Getzoff, D.J. Stueher, and J.A. Tainer. Structure of nitric oxide synthase oxygenase dimer with pterin and substrate. *Science*, 279:2121, 1998.
24. B.R. Crane, A.S. Arvai, S. Ghosh, E.D. Getzoff, D.J. Stueher, and J.A. Tainer. 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.
25. G.J. Davies, S.J. Gamblin, J.A. Littlechild, Z. Dauter, K.S. Wilson, and H.C. Watson. Structure of the adp complex of the 3-phosphoglycerate kinase from bacillus

stearothermophilus at 1.65 c. *Acta Crystallography section D*, 50:202–209, 1994.
26. H. Edelsbrunner, M. Facello, and J. Liang. On the definition and the construction of pockets in macromolecules. *Discrete Applied Mathematics*, 88:83–102, 1998.
27. H. Edelsbrunner and E.P. Mucke. Three-dimensional alpha shapes. *ACM Trans. Graphics*, 13:43–72, 1994.
28. B. Efron. Better bootstrap confidence intervals (with discussion). *J. Amer. Stat. Assoc.*, 82:171, 1987.
29. B. Efron and R. Tibshirani. The bootstrap method for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):1–35, 1986.
30. B Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chappman & Hall, London, 1993.
31. F. Ferré, G. Ausiello, A. Zanzoni, and M. Helmer-Citterich. Surface: a database of protein surface regions for functional annotation. *Nucl. Acid. Res.*, 32:D240–4, 2004.
32. 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.
33. A.C. Hausrath and B.W. Matthews. Thermolysin in the absence of substrate has an open conformation. *Biological Crystallography*, D58:1002–1007, 2002.
34. D.R. Holland, A.C. Hausrath, D. Juers, and B.W. Matthews. Structural analysis of zinc substitutions in the active site of thermolysin. *Protein Science*, 4:1955–1965, 1995.
35. D.R. Holland, D.E. Tronrud, H.W. Pley, K.M. Flaherty, W. Stark, J.N. Jansonius, D.B. McKay, and B.W. Matthews. Structural comparison suggests that thermolysin and related neutral proteases undergo hinge-bending motion during catalysis. *Biochemistry*, 31:11310–11316, 1992.
36. W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, 32:922–923, 1976.
37. E.E. Kim and H.W. Wyckoff. Reaction mechanism of alkaline phosphatase based on crystal structures two-metal ion catalysis. *J. Mol. Biol.*, 218:449–464, 1991.
38. 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.
39. 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.
40. 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.
41. 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.
42. R.A. Laskowski. SURFNET: A program for a program for visualizing molecular surfaces, cavities, and intramolecular interactions. *Journal Molecular Graphics*, 13:321–330, 1995.
43. R.A. Laskowski, N.M Luscombe, M.B. Swindells, and J.M. Thornton. Protein clefts in molecular recognition and function. *Protein Science*, 5:2438–2452, 1996.
44. R.A. Laskowski, J.D. Watson, and J.M. Thornton. Protein function prediction using local 3D templates. *Journal of Molecular Biology*, 351:614–626, 2005.
45. N. Leibowitz, Z.Y. Fligelman, R. Nussinov, and H.J. Wolfson. Automated multiple structure alignment and detection of a common substructural motif. *Prot: Struct. Func. Genet.*, 43:235–245, 2001.

46. N. Leibowitz, R. Nussinov, and H.J. Wolfson. MUSTA a general efficient automated method for multiple structure alignment and detection of common motifs. *J.Comp.Biol*, 8:93–121, 2001.
47. 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.
48. J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Science*, 7:1884–1897, 1998.
49. 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.
50. O. Lichtarge, K.R. Yamamoto, and F.E. Cohen. Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J.Mol.Biol.*, 274:325–7, 1997.
51. D.T. Logan, M.H. Mazaurick, D. Kern, and D. Moras. Crystal structure of glycyl-trna synthetase from thermus thermophilus. *EMBO Journal*, 14(17):4156–4167, 1995.
52. L. Ma, T.T. Tibbitts, and E.R. Kantrowitz. Escherichia coli alkaline phosphatase: X-ray structural studies of a mutant enzyme (his-412 → asn) at one of the catalytically important zinc binding sites. *Protein Science*, 4:1498–1506, 1995.
53. 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.
54. A.G. Murzin, S.E Brenner, T. Hubbard, and C. Chothia. Scop: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*, 247:536–540, 1995.
55. R. Norel, D. Fischer, H.J. Wolfson, and R. Nussinov. Molecular surface recognition by a computer vision-based technique. *Prot. Eng.*, 7:39–46, 1994.
56. R. Norel, D. Petrey, H.J. Wolfson, and R. Nussinov. Examination of shape complementarity in docking of unbound proteins. *Prot: Struct. Funct. Genet.*, 36:307–317, 1999.
57. International Union of Biochemistry. Nomenclature Committee. *Enzyme Nomenclature*. Academic Press: San Diego, California, 1992.
58. C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton. Cath- a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.
59. L. Reshetnikova, N. Moor, O. Laverik, and D.G. Vassylyev. Crystal structures of phenylalanyl-trna synthetase complexed with pheynylalanine and a phenylalanyl-adenylate analogue. *J. Mol. Biol.*, 287:555–568, 1999.
60. M. Rosen, S.L. Lin, H. Wolfson, and R. Nussinov. Molecular shape comparisons in searches for active sites and functional similarity. *Prot. Eng.*, 11(4):263–277, 1998.
61. R.B. Russell. Detection of protein three-dimensional side chain patterns. new examples of convergent evolution. *J. Mol. Biol.*, 279:1211–27, 1998.
62. M. Shatsky, R. Nussinov, and H.J. Wolfson. A method for simultaneous alignment of multiple protein structures. *Proteins*, 56(1):143–56, 2004.
63. 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.
64. 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.
65. R.T. Simpson and B.L. Vallee. Two differentiable classes of metal atoms in alkaline phosphatase of e. coli. *Biochemistry*, 7:4343–4349, 1968.

66. O.S. Smart, J.M. Goodfellow, and B.A. Wallace. The pore dimensions of gramacidin a. *Biophysics Journal*, 65:2455–2460, 1993.
67. 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.
68. G. Verbitsky, R. Nussinov, and H.J. Wolfson. Structural comparison allowing hinge bending. *Prot: Struct. Funct. Genet.*, 34(2):232–254, 1999.
69. A.C. Wallace, N. Borkakoti, and J.M. Thornton. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. application to enzyme active sites. *Prot. Sci.*, 6:2308–2323, 1997.
70. A.C. Wallace, R.A. Laskowski, and J.M. Thornton. Derivation of 3D coordinate templates for searching structural databases. *Prot. Sci.*, 5:1001–13, 1996.
71. H.C. Watson and J.A. Littlechild. Isoenzymes of phosphoglycerate kinase: evolutionary conservation of the structure of this glycolytic enzyme. *Biochem. Soc. Trans.*, 18:187–190, 1990.
72. M.A. Williams, J.M. Goodfellow, and J.M. Thornton. Buried waters and internal cavities in monomeric proteins. *Protein Science*, 3:1224–35, 1994.
73. I.B. Wilson, J. Dayan, and K. Cyr. Some properties of alkaline phosphatase from escherichia coli. transphosphorylation. *J. Biol. Chem.*, 239:4182–4185, 1964.
74. H.J. Wolfson and I. Rigoutsos. Geometric hashing: An overview. *IEEE Comp. Sci. Eng.*, 4(4):10–21, Oct 1997.

Appendix A. Functionally Documented Active Sites

This appendix describes the functionally documented active sites used in this experimentation. Each biochemical mechanism is described in conjunction with the amino acids used, in order to justify their selection in our motifs, citing all papers used to identify these amino acids.

16pk – Phosphoglycerate Kinase Phosphoglycerate kinase (PGK), a flexible glycolytic enzyme [10], is highly conserved across prokaryotic and eukaryotic species [71]. Upon binding both a diphosphate sugar (1,3-bis-phosphoglycerate) and ADP at two different sites [71] separated by a 10 Åto 12 Åcavity, PGK undergoes a drastic hinge bending motion, bringing the two substrates and the N- and C-termini into brief contact for phosphoryl transfer. The transition state is stabilized by the conserved residues Arg-39, Gly-376, and Gly-399 [10]. The absolutely conserved Pro-45 creates a crucial turn that dictates the final 3-dimensional structure of the N-terminal binding site [25].

1ady – Histidyl-TRNA Synthetase 1ADY is a histidyl-aaRS (class II). Glu-130 and Tyr-264 ensure specificity for histidine by hydrogen bonding to histidine's side-chain nitrogens and excluding hydrophobic residues by making the cavity highly polar. Glu-81 and Thr-83 interact with the bound histidine's α-ammonium group and are along with Arg-311 and Arg-112 are conserved class II residues. [9 1]

1ani – Alkaline Phosphatase Alkaline phosphatase is a ubiquitous, non-specific phosphomonoesterase [19] capable of removing an inorganic phosphate (P_i) from a phosphorylated alcohol or transphosphorylation of P_i to a hydroxyl group of an acceptor [73]. Catalysis occurs in a cavity with three bound metal ligands [37]. P_i is coordinated in the cavity by two zinc ions which play a direct role in catalysis [37] and by the two guanidinium nitrogens of Arg-166. His-331 and His-412 are ligated to Zn_1 while Asp-51 is ligated to Zn_2 and Mg. [52,37,65,3] The Zn_1 activated hydroxyl group of Ser-102 is responsible for a nucleophilic

attack on P_i that serves as a transition state in dephosphorylation or transphosporylation [19,37,52].

1b7y – Phenylalanyl TRNA Synthetase A highly hydrophobic environment is created in the active cavity of phenylalanyl-aaRS (1B7Y) by Phe-258 and Phe-260 which flank the binding region, thereby creating a localized hydrophobic environment that sterically and electrostatically complements phenylalanine while excluding hydrophilic amino acid substrates [59]. Gln-218 is crucial in stabilizing the amino acid to aminoacyl-adenylate transition state [15]. The orientation of the bound phenylalanine substrate is stabilized by the hydrogen bonding of Trp-149 and His-178 to the carboxylate oxygens of the C-terminus [59] and of Ser-180 to the α-ammonium group of the substrate [8].

1did – D-xylose Isomerase D-xylose isomerase catalyzes the conversion of a xylose to a xylulose (ex. glucose to fructose). Asp-56 activates His-53, enabling it to act as a monoprotic base and catalyzes the ring opening of the sugar. A coordinated magnesium ion catalyzes isomerization of the sugar by ionization and is stabilized by Lys-182. Phe-25, Phe-93, and Trp-136 together provide a hydrophobic environment in which a hydride shift occurs [20,21].

1dww – Nitric Oxide Synthase The heme-dependent enzyme nitric oxide synthase, as its name implies, catalyzes the synthesis of nitric oxide (NO) from an L-arginine substrate. Synthesis of NO occurs by conversion of L-arginine to N^{ω}-hydroxy-L-arginine (NHA), NHA to L-citrulline, and finally L-citrulline to NO[24]. This multi-step reaction takes place in a deep cavity and involves zinc, tetrahydrobiopterin, and hydride-donating (NADPH or H_2O_2) cofactors [24,2]. The many cofactors involved in this complicated process are bound by active site residues. 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 [23]. Val-346 and Phe-363 create a small hydrophobic cavity within the larger heme-binding cavity allowing dioxide (O_2) to bind end-on to heme without steric interference [24].

1ggm – Glycyl-TRNA Synthetase Because glycine lacks a side-chain, glycyl-aaRS chemical recognition occurs by the α-ammonium group. The active cavity is ligned by Glu-188, Glu-239, Glu-359, and Glu-241, the carboxlate groups of which create a concentration of negative charges which electrostatically complements the positively charged α-ammonium group of glycine. Selectivity is further improved by the rigid active cavity which is able to sterically exclude larger amphiphilic amino acids. Ser-361 and Glu-359 sterically block competing residues, such as the sterically and electrostatically similar alanine, from binding by excluding all side-chains. The class II conserved Arg-220 is also included [4 51].

1lbf – Glycerol Phosphate Syntase Thermolysin, a member of the family of metalloproteases, utilizes an active cavity with possible hinge regions [35]. Two zinc ions are coordinated to His-231, and the side-chains of Met-120, Glu-143, and Leu-144 assume an alternate conformation that opens or closes the active cavity [33]. The active site is known to include the following residues : Met-120, Glu-143, Leu-144, Tyr-157, and His-231 [34].