Appeared in: J. Computational Biology 14(6):791–816, 2007

The MASH Pipeline for Protein Function Prediction and an Algorithm for the Geometric Refinement of 3D Motifs

Brian Y. Chen[1⋆], Viacheslav Y. Fofanov[2⋆], Drew H. Bryant[3], Bradley D. Dodson[1]
David M. Kristensen[4,5], Andreas M. Lisewski[5], Marek Kimmel[2], Olivier Lichtarge[4,5]
Lydia E. Kavraki[1,3,4]

[1] Department of Computer Science, Rice University
[2] Department of Statistics, Rice University
[3] Department of Bioengineering, Rice University
[4] Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine
[5] Department of Molecular and Human Genetics, Baylor College of Medicine
[⋆] = Equal Contribution. Corresponding author: kavraki@rice.edu

Abstract. The development of new and effective drugs is strongly affected by the need to identify drug targets and to reduce side effects. Resolving these issues depends partially on a broad and thorough understanding of the biological function of many proteins, but the experimental determination of protein function is expensive and time consuming. In response, algorithms for computational function prediction have been designed to expand experimental impact by finding proteins with predictably similar function, mapping experimental knowledge onto very similar, unstudied proteins. One approach is to 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.

This paper first summarizes the **MASH** (**M**atch **A**ugmentation with **S**tatistical **H**ypothesis Testing) pipeline for protein function prediction. MASH combines the Match Augmentation algorithm for efficiently identifying matches with a statistical model for assessing the significance of matches found. MASH has been shown capable of identifying statistically significant matches in functional homologs. However, MASH also makes some incorrect predictions when it identifies statistically significant matches in functionally unrelated proteins. Reducing incorrect predictions is critical, because incorrect predictions can lead to costly mistakes when experimentation is used to verify computational predictions.

An effective function predictor requires effective motifs - motifs whose geometric and chemical characteristics are detected by comparison algorithms within functionally homologous targets (*sensitive* motifs), which also are not detected within functionally unrelated targets (*specific* motifs). Designing effective motifs is a difficult open problem. Current approaches select and combine structural, physical, and evolutionary properties to design motifs that mirror functional characteristics of active sites. We present a new approach, Geometric Sieving (GS), which refines candidate motifs into *optimized motifs* with maximal geometric and chemical dissimilarity from all known protein structures. We refer to this property as Geometric Uniqueness. The paper discusses both the usefulness and the efficiency of GS. We show that candidate motifs from 10 well studied proteins, including α-Chymotrypsin, Dihydrofolate Reductase, and Lysozyme, can be optimized with GS to motifs that are among the most sensitive and specific motifs possible for the candidate motifs. For the same proteins, we also report results that relate evolutionarily important motifs with motifs that exhibit maximal geometric and chemical dissimilarity from all known protein structures. Our current observations show that GS is a powerful tool that can complement existing work on motif design and protein function prediction.

1 Introduction

Broad and extensive knowledge of the biological *function* of proteins would have immense practical impact on the identification of novel drug targets, the reduction of potential side effects, and on finding the molecular causes of disease. Unfortunately, the experimental determination of protein function is an expensive and time consuming process. In an effort to accelerate and guide the experimental process, computational techniques have been developed to annotate functional information about well-studied proteins onto predictably similar but less-studied proteins. One approach is to

search for similar active sites. Algorithms like Geometric Hashing [1], JESS [2], pvSOAR [3] and Match Augmentation (MA) [4], search functionally uncharacterized protein structures (*targets*), for substructures with geometric and chemical similarity (*matches*), to known active sites (*motifs*).

All structures share some degree of geometric and chemical similarity, so it is essential to understand the degree of similarity necessary to imply functional similarity. This can be accomplished with statistical models, such as those used with JESS [2], PINTS [5], pvSOAR [3], and MA [4], which assess the statistical significance of matches found. Measuring geometric similarity by least root mean squared distance (LRMSD[1]), these models determine how unusual the LRMSD of a match is, relative to a baseline degree of similarity common among all protein substructures. The identification of a match with statistically significant LRMSD can suggest that the target and motif have a similar active site, implying potentially similar function [2–5].

This paper first summarizes the **MASH** pipeline (**M**atch **A**ugmentation with **S**tatistical **H**ypothesis Testing), which combines a geometric and chemical comparison algorithm with a nonparametric statistical model for assessing the significance of matches. In earlier work [4, 6], MASH has been shown to identify statistically significant matches to cognate active sites in functionally related proteins. Unfortunately, MASH also identifies some statistically significant matches to functionally unrelated proteins. These matches represent incorrect predictions should be reduced as much as possible, since expensive experimental resources could be expended to verify computational predictions.

One way to reduce incorrect predictions is to optimize motifs for geometric comparison. This is because the set of matches identified by a geometric and chemical comparison algorithm is contingent on the geometric and chemical design of the motif being searched for. Currently, many motifs are designed by experts [4, 2], derived directly from biological literature [7] or from databases of condensed active site information [8, 9]. Other methods select motifs based on analysis of structure or sequence data, such as the largest cavity [10], or using evolutionarily significant amino acids close to known ligand binding sites [4, 6]. While biologically derived data is clearly essential for effective motifs, few existing techniques refine motifs based on geometric properties to make them more effective for geometric comparison. In the context of geometric and chemical comparison, ideally effective motifs have geometric and chemical characteristics which have statistically significant matches to functionally homologous targets (*sensitive* motifs). In addition, ideally effective motifs must also have statistically insignificant matches to functionally unrelated targets (*specific* motifs).

This dually constrained problem is an indicator of two initial approaches for designing effective motifs: designing motifs for similarity to functional homologs, and designing motifs for dissimilarity to functionally unrelated proteins. While this geometric approach to motif design has not been extensively studied, one approach to this problem is the seminal algorithm MULTIBIND [11, 12], which identifies binding patterns common to functionally homologous proteins, thereby producing motifs that will retain geometric and chemical similarity to all known functional homologs. In this paper, we seek, alternatively, to refine motifs to have increased geometric dissimilarity to functionally unrelated proteins.

The second half of the paper describes the design and implementation of *Geometric Sieving* (GS), an algorithm for refining candidate motifs into *optimized motifs* before they are used in MASH. As input, GS accepts a selection of candidate motif points, chosen perhaps by another motif design algorithm, called the *input set*, and the number k of motif points desired in the optimized motif. GS outputs an optimized motif: a motif of k candidate motif points with the greatest geometric and chemical dissimilarity to all known protein structures. We refer this property as *Geometric Uniqueness*. As a geometric criteria for motif design, Geometric Uniqueness comple-

[1] LRMSD is the smallest possible root mean square distance (RMSD) between two sets of aligned points in 3D

ments existing methods for motif design with a novel geometric criteria that can be used to further improve existing motifs.

The motivation and inspiration for defining Geometric Uniqueness stems from several observations in our earlier work [4, 13] and the work of other researchers [2, 5], where it has been observed that motifs which are highly representative of protein function do not occur in a large fraction of the known proteins. We used GS to identify 10 *Geometrically Unique* motifs, and tested them in the MASH pipeline. Optimized motifs produced by GS had among the highest sensitivity and specificity among all possible refinements of the input sets.

Measuring and optimizing Geometric Uniqueness is a nontrivial computational problem because numerous structural comparisons must be made between many motifs and many protein structures. Our implementation of GS efficiently distributes this work across clusters of computers, achieving linear speedup with the number of processors. In addition, we have designed an online statistical analysis which refines the data as it is generated. These optimizations make GS a practical preprocessing tool that refines motifs before they are passed to MASH.

In addition to improving systems for function prediction, geometric refinement of motifs can also yield additional insight about active sites. For example, evolutionarily significant amino acids, defined in [14–18], as those most associated with important evolutionary divergences, have been shown to form statistically significant clusters [19] that are often related to active sites [13]. On our limited dataset, we observed that clusters of evolutionarily significant amino acids are more Geometrically Unique than evolutionarily insignificant amino acids.

This paper does not advocate that Geometric Uniqueness should be the sole criterion for defining effective motifs. It argues, rather, that Geometric Uniqueness is an interesting property that seems to be useful for refining existing motifs. It also argues that GS is a novel methodology which can be used to optimize motifs designed by human intuition, or by other motif design methods, such as MULTIBIND [11]. It finally argues that Geometric Uniqueness can be compared with other known criteria for selecting motifs in an effort to better understand and finally attack the difficult problem of protein function prediction.

2 Related Work

Many techniques have been developed that are related to the identification of functional sites and the prediction of protein function. These include methods which analyze individual protein structures and networks of proteins. These methods also include algorithms that compare sequence motifs, whole protein sequences and whole protein structures. While these topics border on the subject of this paper, the comparison of protein substructures for function prediction, in this section, we will focus on describing topics most related to this area.

The most basic problem, in the study of substructural comparison techniques for function prediction, is the need to identify geometric and chemical markers which can unambiguously indicate functional similarity. This problem is manifested in the study of effective motif types. In addition, function prediction tools also require geometric comparison algorithms which efficiently identify matches of geometric and chemical similarity to given motifs. Finally, it is essential to develop statistical models to determine what degree of similarity is necessary to imply functional similarity. This section describes related work on these major subproblems.

2.1 Motif Types and Design

The search for geometric markers of functional similarity has considered many types of motifs. This study could be loosely organized into point-based motifs and volumetric motifs.

Point Based Motifs Point-based motifs are composed of geometric points in three dimensions. One prominent use of point-based motifs has been to represent atom coordinates taken from protein structures and active sites. Point-based motifs have been used to represent amino acid C-alpha atoms [20, 4], sidechain atoms [21, 5], atoms in hinge-bending flexible active sites [20], atoms in catalytic sites [2, 22], catalytic triads [23], and conserved binding patterns [11, 12]. In each of these cases, point-based motifs are used to represent specific atoms or groups of atoms, as a direct representation of atomic structure.

Point-based motifs have also been used to represent more abstract structural data, such as lattice points [24–26] and electrostatic potentials [27] on Connolly surfaces [28]. Here, point-based motifs represent critical topological information, such as the deepest part of a "hole" or the highest part of a "knob", on the protein surface. Another example is the use of pairs of points to represent vectors of sidechain orientation [29]. This abstraction of sidechain orientation permits a higher resolution description of sidechain orientation while preserving the ability to compare different amino acids.

Many data structures have been developed for representing point based motifs. While vectors are the most common representation [23, 26, 20, 2, 27, 4], other representations of points in space include distance matrices [30, 31] and graphs [32–34].

Point-based motifs are easily labeled with biological information. When representing atoms, this natural extension has been used widely to label points with atom and residue information. Points have also been labeled with evolutionary significance and mutation data [4] from the Evolutionary Trace [14, 35], hydrogen donor/acceptor and hydrophobic/hydrophilic properties [12], and electrostatic potential [27].

There are many ways to represent the same active site with motifs of a specific type. For point-based motifs, the choice of atoms and how to label them is critical to successfully finding matches to functionally related portiones. In current work, point-based motifs have been designed using the Evolutionary Trace [14, 35] and proximity to binding sites [13, 4]. Motifs have also been designed using literature search and PSI-BLAST alignments of literature-defined motifs from the Catalytic Site Atlas [8, 9], and manually, by experts [2]. Still other motifs are designed using surface exposure, and algorithms for detecting conserved binding patterns [11]. These methods seek to identify substructures which are involved in biological function. Recent techniques also use geometric analysis to refine point-based motifs. GS [36], presented later in this paper, is one such method. Another excellent example is MULTIBIND [11, 12], which identifies conserved binding patterns by identifying the least common point set among a set of existing motifs.

Volumetric Motifs Another way to represent active sites and function regions is to model the shape of the active cleft or cavity. Volumetric motifs have been represented with spheres [37–40, 7], alpha-shapes [41, 42, 3, 10], and grid-based techniques [43, 38]. The design of volumetric motifs involves the questions of which regions the motif should occupy and what amino acids should border the motif. One example of volumetric motif design is SURFNET-Consurf [44], which modifies the boundaries of computationally identified active clefts, to avoid regions distant from highly conserved amino acids.

2.2 Geometric Comparison Algorithms

A broad range of geometric comparison algorithms have been developed for individual motif types. These algorithms are highly specialized, making performance comparisons difficult.

Algorithms for comparing point-based motifs identify geometric similarity by finding point-to-point correlations between *motif points*, and the points in the target, or *target points*. Point-based motifs have been supported strongly by the seminal Geometric Hashing framework [1, 45], which hashes rotationally and translationally invariant geometric representations for efficiency. Geometric

Hashing has been applied in many different ways: it can search for many point-based motif types [24, 20–22, 11], refine point-based motifs by identifying the largest common point set among a set of similar motifs [12], and simultaneously align multiple [46, 47], even hinge-bent [48], 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 [49], and JESS [2]. Still other point-based comparison algorithms use techniques which find subgraph isomorphisms [34].

pvSOAR [3, 10] compares volumes in protein structure using motifs based on alpha-shapes. Earlier work on volumetric representations found active sites through geometric analysis of a single protein structure. Using varying representations of protein surfaces, studies using grid-based algorithms SURFNET [43] and SURFNET-ConSURF [44], and alpha-shapes technique CASTp [50], observed that ligand binding sites are often the largest "pocket" on the protein surface.

2.3 Statistical Models

One governing assumption is that geometric and chemical identity implies functional similarity. However, protein structures are never perfectly identical. For this reason, understanding the degree of geometric and chemical similarity necessary to imply functional similarity is a critical aspect of function prediction. If a given match indicates similarity that is significantly greater than a baseline degree similarity between functionally unrelated proteins, then we expect that the given match indicates functional similarity. Therefore, a baseline degree of geometric similarity is essential to evaluate the significance of geometric matches.

For some approaches, geometric and chemical similarity is measured differently. Geometric Hashing [1], JESS [2], PINTS [5], and MA [36] measure geometric similarity using LRMSD. pvSOAR [10] uses both LRMSD and *oRMSD*, which is computed by first projecting all points onto the unit sphere at the center of mass, and then computing LRMSD.

Reference Sets To establish a baseline degree of similarity between functionally unrelated proteins, we first require a reference set of functionally unrelated proteins. Any reference set must remain unbiased, so that truly significant matches are identifiable relative to this background. This is a very difficult problem because the space of protein structures is largely unknown, and because the space of known protein structures contains over- and under-represented protein structures.

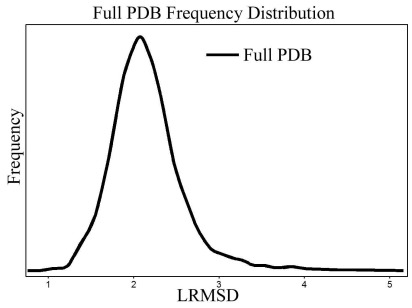


Fig. 1. A typical frequency distribution of matches between a motif and the PDB [51].

Current reference sets are generated from databases and classifications of protein structures, including the Protein Data Bank (PDB) [51], SCOP [52], a classification of protein folds, and CATH [53], a multi-level nested categorization of increasingly specific protein sequence and structure classifications. In an effort to gather an unbiased reference set, recent statistical models have

computed matches to all structures in the PDB [4], and to structurally nonredundant subsets of the PDB [7]. Other statistical models compute matches to fold representatives [5] from SCOP, and non-redundant multi-domain representatives [2] from CATH. The distribution of matches between a motif and proteins in a reference set, such as the PDB, in Figure 1 can be visualized as a frequency distribution, which is essentially a histogram that plots frequency (the number of matches with a particular LRMSD) versus LRMSD.

Measuring Statistical Significance Given a basline degree of similarity, it is then necessary to determine if a specific match LRMSD is statistically significant. This can be determined with several different methods, summarized below.

The PINTS [5] database computes matches between a motif and every protein in a nonredundant subset of SCOP [52]. The tails of the frequency distribution follow the extreme value distribution, with parameters that can be estimated from motif data. Careful calibration of these parameters allow PINTS to generate the extreme value distribution for a wide range of motifs *a priori*. Using this distribution with a given motif and match LRMSD, PINTS can explicity evaluate a p-value, which measures the degree of statistical significance.

JESS [2] uses a set of nonredundant multi-domain representatives from CATH as the basis for generating their reference set. The distributions of matches generated between a motif and this reference set is modeled using a parameteric 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 [3, 54], a method for comparing volumetric motifs, can assess 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 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.

2.4 Systems for protein function prediction

Recent approaches to the problem of function prediction have led to the design of many powerful computational resources, including databases of functional annotations generated by function prediction algorithms, and web servers providing geometric and chemical comparison services. These systems integrate motif types, geometric comparison algorithms, and statistical models to provide the best possible predictions. pvSOAR [3] is provided as part of a web service for identifying similar protein surface regions in protein structures, and CASTp [50] provides an atlas of protein pockets and voids for all structures in the PDB [51]. Finally, the PINTS server [5] provides a rapid database search algorithm coupled with a statistical model of structural similarity. PROFUNC [55], provides numerous sequence and structure analyses in a single package, which include BLAST [56], InterProScan [57], SSM [58], and JESS [2], among many others.

3 MASH: A Pipeline for Protein Function Prediction

In earlier work [4], we presented a prototype pipeline for protein function prediction. In this section we summarize this pipeline and provide critical details needed for the rest of the paper, not found in earlier work. MASH is composed of two major components: MA and a statistical model for analyzing geometric and chemical similarity. As input, MASH accepts a snapshot of the PDB and motifs of a specific type described below. A set of matches to targets in the PDB are computed and

passed to the statistical model, which assigns to each match a p-value that assesses the statistical significance of the match. Using a standard of acceptable statistical significance, α, statistically significant matches with $p < \alpha$ are returned as output.

3.1 Motifs

MASH uses point-based motifs which encode evolutionary data into the labels. A MASH motif S, contains a set of $|S|$ 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*, which is a measure of the functional significance of the motif point. Each s_i also has a set of alternate amino acid *labels* $l(s_i) \subset \{GLY, ALA, ...\}$, which represents 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 [14, 15].

3.2 Matching Criteria

MA compares a motif S to a target T, a protein structure encoded as $|T|$ *target points*: $T = \{t_1, \ldots t_{|T|}\}$, 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 $M = \{(s_{a_1}, t_{b_1}), (s_{a_2}, t_{b_2}) \ldots (s_{a_{|S|}}, t_{|S|})\}$. Referring to the 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 biologically compatible: $l(t_{b_i}) \in l(s_{a_i})$.
> **Criterion 2** LRMSD alignment, via rigid transformation A of S, causes $\forall i, ||A(s_{a_i}) - t_{b_i}|| < \epsilon$, our threshold for geometric similarity.

MA takes as input a motif S and a target T. MA outputs the match with smallest LRMSD among all matches that fulfill the criteria. Partial matches correlating subsets of S to T are rejected. By establishing a threshold for acceptable geometric similarity, the second criterion causes MA to return match LRMSDs bounded above by ϵ. We find that $\epsilon = 7$ Å permits the identification of structurally distant matches when no matches with lower LRMSD exist, while still efficiently identifying matches with high structural similarity.

3.3 Match Augmentation

MA searches for the set of point-to-point correlations which satisfy our criteria, and have the smallest LRMSD among all matches considered. MA takes an algorithmic approach which is distinct from other structural comparison algorithms because it proceeds in a prioritized manner in finding these correlations. Matches are found in two primary phases: *Seed Matching*, and *Augmentation*. Seed Matching first identifies correlations for the three highest ranking motif points, and passes this list of *seed matches* to Augmentation. Augmentation expands each seed match into a set of correlations for all motif points, in order of rank. During this expansion process, Augmentation tracks the match with lowest LRMSD, returning it when all seed matches have been fully expanded.

Seed Matching Given a motif S and target T, seed matching begins by identifying the *seed*, the three highest ranking motif points $S' = \{s_1, s_2, s_3\}$. After identifying the seed, we interpret $T' = \{t_1, t_2, \ldots t_{|T|}\}$ as a graph [59], where each vertex is a target point t_i. We then eliminate all t_i which are not compatible with one of $\{s_1, s_2, s_3\}$. Since S' has exactly three points, there are exactly three interpoint distances between points in S': the distance $||s_1 - s_2||$, $||s_2 - s_3||$, and $||s_1 - s_3||$. We will refer to these distances as *red*, *blue*, and *green*, respectively. Suppose t_i,t_j are compatible with s_1,s_2, respectively. Then, if $-2\epsilon \leq ||t_i - t_j|| - ||s_1 - s_2|| \leq 2\epsilon$, target points t_i,t_j are

at a similar distance and also compatible with s_1,s_2, making them a two point geometric match. We visualize two point geometric matches with s_1,s_2 on the target by inserting red edge between t_i,t_j. An identical process defines blue and green edges between target points compatible with s_1,s_3 and s_2,s_3 respectively, where again inter-point distances are within 2ϵ. Once we complete the search for all colored edges, we search the graph for all three colored triangles. Each triangle identifies three target points which are label compatible with S', and positioned at similar distances. For each triangle, LRMSD with S' is calculated, and if all points are aligned within ϵ, the new seed match is stored. The k lowest LRMSD seed matches are passed to Augmentation, in a stack data structure ordered in ascending LRMSD.

Implementing Seed Matching efficiently requires a range-search data structure like a *kd*-tree [60, 61], which can be used to identify points in a range of distances without checking all points. A target T has at most $\binom{|T|}{3} = O(|T|^3)$ matching triangles, but this worst case requires target points to be very close together. Van der Waals interaction forces make this impossible on biological data, where typical performance has been observed to be close to $O(n^2)$.

Augmentation Augmentation expands a seed match to find correlations between all motif points and a subset of the target. The input seed matches begin on a stack of incomplete matches. Popping off the first seed, augmentation plots the LRMSD alignment of the seed onto the three correlated target points. Relative to this alignment, we calculate the position of the highest ranked unmatched motif point s_i as if it were rigidly aligned with the rest of the seed. We now seek target points which correlate with s_i that do not misalign the match. In the spherical vicinity V of s_i, we identify all t_i within V which are compatible with s_i. We explore only in V because distant points will violate our second match criteria, mentioned earlier. Then, for each compatible t_i, we compute the LRMSD alignment A of the seed match with the addition correlation of s_i to t_i. If $||A(s_i) - t_i|| \geq \epsilon$, the second criteria is violated and the match is discarded. If $||A(s_i) - t_i|| < \epsilon$, the second criteria is not violated, and the seed match with the additional correlation (s_i, t_i), becomes a *partial match*, and is pushed onto the stack of incomplete matches. The use of a stack causes Augmentation to behave like a stack-based depth first search (*DFS*), exhaustively expanding one partial match before continuing on to other seed matches. Once all t_i in V have been considered, we then pop off the first match from the stack of incomplete matches, and repeat this process. Since motifs have a finite number of points, at some point, no unmatched motif points remain. Rather than push these *completed matches* back onto the stack, the match is stored, and the LRMSD is recorded, tracking always the completed match with lowest LRMSD. Eventually, the stack is emptied, completing the Augmentation phase. The final output from Augmentation is the completed match of all s_i to distinct t_i, with lowest LRMSD.

Performance is dependent on the number of motif points $|S|$, and c_r, the number of compatible t_i found in V, giving runtime $O(|S|^2(c_r^{|S|-3}))$. c_r is bounded because repulsive Van der Waals forces limit the number of atoms found in V. The quadratic factor is the aggregate cost of LRMSD calculations, and the exponential is the cost of DFS with c_r possibilities per iteration. With $|S|$ usually 4-13 points, Augmentation is extremely efficient.

3.4 A Nonparametric Statistical Model for Matches

Our statistical model uses a hypothesis testing framework, which detects matches with statistically significant geometric and chemical similarity. Match significance is assessed by comparing the match LRMSD to a baseline degree of geometric and chemical similarity, which is established with a reference set of protein structures. In this section we will first describe the reference set of proteins that we use and then explain the structure of our hypothesis testing framework.

A Reference set of Proteins We refer to our reference set of protein structures as Ω, and for each motif S that we use, our baseline is dependent on the set of matches between S and Ω, which we refer to as the *motif profile* S_Ω. As mentioned earlier, motif profiles are best visualized as frequency distributions (see Figure 1).

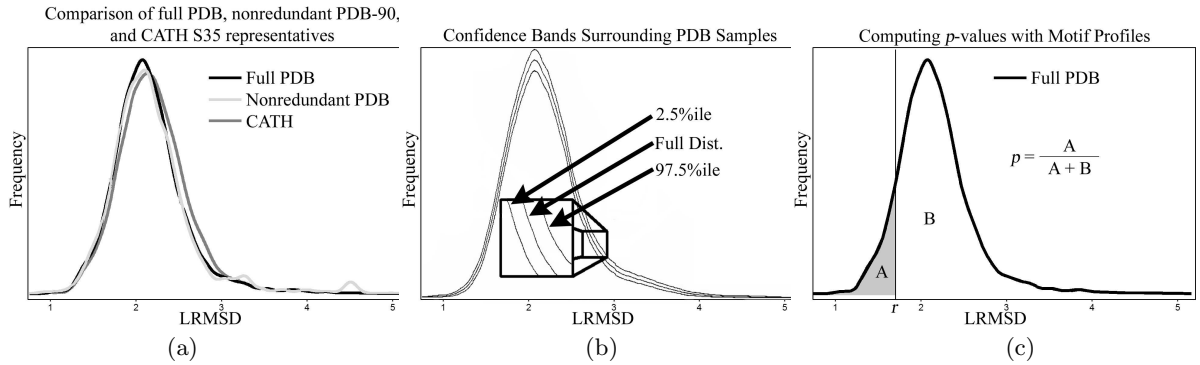


Fig. 2. (a) Comparison of PDB, sequentially nonredundant PDB, and CATH representatives. (b) Confidence band demonstrating the accuracy of samples of the PDB. (c) Volumes measured while computing the p-value. This data computed using the motif C42, H57, C58, D102, D194, S195, S214 from α-Chymotrypsin (1acb).

The purpose of the reference set Ω is to represent the set of all known protein structures. However, we have found that different representations of Ω tend not to have significant effect on the actual shape of motif profiles generated. For the ten motifs optimized for this work, we observed strong similarity between motif profiles calculated with the PDB (Ω_0), and Ω_{nr25} and Ω_{nr90}, which are two sets of sequentially nonredundant PDB structures having no more than 25% (resp. 90%) sequence identity. A similar comparison was true when using the CATH [53] database. We selected a representative of every category at the three most specific levels: Topologies (Ω_T), Homologous Superfamiles (Ω_H), and Sequence Families Ω_S. In our experience, motif profiles on these representatives also resemble Ω_0, in increasing degrees of similarity corresponding to increasingly specific levels of CATH. The similarity between the Ω_0 (black), Ω_{nr25} (light grey) and Ω_S (dark grey) is plotted in Figure 2a. Ω_{nr90}, Ω_T, and Ω_H were excluded for clarity, but are closely related. The similarities between the different reference sets considered here is testament to the high fidelity of structural and sequential classification in CATH [53].

We have also observed that motif profiles on Ω_0 are exceptionally robust to random sampling. Ω_5 is the random 5% sample of PDB structures in Ω_0, and motif profiles with this set are called S_{Ω_5}. In our experience, for any S, S_{Ω_5} resembles S_{Ω_0} with high accuracy. This can be seen in Figure 2b, where we overlayed 5000 distinct S_{Ω_5} samples with a single S_{Ω_0}, the center line in Figure 2b. 95% of the 5000 S_{Ω_5} fell within the upper and lower lines, demonstrating that motif profiles based on Ω_5 retain high similarity to motif profiles based on Ω_0.

Because our observations suggest that motif profiles based on many logical reference sets, including Ω_S, Ω_H, Ω_T, Ω_{nr25}, Ω_{nr90}, differ little from motif profiles based on Ω_5, this paper proceeds by using Ω_5. 5% sampling greatly reduces the number of matches necessary to compute a motif profile, while its simple definition promotes the reproducibility of this work.

Statistical Hypothesis Testing 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. 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[4].

Geometric comparison algorithms operate on 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:

$$H_0\text{:}\quad S \text{ and } T \text{ are structurally dissimilar}$$
$$H_A\text{:}\quad S \text{ and } T \text{ are structurally similar}$$

Our statistical model tests these hypotheses by comparing the given match LRMSD r to the motif profile S_{Ω_5}, which is essentially a large set of functionally unrelated proteins. 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. This is accomplished by computing the volume under the curve to the left of r, relative to the entire volume (see Figure 2c). The probability p, referred to as the p-value, is the measure of statistical significance. Note that when computing p for multiple matches of the same motif to different targets, the motif profile does not need to be recomputed, since it is dependent only on the motif and the reference set.

If p is very low, then we say that m identifies unusually high geometric and chemical similarity, allowing us to follow the implication that this match is significantly similar and thus indicative of functional similarity. Technically speaking, we use a standard of statistical significance α, so that if $p < \alpha$, 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). Under these conditions, we call m statistically significant.

Measuring volumes under motif profile curves, as demonstrated in Figure 2c requires careful numerical treatment. We apply kernel density estimation procedures [62] to estimate population density from the motif profile. Since data is not always evenly spaced, we use Gaussian Kernel smoothing to interpolate between data points, as in previous work [4]. In addition, we avoid under- and over-smoothing by using optimal bin-widths determined by Sheather-Jones method [63, 64].

3.5 Performance of MASH

In earlier work, we successfully demonstrated that MASH can identify cognate active sites in functional homologs [4, 6, 7]. Much like many nontrivial prediction techniques, some incorrect predictions, which identify geometric similarity between functionally unrelated proteins, do occur. In other work, we have shown that integrating volumetric and point-based motifs [36] and eliminating matches to evolutionarily insignificant target residues [6] can reduce incorrect predictions.

In the next section, we describe GS, a geometric analysis for analyzing and refining the selection of atom coordinates used in motif designs. Used in tandem with existing motif design strategies, we will show that our technique can further refine existing motifs.

4 Geometric Sieving

GS accepts an input set, a collection of candidate motif points which could be selected by another motif design method, such as those mentioned in Section 2.1, or provided by a user seeking to

improve a motif. GS also requires k, the number of candidate motif points expected in the output, and, as discussed in the previous section, a geometric comparison algorithm compatible with the motif type used. The output of GS is the subset motif with k points that has highest Geometric Uniqueness.

GS is a refinement process, not a motif discovery algorithm. If no subset motif of size k has geometric and chemical similarity to functionally homologous active sites, then GS cannot select one which does. For this reason, the input set is assumed to contain a subset motif of size k, which has basic geometric and chemical similarity to functional homologs of the input set. By this assumption, matches to functional homologs remain in the low-LRMSD tail at the lower left of the motif profile for many subset motifs, while functionally unrelated proteins, the vast majority of matches in a motif profile, gravitate around the large mode near the median LRMSD. The difference in LRMSD between this low-LRMSD tail and the major mode of the distribution causes matches to functional homologs to be statistically significant relative to the distribution overall [4]. With many different combinations of motif points to choose from, in the form of varying subset motifs, we can select the motif profile which maximizes the LRMSD difference between the low-LRMSD tail and the major mode. As a result, matches to functional homologs will be maximally statistically significant for the input set considered. GS implements this task by analyzing motif profiles.

In this work, between two motif profiles, the motif profile with higher median LRMSD has higher Geometric Uniqueness. Medians are computed on kernel density smoothed motif profiles. While other statistics for quantitative comparison exist, such as the mode, our experimentation shows that comparing the medians of motif profiles is an elegant and effective approach for determining which motif is more Geometrically Unique. In addition, medians are not affected by extreme values at the tails of the distribution. Estimating the true median of the population from a sample is less prone to sampling errors and errors due to incorrect choice of smoothing parameters than mode estimation [65]. In our results, we show the connection between medians and the actual distribution, demonstrating that motif profiles with higher medians are motif profiles with more and/or higher match LRMSDs.

The motif *size*, the number of motif points in a motif, is partially related to Geometric Uniqueness. Larger motifs specify more geometric constraints, and so tend to have higher LRMSD matches than smaller motifs [4]. Thus, we avoid comparing motif profiles from subset motifs of different sizes, ensuring that only the true geometric and chemical differences drive the motif profile comparison. This is why k, the size of the optimized motif, is an input. The operation and success of GS is not affected by k, and our results hold over varying k, as we will demonstrate later. Selecting an ideal k *a priori* remains an open problem, and the subject of continuing research.

4.1 The Geometric Sieving Algorithm

GS has two phases: GATHERING and ANALYSIS, which are described in Algorithms 1 and 2. Ignoring the optimization step in Algorithm 1 for now, the GATHERING phase uses MA to iteratively compute motif profiles (outer loop of Algorithm 1) for every subset motif of size k (inner loop of Algorithm 1). These motif profiles are passed to the ANALYSIS phase, which calculates the medians of each motif profile, and identifies the subset motif with the highest median LRMSD. This subset motif is returned as the optimized motif.

The GATHERING phase is embarrassingly parallel. Given a set of c processors, we can obtain a $(c-1)$-times linear speedup by offloading the task of calculating each match between the current subset motif S', target T_i pair to another processor. This produces a client/server architecture where the server implements GATHERING, and offloads MA problems to the clients.

Further modifications to GS can increase performance. In particular, let us now consider the optimization procedure ELIMINATION (Algorithm 3) which is called from GATHERING. Note that

Algorithm 1. GATHERING

Input: Input Motif S
Input: Optimized motif size k
for each T_i in Ω_5 **do**
  for all subset motifs S' of size k **do**
    Run MA with S' and T_i
    MA returns match M
    Store M in the motif profile S'_Ω
  end for
  ELIMINATION (optimization step)
end for

Algorithm 2. ANALYSIS

Input: all motif profiles S'_Ω from GATHERING phase

Calculate $m(S'_\Omega)$ for all S'_Ω

Find the motif profile S'_Ω with highest $m(S'_\Omega)$

Output: S', the optimized motif

Algorithm 3. ELIMINATION

Input: all motif profiles S_Ω from GATHERING phase

Calculate $r(S_\Omega)$ for all S'_Ω

Among all $r(S_\Omega)$, find l

eliminate all $r(S'_\Omega)$ with $u < l$

return to GATHERING

when we call ELIMINATION during GATHERING, all motif profiles are only partially computed. Eventually ANALYSIS will identify the optimized motif by selecting the motif profile that has the highest median. A closer look at the computations happening during GATHERING revealed that some motif profiles have medians significantly lower than others. Since we are only interested in the motif profile with the highest median, we can stop computing matches for motif profiles that have significantly lower medians, saving computation time. For this reason, in Algorithm 1, we apply ELIMINATION (see outer loop of Algorithm 1), which determines for which motif profiles we can stop computing matches. These motif profiles will be *eliminated* in the next loop through GATHERING. ELIMINATION need not be applied at every iteration of the outer loop of GATHERING, as it will have a limited effect. Instead, we define a parameter called the *step size* and we call ELIMINATION after *step size* iterations of the outer loop of GATHERING.

As we pointed out above, when we call ELIMINATION during GATHERING (see Algorithm 3), all motif profiles are only partially computed. At this point in the algorithm, comparing the medians of these partial motif profiles can be affected by sampling error. For this reason, ELIMINATION computes a 95% Confidence Interval $r(S''_\Omega)$ (see method of Efron and Tibshirani [66–68]), which has 95% probability of containing the median $m(S'_\Omega)$ of S'_Ω. Therefore, for two partially computed motif profiles S'_Ω, S''_Ω, if $r(S'_\Omega) > r(S''_\Omega)$ do not overlap, there is low probability that $m(S'_\Omega) < m(S''_\Omega)$. Since we are interested only in the motif profile with highest median LRMSD, it is thus unnecessary to finish computing S''_Ω because S'' is not the optimized motif with high probability.

We apply this fact during ELIMINATION by finding l, the highest lower bound of all confidence intervals, and eliminate all subset motifs having confidence intervals with upper bound $u < l$. In the next loop through GATHERING, we do not calculate matches for eliminated subset motifs. If only one subset motif remains, or if GATHERING completes, we proceed to the ANALYSIS phase, which identifies the motif profile that has not been eliminated with that highest median. This is returned as the output of GS.

Occasionally, unusual random samplings of Ω can occur, creating motif profiles with medians that differ dramatically from the true median we intend to estimate. While this occurs very rarely, sampling more and more subset motifs exacerbates a multiple testing situation, which eventually leads to an unusual random sampling. Since we use statistical analyses like ELIMINATION to guide program logic, this can lead to accidental elimination of a subset motif. In order to reduce this possibility, ELIMINATION could be applied in a more adaptive manner, such as by running ELIMINATION less often when motif profiles have few samples. We are investigating this for future work.

5 Experimental Results with Geometric Sieving

In previous work, we demonstrated that MASH can be a useful tool for function prediction, showing that MASH identifies cognate active sites in functional homologs [4, 6, 7]. The experimentation detailed in this paper first demonstrates that GS is a practical and efficient tool for motif optimization. Using input sets derived from 10 well-studied proteins, we show that different subset motifs derived from the same input set produce motif profiles which measurably vary in the median. We also demonstrate that estimating medians with a 95% confidence bound and eliminating subset motifs via ELIMINATE strongly reduces the number of calculations necessary to correctly determine the motif profile with highest median. On our small data set, we made two key observations: First, motifs refined by Geometric Sieving, tested in the MASH pipeline, were highly specific and among the most sensitive of all possible refinements. Second, evolutionary significant subset motifs tend to be more Geometrically Unique than motifs containing evolutionarily insignificant amino acids.

5.1 Primary Data

Input Sets The input sets chosen for this work were taken from ten well-studied proteins, listed in Figure 3. Each input set included between 10 and 13 motif points, and the spatial coordinates used for each were derived from the α-carbons of these amino acids. The precise amino acids used are specified and diagrammed in Figure 4, where the "tag" column identifies the amino acid in the diagram, the "AA" column lists the amino acid type, and "#" specifies the residue number. The ET rank ("Rank") is the degree of evolutionary significance, as reported by ET, where lower values are more evolutionarily significant. Diagrams were generated using Pymol [69].

PDB Code	Protein Name	Organism
1acb	α-Chymotrypsin	Bos Taurus
1rx7	Dihydropholate Reductase	Escherichia coli
3lzt	Lysozyme	Gallus gallus
1czf	Endo-polygalacturonase	Aspergillus Niger
1ep0	Dtdp-4-keto-6-deoxy-d-hexulose 3,5-epimerase	Methanobacterium Thermoautotrophicum
1gwz	Tyrosine Phosphatase SHP-1	Homo Sapiens
1juk	Indole-3-Glycerolphosphate Synthase	Sulfolobus Solfataricus
1kpg	Mycolic Acid Cyclopropane Synthase CMAA1	Mycobacterium Tuberculosis
1nsk	Nucleoside Diphosphate Kinase	Homo Sapiens
1ukr	Endo-1,4-Beta-Xylanase C	Aspergillus Niger

Fig. 3. Proteins used in Experimentation.

Selection Criteria Earlier work has produced examples of motifs designed with evolutionarily significant amino acids [4] and amino acids with documented function [8], which were sensitive and specific. Inspired by these approaches, we selected evolutionarily significant (^{E}, in Figure 4) and functionally documented (^{D}, in Figure 4) amino acids for each of our ten input sets, except Lysozyme (3lzt). Functionally documented amino acids are listed in Figure 5. We also included evolutionarily insignificant amino acids (^{I}, in Figure 4), chosen from the same region of the protein. We chose evolutionarily insignificant amino acids by first generating a sphere centered at the centroid of the evolutionarily significant and functionally documented amino acids. The sphere was sized just large enough to contain these amino acids. From the set of all amino acids having at least one atom within this sphere, the most evolutionarily insignificant amino acids were selected. Occasionally this sphere had to be expanded slightly (no more than 10% increase in radius) when no evolutionarily insignificant amino acids intersected it.

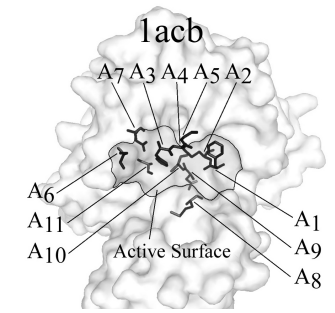
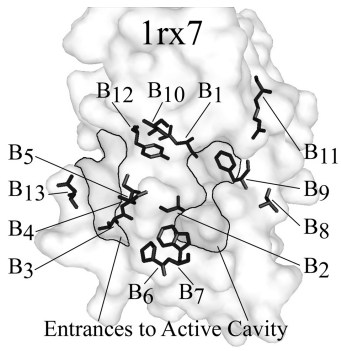
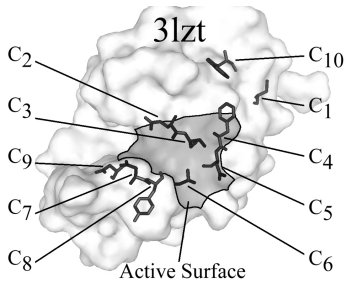
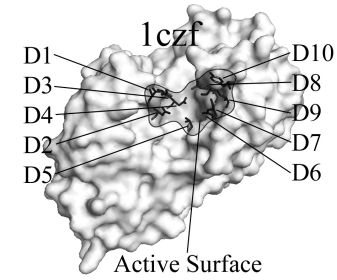
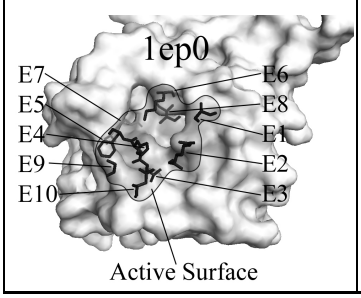
Diagram	tag	AA	#	Rank
	A1	F^I	41	47.91
	A2	C^E	42	3.97
	A3	H^D	57	7.22
	A4	C^E	58	3.97
	A5	G^I	59	38.39
	A6	S^I	96	73.41
	A7	D^D	102	1.90
	A8	M^I	192	29.96
	A9	D^E	194	3.10
	A10	S^D	195	1.93
	A11	S^E	214	2.03
	B1	L^I	4	66.00
	B2	A^E	7	16.00
	B3	V^I	13	63.00
	B4	I^E	14	1.00
	B5	G^D	15	1.00
	B6	P^E	21	27.00
	B7	W^D	22	1.00
	B8	A^I	29	63.00
	B9	F^D	31	34.00
	B10	T^E	46	34.00
	B11	R^E	57	1.00
	B12	Y^E	100	36.00
	B13	D^E	122	3.00
	C1	C^E	6	42.00
	C2	E^E	35	23.00
	C3	S^E	36	1.00
	C4	F^E	38	55.00
	C5	N^E	39	55.00
	C6	A^E	42	31.00
	C7	D^E	52	10.00
	C8	Y^E	53	15.00
	C9	N^E	59	44.00
	C10	W^E	123	42.00
	D1	N^E	178	1.64
	D2	D^D	180	1.00
	D3	D^E	201	1.85
	D4	D^D	202	2.09
	D5	L^I	204	17.69
	D6	H^D	223	5.54
	D7	N^I	253	17.78
	D8	R^D	256	1.61
	D9	K^D	258	1.00
	D10	Y^E	291	1.00
	E1	S^D	53	5.32
	E2	R^D	61	3.71
	E3	L^I	63	14.53
	E4	H^D	64	3.08
	E5	F^I	65	17.47
	E6	K^E	73	1.00
	E7	R^E	90	1.00
	E8	I^I	114	14.60
	E9	G^I	146	19.85
	E10	D^E	172	2.56

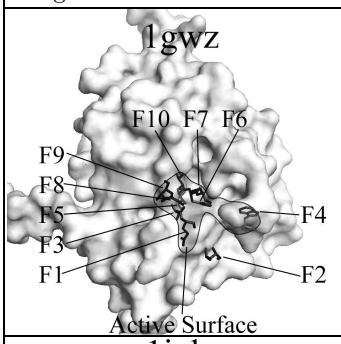
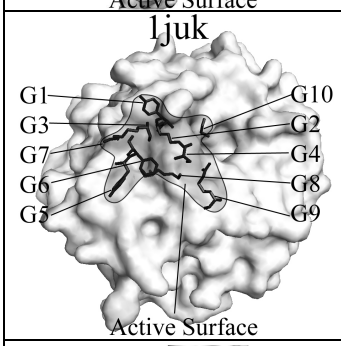
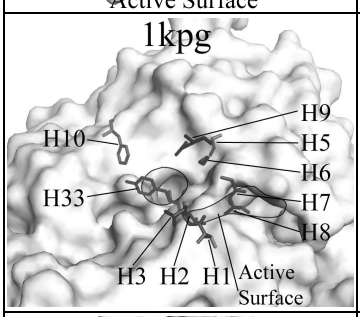
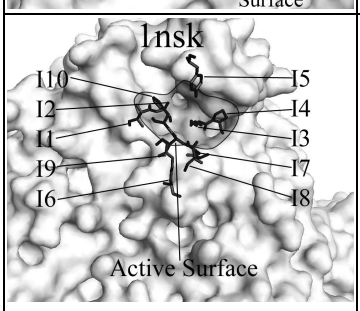
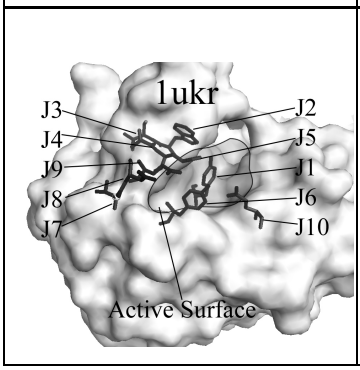
Diagram	tag	AA	#	Rank
	F1	Q^E	327	1.50
	F2	L^I	330	15.10
	F3	S^I	326	11.20
	F4	W^E	367	1.71
	F5	I^I	452	24.69
	F6	H^D	454	2.09
	F7	C^{DE}	455	1.19
	F8	G^E	458	1.00
	F9	I^D	459	11.06
	F10	V^I	453	12.22
	G1	Y^I	52	17.29
	G2	K^D	53	2.43
	G3	K^I	55	11.93
	G4	S^I	58	9.20
	G5	Y^I	88	17.16
	G6	F^E	89	1.04
	G7	G^E	91	1.06
	G8	K^D	110	1.94
	G9	R^D	182	1.91
	G10	G^{DE}	233	1.10
	H1	T^I	30	15.39
	H2	Q^I	31	14.92
	H3	T^I	32	13.66
	H4	Y^D	33	2.20
	H5	G^{DE}	72	1.00
	H6	G^{DE}	74	1.00
	H7	G^E	76	1.00
	H8	A^I	77	16.72
	H9	Q^D	99	2.70
	H10	F^E	200	1.00
	I1	I^I	9	21.28
	I2	A^I	10	21.64
	I3	K^{DE}	12	2.51
	I4	P^E	13	4.16
	I5	Y^D	52	6.57
	I6	R^D	105	3.94
	I7	N^{DE}	115	3.39
	I8	I^I	116	22.74
	I9	I^I	117	19.26
	I10	W^D	118	4.80
	J1	Y^{DE}	70	1.00
	J2	W^{DE}	72	1.00
	J3	V^I	73	10.12
	J4	A^I	78	10.05
	J5	E^{DE}	79	1.00
	J6	Y^{DE}	81	2.21
	J7	T^I	112	16.69
	J8	D^I	113	11.96
	J9	Q^{DE}	129	1.00
	J10	G^{DE}	170	1.79

Fig. 4. Input sets used. "AA": amino acid type; "#": PDB residue number; "Rank": ET rank.

PDB Code	Amino Acids and Citations	EC class	Subset Size (k)
1acb	Ser195 His57 Asp102 [70]	3.4.21.1	7
1rx7	Trp22 [71], and Gly15, Asp27, Phe31, His45, Ile50, Gly96 [72]	1.5.1.3	10
3lzt	Control: Amino acids selected only for Evolutionary Significance.	3.2.1.17	8
1czf	Asp180, Asp202, His223, Arg256, Lys258 [73]	3.2.1.15	6
1ep0	Ser53, Arg61 and His64 [74]	5.1.3.13	6
1gwz	His454, Cys455, Ile459, [75]	3.1.3.48	6
1juk	Lys53, Lys110, Arg182, Gly233 [76]	4.1.1.48	6
1kpg	Gly72, Gly74, GLN99, Tyr33 [77]	2.1.1.79	6
1nsk	Lys12, Tyr52, Arg105, Asn115, His118 [78]	2.7.4.6	6
1ukr	Tyr70, Trp72, Glu79, Tyr81, Gln129, Glu170 [79]	3.2.1.8	6

Fig. 5. Amino acids with documented function (with citations) from each input set. We also provide the EC class this set is derived from, and the size of the subset motifs (k) used when running GS.

Having chosen evolutionarily significant and functionally documented amino acids as part of each input set, we postulated that these "motif-worthy" amino acids, and not the evolutionarily insignificant amino acids, would ultimately result in the most sensitive and specific motifs. For this reason, k, the size of the subset motifs being considered for the optimized motif, was chosen in each case as the total number of evolutionarily significant and functionally documented amino acids in each input set. This guarantees that one subset motif from each input set would contain only evolutionarily significant and functionally documented amino acids. It also guarantees that the other subset motifs will contain all or some of the evolutionarily insignificant amino acids.

As a control, the Lysozyme input set (3lzt) was composed entirely of evolutionarily significant amino acids, to study the effect of having no evolutionarily insignificant amino acids. Conversely, in Endo-polygalacturonase (1czf), there are 8 motif-worthy amino acids, but we chose $k = 6$ to get a broader understanding of the relationship between k and the number of motif-worthy amino acids. For 1gwz, 1juk, 1kpg, 1nsk, and 1ukr, several evolutionarily significant amino acids were also functionally documented (see amino acids labeled ^{DE} in Figure 4).

We will refer to the set of input sets as $\{S_1, S_2, \ldots, S_10\}$, and refer to the subset motifs of each S_i as $S_{i_1}, S_{i_2}, \ldots, S_{i_l}$, where l is the total number of subset motifs for S_i.

Functional Homologs In order to measure sensitivity and specificity, it is essential to fix a set of functional homologs for benchmarking. For this work, we use the functional classification of the Enzyme Commission [80] (EC), which identifies families of functional homologs for each input set used (see Figure 5). Input sets were chosen from distinct EC families. Proteins with PDB structures in each family form the set of functional homologs we search for. Structure fragments, mutants, and structures with artificially induced long distance conformational changes, were removed. We will refer to the set of functional homologs for any input set S_i as $H(S_i)$.

The Protein Data Bank In this paper, we use Ω_5, as mentioned in Section 3.4, which is sampled from the set of crystallographic protein structures in the PDB on Sept 1, 2005. PDB entries with multiple chains were divided into separate structures, producing 79322 structures. While this could prevent the identification of matches to active sites that span multiple chains, it is not clear from the PDB file format how to determine which chains are intended to be in complex. Incorrectly combining chains can lead to searches within physically impossible colliding molecules. Since none of the active sites used in this study span multiple chains, separation was the most reproducible and well defined policy.

Implementation Specifics GS was implemented in C/C++ using the Message Passing Interface [81] (MPI) protocol for interprocess communication, and prototyped on a 16-node dual Athlon 1900MP cluster. Final data was run on the Rice TeraCluster (`http://www.rtc.rice.edu/`), a cluster of 272 800Mhz Intel Itanium2 processors, and on Ada, an experimental 28 chassis Cray XD1 with 672 2.2Ghz AMD Opteron cores. The parameter ϵ, described in Section 2.2 was set to 7Å.

5.2 Median LRMSD Differentiates Motif Profiles

As mentioned in Section 5.1, our input sets were defined on both evolutionarily significant and insignificant amino acids, as well amino acids with documented function. Since GS calculates motif profiles for every possible subset motif, we hypothesized that the diversity of these input sets would present a spectrum of motif profile medians, and that medians within this spectrum would vary sufficiently to justify motif profile comparison by measuring median LRMSD.

Experiment Each of our ten input sets has between 10 and 13 motif points, and a specific k for each input set. GS computed motif profiles for every combination of k motif points in each input set. For example, α-Chymotrypsin and DHFR each contained, respectively, 7 and 10 amino acids which were either evolutionarily significant or functionally documented, out of the 11 and 13 amino acids total. Running GS with $k = 7$ and $k = 10$, respectively, GS exhaustively analyzed all combinations of 7 and 10 (resp.) amino acids as the subset motifs considered. We expected the differences between subset motifs to create a spectrum of median LRMSDs from the motif profiles calculated. The Lysozyme input set, a control composed entirely of evolutionarily significant amino acids, lacked evolutionarily insignificant amino acids. Running with $k = 8$ out of 10 amino acids in the input set, we expected Lysozyme's input set to also lack a broad spectrum of median LRMSDs.

Observations The medians of the motif profiles generated (vertical hashes on the x-axes in Figure 6) from α-Chymotrypsin, DHFR, and Lysozyme, occurred in ranges of .9 LRMSD, .7 LRMSD and .4 LRMSD, respectively. This behavior was typical of the 7 remaining input sets. Motif profiles corresponding to the highest medians clearly had more matches at higher LRMSDs than motif profiles at the lowest medians, and thus higher Geometric Uniqueness. This is demonstrated by darkened hashes and darkened curves in Figure 6, where the biggest differences in medians (darkened hashes) correlated to obvious differences in motif profiles (darkened curves). Differences in medians in α-Chymotrypsin and DHFR were greater than in Lysozyme, which did not contain a spectrum of evolutionarily insignificant and significant amino acids. Higher median LRMSD in this application is clearly directly associated with more and higher match LRMSDs, showing on these examples that medians can be used to measure Geometric Uniqueness.

5.3 Median Estimation Accelerates Performance With Minor Loss of Accuracy

Our implementation of GS uses online estimation of motif profile medians, reducing the number of matches which need to be calculated before the optimized motif is identified. Using input sets from Section 5.2, we first generated matches without using the ELIMINATION optimization, mentioned in Section 4. Next, we repeated this calculation with the ELIMINATION optimization, with step sizes of 100 and 500, to stop sampling on motif profiles which clearly did not have the highest median LRMSD, thereby reducing the number of matches necessary.

Observations Median estimation substantially reduces running time necessary to determine the optimized motif. Using exhaustive sampling, the seven input sets run in Ada took an average of 1556:57:46 (hrs:mins:secs) of distributed computing time to complete, taking 2-3 hours to complete on 600 Opteron cores. Using a step size of 500 matches, these seven sets took an average of 113:31:54,

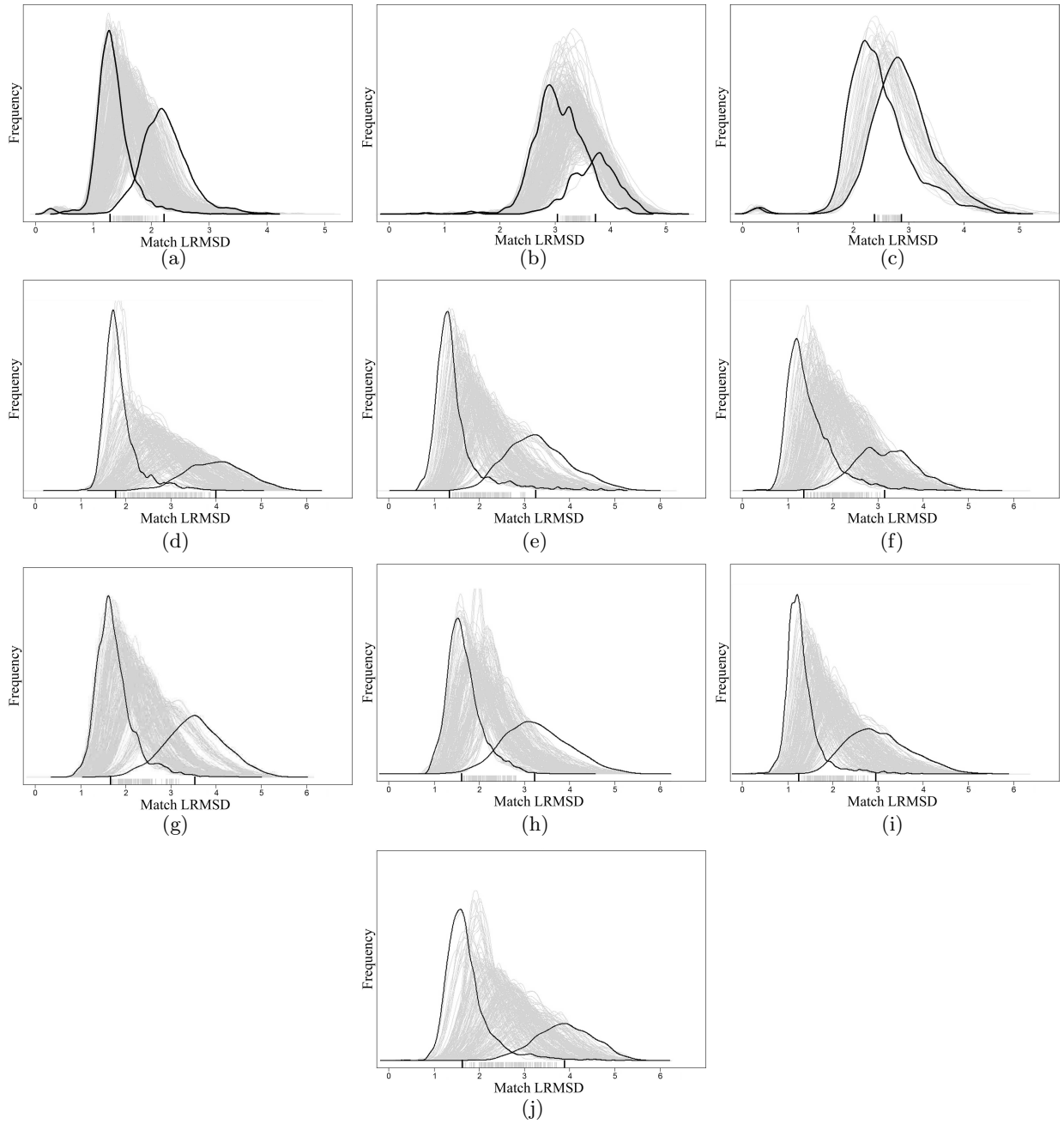


Fig. 6. Motif profile examples from (a) 1acb, (b) 1rx7, (c) 3lzt, (d) 1czf, (e) 1ep0, (f) 1gwz, (g) 1juk, (h) 1kpg, (i) 1nsk, (j) 1ukr. In each picture, the motif profile with highest and lowest median are darkened. These correspond to the rugplot on the horizontal axis, where the darkened hashes plot the highest and lowest median LRMSD.

and at a step size of 100 matches, took an average of only 30:14:31, or about 3 minutes on 600 cores. Similar performance increases occurred for input sets run on the Rice TeraCluster, but relative runtime was longer because of differences in processor speed. GS operating on step sizes of 100 can identify the optimized motif an average of 10 times faster than GS without median estimation.

The reason for this speedup follows directly from the early elimination of motifs which, with high probability, do not have the highest median. This is apparent in the number of matches necessary:

Input Set	Time-Full	Matches-Full	Time-500	Matches-500	Time-100	Matches-100
1acb*	12545:33:20	1,322,230	2683:07:40	186,883	1424:13:20	97,836
1rx7*	10826:50:00	1,211,266	915:20:40	203,356	554:56:40	107,657
3lz7*	1204:52:00	184,395	227:56:00	97,593	942:00:00	92,099
1czf	2678:24:24	1,068,902	156:46:40	179,020	39:43:20	91,107
1ep0	1239:13:20	1,107,251	76:06:40	181,800	25:16:40	76,864
1gwz	1167:40:00	1,109,775	103:26:40	187,627	25:23:20	80,708
1juk	1059:06:40	1,100,452	100:33:20	183,086	22:13:20	87,098
1kpg	1224:53:20	1,092,748	80:26:40	179,721	22:46:40	78,014
1nsk	1499:00:00	1,126,496	127:10:00	177,201	41:00:00	69,145
1ukr	2030:26:40	1,063,797	150:13:20	110,043	35:40:00	74,613

Fig. 7. Computational Speedups from Median Estimation. Here we show the differences, in execution time and number of matches computed, between step sizes of 100, 500, and full sampling. * = These runs were done on the Rice TeraCluster. Remaining runs were done on Ada.

For exhaustive sampling, the ten input sets computed an average of 1,095,631 matches. But at a step size of 500, only 171,214 matches were computed, on average, before determining the motif with the highest median LRMSD. At a step size of 100, an average of 79,649 were computed before finding the optimized motif. GS operating on step sizes of 100 can identify the optimized motif with an average of 10 times less matches than GS without median estimation. Figure 7 describes the precise number of matches and time consumed.

Median estimation is very accurate. In every case described in Figure 7, median estimation identified the same optimized motif as GS using full sampling. However, at step size 100, GS also identifies an alternative subset motif for 3lzt and 1gwz. GS was unable to eliminate the alternative subset motif because overlapping confidence intervals (see Section 4.1) did not separate by the time sampling was complete. The same was true at a step size of 500 for 3lzt, 1gwz, and 1ukr. This suggests that for some motifs, achieving certainty of the optimized motif beyond 95% confidence can require sampling more than 5% of the PDB. Given the large computational advantages of this approach, additional sampling on alternative optimized motifs is only a minor computational cost. Furthermore, the presence of alternative optimized motifs provides additional information to the user, who may consider both of them, in practice. It was particularly interesting that GS identified alternative optimized motifs on the input sets which had either no sensitive and specific subset motifs (1gwz and 1ukr), or were entirely composed of sensitive and specific motifs (3lzt, see Section 5.4). Ultimately, the ability to identify alternative optimized motifs is an advantage in the search for effective motifs, but more careful study is required to understand the circumstances under which alternative optimized motifs occur. Median estimation strongly accelerates the determination of the optimized motif with minor sacrifices in accuracy.

5.4 Optimizing Geometric Uniqueness Improves Motif Effectiveness

GS was designed for the purpose of improving the sensitivity and specificity of motifs by identifying the subset motif with highest median LRMSD, our measure of Geometric Uniqueness. We demonstrate that optimized motifs on our ten input sets are among the most sensitive and specific of all possible motifs definable from the input sets.

Experiment Beginning with each S_i of our input sets $S_1, S_2, \ldots, S_{10}$, we generate all possible subset motifs $S_{i_1}, S_{i_2}, \ldots, S_{i_l}$. We then apply MASH to compute matches and p-values between every subset motif S_{i_j} and every protein structure in $\Omega_5 \cup H(S_i)$.

For any motif S_i, a true positive match is a match to a member of $H(S_i)$ with a p-value below α, our standard for statistical significance. A false positive match is a match with a protein outside

$H(S_i)$, but with p-value less than α. True negative matches are matches to a protein outside $H(S_i)$ with a p-value above α, and false negative matches are matches to a member of $H(S_i)$ with a p-value below α. For every subset motif generated, these values allow us to calculate sensitivity and specificity. Holding α at .02, specificity was always slightly above 98%.

Observations In exhaustive comparison to all possible motifs definable from the input sets at their respective subset sizes, GS identified optimized motifs which, used with the MASH pipeline, were quite sensitive at a high level of specificity (see Figure 8). From each of the 10 input motifs we tested, GS produced 8 optimized motifs with greater sensitivity than the average subset motif from the same input set. 5 of these optimized motifs had perfect sensitivity. Figure 8 demonstrates the spectrum of sensitivity among the subset motifs observed.

We provide maximum and average sensitivity of every subset motif derived from our input sets, as well as the sensitivity of the optimized motif identified by GS, in Figure 8. The two input sets which did not perform well, 1gwz and 1ukr, displayed no subset motifs with high sensitivity. While these input sets were created with the same criteria as the other input sets, it is clear that highly sensitive motifs are not within these two input sets. Overall, GS performed well, identifying optimized motifs among the most sensitive of 8 out of 10 input sets, except where no effective motif could be found.

5.5 Geometric Uniqueness Correlates with Evolutionary Significance

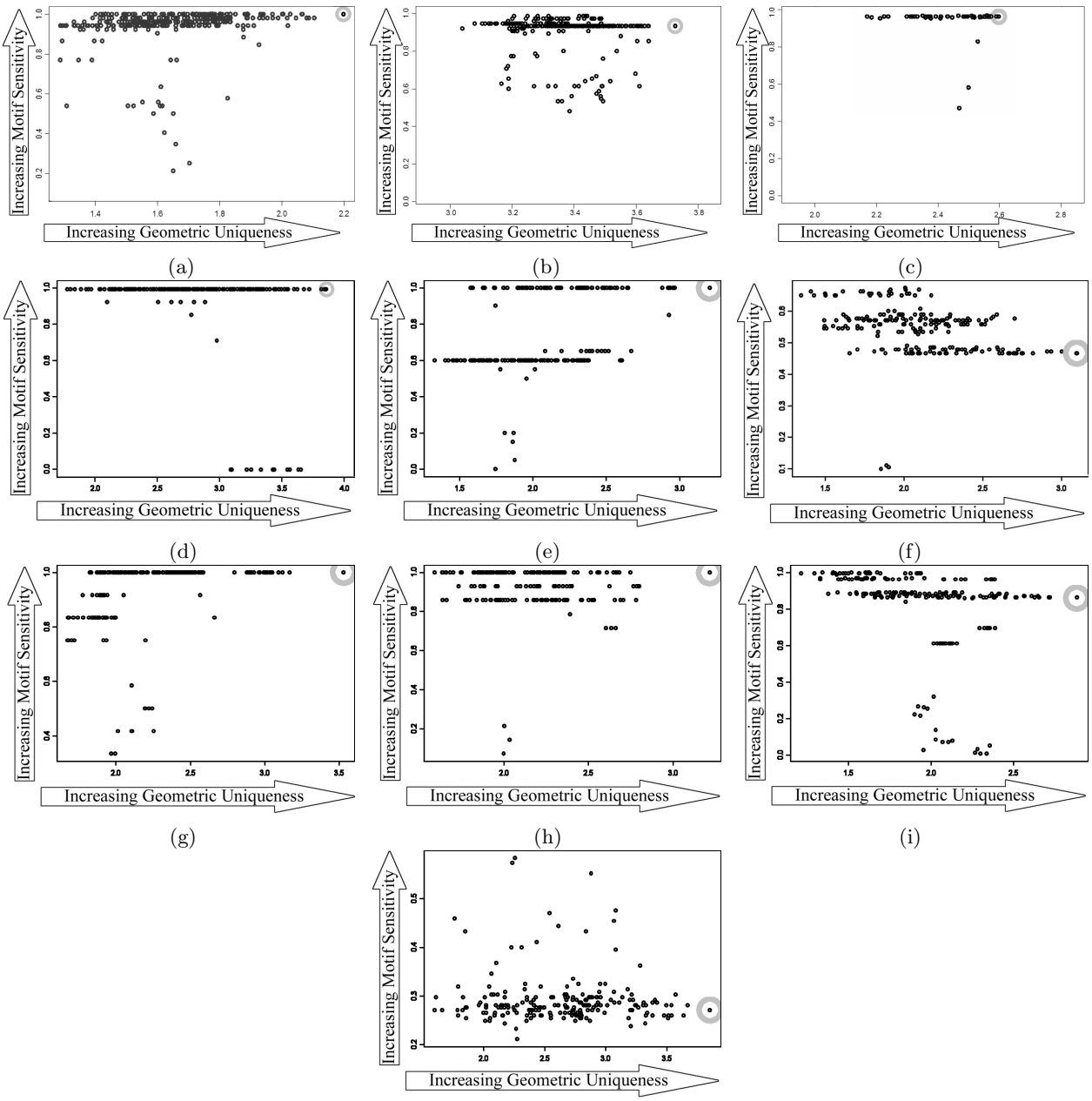
In this section, we investigate if evolutionarily significant amino acids are also structurally dissimilar from all known protein structures, or Geometrically Unique.

Experiment Using the motif profiles calculated over Ω_5, we have a representation of the median LRMSD of every subset motif in our input sets. Since we also have the evolutionary significance of every amino acid in our input sets, we can evaluate the evolutionary significance of every subset motif relative to its Geometric Uniqueness. We represent the total evolutionary significance of a subset motif as the sum of the ET ranks of its elements. Increasing sums relate to decreasing evolutionary significance, displayed on the vertical axis in Figure 9. Median LRMSD was plotted on the horizontal axis.

Observations Motif profiles with the highest median corresponded to the subset motif with the most evolutionarily significant amino acids (grey circles in Figure 9). In all cases but Lysozyme (3lzt), the input sets used demonstrate how evolutionary significance increases proportionately to decreasing median LRMSD. In Lysozyme, a control set where every candidate motif point was evolutionarily significant, no apparent trend is visible. Banding and grouping, apparent in some input sets, seems to be related to the amino acid composition of subset motifs involved. For example, subset motifs with one evolutionarily insignificant amino acid tend to group together, at higher median LRMSDs than subset motifs with two evolutionarily insignificant amino acids. While this is only a small experiment with 10 examples, the existence of this apparent trend suggests that Geometric Uniqueness may be tied to evolutionary conservation.

6 Conclusions

We have presented MASH, an efficient pipeline for identifying statistically significant matches. MASH combines Match Augmentation, an efficient hierarchical geometric and chemical comparison algorithm, with a nonparametric statistical model. As input, MASH accepts 3D motifs labeled with evolutionary information, and, using a snapshot of the PDB, MASH finds geometric matches and measures their statistical significance. Inspired by the success of MASH in preliminary tests for



	1acb	1rx7	3lzt	1czf	1ep0	1gwz	1juk	1kpg	1nsk	1ukr
Max Sensitivity	100.0%	98.7%	96.7%	100.0%	100.0%	67.4%	100.0%	100.0%	100.0%	58.4%
Avg Sensitivity	94.2%	90.4%	93.4%	93.8%	75.5%	51.2%	93.9%	93.4%	81.7%	29.2%
GS Sensitivity	100.0%	93.3%	96.3%	100.0%	100.0%	46.6%	100.0%	100.0%	86.3%	27.0%

Fig. 8. Sensitivity of (a) 1acb, (b) 1rx7, (c) 3lzt, (d) 1czf, (e) 1ep0, (f) 1gwz, (g) 1juk, (h) 1kpg, (i) 1nsk, (j) 1ukr, vs median LRMSD. The table below specifies the sensitivity of the most sensitive subset motif, the average sensitivity of all subset motifs, and the sensitivity of the optimized motif identified by GS. All data represents sensitivity while specificity is held at 98%.

protein function prediction and the efficiency of the overall pipeline, we developed GS, a novel distributed algorithm for exhaustively refining input sets of candidate motif points into optimized motifs that can be used in MASH. We have implemented GS with techniques and optimizations

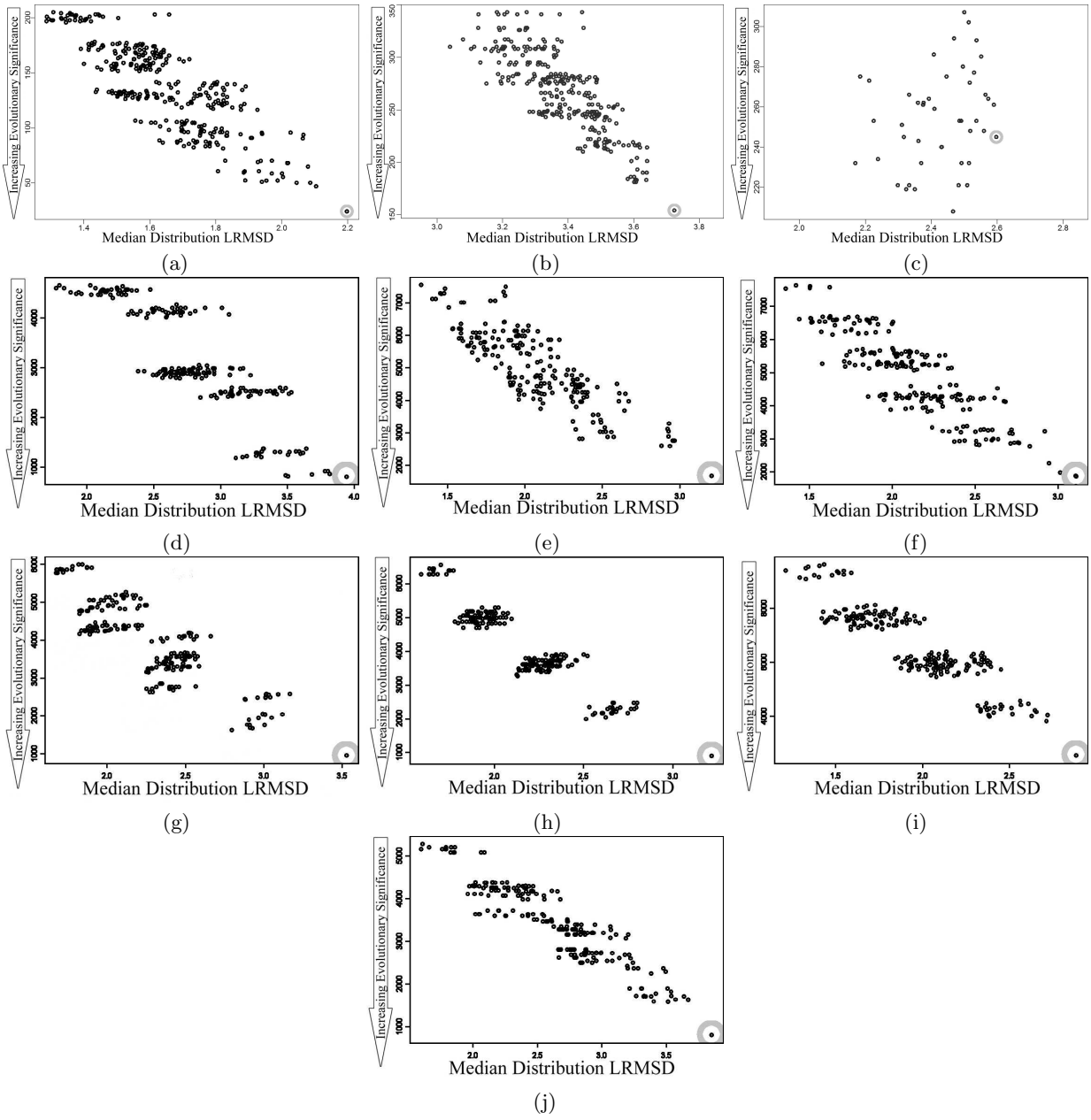


Fig. 9. Relationship of Geometric Uniqueness to evolutionary significance in (a) 1acb, (b) 1rx7, (c) 3lzt, (d) 1czf, (e) 1ep0, (f) 1gwz, (g) 1juk, (h) 1kpg, (i) 1nsk, (j) 1ukr.

suitable for large scale distributed systems, testing it successfully on a cluster with more than 600 CPUs. We demonstrated the refinement of ten well studied input sets using GS. Using the MASH pipeline, these optimized motifs functioned at a very high level of specificity and were among the most sensitive of all motifs definable from these input sets. In addition, using GS in conjunction with the Evolutionary Trace permitted us to demonstrate examples where amino acids that are evolutionarily significant are also Geometrically Unique. Our current observations show that GS is a powerful motif refinement algorithm which can be used in conjunction with other motif design techniques in an effort to create sensitive and specific motifs. GS can thus be used

as an improvement for MASH and other pipelines in the form of a preprocessing step. In the future, we hope to accomplish larger-scale investigations to help clarify the problem of selecting the appropriate motif size, which remains an open problem, and also to understand how Geometric Uniqueness can be combined with other motif design principles to produce more effective motifs.

Acknowledgements

This work is supported by a grant from the National Science Foundation NSF DBI-0318415. Additional support is gratefully acknowledged from training fellowships the Gulf Coast Consortia (NLM Grant No. 5T15LM07093) to B.C. and D.K.; from March of Dimes Grant FY03-93 to O.L.; from a Whitaker Biomedical Engineering Grant and 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. B.D. has been partially supported by the Rice Century Scholar Program and by the W.M. Keck Center. The authors are exceptionally grateful for the assistance of Anand P. Dharan, Colleen Kenney, Amanda Cruess and Jessica Wu.

References

1. Lamdan Y. and Wolfson H.J. Geometric hashing: A general and efficient model based recognition scheme. *Proc. IEEE Conf. Comp. Vis.*, pages 238–249, Dec 1988.
2. 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.
3. 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.
4. 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.
5. Stark A., Sunyaev S., and Russell RB. A model for statistical significance of local similarities in structure. *J. Mol. Biol.*, 326:1307–1316, 2003.
6. 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.
7. Chen B.Y., Bryant D.H, Fofanov V.Y., Kristensen D.M., Cruess A.E., Kimmel M., Lichtarge O., and Kavraki L.E. Cavity-aware motifs reduce false positives in protein function prediction. *Proceedings of the 2006 IEEE Computational Systems Bioinformatics Conference (CSB 2006), accepted*, August 2006.
8. 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.
9. Porter C.T., Bartlett G.J., and Thornton J.M. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32:D129–D133, 2004.
10. 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.
11. 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.
12. Shatsky M., Shulman-Peleg A., Nussinov R., and Wolfson H.J. The multiple common point set problem and its application to molecule binding pattern detection. *J. Comp. Biol.*, 13(2):407–28, 2006.
13. Yao H., Kristensen D.M., Mihalek I., Sowa M.E., Shaw C., Kimmel M., Kavraki L., and Lichtarge O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, 326:255–261, 2003.
14. 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.
15. 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.
16. Sowa M.E., He W., Slep K.C., Kercher M.A., Lichtarge O., and Wensel T.G. Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat. Struct. Biol.*, 8:234–237, 2001.
17. Lichtarge O. and Sowa M.E. Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.*, 12(1):21–27, 2002.

18. Lichtarge O., Sowa M.E., and Philippi A. Evolutionary traces of functional surfaces along g protein signaling pathway. *Meth. Enzymol.*, 344:536–556, 2002.
19. Madabushi S., Yao H., Marsh M., Kristensen D.M., Philippi A., Sowa M.E., and Lichtarge O. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, 316:139–154, 2002.
20. Verbitsky G., Nussinov R., and Wolfson H.J. Structural comparison allowing hinge bending. *Prot: Struct. Funct. Genet.*, 34(2):232–254, 1999.
21. 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.
22. Wallace A.C., Borkakoti N., and Thornton J.M. 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.
23. 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.
24. 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.
25. Norel R., Fischer D., Wolfson H.J., and Nussinov R. Molecular surface recognition by a computer vision-based technique. *Prot. Eng.*, 7:39–46, 1994.
26. Norel R., Petrey D., Wolfson H.J., and Nussinov R. Examination of shape complementarity in docking of unbound proteins. *Prot: Struct. Funct. Genet.*, 36:307–317, 1999.
27. 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.
28. Connolly M.L. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221:709–713, 1983.
29. Ferré F., Ausiello G, Zanzoni A, and Helmer-Citterich M. Surface: a database of protein surface regions for functional annotation. *Nucl. Acid. Res.*, 32:D240–4, 2004.
30. Rhodes N., Clark D.E., and Willett P. Similarity searching in databases of flexible 3D structures using autocorrelation vectors derived from smoothed bounded distance matrices. *J Chem Inf Model.*, 46(2):615–9, 2006.
31. Holm L. and Sander C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233: 123–138, 1990.
32. Grindley H.M., Artymiuk P.J., Rice D.W., and Willett P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.*, 229:707–21, 1993.
33. Brint A.T., Davies H.M., Mitchell E.M., and Willett P. Rapid geometric searching in protein structure. *J. of Mol. Graph.*, 9:48–53, 1989.
34. 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.
35. Mihalek I., Res I., and Lichtarge O. A family of evolution-entropy hybrid methods for ranking of protein residues by importance. *J. Mol. Biol.*, 336(5):1265–82, 2004.
36. Chen B.Y., Fofanov V.Y., Bryant D.H., Dodson B.D., Kristensen D.M., Lisewski A.M., Kimmel M., Lichtarge O., and Kavraki L.E. 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.
37. 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.
38. 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, Dec 1992.
39. Smart O.S., Goodfellow J.M., and Wallace B.A. The pore dimensions of gramacidin a. *Biophysics Journal*, 65: 2455–2460, 1993.
40. Williams M.A., Goodfellow J.M., and Thornton J.M. Buried waters and internal cavities in monomeric proteins. *Protein Science*, 3:1224–35, 1994.
41. Edelsbrunner H. and Mucke E.P. Three-dimensional alpha shapes. *ACM Trans. Graphics*, 13:43–72, 1994.
42. Liang J. Edelsbrunner H., Facello M. On the definition and the construction of pockets in macromolecules. *Discrete Applied Mathematics*, 88:83–102, 1998.
43. Laskowski R.A. SURFNET: A program for a program for visualizing molecular surfaces, cavities, and intramolecular interactions. *Journal Molecular Graphics*, 13:321–330, 1995.
44. Glaser F, Morris R.J., Najmanovich R.J., Laskowski R.A., and Thornton J.M. A method for localizing ligand binding pockets in protein structures. *Proteins*, 62(2):479–88, 2006.
45. Wolfson H.J. and Rigoutsos I. Geometric hashing: An overview. *IEEE Comp. Sci. Eng.*, 4(4):10–21, Oct 1997.
46. Leibowitz N., Nussinov R., and Wolfson H.J. MUSTA a general efficient automated method for multiple structure alignment and detection of common motifs. *J.Comp.Biol*, 8:93–121, 2001.

47. Leibowitz N., Fligelman Z.Y., Nussinov R., and Wolfson H.J. Automated multiple structure alignment and detection of a common substructural motif. *Prot: Struct. Func. Genet.*, 43:235–245, 2001.
48. Shatsky M., Nussinov R., and Wolfson H.J. A method for simultaneous alignment of multiple protein structures. *Proteins*, 56(1):143–56, 2004.
49. Russell R.B. Detection of protein three-dimensional side chain patterns. new examples of convergent evolution. *J. Mol. Biol.*, 279:1211–27, 1998.
50. 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.
51. 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, Sept 2000.
52. Murzin A.G., Brenner S.E., Hubbard T., and Chothia C. Scop: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*, 247:536–540, 1995.
53. Orengo C.A., Michie A.D., Jones S., Jones D.T., Swindells M.B., and Thornton J.M. Cath- a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.
54. Binkowski T.A., Freeman P., and Liang J. pvsoar: Detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucl. Acid. Res.*, 32:W555–8, 2004.
55. Laskowski R.A., Watson J.D., and Thornton J.M. Profunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, 33:W89–93, 2005.
56. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., and Lipman D.J. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids. Res.*, 25(17):3389–3402, Sept 1997.
57. Zdobnov E.M. and Apweiler R. Interproscan: an integration platform for the signature-recognition methods in inter pro. *Bioinformatics*, 17:847848, 2001.
58. Krissinel E. and Henrick K. Protein structure comparison in 3D based on secondary structure matching (ssm) followed by ca alignment, scored by a new structural similarity function. *Kungl,A.J. and Kungl,P.J. (eds), Proceedings of the 5th International Conference on Molecular Structural Biology*, page 88, 2003.
59. Diestel R. *Graph Theory*. Springer, New York, USA, 1997.
60. Freidman J.H., Bentley J.L., and Finkel R.A. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. on Mathematical Software*, 3(3):209–226, 1977.
61. de Berg M., van Kreveld M., and Overmars M.H. *Computational Geometry: Algorithms and Applications*. Springer, Berlin, Germany, 1997.
62. Silverman B.W. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall: London, 1986.
63. Jones M.C., Marron J.S., and Sheather S.J. A brief survey of bandwidth selection for density estimation. *J. Amer. Stat. Assoc.*, 91:401–407, Mar 1996.
64. Sheather S.J. and Jones M.C. A reliable data-based bandwidth selections method for kernel density estimation. *J. Roy. Stat. Soc.*, 53(3):683–690, 1991.
65. Cassella G. and Berger R.L. *Statistical Inference*. Brooks/Cole Publishing Co., New York, USA, 1990.
66. Efron B. and Tibshirani R. The bootstrap method for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):1–35, 1986.
67. Efron B. Better bootstrap confidence intervals (with discussion). *J. Amer. Stat. Assoc.*, 82:171, 1987.
68. Efron B. and Tibshirani R.J. *An Introduction to the Bootstrap*. Chappman & Hall, London, 1993.
69. Delano W.L. The PyMol molecular graphics system (2002), on world wide web: *http : //www.pymol.org*, 2002.
70. Blow D.M., Birktoft J.J., and Hartley B.S. Role of a buried acid group in the mechanism of action of chymotrypsin. *Nature*, 221(178):337–40, Jan 1969.
71. Reyes V.M., Sawaya M.R., Brown K.A., and Kraut J. Isomorphous crystal structures of *Escherichia coli* dihydrofolate reductase complexed with folate, 5-deazafolate, and 5,10-dideazatetrahydrofolate: mechanistic implications. *Biochemistry*, 34:2710–2723, 1995.
72. Bystroff C., Oatley S.J., and Kraut J. Crystal structures of *Escherichia coli* dihydrofolate reductase: the $nadp^+$ holoenzyme and the folate-$nadp^+$ ternary complex. substrate binding and a model for the trasition state. *Biochemistry*, 29:3263–3277, 1990.
73. van Santen Y., Benen J.A., Schroter K.H., Kalk K.H., Armand S., Visser J., and Dijkstra B.W. 1.68-a crystal structure of endopolygalacturonase ii from aspergillus niger and identification of active site residues by site-directed mutagenesis. *J. Biol. Chem.*, 274(43):30474–30480, Oct 1999.
74. Christendat D., Saridakis V., Dharamsi A., Bochkarev A., Pai E.F., Arrowsmith C.H., and Edwards A.M. Crystal structure of dtdp-4-keto-6-deoxy-d-hexulose 3,5-epimerase from methanobacterium thermoautotrophicum complexed with dtdp. *J. Biol. Chem.*, 275:24608–24612, 1999.
75. Yang J., Liu L., He D., Song X., Liang X., Zhao Z.J., and Zhou G.W. Crystal structure of the catalytic domain of protein-tyrosine phosphatase shp-1. *J. Biol. Chem.*, 273:28199–28207, 1999.
76. Knochel T.R., Hennig M., Merz A., Darimont B., Kirschner K., and Jansonius J.N. The crystal structure of indole-3-glycerol phosphate synthase from the hyperthermophilic archaeon sulfolobus solfataricus in three different crystal forms: effects of ionic strength. *J. Mol. Biol.*, 262:502–515, 1996.

77. Huang C.C., Smith C.V., Glickman M.S., Jacobs W.R. Jr., and Sacchettini J.C. Crystal structures of mycolic acid cyclopropane synthases from mycobacterium tuberculosis. *J. Biol. Chem.*, 277:11559–11569, 2002.
78. Webb P.A., O. Perisic, Mendola C.E., Backer J.M., and R.L. Williams. The crystal structure of a human nucleoside diphosphate kinase, nm23-h2. *J. Mol. Biol.*, 251:574–587, 1995.
79. Krengel U. and Dijkstra B.W. Three-dimensional structure of endo-1,4-beta-xylanase i from aspergillus niger: Molecular basis for its low ph optimum. *J. Mol. Biol.*, 263:70–78, 1996.
80. International Union of Biochemistry. Nomenclature Committee. *Enzyme Nomenclature*. Academic Press: San Diego, California, 1992.
81. Snir M. and Gropp W. *MPI: The Complete Reference (2nd Edition)*. The MIT Press, 1998.