

Geometric Sieving: Automated Distributed Optimization of 3D Motifs for Protein Function Prediction

Brian Y. Chen^{1,*}, Viacheslav Y. Fofanov^{2,*}, Drew H. Bryant⁵,
Bradley D. Dodson¹, David M. Kristensen^{3,4}, Andreas M. Lisewski⁴,
Marek Kimmel², Olivier Lichtarge^{3,4}, and Lydia E. Kavraki^{1,3,5,**}

¹ Department of Computer Science, Rice University, Houston, TX 77005, USA

² Department of Statistics, Rice University

³ Structural and Computational Biology and Molecular Biophysics,
Baylor College of Medicine, Houston, TX 77005, USA

⁴ Department of Molecular and Human Genetics, Baylor College of Medicine

⁵ Department of Bioengineering, Rice University

kavraki@cs.rice.edu

Abstract. Determining the function of all proteins is a recurring theme in modern biology and medicine, but the sheer number of proteins makes experimental approaches impractical. For this reason, current efforts have considered in silico function prediction in order to guide and accelerate the function determination process. One approach to predicting protein function is to search functionally uncharacterized protein structures (*targets*), for substructures with geometric and chemical similarity (*matches*), to known active sites (*motifs*). Finding a match can imply that the target has an active site similar to the motif, suggesting functional homology.

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. The paper discusses both the usefulness and the efficiency of GS. We show that candidate motifs from six 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.

* Equal Contribution.

** Corresponding author.

Our current observations show that GS is a powerful tool that can complement existing work on motif design and protein function prediction.

1 Introduction

The determination of protein function is an important goal in biology, but experimental techniques for determining function are expensive and time consuming. One way to accelerate this process is to use computational techniques to search the structure of functionally uncharacterized proteins (*targets*), for *matches* of geometric and chemical similarity to known functional sites (*motifs*). To achieve this, algorithms like Geometric Hashing [1], JESS [2], and Match Augmentation [3] identify a subset of a target with the greatest geometric and chemical similarity to the motif. Typically, geometric similarity is measured by least root mean squared distance (LRMSD¹) and chemical similarity is ensured by examining the chemical compatibility of corresponding matches. The identification of a match with statistically significant LRMSD can suggest that the target and motif have similar function [2, 3, 4].

Designing effective motifs is a two-sided open problem: The geometric configuration and chemical makeup of effective motifs must be similar to functionally related proteins (*sensitive*), as well as dissimilar to functionally unrelated proteins (*specific*). For this reason, it is difficult to select *motif points*, the points in space with chemical labels which comprise motifs, so that sensitivity and specificity are simultaneously maximized. Many methods for designing motifs exist, and we are only able to include a partial list here. Motifs have been designed using evolutionary significance and proximity to binding sites [5]. Motifs have also been designed using literature search and PSI-BLAST alignments of literature-defined motifs from the Catalytic Site Atlas [6, 7]. Still other motifs are designed using surface exposure, and algorithms for detecting conserved binding patterns [8]. The work presented in this paper complements these methods with a novel criteria for motif design and an algorithm that can be used to further improve existing motifs.

Contributions and Outline. We begin by describing the design and implementation of *Geometric Sieving* (GS), an algorithm for refining candidate motifs into *optimized motifs*. 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*.

The motivation and inspiration for defining Geometric Uniqueness stems from several observations in our earlier work [3, 5] and the work of other researchers [2, 4], where it has been observed that motifs which are highly representative

¹ LRMSD is the root mean square distance (RMSD) between two sets of points in 3D, aligned with smallest RMSD.

of protein function do not occur in a large fraction of the known proteins. One question that we posed is whether geometric and chemical dissimilarity of a motif to all other known proteins (a.k.a. Geometric Uniqueness) can be computed in a reasonable amount of time and whether Geometric Uniqueness can be used to identify sensitive and specific motifs. After we obtained a positive answer to the above question for a limited but well-designed set of experiments, we proceeded to investigate a second question which is whether Geometric Uniqueness correlates with other characteristics of active sites. For example, evolutionarily significant amino acids, those most associated with important evolutionary divergences, as defined in [9, 10], are often related to active sites [5]. We observed, on our limited set of examples, a correlation between Geometric Uniqueness and evolutionary significance.

Measuring and optimizing Geometric Uniqueness is a nontrivial computational problem because numerous structural comparisons must be made between many motifs and many protein structures. In Section 2, we present recent advances in the field of motif comparison algorithms that enabled the development of GS. In Section 3, we detail the GS algorithm, a distributed algorithm coupled with on-line statistical optimization, which measures Geometric Uniqueness to optimize motifs. Our experimental results are shown in Section 4. Targeting our first question, we optimized input sets derived from six well-studied proteins. On these examples, optimized motifs computed by GS had among the highest sensitivity and specificity of every subset motif definable from the input sets. Using information from the Evolutionary Trace (ET) [5, 9] we observed, on our examples, that evolutionarily significant motifs exhibited higher Geometric Uniqueness.

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 the milestone algorithm MultiBind [8]. 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

Motif Types. The many approaches to designing effective motifs have created different types of motifs: motifs have been composed of points on the Connolly surface [11] representing electrostatic potentials [12], of hinge-bending sets of points in space [13], of sets of “pseudo-centers” representing protein-ligand interactions [8], or of points taken from atom coordinates with evolutionary data [3, 9], to name a few. Depending on how motif points are defined, they have different labels associated with them and these labels need to be taken into account when comparing motifs. GS is orthogonal to the choice of motif type and could be applied with any of the motif types above.

In this work, a motif S is a set of m points $\{s_1, \dots, s_m\}$ 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* $p(s_i)$, a measure of the functional significance of the motif point. Each s_i also has a set of alternate amino acid *labels* $l(s_i) \subset \{GLY, ALA, \dots\}$, which represent residues this amino acid has mutated to 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 ET [9, 10].

Motif Comparison Algorithms. GS requires a geometric and chemical comparison algorithm to compare motifs to targets. Many such algorithms exist, but differ fundamentally in that they are optimized for comparing different types of motifs. There are algorithms for comparing graph-based motifs [14], algorithms for finding catalytic sites [2], and the seminal Geometric Hashing framework [1] which can search for many types of motifs, including motifs based on atom position [15], points on Connolly face centers [16], catalytic triads [17], and flexible protein models [13]. The comparison algorithm we use in this work is Match Augmentation (MA) [3], because of its availability and compatibility with our selected motif type. GS is independent of MA, and adapting another comparison algorithm to use our motifs could be equally successful.

MA compares a motif S to a target T , a protein structure encoded as n *target points*: $T = \{t_1, \dots, t_n\}$, where each t_i is taken from atom coordinates, and labeled $l(t_i)$ for the amino acid t_i belongs to. A match M , is a bijection correlating all motif points in S to a subset of T of the form $M = \{(s_{a_1}, t_{b_1}), (s_{a_2}, t_{b_2}) \dots (s_{a_m}, t_{b_m})\}$. 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 biologically compatible: $l(t_{b_i}) \in l(s_{a_i})$.

Criterion 2. LRMSD alignment, via rigid transformation A of S , causes

$$\forall i, \|A(s_{a_i}) - t_{b_i}\| < \epsilon, \text{ our threshold for geometric similarity.}$$

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 by ϵ , even if the smallest LRMSD is not very low. This allows us to generate a spectrum of matches ranging from high to low geometric and chemical similarity, which we refer to as a motif profile.

Obtaining Motif Profiles. The basic object of comparison used by GS is the *motif profile*, a set of matches S_Ω between a single motif S and a very large set of targets, Ω . We compute these matches with MA. Motif profiles are best visualized as frequency distributions (see Figure 1(a)), which are essentially histograms that plot frequency (the number of matches with a particular LRMSD) versus LRMSD. We apply kernel density estimation procedures [18] to estimate population density from the motif profile, using Gaussian Kernel smoothing to interpo-

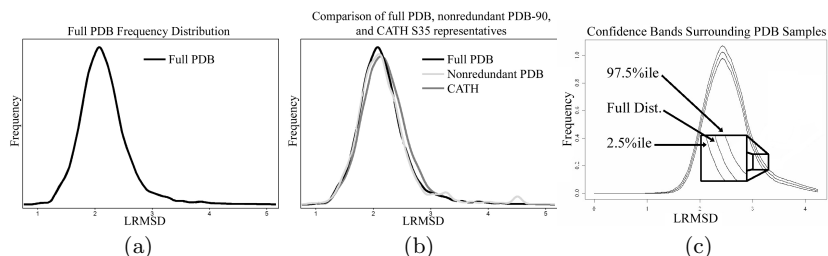


Fig. 1. (a) Typical frequency distribution of matches between a motif and the PDB [21]. (b) Comparison of PDB, sequentially nonredundant PDB, and CATH representatives. (c) Confidence band demonstrating the accuracy of samples of the PDB. This data computed using the motif C42, H57, C58, D102, D194, S195, S214 from α -Chymotrypsin (1acb).

late between data points, as in previous work [3]. Optimal bin-widths determined by Sheather-Jones method [19, 20] were used to avoid under- and over-smoothing.

The purpose of Ω is to represent the set of all known protein structures. We have found, however, that different representations of Ω tend not to have significant effect on the actual shape of motif profiles generated. For the 6 motifs optimized for this work, as well as 12 motifs used in previous work [3], we observed strong similarity between motif profiles calculated with the PDB (Ω_0), and Ω_{nr25} and Ω_{nr90} , 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 [22] database, a multi-level nested categorization of increasingly specific protein sequence and structure classifications. We selected a representative of every category at the three most specific levels: Topologies (Ω_T), Homologous Superfamilies (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 1(b). Ω_{nr90} , Ω_T , and Ω_H were excluded for clarity, but are closely related.

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 1(c), where we overlaid 5000 distinct S_{Ω_5} samples with a single S_{Ω_0} , the center line in Figure 1(c). 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 . This is a result of sampling a largely unimodal distribution.

GS is not dependent on the selection of Ω , but because our observations suggest that motif profiles based on many logical representations of Ω , including Ω_S , Ω_H , Ω_T , Ω_{nr25} , and 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. Other investigations could use alternate selections of Ω .

Motif profiles are especially useful for determining the statistical significance of matches with a given motif S . In previous work, we showed that nonparametric density estimation of motif profiles generated with S can be used to calculate p -values, which measure the statistical significance of any match of S [3]. Matches with low p -values, which correspond to high statistical significance, seem to correlate with functional similarity [3]. This result corroborates earlier work which applied parametric approaches [2, 4] to generate other measures of statistical significance which also correlate with functional homology.

3 Geometric Sieving

GS accepts an input set, a collection of candidate motif points which could be selected by another motif design method, 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 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 [3]. 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. Geometric Sieving 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. Confidence bounds about the median, an integral part of our approach, are better studied than con-

fidence bounds about the mode. Finally, 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 [3]. 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.

3.1 The Geometric Sieving Algorithm

GS has two phases: GATHERING and ANALYSIS, which are described in Algorithms A1 and A2. Ignoring the ELIMINATION step (* in Algorithm A1) for now, the GATHERING phase uses MA to iteratively compute motif profiles (outer loop of Algorithm A1) for every subset motif of size k (inner loop of Algorithm A1). 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.

A1 GATHERING	A2 ANALYSIS	A3 ELIMINATION
Input: Input Motif S Input: Desired 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 profile S'_Ω end for ELIMINATION* end for	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.	Input: all motif profiles S_Ω from GATHERING phase $\forall S'_\Omega$, compute $r(S_\Omega)$ $\forall r(S_\Omega)$, find l eliminate all $r(S'_\Omega)$ with $u < l$

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 A3) which is called from GATHERING. Note that 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 A1, we apply ELIMINATION (see outer loop of Algorithm A1), 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 A3), 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 [23, 24, 25]), 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 the highest median. This is returned as the output of GS.

4 Experimental Results

We begin our experimentation by demonstrating that GS is a practical and efficient tool for motif optimization. Using input sets derived from 6 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 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, sensitive and specific optimized motifs can be identified by Geometric Uniqueness. Second, evolutionary significant subset motifs tend to be more Geometrically Unique than evolutionarily insignificant amino acids. Full details can be found at:

<http://www.cs.rice.edu/~briany/papers/RECOMB2006/>.

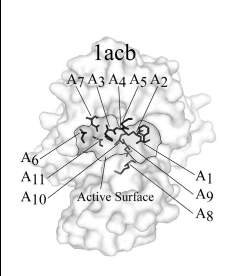
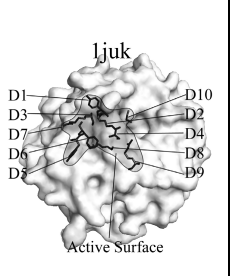
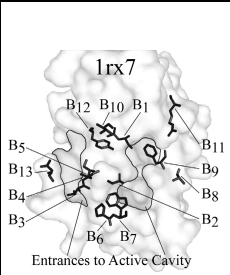
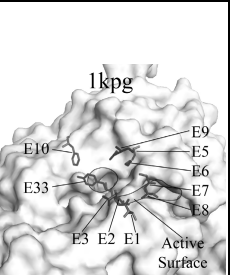
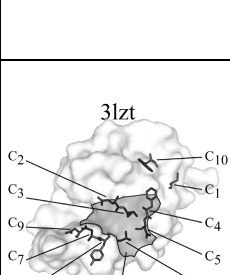
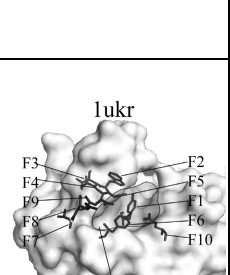
Diagram	tag	AA	#	Rank	Diagram	tag	AA	#	Rank	
 <p>1acb</p> <p>A7 A3 A4 A5 A2</p> <p>A6 A11 A10 Active Surface A1 A9 A8</p>	A1	F ^I	41	47.9	 <p>1juk</p> <p>D1 D3 D7 D6 D8 D9 D10 D2 D4 D5</p> <p>Active Surface</p>	D1	Y ^I	52	17.2	
	A2	C ^E	42	3.97		D2	K ^D	53	2.4	
	A3	H ^D	57	7.22		D3	K ^I	55	11.9	
	A4	C ^E	58	3.97		D4	S ^I	58	9.2	
	A5	G ^I	59	38.3		D5	Y ^I	88	17.1	
	A6	S ^I	96	73.4		D6	F ^E	89	1.0	
	A7	D ^D	102	1.90		D7	G ^E	91	1.0	
	A8	M ^I	192	29.9		D8	K ^D	110	1.9	
	A9	D ^E	194	3.10		D9	R ^D	182	1.9	
	A10	S ^D	195	1.93		D10	G ^D	233	1.1	
	A11	S ^E	214	2.03						
 <p>1rx7</p> <p>B12 B10 B1</p> <p>B5 B11 B9 B8 B4 B3 B6 B7</p> <p>Entrances to Active Cavity</p>	B1	L ^I	4	66.0	 <p>1kpg</p> <p>E9 E5 E6 E7 E8 E3 E2 E1</p> <p>Active Surface</p>	E1	T ^I	30	15.3	
	B2	A ^E	7	16.0		E2	Q ^I	31	14.9	
	B3	V ^I	13	63.0		E3	T ^I	32	13.6	
	B4	I ^E	14	1.00		E4	Y ^D	33	2.20	
	B5	G ^D	15	1.00		E5	G ^D	72	1.00	
	B6	P ^E	21	27.0		E6	G ^D	74	1.00	
	B7	W ^D	22	1.00		E7	G ^E	76	1.00	
	B8	A ^I	29	63.0		E8	A ^I	77	16.7	
	B9	F ^D	31	34.0		E9	Q ^D	99	2.70	
	B10	T ^E	46	34.0		E10	F ^E	200	1.00	
	B11	R ^E	57	1.00						
	B12	Y ^E	100	36.0						
	B13	D ^E	122	3.00						
 <p>3lzt</p> <p>C2 C10 C3 C1 C9 C4 C7 C5 C8 C6</p> <p>Active Surface</p>	C1	C ^E	6	42.0	 <p>1ukr</p> <p>F3 F4 F9 F8 F7 F2 F5 F1 F6 F10</p> <p>Active Surface</p>	F1	Y ^D	70	1.00	
	C2	E ^E	35	23.0		F2	W ^D	72	1.00	
	C3	S ^E	36	1.00		F3	V ^I	73	10.1	
	C4	F ^E	38	55.0		F4	A ^I	78	10.0	
	C5	N ^E	39	55.0		F5	E ^E	79	1.00	
	C6	A ^E	42	31.0		F6	Y ^D	81	2.21	
	C7	D ^E	52	10.0		F7	T ^I	112	16.6	
	C8	Y ^E	53	15.0		F8	D ^I	113	11.9	
	C9	N ^E	59	44.0		F9	Q ^D	129	1.00	
	C10	W ^E	123	42.0		F10	G ^D	170	1.79	

Fig. 2. Input sets used. “AA”: amino acid type; “#”: PDB residue number; “Rank”: ET rank.

4.1 Primary Data

Input Sets. Earlier work has produced examples of motifs designed with evolutionarily significant amino acids [3] and amino acids with documented function [6], which were sensitive and specific. Inspired by these approaches, we selected evolutionarily significant (^E, Figure 3) and functionally documented (^D, Figure 3) amino acids for each of our six input sets, except Lysozyme (3lzt). We also included evolutionarily insignificant amino acids (^I, Figure 3), chosen from the same region of the protein.

PDB Code	Amino Acids and Citations	EC class	size	k
1acb	S195 H57 D102 [26]	3.4.21.1	11	7
1rx7	W22 [27], G15, D27, F31, H45, I50, G96 [28]	1.5.1.3	13	10
3lzt	Control	3.2.1.17	10	8
1juk	Lys53, Lys110, Arg182, Gly233 [29]	4.1.1.48	10	6
1kpg	G72, G74, Q99, Y33 [30]	2.1.1.79	10	6
1ukr	Y70, W72, E79, Y81, Q129, E170 [31]	3.2.1.8	10	6

Fig. 3. Functionally documented amino acids used in our input sets (cited), with protein EC class, input set size (“size”), and subset motif size (k)

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 create the most sensitive and specific motifs. For this reason, k 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, while the other subset motifs must contain evolutionarily insignificant amino acids. As a control, the Lysozyme input set (3lzt) was composed entirely of evolutionarily significant amino acids.

Functional Homologs. Measuring sensitivity and specificity requires a benchmark set of functional homologs. We use the functional classification of the Enzyme Commission [32] (EC), which identifies families of functional homologs for each input set (see Figure 3). Structure fragments and mutants were removed.

The Protein Data Bank. In this paper, we use Ω_5 , as mentioned in Section 2, which is sampled from the set of crystallographic protein structures in the Protein Data Bank 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 uses the Message Passing Interface [33] (MPI) protocol for interprocess communication, and was tested on a 16-node Athlon 1900MP cluster. The Rice TeraCluster, a cluster of 272 800Mhz Intel Itanium2s, and Ada, a Cray XD1 with 672 2.2Ghz AMD Opteron cores, computed final data. ϵ (see Section 2) was set to 7Å.

4.2 Median LRMSD Differentiates Motif Profiles

As mentioned in Section 4.1, our input sets were defined on both evolutionarily significant and insignificant amino acids, as well amino acids with documented

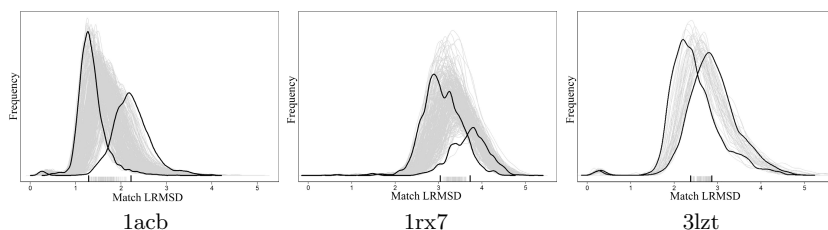


Fig. 4. Motif profiles generated using GS

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 six 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 amino acids as the subset motifs considered. We expected the Lysozyme input set, a control composed entirely of evolutionarily significant amino acids, to have a narrower spectrum of median LRMSDs, relative to the other sets of motif profiles.

Observations. The medians of the motif profiles generated (vertical hashes on the x-axes in Figure 4), occurred in ranges of approximately 1 Å LRMSD. 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 4, where the biggest differences in medians (darkened hashes) correlated to obvious differences in motif profiles (darkened curves). Lysozyme, which did not contain a spectrum of evolutionarily insignificant and significant amino acids, had a smaller range of medians. 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.

4.3 Median Estimation Cuts Runtime, Minor Accuracy Loss

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 4.2, we first generated matches without using the ELIMINATION optimization, mentioned in Section 3. 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

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
3lzt*	1204:52:00	184,395	227:56:00	97,593	942:00:00	92,099
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
1ukr	2030:26:40	1,063,797	150:13:20	110,043	35:40:00	74,613

Fig. 5. Speedups from Median Estimation: Execution time and number of matches computed, using step sizes of 100, 500, and exhaustive sampling. * = Run on the Rice TeraCluster. Remaining runs were done on Ada.

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. Operating at step sizes of 100, GS can identify the optimized motif an average of 10 times faster than GS without median estimation. This speedup follows directly from the early elimination of motifs which, with high probability, do not have the highest median. At step sizes of 100, GS can identify the optimized motif with an average of 10 times less matches than GS without median estimation. Figure 5 describes the precise number of matches and time consumed.

Median estimation is very accurate. In every case described in Figure 5, 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. GS was unable to eliminate the alternative subset motif because overlapping confidence intervals (see Section 3.1) did not separate by the time sampling was complete. The same was true at a step size of 500 for 3lzt, 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. Median estimation strongly accelerates the determination of the optimized motif with minor sacrifices in accuracy.

4.4 Geometric Uniqueness Identifies Effective Motifs

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 six input sets, are among the most sensitive and specific of all possible motifs definable from the input sets.

Experiment. For each input set, we computed a match between every possible subset motif and every functional homolog in the corresponding EC class, except for the identical structure. Then, for each match, we accessed the p -value, a measure of statistical significance determined using a method from previous work [3]. Using $\alpha = .02$, our standard of statistical significance, we determined

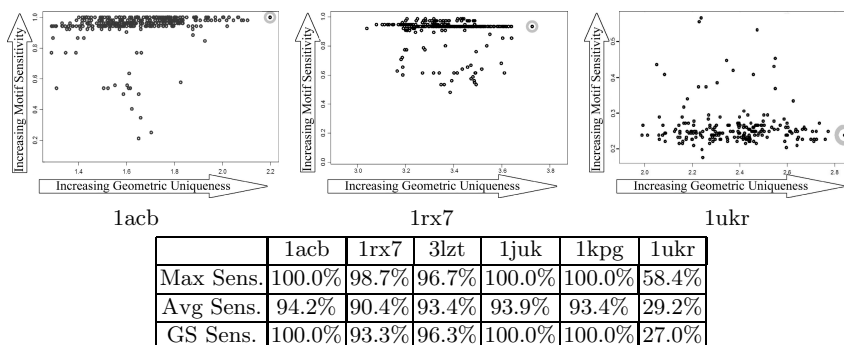


Fig. 6. Sensitivity of 1acb, 1rx7, 1ukr vs median LRMSD (above), and sensitivity per input set: the most sensitive subset motif, the average sensitivity, and the sensitivity of the optimized motif from GS (table)

the number of matches with p -values below α - the true positives. The proportion of true positives relative to the total number of functional homologs is the sensitivity of the motif. With 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 k values, GS identified optimized motifs which were quite sensitive, at a high level of specificity. From the 6 input motifs, GS produced 5 optimized motifs with greater sensitivity than the average subset motif from the same input set (see Figure 6). The exception, 1ukr, displayed no subset motifs with high sensitivity, even though it was created with the same criteria as the other input sets. Overall, Geometric Sieving performed well, identifying optimized motifs among the most sensitive of 5 out of 6 input sets, except where no effective motif could be found.

4.5 Geometric Uniqueness Correlates with Evolutionary Significance

Using the motif profiles calculated over Ω_5 , we have the median LRMSD of every subset motif. Since we also have the evolutionary significance of every amino acid

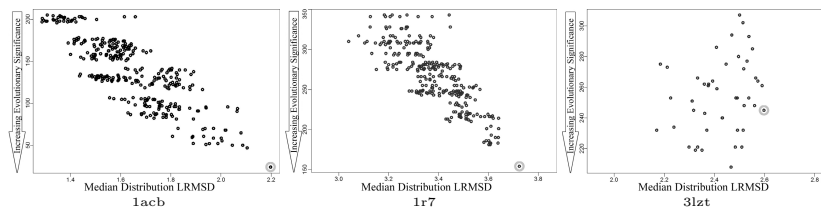


Fig. 7. Geometric Uniqueness vs. Evolutionary Significance

in our input sets, we can evaluate the evolutionary significance of every subset motif relative to its Geometric Uniqueness. In this experiment, we represented 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 7. Median LRMSD was plotted on the horizontal axis.

Observations. Motif profiles with high medians corresponded to subset motifs with evolutionarily significant amino acids (grey circles in Figure 7). In all cases but Lysozyme (3lzt), the input sets used demonstrate how evolutionary significance increases proportionately with increasing median LRMSD. In Lysozyme, a control set where every candidate motif point was evolutionarily significant, no apparent trend is visible. The existence of this apparent trend suggests that Geometric Uniqueness may be tied to evolutionary conservation.

5 Conclusions

We have presented GS, a novel distributed algorithm for exhaustively refining input sets of candidate motif points into optimized motifs. We have implemented GS with techniques and optimizations suitable for large scale distributed systems, and tested it on a cluster with more than 600 CPUs. By demonstrating refinement on 6 well-studied input sets, we show that, at a very high level of specificity, the optimized motifs from these examples were among the most sensitive of all motifs definable from these input sets. 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. 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 E. Cruess and Yi-Chieh J. Wu.

References

1. Wolfson H.J. and Rigoutsos I. Geometric hashing: An overview. *IEEE Comp. Sci. Eng.*, 4(4):10–21, Oct 1997.
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. Chen B.Y. et al. Algorithms for structural comparison and statistical analysis of 3d protein motifs. *Proceedings of Pacific Symposium on Biocomputing 2005*, pages 334–45, 2005.
4. Stark A., Sunyaev S., and Russell R.B. A model for statistical significance of local similarities in structure. *J. Mol. Biol.*, 326:1307–1316, 2003.
5. Yao H. et. al. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, 326:255–261, 2003.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. Connolly M.L. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221:709–713, 1983.
12. 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.
13. Shatsky M., Nussinov R., and Wolfson H.J. Flexprot: Alignment of flexible protein structures without a predefinition of hinge regions. *Journal of Computational Biology*, 11(1):83–106, 2004.
14. Artymuik P.J. et. al. 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.
15. Bachar O. et. al. A computer vision based technique for 3-d sequence independent structural comparison of proteins. *Prot. Eng.*, 6(3):279–288, 1993.
16. Rosen M. et. al. Molecular shape comparisons in searches for active sites and functional similarity. *Prot. Eng.*, 11(4):263–277, 1998.
17. 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.
18. Silverman B.W. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall: London, 1986.

19. 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.
20. 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.
21. Berman H.M. et. al. The protein data bank. *Nucleic Acids Research*, 28:235–242, Sept 2000.
22. 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.
23. 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.
24. Efron B. Better bootstrap confidence intervals (with discussion). *J. Amer. Stat. Assoc.*, 82:171, 1987.
25. Efron B. and Tibshirani R.J. *An Introduction to the Bootstrap*. Chappman & Hall, London, 1993.
26. 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.
27. Reyes V. et. al. Isomorphous crystal structures of *Escherichia coli* dihydrofolate reductase complexed with folate, 5-deazafolate, and 5,10-dideazatetrahydrofolate: mechanistic implications. *Biochemistry*, 34:2710–2723, 1995.
28. Bystroff C. et. al. Crystal structures of *Escherichia coli* dihydrofolate reductase: the nadp⁺ holoenzyme and the folate-nadp⁺ ternary complex. substrate binding and a model for the transition state. *Biochemistry*, 29:3263–3277, 1990.
29. Knochel T.R. et al. 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.
30. Huang C.-C. et al. Crystal structures of mycolic acid cyclopropane synthases from *Mycobacterium tuberculosis*. *J. Biol. Chem.*, 277:11559–11569, 2002.
31. 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.
32. International Union of Biochemistry. Nomenclature Committee. *Enzyme Nomenclature*. Academic Press: San Diego, California, 1992.
33. Snir M. and Gropp W. *MPI: The Complete Reference (2nd Edition)*. The MIT Press, 1998.