RICE UNIVERSITY

Geometry-based Methods for Protein Function Prediction

by

Brian Y. Chen

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Lydia Kavraki

Lydia E. Kavraki, Chair,
Professor
Computer Science, Rice University

Olivier Lichtarge

Olivier Lichtarge,
Professor
Molecular and Human Genetics,
Baylor College of Medicine

Moshe Vardi

Moshe Y. Vardi,
Professor
Computer Science, Rice University

John Mellor

John Mellor-Crummey,
Associate Professor
Computer Science, Rice University

Kevin R. MacKenzie,
Assistant Professor
Biochemistry and Cell Biology,
Rice University

Houston, Texas
September, 2006

Geometry-based Methods for Protein Function Prediction

Brian Y. Chen

Abstract

The development of new and effective drugs is strongly affected by the need to identify drug targets and to reduce side effects. Unfortunately, resolving these issues depends partially on a broad and thorough understanding of the biological function of many proteins, and the experimental determination of protein function is expensive and time consuming. In response to this problem, 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. This thesis seeks to develop one method that can identify useful geometric and chemical similarities between well studied and unstudied proteins. Our approach is to identify *matches* of geometric and chemical similarity between *motifs*, representing known functional sites, and substructures of functionally uncharacterized proteins (*targets*). It is commonly hypothesized that the existence of a match could imply that the target contains an active site similar to the motif.

We have designed the **MASH** (**M**atch **A**ugmentation with **S**tatistical **H**ypothesis Testing) pipeline, a software tool for computing matches. MASH is the first method to match *point-based* motifs, developed in earlier work, that represent functional sites as points in space with ranked priorities and alternative chemical labels. MASH is also first to match *cavity-aware* motifs, a novel contribution of this work, that extend point-based motifs with volumetric information describing active clefts critical to protein function. Controlled experiments demonstrate that matches for both types of motifs can identify cognate active sites.

However, motifs can also identify matches to functionally unrelated proteins. For this reason, we developed *M*otif Profiling (MP), the first method for motif refinement that reduces geometric similarity to functionally unrelated proteins. MP is implemented in two forms: Geometric Sieving (GS) refines point-based motifs and Cavity Scaling (CS) refines cavity-aware motifs. Controlled experimentation demonstrates that GS and CS identify motif refinements that have more matches to functionally related proteins and less matches to functionally unrelated proteins.

This thesis demonstrates the importance of computational tools for matching and refining motifs, emphasizing the applicability of large-scale geometric and statistical analysis for functional annotation.

Acknowledgments

I would like to deeply thank my advisor, Dr. Lydia Kavraki, for her guidance, patience, and confidence in me. Above everything, I am most grateful to her for my own growth as a researcher, a communicator, a mentor, and a collaborator. I am also greatly indebted to Dr. Olivier Lichtarge for his clear vision and truly unflagging patience in helping me to better understand the methods, substance, and culture of biological research. In any path of learning and self-improvement there will always be obstacles, but with Dr. Kavraki and Dr. Lichtarge's guidance, even the most insurmountable difficulties seem surpassable. If I have not seen further than others on the shoulders of these giants, then it is because I am not yet ready to see so far.

I would also like to thank Viacheslav Y. Fofanov and David M. Kristensen, whom I have worked with very closely with during my stay at Rice. Viacheslav has shown me a tiny glimpse into statistics that will forever change how I think about algorithms and data, and I am continually inspired by the example he sets in combining quantitative and biological thinking. I am also very grateful to David, who has helped so many times to renew my perspective and to broaden my biological horizons. I wish them both the best in the completion of their degrees.

I am also deeply indebted to the undergraduates whom I have had the opportunity to collaborate with and train. Anne E. Christian, Drew H. Bryant, Anand P. Dharan, Bradley D. Dodson, Amanda E. Cruess, Jessica Y. Wu, and Joseph Bylund have been fundamental in helping me to extend the reach of my ideas beyond my limits, in helping me to see beyond my own perspectives on research problems, and in teaching me how to train young researchers to be self-sustaining thinkers. Collaborating with these bright individuals has focussed

my vision and peerlessly enriched my experience as a graduate student.

I am tremendously grateful to my family and friends for their emotional and social support in this difficult endeavor. I am forever in debt to Andrew Ladd for his constant encouragement and his insightful advice. I am also deeply indebted to Kostas Bekris, for his continual support, his inspiring work ethic, and his patience. I am also especially indebted to Algis Rudys, whose knowledge of Unix esoterica and amazing familiarity with the intricacies of LaTeX have come to my rescue countless times. I would be much less of a person if I had never met these people.

This work has been supported in part by NSF DBI-0114796, NSF DBI-0318415, a training fellowship from the Gulf Coast Consortia and the National Library of Medicine (NLM Grant No. 5T15LM07093), and Rice University. Our experiments were made possible with equipment funded by AMD and EIA-0216467, and NSF CNS-0523908. Large-scale experimentation on high-performance computing machinery was made possible with equipment supported by NSF CNS-042119, Rice University, and partnership with AMD and Cray.

Contents

Abstract ii

Acknowledgments iv

List of Illustrations x

1 Introduction 1

1.1 The General Protein Function Prediction Problem 2

1.2 Specific Problems Targeted in this Work 3

1.3 Statement of Thesis 5

1.4 The **MASH** Pipeline for Identifying Geometric Matches 5

1.4.1 Point-based MASH 6

1.4.2 Cavity-aware MASH 7

1.5 Motif Profiling (MP): A Method for Automated Motif Refinement 8

1.5.1 Geometric Sieving 9

1.5.2 Cavity Scaling 10

1.6 Contributions 10

1.7 Thesis Road Map 11

2 Related Work 13

2.1 Motif Types and Design 13

2.1.1 Point-Based Motifs 14

2.1.2 Volumetric Motifs 15

2.2 Geometric Comparison Algorithms 16

2.2.1 Point-based Comparison Algorithms 17

2.2.2 Volume-based Comparison Algorithms 19

2.3 Theoretical Foundations 20

2.3.1 Exact Pattern Matching 20

2.3.2 Inexact Pattern Matching 21

2.3.3 Approximate Inexact Pattern Matching 24

2.3.4 Complexity of Pattern Matching Used in This Work . . 26

2.4 Statistical Models 26

2.4.1 Reference Sets 27

2.4.2 Measuring Statistical Significance 28

2.5 Systems for Protein Function Prediction 29

3 MASH: A Pipeline for Protein Function Prediction 30

3.1 Point-Based MASH 31

3.1.1 Motifs 31

3.1.2 Matching Criteria for MA 31

3.1.3 Match Augmentation 32

3.1.4 A Nonparametric Statistical Model 35

3.2 Cavity-Aware MASH 38

3.2.1 Cavity-Aware Motifs 39

3.2.2 Cavity-Aware Matching Criteria 41

3.2.3 Cavity Aware Match Augmentation 41

3.2.4 Statistical Model 43

3.3 Discussion and Contributions 45

3.3.1 Contributions 46

4 Testing the MASH pipeline 47

4.1 Testing point-based MASH 47

4.1.1 Input Data 47

4.1.2 MA Identifies Cognate Active Sites 48

4.1.3 Cognate Active Sites are Statistically Significant 50

4.1.4 Discussion 52
4.1.5 A Performance Comparison of MA and Geometric Hashing 52
4.2 Testing Cavity-Aware MASH 53
4.2.1 Input Data 53
4.2.2 C-Spheres Eliminate FPs, Preserve TPs 56
4.3 Discussion 59

5 Motif Profiling: A Motif Refinement Framework 60

5.1 The Motif Profiling Method 61
5.2 Geometric Sieving 62
5.2.1 The Geometric Sieving Algorithm 63
5.2.2 Accelerating GS 64
5.2.3 Discussion 66
5.3 Cavity Scaling 66
5.3.1 Markers of High-impact C–spheres 67
5.3.2 The Cavity Scaling Algorithm 68
5.4 Discussion and Contributions 69
5.4.1 Contributions 70

6 Experimentation on Motif Profiling Methods 71

6.1 GS Identifies Sensitive and Specific Motifs 72
6.1.1 Primary Data 73
6.1.2 Median LRMSD Differentiates Motif Profiles 78
6.1.3 Median Estimation Accelerates Performance with Minor Loss of Accuracy 81
6.1.4 Optimizing Geometric Uniqueness Improves Motif Effectiveness 83
6.1.5 Geometric Uniqueness Correlates with Evolutionary Significance 85

6.1.6 Discussion 87

6.2 Cavity Scaling Identifies Effective Cavity-Aware Motifs 88

6.2.1 Analysis of Individual C–spheres 88

6.2.2 Automatically Refined Cavity-aware Motifs 90

6.3 Discussion 91

7 Conclusions 93

Bibliography 97

Illustrations

2.1 A Diagram of a Geometric Hashing Invariant 17

2.2 Complexity of Exact Pattern Matching and Relaxations 20

2.3 Complexity of Inexact Pattern Matching 24

2.4 Complexity of Approximate Pattern Matching 25

2.5 Distribution of matches between a motif and all structures in the Protein Data Bank 27

3.1 The MASH pipeline 31

3.2 A study of protein reference sets 36

3.3 A diagram of a cavity-aware motif. 40

3.4 Two cases of cavity-aware matching. 42

3.5 Tree of partial matches considered in CAMA. The tree branches on alternative correlations between the highest ranked unmatched motif point and an unmatched target point. For example, the three branches from the seed match illustrate that there are three target points that the highest unranked motif point can be correlated with. If optimal alignment of the motif with the correlated target points causes a target point to fall within one or more C–spheres, we can immediately eliminate the match without considering further correlations. 44

4.1 Families used in MASH experimentation. 48

4.2 Experimental Results: 12 motifs and 73 targets plotted by LRMSD 50

4.3 p-values of LRMSDs from Figure 4.2 (log scale) 51

4.4 Motifs used, with example diagrams below. Starred (*) motifs use functionally documented amino acids. The column marked "#C" denotes the number of C–spheres in each motif. "R" denotes the range of C–sphere maximum diameters (in Å) for the motif. 54

4.5 Chemical Names of Ligands Used in Experimentation 54

4.6 Average effect of cavity-aware motifs on TP and FP matches, over all motifs. The horizontal axis charts C–sphere radius, where the radius of all C–spheres scales simultaneously from zero to individual maximum radius (see Section 4.2.2). The vertical axis charts the average percentage, per motif, of TP and FP matches remaining, relative to their respective point-based motifs. The number of homologs, and the number of TP and FP matches for each point-based motif is shown at left. FP matches are dramatically reduced while most TP matches are preserved, for both C–sphere designs. However, space-filling C–sphere designs tend to eliminate more FP matches and reject more TP matches, while ligand-based C–sphere designs tend to eliminate less FP matches, but preserve more TP matches. 58

5.1 Motif profiles for a low-impact C–sphere (a), and a high-impact C–sphere (b), as radius increases. For clarity, we provide 20 motif profiles for each C–sphere, showing how much the motif profile changes for a high-impact C–sphere. CS normally inspects only the motif profile with no C–spheres (the profile at the furthest left in both (a) and (b), and the motif profile corresponding to the C–sphere at maximum radius, at the furthest right in both (a) and (b). 68

5.2 How CS detects low-impact C–spheres (a) and high-impact C–spheres (b). Motif profiles corresponding to high-impact C–spheres vary significantly in their medians as C–sphere radius increases. Medians for low-impact C–spheres vary little. 69

6.1 Proteins used to test GS. 73
6.2 Input sets used to test GS I 74
6.3 Input sets used to test GS II 75
6.4 Input sets used to test GS III 76
6.5 Functionally documented amino acids 76
6.6 Motif profile variation among different subset motifs 79
6.7 Computational speedups from Median Estimation. 82
6.8 Comparison of subset motifs sensitivity I 84
6.9 Comparison of subset motifs sensitivity II 85
6.10 Relationship between geometric uniqueness and evolutionary significance 86
6.11 Effect of individual C–spheres on motif specificity 89
6.12 TP/FP matches preserved when using automatically refined cavity-aware motifs. 91

Chapter 1

Introduction

Broad and extensive knowledge of the biological *function* of proteins would have immense impact on medical and biological research. Using this knowledge, practical goals such as the identification of novel drug targets, the reduction of potential side effects, and the development of treatments affecting biological mechanisms, could be broadly advanced. In addition, knowledge of individual protein functions could accelerate many studies at the broader scope of protein-protein interactions and protein networks, by providing supporting information about single proteins.

Unfortunately, the function of many proteins is not known because experimental determination of protein function is an expensive and time-consuming process. Furthermore, it is currently infeasible to automate the collection of protein function information because elucidating the function of even a single protein currently depends on the insight of skilled investigators, and can require a broad range of empirical experimentation. This substantial barrier to automation inspires the design of computational methods that can provide investigators with useful information and analysis that may assist the determination of protein function. These algorithms are frequently called methods for *function prediction* or *functional annotation.* For simplicity, we use the former.

1.1 The General Protein Function Prediction Problem

Algorithms for Function Prediction approach the abstract problem of Protein Function Prediction, formulated as follows:

The Problem of Protein Function Prediction

Input:	A specific protein **P** and **I**, any known biological information useful to determine the output.
Output:	Specific information hinting at the biological function of the protein.

This description is an abstraction of many formulations of the problem. Varying formulations differ in how **P** is represented, such as by the sequence [1, 2, 3] or the structure [4, 5, 6, 7] of the protein, and may also use additional information (**I**), such as a set of sequence homologs [1, 8, 9, 10, 11, 12, 13], a network of proteins which **P** interacts with [14, 15, 16], or a set of protein structures [17, 18, 19, 20, 21, 22, 23]. Finally, many existing formulations of the problem yield information about protein function in many different ways. Some examples identify distant evolutionary relationships between proteins [2, 3], which may suggest functional similarities. Some formulations identify cavities and pockets in protein structures [24, 4, 7] which tend to be located at *functional sites* on the protein structure. Functional sites, also known as *active sites*, are regions within protein structure, frequently on the surface, believed to be most significant to the biological function of proteins. Still other formulations analyze networks of proteins [14, 15, 16] to deduce which proteins have important interactions with other proteins. All of these methods analyze existing data about proteins in an effort to gather small hints about protein functions, maximizing the impact of data collection efforts and potentially yielding information that might accelerate the study of proteins in the laboratory.

1.2 Specific Problems Targeted in this Work

Within the broader class of methods for Protein Function Prediction, this document studies the rigid geometric comparison of protein substructures. These techniques makes the governing assumption that geometric and chemical identity implies functional similarity. Within this subclass, one popular approach is to determine if a *target* protein structure contains a substructure, or *match*, that resembles a well documented active site structure, or *motif*. It has been widely hypothesized that rigid geometric similarity in active site geometry, with similar chemical labels, might identify active sites with similar biological function [17, 13, 18, 25, 20, 21, 22, 23].

In an effort to develop the most effective approach for identifying similar active sites, it is essential to answer two central questions:

> **1)** How do we efficiently determine if a substructure of the target is geometrically and chemically similar to the motif?
>
> **2)** What is the best geometric and chemical representation of the motif and target protein?

Producing an answer to these questions is difficult because any answer to one question requires an answer to the other. In order to identify matches of geometric and chemical similarity, it is essential to fix a representation of protein structure that is adequate to capture characteristics relevant to protein function. However, in order to develop and test adequate representations of protein structure, it is essential to have geometric comparison algorithms for controlled experiments.

The Geometric and Chemical Matching Problem

Input:	A motif structure representing a known functional site A target protein structure
Output:	A match (or no match if none found) of geometric and chemical similarity.

One part of this thesis explores question 1 in an effort to target the *Geometric and Chemical Matching Problem.* For two representations of motifs, called *point-based* and cavity-aware motifs, defined later, a solution to this problem is an algorithm that accepts a motif and a target as input, and returns matches of geometric and chemical similarity, if they exist, as output.

An effective solution to this problem requires effective motifs and an effective algorithm. Effective motifs have geometric and chemical similarity to functionally related proteins while maintaining geometric and chemical dissimilarity to functionally unrelated proteins. An effective algorithm identifies matches to effective motifs, when they exist.

The Motif Refinement Problem

Input:	A motif structure representing a known functional site
Output:	A refined version of the input motif with greater sensitivity and/or specificity.

Point-based and cavity-aware motifs are two ways represent a functional site with a motif. However, the question of how best to represent specific active sites is a very broad one. Various characteristics of any active site may be chosen, regardless of how they are represented, so that the number of matches to functionally related proteins is maximized, while the number of matches to functionally unrelated proteins is minimized. Selecting the set of characteristics to be represented can be formulated as the *Motif Refinement Problem.* This problem accepts, as input, a motif structure which represents several characteristics of an active site. The output desired is the subset of characteristics which maximize geometric and chemical similarity to functionally related proteins, while minimizing similarity to functionally unrelated proteins.

Together, our approaches to the Geometric and Chemical Matching problem and to the Motif Refinement Problem provide a comprehensive approach to the problem of using rigid geometric and chemical similarity to identify potentially similar active sites represented with point-based and cavity-aware motifs.

1.3 Statement of Thesis

This thesis shows that an efficient matching algorithm and large scale geometric comparison can yield an effective approach to the Geometric and Chemical Matching Problem and the Motif Refinement Problem. Large scale geometric comparison enables accurate, data-driven assessments of statistical significance not possible in earlier work. In Section 1.4, we will outline how efficient geometric matching algorithms can drive statistical models that can identify matches to functionally related active sites. Large scale geometric comparison also enables us to refine motifs using criteria other than the criteria used to design the motifs in the first place. In Section 1.5 we describe a novel technique that uses thousands of matches to identify motif refinements that have minimized geometric similarity to all known protein structures. As one of the first approaches to the Motif Refinement Problem, our technique would not be possible without large scale geometric comparison.

1.4 The MASH Pipeline for Identifying Geometric Matches

We have designed two variations of a pipeline called MASH (**M**atch **A**ugmentation with **S**tatistical **H**ypothesis Testing) that identify geometric matches using two types of motifs. Point-based MASH uses motifs represented as sets of chemically labeled and prioritized points in space, *motif points*, taken from atom positions in crystallographic protein structures. Motifs used for Cavity-aware MASH use motif points as well, but also incorporate *C–spheres* that represent geometric volumes essential for protein function.

1.4.1 Point-based MASH

Many existing methods [17, 18, 19, 20] have studied the possibility of representing protein structures as chemically labeled points in space, in an effort to compare protein structures. Our point-based motifs follow this basic blueprint; however, they also incorporate expanded labeling definitions for additional matching criteria. Section 3.1.1 explains that our point-based motifs follow this basic blueprint, and also support more flexible labellings, including multiple labels per point and priority ranks as those defined and computed in the work of [26, 27, 28, 1, 29, 8, 11].

We designed an algorithm called *Match Augmentation* (MA) [21]. MA identifies matches within the target that have substructural and chemical similarity to the motif. On a small data set, we demonstrated that MA is capable of finding 96.5% of target active sites cognate to a given motif. However, MA also identifies matches with many functionally unrelated proteins. We call these *false positive* matches. In the context of function prediction, where expensive experimental resources could be applied to verify functional similarity, false positive matches must be reduced as much as possible. For this reason, it is essential to understand the degree of similarity necessary to imply functional similarity.

The degree of geometric and chemical similarity associated with functional similarity can be studied with statistical models that assess the statistical significance of matches found. We measure geometric similarity by least root mean squared distance (LRMSD*), and enforce chemical similarity using chemical labels. Each time we find a match, our statistical model determines how *statistically significant* the LRMSD of a match is, relative to a baseline degree of similarity common among all protein substructures. Using our statistical

*LRMSD is the smallest possible root mean square distance (RMSD) between two sets of aligned points in 3D model, we showed that the identification of a match with statistically significant similarity can suggest that the target and motif have a similar active site, a result that concurs with existing findings using other motif designs and statistical models [25, 30, 21, 20].

MA, combined with our statistical model, forms the point-based MASH pipeline. As input, MASH takes a motif, and target protein of interest, and a representative set of all known protein structures. As output, MASH provides a match to the target provided, as well as a p-value that quantifies the statistical significance of the match.

1.4.2 Cavity-aware MASH

It is hypothesized that ligand binding proteins often contain active clefts or cavities which create chemical microenvironments essential for biological function. In several instances, large surface concavities have been associated with protein function [6, 31]. For this reason, existing methods have also studied motif designs which represent the volumes of clefts and cavities essential for protein function [32, 24, 7, 33]. Inspired by seminal work in the modeling and search for protein cavities [6, 34, 30], we explored an expanded definition of motifs and targets by combining a representation of cavities in protein structures with our own point-based motifs. We use geometric representations of cavities to eliminate false positive matches: If the matching atoms of the target truly form a cognate active site with similar function, the matching atoms of the target should surround an empty cavity with similar shape.

One of the strengths of MA is that it can be adapted for compatibility with certain motif and target variations. We developed a modified version of MA called Cavity-Aware Match Augmentation (CAMA) that searches for motifs built from motif points, such as those used above, while requiring specific geometric volumes, represented with sets of *C–spheres*, to remain empty. These *cavity-aware* motifs represent active sites as a combination of protein structure and functionally significant protein volumes, simultaneously. C-spheres also accelerate CAMA, because the search tree applied in CAMA, described later, can be pruned using the principle of maintaining empty C-spheres. Matches under consideration can be eliminated before computationally intensive analysis is spent, reducing computation time by two thirds while maintaining high accuracy. As we will demonstrate in our experiments, in comparison to point-based motifs, cavity-aware motifs have many fewer matches to false positives, while preserving most matches to functionally related proteins.

1.5 Motif Profiling (MP): A Method for Automated Motif Refinement

Currently, many motifs are designed by experts [21, 25], derived from biological literature [35], built from evolutionary analysis of families of homologous proteins [8, 11, 13], or from databases of active site information [19, 36]. Other methods select motifs based on analysis of structure or sequence data, such as the largest cavity [32] or using evolutionarily significant amino acids close to known ligand binding sites [21, 13]. 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, except MULTIBIND [37, 17] and our method, MP.

In the design of MP, we observed that the set of proteins with known structure have a very diverse set of functions. No significant fraction of the set of proteins with known structure share functional similarity with any given protein. For this reason, computing a *motif profile*, the set of matches between a given motif and the set of known protein structures, yields a close approximation to the set of matches to all functionally unrelated proteins. Given an input motif for refinement, MP compares motif profiles to find the motif refinement that maximizes geometric dissimilarity to the set of functionally unrelated proteins, a property we call *Geometric Uniqueness*. The motif refinement with maximal Geometric Uniqueness is returned as output.

We have applied the measurement of Geometric Uniqueness to the problem of refining the selection of motif points in point-based motifs, as well as the selection of C–spheres in cavity-aware motifs. These two applications of MP are GS [38] and CS [35].

1.5.1 Geometric Sieving

GS refines candidate motifs into *optimized motifs* before they are used in point-based MASH [38]. 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 Uniqueness. We used GS to identify 10 Geometrically Unique motifs, and tested them in the point-based 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 a cluster of computers and achieves speedup that is linear in the number of processors. In addition, we have designed an online statistical analysis that refines the data as it is generated. These optimizations make GS a practical preprocessing tool for refining motifs before they are passed to point-based MASH. This reduces the dependence on human experts for motif design by automatically refining more broadly defined motifs.

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 [26, 27, 1, 8, 10], as those most associated with important evolutionary divergences, have been shown to form statistically significant clusters [9] that are often related to active sites [11].

On our limited dataset, we observed that clusters of evolutionarily significant amino acids are more Geometrically Unique than evolutionarily insignificant amino acids.

1.5.2 Cavity Scaling

We observed that certain *high-impact* C–spheres contribute more to the elimination of false positive matches than others. In particular, when computing motif profiles on cavity-aware motifs, we observed that high-impact C–spheres force cause matches with all known protein structures to have greater geometric dissimilarity. For this reason, measuring the Geometric Uniqueness of some cavity-aware motifs, in comparison to identical point-based motifs, can identify high-impact C–spheres. We call this process CS [35].

CS allowed us to automatically refine our existing motifs to contain only high-impact C-spheres, guiding the design of cavity-aware motifs that eliminate many false positive matches, and reducing reliance on human experts. Applying CS to a set of cavity-aware motifs, we tested these refined motifs with cavity-aware MASH. Motifs using only high-impact C–spheres identified additional matches to functionally related proteins, while still eliminating many matches to false positives.

1.6 Contributions

MASH is a novel procedure that accepts two kinds of input motifs. The first, point based motifs, contain geometric, chemical and priority rank information. The development of motifs that fall in this class is a very active area of research [26, 39, 27, 1, 29, 9, 10, 12, 21, 38, 13, 35]. The second class of motifs, cavity-aware motifs, which first appear in this work, consists of a combination of point based motifs and volumetric information. MASH is the first procedure to accept these two classes of motifs.

In addition to developing an algorithm for matching the above motifs, MASH is also the first procedure that we know of that applies a non-parametric model to the measurement of the statistical significance of matches. Our statistical model does not require calibration to representative sets of protein structures, unlike earlier parametric models, and still identifies statistically significant matches to functionally related proteins.

Having developed a platform capable of identifying matches to similar functional sites, we developed MP in an effort to refine existing motifs and improve the accuracy of our platform. MP is among the first methods to introduce the idea of motif refinement based on non-biological criteria. MP is also the first method to formulate the concept of Geometric Uniqueness and demonstrate that Geometric Uniqueness can refine point-based motifs as well as cavity-aware motifs. MP is a valuable contribution because it does not depend on expert knowledge, thereby making first steps towards the automated design of motifs. While other methods have been motivated to refine motifs, MP is the first method that yields demonstrable improvements in accuracy, as we will demonstrate in our experimental results.

1.7 Thesis Road Map

Chapter 2 includes a discussion of the different types of motifs that have been considered in the past, summaries of other geometric comparison algorithms and other statistical models of geometric similarity, and a survey of the computational complexity of the geometric matching problem in several relevant formulations.

Chapter 3 then describes point-based MASH and cavity-aware MASH. First, we provide an in-depth description of point-based motifs, the Match Augmentation algorithm, and our statistical model. We then describe how cavity-aware MASH modifies point-based MASH to produce a cavity-aware version of the MASH pipeline.

Chapter 4 provides experimentation which demonstrates that point-based MASH and cavity-aware MASH are capable of identifying matches to functionally related proteins. We also demonstrate that cavity-aware MASH substantially reduces matches to functionally unrelated proteins.

Chapter 5 describes MP and the concept of Geometric Uniqueness. We then applied MP to the design of Geometric Sieving and Cavity Scaling. We also describe how statistical analysis can be used to accelerate a distributed version of Geometric Sieving.

Chapter 6 provides experimentation demonstrating that GS and CS identify sensitive and specific refinements of point-based and cavity-aware motifs. We also demonstrate the efficiency of our distributed implementation of GS.

Finally, Chapter 7 summarizes the contributions and results presented in this document.

Chapter 2

Related Work

A vast space of techniques are related to the identification of functional sites and the prediction of protein function. These methods study a wide variety of data types, ranging from protein sequences to protein structures to networks of proteins, varying in scope from analyzing a single protein to comparing several proteins, to hundreds of related proteins. While these topics border on the subject of this document, they are too numerous and diverse for the scope of this document. Instead, this section focuses on describing core topics directly related to the identification of potential functionally similar active sites through the comparison of protein substructures.

We begin by describing recent approaches to the critical subproblems of our function prediction approach, mentioned in Section 1.1. First, we describe different ways of representing active sites used in recent work. We then describe algorithms used to compare these motifs to target protein structures. One issue for these comparison methods is to understand what degree of similarity is necessary to imply functional similarity. For this reason, we also reserve a section for statistical models which help establish the degree of similarity associated with functional similarity. Finally, we review the algorithmic complexity of the geometric comparison problem that we seek to solve.

2.1 Motif Types and Design

The search for geometric markers of functional similarity has considered many different ways to represent active sites. In each study, it is hypothesized that one particular representation of an active site, or *motif type*, can be successful in identifying similar-functioning sites. The motif types considered can be loosely organized into point-based motifs and volumetric motifs.

In the future, it could be possible to study the differences in sensitivity and specificity common to each motif type, and perhaps arrive at a single type of motif that is useful for many proteins, or a well defined set of proteins. However, protein active sites vary greatly in chemical properties and geometric shape, and so it is likely that some motif types are more effective for some types of active sites, and that few are universally effective.

2.1.1 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 [40, 21], sidechain atoms [41, 20], atoms in hinge-bending flexible active sites [40], atoms in catalytic sites [25, 42], catalytic triads [43], and conserved binding patterns [37, 17]. 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 [44, 45, 46] and electrostatic potentials [18] on Connolly surfaces [47]. 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 [48]. 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 [43, 46, 40, 25, 18, 21], other representations of points in space include distance matrices [49, 50] and graphs [51, 52, 53].

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 [21] from the Evolutionary Trace (ET) [26, 12], hydrogen donor/acceptor and hydrophobic/hydrophilic properties [17], and electrostatic potential [18].

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 [26, 12] and proximity to binding sites [11, 21]. Motifs have also been designed using literature search and PSI-BLAST alignments of literature-defined motifs from the Catalytic Site Atlas [19, 36], and manually, by experts [25]. Still other motifs are designed using surface exposure, and algorithms for detecting conserved binding patterns [37]. These methods seek to identify substructures that are involved in biological function. Recent techniques also use geometric analysis to refine point-based motifs. GS [38], presented later in this paper, is one such method. Another excellent example is MULTIBIND [37, 17], an algorithm that identifies conserved binding patterns by identifying the least common point set among a set of existing motifs.

2.1.2 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 [54, 7, 55, 56, 35], alpha-shapes [57, 34, 30, 32], and grid-based techniques [24, 7, 33]. 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 [33], an algorithm that modifies the boundaries of computationally identified active clefts, to avoid regions distant from highly conserved amino acids.

Alpha shapes are one especially interesting way to identify and describe pockets and voids in protein structures through a natural geometric construction. While this construction is applied in three dimensions, we explain it in two dimensions for clarity. We begin with coordinates for each atom in the protein. Centered at each coordinate, we place a constant-radius circle. We also compute the Voronoi diagram on these coordinates, generating a set of Voronoi edges. Some Voronoi edges are completely outside the circles surrounding each point, while others touch at least one circle. All Voronoi diagrams are dual to a Delauney trianglization, because all Delauney edges correspond to adjacent Voronoi cells. In this case, we color Delauney edges red if the Voronoi edge that it crosses does not touch any circles. We color the remaining edges green. The alpha shape is said to be the set of green edges, the points they connect, and the volume enclosed within the green edges.

One interesting result of defining alpha shapes is that voids and pockets on protein surfaces can be identified easily. From the previous description, triangles in the Delauney trianglization which have at least one red edge describe pockets on the exterior of the protein, or are interior voids. This is critical in the identification of voids found by CASTp [31] and compared by pvSOAR [30].

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. For clarity, we loosely organize these comparison algorithms into point-based methods that search for point-based motifs, and volumetric methods that deal with volumetric motifs.

*For some approaches, geometric and chemical similarity is measured differently. Geometric Hashing [58], JESS [25], PINTS [20], and MA [38] measure geometric similarity using LRMSD. pvSOAR [32] uses both LRMSD and *oRMSD*. oRMSD is computed by first projecting all points onto the unit sphere at the center of each pocket, and then computing LRMSD.

2.2.1 Point-based Comparison Algorithms

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 [58, 59], a paradigm that 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 [44, 40, 41, 42, 37], refine point-based motifs by identifying the largest common point set among a set of similar motifs [17], and simultaneously align multiple [60, 61], even hinge-bent [62], protein structures.

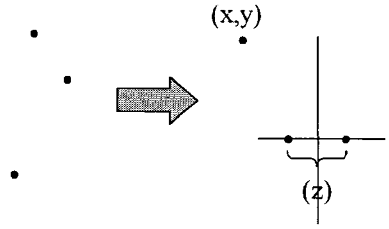


Figure 2.1 : A Diagram of a Geometric Hashing Invariant

Three points in space can be stored in an invariant manner. For example, the two closest points can be used to define a two dimensional axis on which the last point is the two dimensional vector (x, y). The distance between the closest two points provides the last aspect of the invariant, z. This type of geometric representation can be used to generate initial alignments for geometric comparison.

The Geometric Hashing paradigm hinges on the efficient application of geometrically invariant representations of points in space. Since many variations on geometrical invariants exist, we describe one example here. Given a set of 3 points in space, we can define a procedure which generates a vector that represents all three points, as well as perhaps several symmetric reflections of the points. As in Figure 2.1, for example, two of the three points can define a coordinate axis, with the perpendicular bisector of the segment defining a perpendicular axis. The distance between the first two points, and the 2D vector describing the position of the third point relative to the orthogonal axes described, generates a 3D vector that describes these points in space. We refer to this vector as an *invariant*. Similar invariants describe similar point triplets. Therefore, given two structures to compare, we can iterate through all triplets of both structures, generating invariants.

Given two triplets, we can compute an LRMSD alignment that provides a transformation of one triplet onto another. For this reason, when we observe a pair of similar invariants from the motif and target, we can compute a transformation between these two invariants, and store the transformation as a vector in the space of transformations. When we exhaustively compare all invariants from the motif and target, plotting the transformations corresponding to all similar pairs of invariants, we can cluster similar transformations. Averaging the largest cluster of transformations yields a single geometric alignment between the most triplets in the motif and target — a match. Geometric Hashing facilitates the rapid comparison of invariants by hashing all motif invariants, so that comparison of all motif invariants against all target invariants is dependent on generating invariants only for the target.

Other point-based comparison algorithms test possible point-to-point correlations in a depth-first-search (DFS) manner, such as the database search algorithm used in PINTS [63], JESS [25], and the Augmentation phase of MA [21]. These approaches to identifying matches begin with a partial correlation between the motif and the target, and expand the number of correlated pairs with a DFS. JESS and PINTS use multiple starting points from different atoms in the motif. MA uses seed matches, described later, of three points. Given the initial correlations, each method identifies a motif point and a target point that could be added as the next correlation, repeating this process until either no more motif points remain uncorrelated, or no more target points can be correlated with a motif point. The resulting partial match is saved, the last added correlation is removed, and a new target point is found to be correlated with the last unmatched motif point. This DFS process continues exhaustively, always remembering the match with lowest LRMSD that also fulfills all matching criteria. This match is returned once DFS iteration is complete.

Finally, other point-based comparison algorithms use techniques which find subgraph isomorphisms [53]. These approaches represent motifs as geometric graphs, where each atom is a vertex of the graph, and edge is weighted by its geometric distance between two points. Here, existing algorithms for edge-weighted subgraph isomorphism [64] were used to identify matches between motifs and targets.

2.2.2 Volume-based Comparison Algorithms

pvSOAR [30, 32] compares volumes in protein structure using motifs based on alpha-shapes. Earlier work on volumetric representations features analysis of only a single protein without comparison. Using varying representations of protein surfaces, these studies, using grid-based algorithms SURFNET [24] and SURFNET-ConSURF [33], and alpha-shapes technique CASTp [31], observed that ligand binding sites are often the largest "pocket" on the protein surface.

Among all earlier work, pvSOAR is the most closely related to our own approach, being the only algorithm that compares protein volumes between two proteins. pvSOAR leverages a sequence comparison of amino acids that border on protein pockets, as well as a scale-independent geometric comparison of amino acids relative to the pocket. Beginning with two protein structures to compare, all pockets in both structures are first identified. Then the set of amino acids adjacent to all pockets are identified and compared. If two pockets have high sequence similarity, then the geometry of the pockets are compared by oRMSD. pvSOAR uses both the sequence similarity score and unit RMSD to evaluate geometric and chemical similarity.

2.3 Theoretical Foundations

In our representation of protein structures, identifying matches between a motif and target generalizes to the problem of identifying the subset of a point set in 3D with smallest LRMSD to another set of points. While a polynomial solution for this problem remains unknown, as we explain later, many variations of this pattern matching problem have been studied, providing several bounds on the complexity of the problem. We survey these results in this section.

2.3.1 Exact Pattern Matching

The exact case of the pattern matching problem is well studied in $\mathbb{R}^d$ for all d. Exact pattern matching searches for a *congruence* between point sets A and B in $\mathbb{R}^d$ with sizes $|A|$ and $|B|$. A congruence is defined as an isometric mapping $T \in \mathcal{T}$, the space of all rigid rotations and translations in $\mathbb{R}^d$. A and B are said to be *congruent* if there exists a congruence T such that $\forall a_i \in A$, $T(a_i) = b_j \in B$, where j are distinct for all i. The asymptotic complexity of determining congruence, and several relaxations of the congruence problem, is summarized in Figure 2.2.

Exact Pattern Matching Performance		
Congruence	CCD	LCP
$O(n^{\frac{d-1}{2}} \log n)$	$O(n^{1.89}m^{.8} + min\{n^{2.5}, n^3m^{-2}\})$	$O(n^{1.89}m^{2.8} + n^3)$

Figure 2.2 : Complexity of Exact Pattern Matching and Relaxations

We survey here the complexity of exact pattern matching (furthest left), congruent copy detection (center), and largest common pointset (right) problems. While complexity is known for exact pattern matching in all dimensions d ($\mathbb{R}^d$), this is not the case for CCD and LCP. For CCD and LCP above, we state complexity for 3D (three dimensions), which is most relevant to the work in this document.

In their seminal publication, which initiated much of the study of exact and inexact, pattern matching, Alt, Mehlhorn, Wagener, and Welzl [65] demonstrated that for $|A| = |B| = n$, identifying a congruence (or returning failure when none exists) is possible in $O(n^{d-2} \log n)$ time. This was later improved by Akutsu [66] to $O(n^{\frac{d-1}{2}} \log n)$. Akutsu also added that the problem of identifying congruence for unbounded d is NP-hard, by demonstrating that if a polynomial time algorithm existed for congruence in unbounded d, there also exists a polynomial time algorithm for graph isomorphism [66].

The exact pattern matching problem is very restrictive. In many applications, we seek to determine if one pattern A exists within a field of observed data B, which is not the same size as A. Thus, one well studied relaxation of the exact pattern matching problem is when $|A| = m < n = |B|$. Here, the problem is to determine if A is congruent to a subset of B, and the problem is called the *Congruent Copy Detection* (CCD) problem. In 1998, Akutsu, Tamaki, and Tokuyama [67], demonstrated that for $d = 2$, the complexity of CCD is $O(min\{n^{1.43}m^{.77}, n^{\frac{4}{3}}m\})$, for $d = 3$, CCD is $O(n^{1.89}m^{.8} + min\{n^{2.5}, n^3m^{-2}\})$, for $d = 4$, CCD is $O(n^{2.87}m + n^{3.83})$, and for $d \geq 5$, CCD is $O(n^{d-1}m + n^d)$.

In other applications, we seek to determine of A exists within B, but it can be possible that only a portion of A is detectable. Thus, a further relaxation of the exact pattern matching problem is when $|A| = m < n = |B|$, but we seek the largest $p < m$ such that a subset $A' \subset A$ has $|A'| = p$ and $|A'|$ is congruent to a subset of B. This problem is called the *Largest Common Pointset* (LCP) problem. In 1998, Akutsu, Tamaki, and Tokuyama [67] demonstrated that for $d = 2$, LCP is $O(n^{1.43}m^{1.77} + n^2)$, for $d = 3$, LCP is $O(n^{1.89}m^{2.8} + n^3)$, for $d = 4$, LCP is $O(n^{2.87}m^4 + n^4)$, and for $d \geq 5$, LCP is $O(n^{d-1}m^d + n^d)$.

2.3.2 Inexact Pattern Matching

Exact pattern matching has difficulties in application, because in practice, imprecise data and inaccurate sensors produce data that is rarely identical. For this reason, one important relaxation of the exact pattern matching problem is the inexact pattern matching problem. The inexact pattern matching problem seeks to determine how similar two non-identical point-sets A and B are, by identifying the smallest possible value of a given distance measure between point sets. These distance measures vary based on the alignment of the two point sets, or the mapping between points in A and B. For this reason, identifying the minimum distance can be very time consuming.

Distance Measures The most well studied point set distance measures include the *Hausdorff* measure δ_H, the *bottleneck* measure δ_B, and *Root Mean Squared Distance* (RMSD), δ_{RMSD}. These distance measures are frequently defined on underlying point-to-point distance metrics such as the L_2 norm and the L_∞ norm. In this document, we report results for the L_2 norm, referred to as $d(a,b)$ between points a and b, because it is far better studied and is most closely related to the work in this document.

Definition 1 (Hausdorff Distance) *Given point sets A and B, the asymmetric Hausdorff distance from A to B is*

$$\overrightarrow{\delta_H}(A,B) = \max_{a \in A} \min_{b \in B} d(a,b)$$

The symmetric Hausdorff distance from A to B is

$$\delta_H(A,B) = \max(\overrightarrow{\delta_H}(A,B), \overrightarrow{\delta_H}(B,A))$$

Hausdorff distance differs from the other two measures in that for $A, B \subset \mathbb{R}^d$, $\delta_H(A,B)$ depends solely on an alignment $T \in \mathcal{T}$, the space of all rigid rotations in $\mathbb{R}^d$. Thus, the smallest possible value of $\delta_H(A,B)$ is $\min_{T \in \mathcal{T}} \delta_H(A,B)$.

Definition 2 (Bottleneck Distance) *Given point sets A and B, let $\mathcal{M}$ be the set of all matchings (one-to-one relations) M from A to B. The symmetric bottleneck distance from A to B is*

$$\delta_B(A,B) = \min_{M \in \mathcal{M}} \max_{(a,b) \in M} d(a,b)$$

Definition 3 (Root Mean Squared Distance) *Given point sets A and B, let $M \in \mathcal{M}$ be a one-to-one relation in the set of all one-to-one relations $\mathcal{M}$ from A to B. The symmetric RMSD for M is*

$$\delta_{RMSD}(A,B) = \sqrt{\frac{\sum\limits_{(a,b)\in M} (d(a,b))^2}{2}}$$

The Least RMSD, or LRMSD, from A to B is

$$\delta_{LRMSD}(A,B) = \min_{M\in\mathcal{M}} (\delta_{RMSD}(A,B))$$

Unlike the Hausdorff measure, the bottleneck and RMS measures depend on one-to-one relations between the point sets under comparison. While one way to generate these one-to-one relations is to generate an alignment and then carefully select a one-to-one relation based on proximities produced by the alignment, both measures are entirely independent of the alignment. In addition, given A, B and a one-to-one relation M from A to B, there is exactly one alignment $T \in \mathcal{T}$ for which $\delta_{RMSD}(T(A), B) = \delta_{LRMSD}(A, B)$. However, given M, a transformation T which satisfies $\delta_B(T(A), B) = \min\limits_{M\in\mathcal{M}} \delta_B(A, B)$ is clearly not unique.

Determining the Hausdorff similarity between given A and B is to determine the minimum of $\delta_H(A, B)$ for all rotations and translations of A. For $A, B \subset \mathbb{R}^2, |A| = m < n = |B|$, the minimum $\delta_H(A, B)$ can be determined in $O((m+n)^5 \log^2 mn)$ using a method by Huttenlocher, Kedem, and Kleinberg [68], and was improved to $O((m+n)^5 \log^2 mn)$ by Chew et. al. [69]. Later, for A and B ($|A| = |B| = n$) in $\mathbb{R}^d$, under the δ_H, Chew, Dor, Efrat, and Kedem [70] showed that the minimum of $\delta_H(A, B)$ could be computed $O(n^{\lceil 3d/2 \rceil} \log^3 n)$. Minimizing $\overrightarrow{\delta_H}(A, B)$,$A, B \subset \mathbb{R}^2$ is possible in $O(m^3 n^2 \log mn)$ time [69].

Minimizing δ_B and δ_{RMSD} is much more difficult. Restricting the problem to determining the Given $A, B \subset \mathbb{R}^2$, $|A| = |B| = n$, Alt, Mehlhorn, Wa-

Complexity of Determining Optimal Matching Distance		
Symmetric Hausdorff	Bottleneck*	RMSD
$O(n^{\lceil 3d/2 \rceil} \log^3 n)$	$O(n^6 \log n)$	Unknown

Figure 2.3 : Complexity of Inexact Pattern Matching

We survey here the complexity of the inexact pattern matching, where we seek to determine the minimum Hausdorff (left), Bottleneck (center), or RMS (right) distance between point sets A and B. While complexity is known for Hausdorff pattern matching in all dimensions d ($\mathbb{R}^d$), this is not the case for Bottleneck and RMSD. For Bottleneck and RMSD, we state complexity for 2D (two dimensions), which is closest to the work in this document. $\star$ = The only known complexity bound for Bottleneck measure is for translation only.

gener, and Welzl [65] demonstrated that computing the minimum of δ_B under translation only is $O(n^6 \log n)$. No polynomial algorithm is known to minimize δ_{RMSD} for all rotations and translations in the plane [71], or in higher dimensions.

2.3.3 Approximate Inexact Pattern Matching

The difficulty of determining the minimum distance via δH δ_B and δ_{RMSD} led to the development of techniques for approximating the minimum distance. Under the Hausdorff measure, these algorithms seek to determine, for ϵ, β, A and B given, if there exists a transformation $T \in \mathcal{T}$, the space of rigid transformations, such that $\delta H(T(A), B) < (1+\beta)\epsilon$. Under the Bottleneck and RMSD measures, given A, B, and ϵ, many approximations seek to determine if there exists $M \in \mathcal{M}$ such that $\delta_B(A, B) < \epsilon$ or $\delta_{RMSD}(A, B) < \epsilon$, respectively.

In $\mathbb{R}^2$, Goodrich, Mitchell and Orletsky [72] demonstrated that identifying a rigid motion such that $\delta H(T(A), B) < (1+\beta)\epsilon$ is possible in $O(n^2 m \log n)$, and that in $\mathbb{R}^3$, a similar result is possible in $O(n^3 m \log n)$. Recognizing that applications of geometric pattern matching are frequently applied to data with well understood geometric properties, Indyk, Motwani, and Venkatasubramanian [73] later developed an approximation scheme that performs better when

Complexity of Approximating Minimum Pointset Distance		
Symmetric Hausdorff	Bottleneck	RMSD
$O(n^3 m \log n)$	$O(m^{16} n^{16} \sqrt{m+n})$	$O(n^4 \epsilon^{-5/2} \log^6 n)$

Figure 2.4 : Complexity of Approximate Pattern Matching

We survey here the complexity of the approximate pattern matching, where we seek, under the Hausdorff (left), Bottleneck (center), or RMS (right) measures, to identify a rigid transformation $T \in \mathcal{T}$ such that distance between A and B is less than a given value. Some approximation techniques that also take advantage of data-specific properties are described below but not categorized here.

the ratio Δ of the maximum distance between points in B relative to the minimum distance between points in B is small. Under the Hausdorff measure in $\mathbb{R}^2$, this approximation scheme results in $O(\min(k(n^4\Delta)^{1/3}, n(\Delta+n)))$, and in $\mathbb{R}^3$ the scheme results in $O(\min(k \max(n^{2.25}\sqrt{\Delta}, n^{2.5}), n\Delta(\Delta^2+n), n^2(n+\Delta)))$.

Approximating Bottleneck distance is very difficult. In 2000, Ambuhl, Chakraborty, and Gartner demonstrated that, given small ϵ and point sets $A, B \in \mathbb{R}^3$, $|A| = m < n = |B|$, it is possible to determine if there exists a rigid transformation such that

$$\forall a_i \in A \quad \exists b_j \in B, i \neq j \mid d(a_i, b_j).$$

The algorithm Ambuhl et. al. provided was $O(m^{16} n^{16} \sqrt{m+n})$. Indyk, Motwani, and Venkatasubramanian [73] also developed a density-based approximation scheme under the Bottleneck measure. In $\mathbb{R}^2$, this approximation scheme results in $O(k^{3/2}(n^4\Delta)^1/3)$. In $\mathbb{R}^3$, this scheme results in $O(k^{1.5} max(n^{2.25}\sqrt{\Delta}, n^{2.5}))$.

In 2006, Phillips and Agarwal showed that, given the parameter $\epsilon > 0$ and $A, B \subset \mathbb{R}^2$ with $|A| = |B| = n$, if $(\delta_{LRMSD}(A, B) = \epsilon^\star)$, it is possible to determine if $\delta_{RMSD}(A, B) < (1+\epsilon)\epsilon^\star$ in time $O(n^4 \epsilon^{-5/2} \log^6 n)$.

2.3.4 Complexity of Pattern Matching Used in This Work

The pattern matching algorithm used in the work, MA, seeks to determine the LRMSD in three dimensions, between A and B with $|A| = n < m = |B|$. Unfortunately, there is no known polynomial-time algorithm for this precise problem. In fact, there is also no known polynomial-time algorithm for the 2D optimization problem, even for $|A| = |B|$ [71]. In addition, given a match m that claims to have lowest LRMSD among all possible 3D matches between A and B, it is unclear that it is even possible to verify that m is minimal in polynomial time.

Even though identifying the single match with lowest LRMSD is clearly a difficult problem, many existing techniques like Geometric Hashing and MA, operating without a constant of approximation, can identify a useful match in a fraction of a second on modern desktop computers. In fact, we will demonstrate in Chapter 4 that MA can rapidly identify amino acids cognate to the amino acids of the motif. It remains unclear at this point if biological characteristics reduce the complexity of the pattern matching problem on biological data.

2.4 Statistical Models

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.

2.4.1 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.

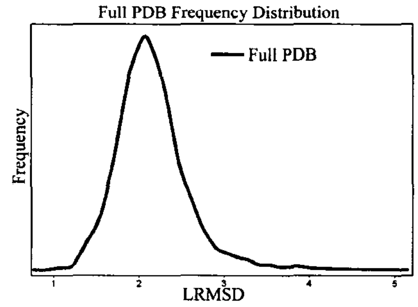


Figure 2.5 : Distribution of matches between a motif and all structures in the Protein Data Bank

Current reference sets are generated from databases and classifications of protein structures, including the Protein Data Bank (PDB) [74], SCOP [75], a classification of protein folds, and CATH [76], 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 [21], and to structurally nonredundant subsets of the PDB [35]. Other statistical models compute matches to fold representatives [20] from SCOP, and non-redundant multi-domain representatives [25] from CATH. The distribution of matches between a motif and proteins in a reference set, such as the PDB, in Figure 2.5 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.

2.4.2 Measuring Statistical Significance

Given a baseline 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 [20] database computes matches between a motif and every protein in a nonredundant subset of SCOP [75]. 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 explicitly evaluate a p-value that measures the degree of statistical significance.

JESS [25] 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 parametric model of mixtures of normal distributions. JESS applies this approach to comparatively evaluate the significance of matches between a library of motifs and a given target structure. The most significant match in the library provides evidence of functional similarity between the given target and the matching motif.

pvSOAR [30, 22], 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.

2.5 Systems for Protein Function Prediction

In this chapter, we have described many techniques relevant to the problem of identifying statistically significant matches of geometric and chemical similarity to functionally related proteins. However, combining these tools into systems useful for making function predictions is also an important problem. For example, pvSOAR [30] combines a cavity-based matching algorithm and an empirical statistical model as part of a web service for identifying similar protein surface regions in protein structures. CASTp [31] provides an atlas of protein pockets and voids for all structures in the PDB [74]. The PINTS server [63, 20] provides a rapid database search algorithm coupled with a statistical model of structural similarity. PROFUNC [77], provides numerous sequence and structure analyses in a single package, including BLAST [3], InterProScan [78], SSM [79], and JESS [25], among many others. These integrated systems inspired our own integrated approach to the design of MASH and MP, described in the next chapters.

Chapter 3

MASH: A Pipeline for Protein Function Prediction

Our first approach to the problem of identifying similar functional sites was to fix a representation of protein structures, and to design an algorithm, MA, for identifying matches of geometric and chemical similarity between given motifs and targets. The design of MA is highly modular, and compatible with several different types of motif. We will begin by describing point-based MASH, a pipeline for identifying matches based on representing protein structures as sets of points in three dimensions. Next, we describe cavity-aware MASH, which extends point-based MASH by also modeling active clefts critical to protein function. Both variations contain three critical components: a representation of motifs and targets, a matching algorithm, and a model for evaluating the statistical significance of matches found. These components complete the MASH pipeline.

Input/Output Requirements for MASH

Input:	A motif, A target protein structure All Protein Structures in the PDB
Output:	A match (or no match if none found) and p-value measuring the Statistical Significance of the match.

As input MASH accepts either a point-based or cavity-aware motif, as well as a target protein structure and has access to the set of all protein structures in the PDB. MASH uses this input to determine if a match exists between the motif and the target, and measure the statistical significance of the match, returning this information as output.

3.1 Point-Based MASH

In this section, the term "MASH" refers to point-based MASH. As input, MASH accepts a motif and a list of target proteins in which to search for matches. After applying MA to identify matches in each target, MASH then uses a snapshot of the PDB to compute a p-value that assesses the statistical significance of each match. Using a standard of acceptable statistical significance, α, statistically significant matches, where $p < \alpha$, and statistically insignificant matches, where $p \geq \alpha$ are returned as output.

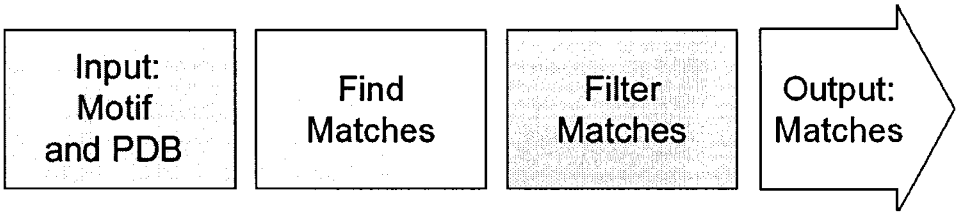


Figure 3.1 : The MASH pipeline

3.1.1 Motifs

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* that measures the functional significance of the motif point. Each s_i also has a set of alternate amino acid *labels* $l(s_i) \subset \{GLY, ALA, ...\}$, that 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 work, we obtain labels and ranks using the Evolutionary Trace [26, 27].

3.1.2 Matching Criteria for MA

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.

3.1.3 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 that 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 $T' = \{t_1, t_2, \dots t_{|T|}\}$ as a graph [80], where each vertex is a target point t_i. We then eliminate all t_i that 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 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 that 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 [81, 82], 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 that 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 that 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.

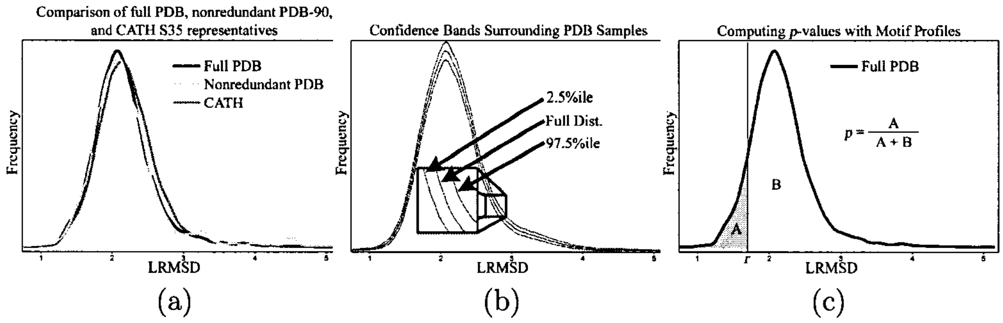
3.1.4 A Nonparametric Statistical Model

In collaboration with Viacheslav Y. Fofanov and his advisor Marek Kimmel at Rice University, we have developed a statistical model that uses a hypothesis testing framework. We use our statistical model to detect matches with statistically significant geometric and chemical similarity. Statistical 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, a motif profile, S_{Ω}. As mentioned earlier, motif profiles are best visualized as frequency distributions (see Figure 2.5).

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%) amino acid sequence identity. A similar comparison was true when using the CATH [76] 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 3.2a. 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 [76].



(a)

(b)

(c)

Figure 3.2 : A study of protein reference sets

(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).

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 3.2b, where we overlayed 5000 distinct S_{Ω_5} samples with a single S_{Ω_0}, the center line in Figure 3.2b. 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. Kolmogorov-Smirnoff [83] tests confirmed a lack of statistically significant differences between sampled distributions and D_{S_i}.

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 [21].

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: S and T are structurally dissimilar

H_A: S and T 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 3.2c). 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 3.2c requires careful numerical treatment. We apply kernel density estimation procedures [84] 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 [21]. In addition, we avoid under- and over-smoothing by using optimal bin-widths determined by Sheather-Jones method [85, 86].

3.2 Cavity-Aware MASH

We hypothesized that the sensitivity and specificity of our motifs could be further optimized by incorporating volumetric information that represents active clefts and cavities that must remain vacant for biological activities like ligand binding. For example, even though the motifs we defined exist on the surface of the protein around well studied active clefts and cavities, we have observed that matches are sometimes found in the interior of the protein. By representing active volumes in motifs, and insisting that they remain empty for biological activity in the target, we can eliminate matches that do not fulfill this property. If the matching atoms of the target truly form a cognate active site with similar function, the matching atoms of the target should surround an empty cavity with similar shape.

We begin by first describing how we modified our existing motif representation to include volume information describing active cavities, producing cavity-aware motifs. We then explain how we modified MA to produce CAMA, demonstrating how cavity-aware motifs can be used to also accelerate MA. Finally, we describe how we adapted our statistical model, used in the previous section, to address matches to cavity-aware motifs.

3.2.1 Cavity-Aware Motifs

The cavity-aware motifs developed and used in this work are an integration of a point-based motif and a cavity-based motif. Cavity-aware motifs contain motif points taken from atom coordinates labeled with evolutionary data [26, 27, 21, 13]. A motif S contains a set of m motif points $\{s_1, \ldots, s_m\}$ 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* $l(s_i) \subset \{GLY, ALA, ...\}$, that represent residues to which this amino acid has mutated during evolution. Labels permit our motifs to simultaneously represent many homologous active sites with slight mutations, not just a single active site. In this work, we obtain labels and ranks using the Evolutionary Trace [26, 27].

Cavity-aware motifs also contain a set of C–spheres $C = \{c_1, c_2, \ldots c_k\}$ with radii $r(c_1)$, $r(c_2)$, ..., $r(c_k)$ that are rigidly associated with the motif points. C–spheres are a loose approximation of solvent exposed volumes essential for ligand binding. C–spheres can have arbitrary radius, and can be centered at arbitrary positions. While this work targets the functional prediction of active sites that bind small ligands, this representation could be used to represent protein-protein interfaces and other generalized interaction zones.

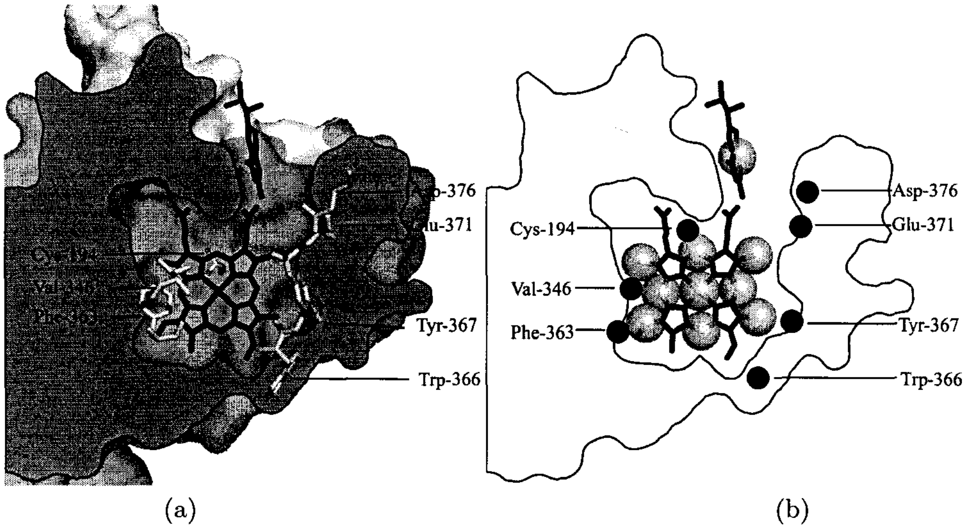


Figure 3.3 : A diagram of a cavity-aware motif.

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

C–sphere positions in this work were selected based on the coordinates of atoms in bound ligands. For example, in Figure 3.3, we modeled the heme-dependent enzyme nitric oxide synthase, a protein that catalyzes the synthesis of nitric oxide (NO) from an L-arginine substrate. This multi-step reaction takes place in a deep cleft and involves zinc, tetrahydrobiopterin, and hydride-donating (NADPH or H_2O_2) cofactors [87, 88]. Using PDB structure 1dww, we centered C–spheres at several atom coordinates on the heme, in order to fill the heme-binding cavity, and placed one C–sphere to represent tetrahydrobiopterin, which was further outside from the main cavity, as was shown in Figure 3.3. Future work could explore the generalized positioning of C–spheres.

3.2.2 Cavity-Aware Matching Criteria

Matching Criteria

CAMA compares a cavity-aware motif S to a target T, a protein structure encoded as n *target points* referred to as $T = \{t_1, \ldots t_n\}$, 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 $M = \{(s_{a_1}, t_{b_1}), (s_{a_2}, t_{b_2}) \ldots (s_{a_m}, t_{b_m})\}$. 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 label compatible: $l(t_{b_i}) \in l(s_{a_i})$.
>
> **Criterion 2** $\forall i$, $||A(s_{a_i}) - t_{b_i}|| < \epsilon$, our threshold for geometric similarity.
>
> **Criterion 3** $\forall t_i \forall c_j$ $||t_i - A(c_j)|| > r(c_j)$

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

3.2.3 Cavity Aware Match Augmentation

CAMA is a two stage hierarchical matching algorithm, based on MA, that identifies correlations for motif points in order of rank. The first stage, *Seed Matching* is a hashing technique that exploits pairwise distances between motif points to rapidly identify correlations between the three highest ranking motif

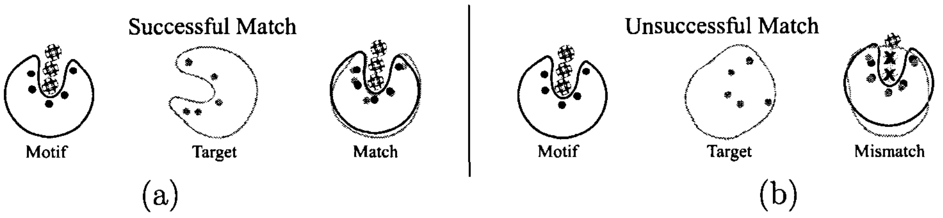


Figure 3.4 : Two cases of cavity-aware matching.

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

points and triplets of target points. Seed matching in CAMA is identical to seed matching in MA, and is not repeated here. These triplets are passed to the second stage, *Augmentation*, that expands seed matches to full correlations of all motif points. As an improvement over our method from earlier work [35], as correlations are being expanded, we insist that C–spheres remain empty. The final output is the correlation with the smallest LRMSD, satisfying all matching criteria.

Seed Matching Seed Matching generates three-point correlations between the 3 highest ranking motif points and three distinct points in the target. These triplets are sorted in LRMSD and passed to Augmentation. See Section 3.1.3 for a detailed description.

Augmentation As we described in Section 3.1.3, Augmentation applies DFS to exhaustively identify target points that can be correlated to the highest ranked unmatched motif point while maintaining our matching criteria. To adapt MA to the comparison of C–spheres, each time we generate a potential correlation, we plot the positions of the C–spheres in rigid alignment with the motif. Then, for each C–sphere, we check if a target point exists within the C–sphere. If any target point is found within any C–sphere, the match is discarded.

As in Section 3.1.3, if there are no more unmatched motif points, we put this match into a heap that maintains the match with smallest LRMSD. If unmatched motif points remain, we put this partial match back onto the stack. Finally, we return to the stack, pop off the first match on the stack, and repeat this process until the stack is empty.

C–spheres Accelerate MA In addition to eliminating matches that do not satisfy our matching constraints, C–Spheres can also eliminate some potential matches being considered by CAMA, increasing algorithmic efficiency. This is because the Augmentation stage is a depth first search that can be represented as a branching search tree. Correlations of motif points and target points represent nodes in this tree, where seed matches represent root nodes. An edge between a parent node and child node represents an instance where the highest ranking unmatched motif point can be aligned with a target point, generating an expanded partial match with an additional correlated pair. Since multiple target points may be available to expand a partial match, the tree can branch from a parent node to several child nodes. This is depicted in the left of Figure 3.5, while the next unmatched motif points s_3, s_4, and s_5, are shown on the right.

When testing an alignment, if the C–spheres contain a target point, then the children of this node, having correlations with only one additional motif-target pair, will have similar alignments and are likely to have C–spheres which also contain the same target point. Heuristically, we can eliminate the parent node, rather than continue to test additional partial matches. Pruning the tree in this manner reduces the number of comparisons necessary.

3.2.4 Statistical Model

Our method for measuring the statistical significance of matches computed with cavity-aware motifs is very similar to measuring the statistical significance of matches to a point-based motif. Again, for a given match m with LRMSD

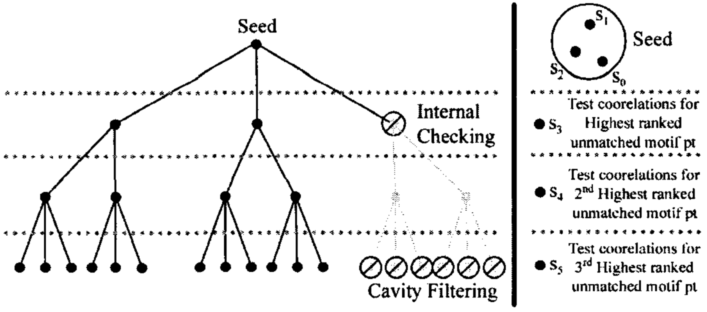

Seed
Internal Checking
Cavity Filtering
S_1
S_2
S_0
Seed
S_3
Test coorelations for Highest ranked unmatched motif pt
S_4
Test coorelations for 2nd Highest ranked unmatched motif pt
S_5
Test coorelations for 3rd Highest ranked unmatched motif pt

Figure 3.5 : Tree of partial matches considered in CAMA. The tree branches on alternative correlations between the highest ranked unmatched motif point and an unmatched target point. For example, the three branches from the seed match illustrate that there are three target points that the highest unranked motif point can be correlated with. If optimal alignment of the motif with the correlated target points causes a target point to fall within one or more C–spheres, we can immediately eliminate the match without considering further correlations.

r between motif S and target T, our earlier model assessed the probability p of observing a match with similar LRMSD r', when comparing the same motif and any protein with known structure. First, a match is computed between S and every member of a representative set of proteins, in order to establish a baseline degree of geometric similarity between S and the space of known protein structures. This set of matches is depicted as a frequency distribution, or motif profile, in Figure 3.2c. Figure 3.2c indicates how p, or the p-value, our measure of statistical significance, is computed. Given a standard of statistical significance α, we say that m is statistically significant if $p < \alpha$.

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

The impact of C–spheres on predictions made by cavity-aware MASH is a direct result of how C–spheres eliminate matches. During the augmentation phase, C–spheres eliminate partial matches that a point-based motif would not have eliminated. If the lowest LRMSD match is eliminated in this fashion, then the ultimate output of CAMA will have a higher LRMSD than the ultimate output of MA, which does not eliminate matches based on C–spheres. Matches from CAMA thus have equal or greater LRMSD than matches from MA, when using the same motif (MA disregards C–spheres). By our statistical model, greater LRMSDs generate greater p-values, and, if p becomes greater than α, a statistically significant match for a point-based motif becomes statistically insignificant when C–spheres are added. Thus, C–spheres convert FP matches under point-based motifs into TN matches with cavity-aware motifs by making some FP matches statistically insignificant.

Due to variations in active site structure, some functional homologs have atoms that occupy C–spheres, when the match and the motif are optimally superimposed. In our earlier experimentation, which we review in Section 4.2, we measure both the number of FP matches eliminated, as well as the number of TP matches lost by adding C–spheres. Given effective motifs, the number of TP matches lost is small in comparison to the number of FP matches eliminated.

3.3 Discussion and Contributions

One of the major strengths of MA is the modularity of the depth first search. By repeatedly testing and aligning partial matches, MA permits additional biological information to be integrated, such as priority ranks of functional significance and the C–sphere emptiness criteria. In the future, this modularity could permit additional geometric criteria and analyses to be tested, allowing for a wide range of additional modifications.

The geometric variation threshold ϵ affects the number of matches that

Seed Matching and Augmentation consider. If the match with lowest LRMSD has an alignment where the correlated motif and target point fall further than ϵ, then MA and CAMA cannot observe this match. As a result, as long as ϵ is set sufficiently high, most reasonable matches can be observed and considered, in the search for the single match with smallest LRMSD. This work uses $\epsilon = 7\text{Å}$. Setting ϵ to very large values obviously increases the number of matches which must be considered, and lengthens the runtime of CAMA or MA. However, this additional expense can contribute to making MA robust to variations in the input data, such as geometric variations in matching structures, that could occur because of protein flexibility, as well as variations in priority ranking.

3.3.1 Contributions

Point-based MASH and cavity-aware MASH present several novel contributions to the general problem of identifying matches for motifs representing known active sites. MASH is the first method to identify matches for motifs that combine point-based structure representations with priority ranking information. In this work, ranks were obtained using data from the Evolutionary Trace [26, 12]. In addition, our cavity-aware motifs are the first motifs to combine point-based and volumetric representations of protein structure, and MASH is also the first method for finding matches for these motifs. In addition, we have also contributed a method for integrating priority rankings and volumetric representations into depth-first-search comparisons of protein structure.

We have also presented a novel application of nonparametric statistical modeling to measuring the statistical significance of matches. Our data-driven model can be used to compute p-values which are specific to the motif used, and can be specialized to varying representative sets of protein structures. In the next chapter, we demonstrate that significant p-values can identify matches to functionally related proteins.

Chapter 4

Testing the MASH pipeline

This chapter demonstrates that both the point-based and the cavity-aware variations of the MASH pipeline are capable of identifying statistically significant matches to functionally related targets. In addition, we demonstrate that cavity-aware MASH transforms many statistically significant matches to functionally unrelated proteins into statistically insignificant matches.

4.1 Testing point-based MASH

To demonstrate the accuracy of MASH, we searched for motifs within structures of evolutionarily related proteins, and then verified that matches were found to actual cognate amino acids. We then measured the statistical significance of each match, and demonstrate that matches to cognate active sites tend to be statistically significant.

4.1.1 Input Data

Motif Points Our primary data (Figure 4.1) is 12 families of enzymes with known active sites. Each family is composed of a set of homologous sequences identified by BLAST, some of which have known structures in the Protein Data Bank [74] (PDB). Of the structures found, each family is assigned a *major* structure; the rest are *minor*. ET is applied on each family of sequences, and the significance ranks and labels generated are mapped onto the major structure for each family. Between 4 and 9 of the most functionally significant residues surrounding the active site on the major protein are selected, and their alpha carbons (C_α) become the points in the motif. (C_α) atoms were used in our motifs as preliminary data. Rather than debate the adequacy of

C_α atoms to represent function, we seek only to document the correctness of our techniques.

Functional Homologs We use targets identified by sequence similarity because each residue in the motif has a cognate residue in the target: we know what match to expect beforehand. Using functional analogs may seem more relevant for functional annotation, but successfully matching analogs would only demonstrate how well our motifs represent function. Because analogs lack easily identifiable cognate residues, using analogs would sacrifice precise verifiability.

{**16pk**, 1vpe, 1php} {**1bqk**, 8paz, 1aaj, 1aan, 1ag6, 1b3i, 1baw, 1bxa, 1bxv, 1paz, 1pza, 1pzb, 1pzc, 1zia, 1zib, 2plt, 2rac, 3paz, 1aac} {**1amk**, 1tpe} {**1aky**, 5ukd, 1qf9, 1uke, 1zin, 1zio, 1zip, 2ak2, 2ukd, 3ukd, 4ukd, 1ak2} {**1a6m**, 1ymc, 1dwr, 1dws, 1dwt, 1m6c, 1mbs, 1mno, 1mwd, 1myg, 1pmb, 1wla, 1ymb, 1azi} {**1a3k**, 1slt, 1sla, 1slc, 1qmj} {**1finA**, 1hcl, 1hck, 1b38} {**1ukrA**, 1xyn, 1xnb, 1yna} {**3lzt**, 2ihl, 2lz2, 1jhlA, 1ghlA, 1fbiX, 1lz3, 1hhl, 1jug, 2eql, 1gd6A, 1f6rA, 1hfx} {**7a3hA**, 1g01A,1egzA} {**1juk**, 1j5tA, 1i4nA} {**1f8eA**, 1nn2, 1nsbA}

Figure 4.1 : Families used in MASH experimentation.

PDB IDs of each family of homologs, bracketed to indicate membership. Bolded proteins are the major proteins, whose structures were used to construct motifs.

Implementation Specifics MA was implemented in C/C++. Code was prototyped and run on a desktop Athlon 1900MP. Experiments were run on Athlon 1900+ CPUs. GH and MA memory footprints varied between 5 and 20 megabytes, depending on input.

4.1.2 MA Identifies Cognate Active Sites

Our first goal is to demonstrate that MA is capable of identifying matches with cognate active sites. We search for each motif in the minor structures of the same family. These are homologous proteins (HPs). ET uses multiple sequence alignments, so a functional residue in one sequence correlates with cognate residues of related function, at the same position, in all sequences of the family. Thus we can verify MA: if we find a *cognate match* where the target points are cognate to the motif points, we have a correct match, residue by residue. For comparison, we also searched for each motif in the minor proteins of the other families. These proteins are not homologous (NHPs).

Observations In 69 out of the 73 motif-HP pairs (95.4%), MA matches 100% of the source motif with cognate residues in the target. Of the remaining four cases, two of the target structures (1m6c and 1mno) were experimental structures that had a point mutation that changed the label of residue 68 (in both cases) from a valine to an asparagine in order to over-stabilize oxygen binding in myoglobin (1a6m). As a result, the labels of the points corresponding to residue 68 in both 1m6c and 1mno were incompatible, and, correctly, the points were not matched. While this was not intended, it demonstrates the ability of our algorithm to eliminate potential matches with incorrect labels. In the other two cases, a match existed with lower LRMSD than the cognate match. These occurred between major protein 1amk with target 1tpe, and 1f8eA with 1nsbA. In each case the cognate match had a higher LRMSD (approx. .5Å) than the match MA identified. This is no fault of MA. Instead, it suggests that 1amk and 1f8eA are sub-optimal motifs that bear accidental similarity to functionally unrelated structures: Ideally, motifs should have structural similarity only with proteins with functional similarity. True failures of MA would be the opposite: We would return a match with LRMSD higher than the cognate match, showing that the cognate match was overlooked. This never occurs. From our experiments, we found that MA is accurate and efficient on biological data, identifying cognate residue correspondences, except when the motif bears incidental structural similarity to unrelated residues.

Matches between motifs and HPs tended to have lower LRMSDs than between the same motif and NHPs. This is apparent in Figure 4.2, which plots LRMSD for all matches found. 9 out of 12 motifs considered had matches

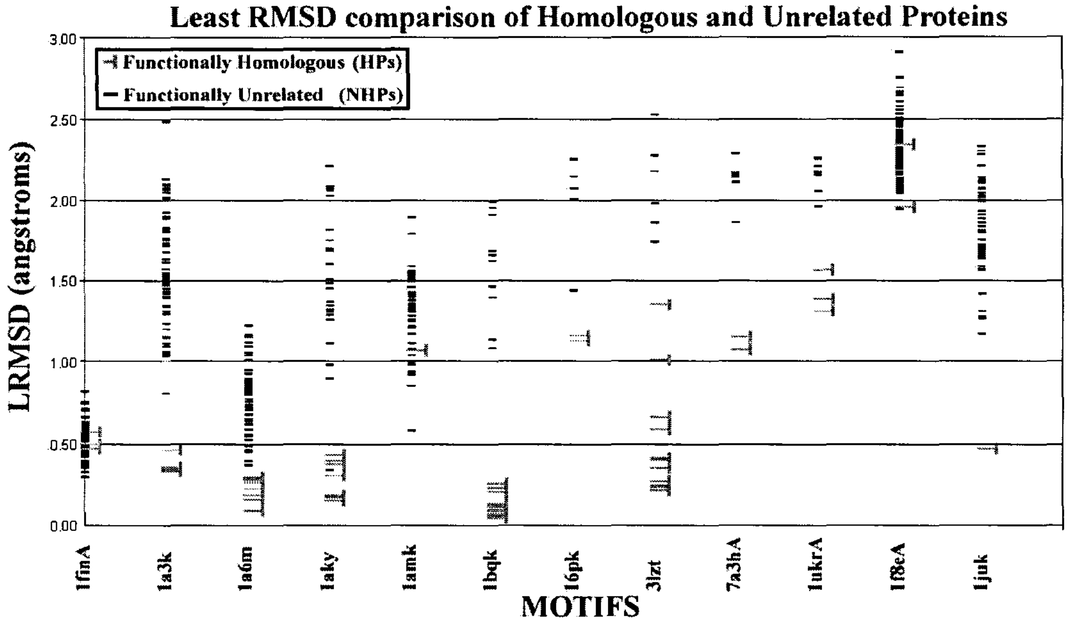


Figure 4.2 : Experimental Results: 12 motifs and 73 targets plotted by LRMSD

of HPs (Blue, Fig. 4.2) with LRMSD lower than most matches of NHPs (Red, Fig. 4.2). Two of the motifs breaking this trend were lamk and 1f8eA, motifs which had incidental similarity with functionally unrelated residues, suggesting again that these motifs are not specific representatives of function. The remaining motif, lfinA, was defined on a flexible active site, so cognate active sites, flexible themselves, had less geometric similarity.

4.1.3 Cognate Active Sites are Statistically Significant

Our second goal is to demonstrate that matches to cognate active sites can be statistically significant. We accomplished this by computing a motif profile for all motifs used, with a snapshot of the PDB from 8.17.2003. PDB files with multiple chains were divided into individual files, generating 55,305 structures. A handful of unparseable files were removed, and certain degeneracies were fixed, such as negatively indexed residues. We then calculated p-values for each match we found in the previous paragraph, to understand the relationship between statistically significant matches and matches to cognate active sites.

Observations The majority of p-values generated for HPs were between 1% and 0.01%. This is apparent in Figure 4.3, where we plotted p-values for all matches found. Most p-values generated for NHPs are above 10%. Notable exceptions are the p-values for matches of motifs 1amk and 1f8eA, which had accidental similarity to functionally unrelated structures. These had p-values above 10%. This verifies on a PDB-scale that 1amk and 1f8eA poorly represent functional sites: they have geometric and chemical similarity to 10% of all PDB proteins. The motif defined on 1finA, which had a flexible active site, also lacks statistical significance in its matches, because the geometry of functional residues may change relative to the motif. Matches of HPs represent identifiably significant structural similarity, except where the motif itself poorly represents protein function.

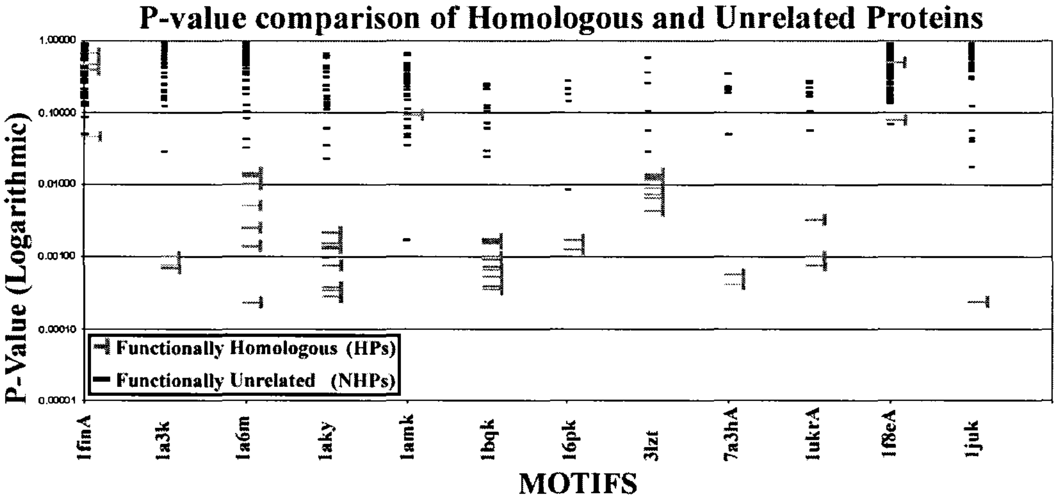


Figure 4.3 : p-values of LRMSDs from Figure 4.2 (log scale)

We also observed that computing motif profiles using random sampling directly improves performance. On average, brute force computation time was 12:48 (hrs:mins), while sampling took 0:38 on average. The best case fell from 2:40 to 0:08 and the worst case from 631:41 to 31:30. Sampling cuts runtime by almost exactly 95%. Sampling does efficiently estimate D_{S_i} without statistically significant loss of accuracy.

4.1.4 Discussion

On our data set of evolutionarily related proteins, our results show a correlation of statistically significant structural similarity to evolutionary relatedness between proteins, as long as the motifs properly represent function. This correlation indicates that statistically significant geometric and chemical similarity can be markers of cognate active sites. However, it should also be noted that there exist some statistically significant matches to functionally unrelated proteins. In particular, in a PDB-scanning scenario, where, given a motif, we seek to find all functional homologs based on statistically significant matches, 1% of the matches to the PDB will always be statistically significant. In most cases, the set of functional homologs is far fewer than 1% of the PDB, and in the cases of highly overrepresented protein families in the PDB, the set of functional homologs is greater than % of the PDB. The usage of statistical significance to identify functional homologs produces incorrect predictions, but can be very successful in dramatically narrowing the set of possible homologs.

4.1.5 A Performance Comparison of MA and Geometric Hashing

We compared performance to our implementation of Geometric Hashing (GH), as described by Rosen [44], because the source code is not available. All published heuristics compatible with our data were implemented. GH has been applied many times [41, 40, 43, 61], but cannot be prioritized as is the case with MA. Our performance comparison was run by comparing the amount of time necessary to compute matches between all motifs and all target structures.

Observations GH identified identical HP matches and similar NHP matches, but on our motifs of 4 to 9 motif points, and targets with 123 to 398 target points, MA was about 60 times faster. Average execution time was 6.195 seconds for GH, and only 0.103 seconds for MA using identical thresholds. Without loss of accuracy, Seed Matching narrows the search to matches of the highest ranking motif points, whereas GH considers all points equally.

Evolutionary prioritization seems to strongly improve performance.

4.2 Testing Cavity-Aware MASH

We begin by demonstrating that cavity-aware motifs using C–spheres centered on atoms of bound ligands and cofactors eliminate many FPs while preserving most TPs. In controlled experiments, we compared the ability of these two types of cavity-aware motifs to eliminate FPs and preserve TPs, relative to identical motifs without C–spheres.

4.2.1 Input Data

Motif Points The motifs used in this experimentation begin as 18 point-based motifs designed to represent a range of unrelated active sites in unmutated protein structures with biologically occurring bound ligands. These are documented in Figure 4.4. Earlier work has produced examples of motifs designed with evolutionarily significant amino acids [21, 13] and amino acids with documented function [19], so these principles were followed in the design of our point-based motifs. Amino acids for use in 10 of the motifs were selected by evolutionary significance, and are taken directly from earlier work [13], and the remaining 8 motifs were identified by functionally active amino acids documented in the literature (marked * in Figure 4.4).

The selection of motif points strongly influences motif sensitivity and specificity. In this work, we seek to demonstrate that adding C–spheres can improve point-based motifs. For this reason, we take the selection of motif points and the number of TP and FP matches found, for each point-based motif, as given. These values are provided in Figure 4.6.

C–Spheres C–spheres used in our experimentation were generated in two ways for each set of motif points, generating two variations we refer to as *ligand-based* and *space-filling* C–sphere designs. In general, a small number

Motifs Used in Experimentation

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

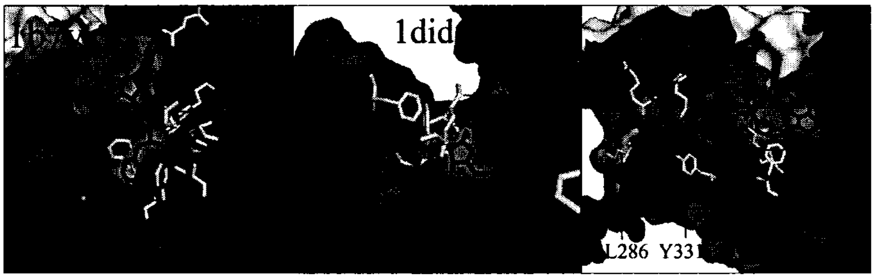


Figure 4.4 : Motifs used, with example diagrams below. Starred (*) motifs use functionally documented amino acids. The column marked "#C" denotes the number of C–spheres in each motif. "R" denotes the range of C–sphere maximum diameters (in Å) for the motif.

Ligands Used in Experimentation

PDB id	Ligand Used
16PK	Tetrafluorophosphopentylphosphonic Acid Adenylate Ester
1ADY	Histidyl-Adenosine Monophosphate
1ANI	Zinc Ion, Phosphate Ion
1AYL	Oxalate Ion, ATP
1B7Y	AMP
1CZF	N-Acetyl-D-Glucosamine, Zinc Ion
1DID	Manganase Ion, 2,5-Dideoxy-2,5-Imino-D-Glucitol
1DWW	Protoporphyrin Ix Containing Fe, NHA
1GGM	Glycyl-Adenosine-5'-Phosphate
1JA7	N-Acetyl-D-Glucosamine
1JG1	S-Adenosyl-L-Homocysteine
1KP3	ATP
1KPG	Selenomethionine
1LBF	1-(O-Carboxy-Phenylamino)-1-Deoxy-D-Ribulose- 5-Phosphate
1UCN	Phosphate Ion, Calcium Ion, ADP
2AHJ	Iron (Iii) Ion, Nitrogen Oxide, 1,4-Diethylene Dioxide, Zinc Ion
7MHT	S-Adenosyl-L-Homocysteine
8TLN	Dimethyl Sulfoxide, Zinc Ion, Calc Ion

Figure 4.5 : Chemical Names of Ligands Used in Experimentation

(usually 10) C–spheres were selected for each motif used in our experimentation. Ligand-based C–sphere designs center C–spheres at the coordinates of atoms in bound ligands and cofactors. For example, in Figure 3.3, we modeled the heme-dependent enzyme nitric oxide synthase, which catalyzes the synthesis of nitric oxide (NO) from an L-arginine substrate. This multi-step reaction takes place in a deep cleft and involves zinc, tetrahydrobiopterin, and hydride-donating (NADPH or H_2O_2) cofactors [87, 88]. Using PDB structure 1dww, we centered C–spheres at several atom coordinates on the heme, in order to fill the heme-binding cavity, and placed one C–sphere to represent tetrahydrobiopterin, which was further outside from the main cavity, as shown in Figure 3.3. In some cases, not all atoms of the ligand were used, such as in heme in Figure 3.3, but selections were made to approximate the shape of the ligand binding cavity based on the position of atom coordinates available.

Space-filling C–sphere designs seek to optimally represent the shape of the active cleft. This is accomplished by computing a Voronoi tessellation on the atom coordinates of the protein structure used to generate each motif. We use points in the Voronoi tessellation at the intersection of several planes, because they are equidistant to several atoms in the protein structure. This allows a sphere centered at these points to occupy maximum volume between points of the protein structure. For each atom in bound ligands and cofactors, we found the largest space-filling sphere containing it, and used that as a C–sphere.

For both C–sphere designs, the maximum radius of any C–sphere was the distance to the nearest atom in the protein structure used to generate the motif. C–spheres can have any radius smaller than the *maximum size.*

Functional Homologs In order to count TP and FN matches, it is essential to fix a benchmark set of functional homologs. We use the functional classification of the Enzyme Commission [89] (EC), which identifies distinct families of functional homologs for each motif used. Proteins with PDB structures in these families form the set of functional homologs we search for. Structure fragments and mutants were removed to ensure accuracy.

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

Implementation Specifics CAMA was implemented in C/C++. Code was prototyped on a 16-node Athlon 1900MP cluster and the Rice TeraCluster, a cluster of 272 800Mhz Intel Itanium2 processors. Final production runs ran on Ada, a 28 chassis Cray XD1 with 672 2.2Ghz AMD Opteron cores.

Empirical testing indicates that eliminating partial matches in the Augmentation phase causes CAMA to be approximately 3 times faster than testing C–spheres in complete matches.

4.2.2 C-Spheres Eliminate FPs, Preserve TPs

In this section, we demonstrate that C–spheres eliminate many FP matches while preserving most TP matches. It should also be noted that for any given point-based motif, infinite numbers of possible C–sphere centers and radii could be considered, and among these possibilities, some C–sphere configurations undoubtedly lead to more sensitive and specific cavity-aware motifs. Unfortunately, we cannot consider all possibilities. However, one natural question stands out: If C–spheres are made as large as possible within the active cleft, what is the effect on the elimination of FP matches?

We first demonstrate that C–spheres affect the elimination of FP matches and the retention of TP matches. We compared the number of TP and FP matches found with 18 point-based motifs to cavity-aware versions of the same motifs. The first cavity-aware version of these motifs uses ligand-based C–spheres, while the second version uses a space-filling C–sphere design, as mentioned in Section 4.2.1. Scaling C–sphere radii between 0 and maximum radius in 20 increments, we have a comparison experiment between our two C–sphere designs and point-based motifs.

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

Since matches to $S_{i_1}, S_{i_2}, \ldots, S_{i_{19}}$ have p-values greater than or equal to S_{i_0}, because they have C–spheres with non-zero radii, the number of FP and TP matches identified among $S_{i_1}, S_{i_2}, \ldots, S_{i_{19}}$ is less than or equal to that of S_{i_0}. The number of homologs matched by each point-based motif, S_{i_0}, is listed in the left of Figure 4.6. The number of TP and FP matches eliminated is calculated relative to the number matched by the point-based motif, and thus all S_{i_0} have 100% of TP and FP matches, as in the leftmost point of the graph in Figure 4.6. Second from the left, we plot the percentage of TP and FP matches retained among S_{i_1}, relative to S_{i_0}, for all i, and then average these percentages over all S_{i_1}. Continuing from left to right, we compute the average percentage of TP and FP matches, over all S_{i_2}, then all S_{i_3}, etc., again relative to S_{i_0}.

Observations Demonstrated in Figure 4.6, as C–sphere radius increases, for both C–sphere designs, the number of FP matches are reduced dramatically. C–spheres based on ligand and cofactor atoms eliminated very few matches until C–sphere radius increased to approximately 80% of maximum

Point-based Motif Perf.

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

Average Percentage of TP and FP Matches

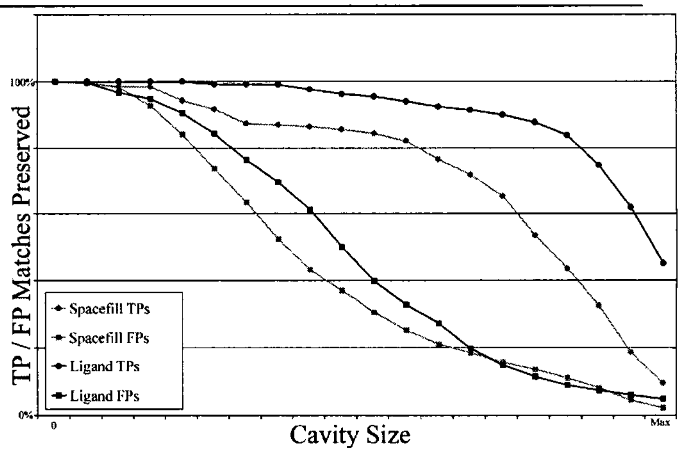


Figure 4.6 : Average effect of cavity-aware motifs on TP and FP matches, over all motifs. The horizontal axis charts C–sphere radius, where the radius of all C–spheres scales simultaneously from zero to individual maximum radius (see Section 4.2.2). The vertical axis charts the average percentage, per motif, of TP and FP matches remaining, relative to their respective point-based motifs. The number of homologs, and the number of TP and FP matches for each point-based motif is shown at left. FP matches are dramatically reduced while most TP matches are preserved, for both C–sphere designs. However, space-filling C–sphere designs tend to eliminate more FP matches and reject more TP matches, while ligand-based C–sphere designs tend to eliminate less FP matches, but preserve more TP matches.

radius, whereas C–spheres intended to maximally occupy the active cleft eliminated TPs more rapidly.

One motif, Phenylalanyl-tRNA Synthetase (1b7y), exhibited 0 sensitivity. The point-based version of 1b7y matched no functional homologs, so no cavity-aware motifs based on 1b7y matched any functional homologs either. For this reason, the percentage of TP matches eliminated by cavity-aware variations of 1b7y is undefined, and therefore no TP and FP data (for consistency) is included in the averages plotted in Figure 4.6. Cavity-aware variations on 1b7y still rejected more FPs as C–sphere radius increased. Point-based motifs from 1ja7 and 2ahj exhibited low sensitivity, identifying less than 20% of the total number of true positives. Having a flexible active site, cavity-aware variations of 16pk were significantly less sensitive than its point-based counterparts. Overall, cavity-aware motifs eliminate many FP matches, while preserving most TP matches.

C–spheres designed using ligand atom coordinates seemed to eliminate less matches in general than C–spheres designed to occupy maximum space in active cavities. These observations suggest that the latter C–sphere design was more strongly affected by the natural variation of active site structures in functionally related proteins. In combination with the earlier observation that C–spheres that preserved the most TP matches while eliminating the most FP matches were not the largest C–spheres, but instead around 80% of maximum radius, these observations emphasize the point that selecting C–spheres that occupy the most volume within the active cleft do not necessarily produce cavity-aware motifs that eliminate the most FPs and preserve the most TPs.

4.3 Discussion

The controlled experiments presented in this chapter, as well as in our earlier work [21, 38, 35, 13], demonstrate that point-based MASH and cavity-aware MASH are capable of identifying matches to cognate active sites, using motifs designed by hand. In the next chapter, we will demonstrate how we used MASH and cavity-aware MASH as a platform for developing techniques for refining motifs.

One way to improve matching predictions is to use evolutionary information, mapped onto target protein structures, to eliminate matches that correlate motif points to evolutionarily insignificant amino acids. By using matches to evolutionarily significant amino acids alone, many FP matches can be eliminated [13]. Another approach is to refine motifs so that they have minimum geometric and chemical similarity to functionally unrelated proteins, while maintaining similarity to functionally related proteins. We target this problem in the next section.

Chapter 5

Motif Profiling: A Motif Refinement Framework

The point-based and cavity-aware versions of the MASH pipeline provide a tool for identifying statistically significant matches to motifs designed with expert knowledge. However, experts select amino acids for motifs based on documented roles in biochemical function, disregarding the possibility that some geometric configurations of certain amino acids may recur frequently in the space of protein structures. Humans are simply unable to precisely perceive subtle recurring trends in thousands of protein substructures. While this approach guarantees that matches to the motif maintain geometric similarity to an active site structure with documented function, a property essential for motif sensitivity, expert design does not guarantee that many matches do not also exist to many functionally unrelated proteins, thus lacking motif specificity.

MP is an abstract method for refining motifs with distinct geometric and chemical variations. To our knowledge, MP is a completely unique method for refining motifs. MP uses an underlying algorithm for computing matches to implement the abstract algorithm, requiring only that the similarity of matches is measured with LRMSD. In this document, variations of MA serve this purpose very effectively, but other efficient algorithms, like Geometric Hashing [58] or JESS [25], could be used as well. MP is not dependent on any specific comparison algorithm, requiring only efficient implementations and compatibility with motifs types used.

5.1 The Motif Profiling Method

As input, algorithms implementing MP begin with a list of motifs $\{S_1, S_2, \ldots, S_n\}$ as input, and the set of known protein structures, Ω. We then use MA to compute matches between each input motif, and the set of known protein structures, computing motif profiles $\{P_{S_1}, P_{S_1}, \ldots, P_{S_n}\}$. Measuring the medians of each profile $\{med(P_{S_1}), med(P_{S_1}), \ldots, med(P_{S_n})\}$, we determine the motif with highest median S_a. S_a is the output of MP, because we consider S_a to have the greatest geometric and chemical dissimilarity to Ω – a property we call *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 [90]. 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. In the context of two applications of MP, we will also demonstrate that MP can be extremely effective in identifying motif refinements that yield high sensitivity and specificity.

Earlier in this document, we demonstrated that point-based and cavity-aware motif types can be effective for identifying cognate active sites. In both cases, the motifs used were based on expert knowledge from biochemical literature. In the remainder of this chapter, we describe two methods for refining these motifs. GS, which refines sections of amino acids for point-based motif design, and CS, which refines C–sphere definitions for cavity-aware motif design.

5.2 Geometric Sieving

We hypothesized that selecting functionally active amino acids that also exist in uncommon geometric configurations can yield sensitive and specific motifs. To test this hypothesis, we first designed GS, a geometric analysis that identifies patterns in uncommon geometric configurations by measuring Geometric Uniqueness. We will apply GS to automatically refine motif designs, reducing the dependence of point-based MASH on experts, and simultaneously improving the design of motifs.

GS accepts an input set, a collection of candidate motif points that 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. The output of GS is the subset motif with k points that has highest Geometric Uniqueness. Combined with point-based MASH, GS provides a pre-processing stage for motif refinement that improves sensitivity and specificity.

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 that 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 [21]. With many different combinations of motif points to choose from, in the form of varying subset motifs, we can select the motif profile that 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.

5.2.1 The Geometric Sieving Algorithm

GS has two phases: GATHER and ANALYZE, which are described in Algorithms 1 and 2. Ignoring the optimization step in Algorithm 1 for now, the GATHER 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 ANALYZE phase. This phase 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.

Algorithm 1 Gather

```
Input: Input Motif $S$
Input: Optimized motif size $k$
for each $T_i$ in $\Omega_5$ 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 $S'_\Omega$
  end for
  ELIMINATE (optimization step)
end for
```

Algorithm 2 Analyze

```
Input: all motif profiles $S'_\Omega$
from GATHER phase
Calculate $m(S'_\Omega)$ for all $S'_\Omega$
Find the motif profile $S'_\Omega$
with highest $m(S'_\Omega)$
Output: $S'$, the optimized motif
```

The GATHER 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 GATHER, and offloads MA problems to the clients.

One problem with offloading MA problems to client machines is the problem of data locality. We are constantly computing matches between different motifs and different targets, on each client. When we originally implemented this software, all motifs were stored locally in client system memory, but targets were stored as individual files on a globally accessible file server. File server load on larger runs with hundreds of simultaneous clients bottlenecked performance at the file server. Our initial solution was to store the files locally on each client in temporary disk space, but, given that each PDB structure is represented as a single file, simple act of copying the PDB from the file server to all client machines caused the file server to crash. We resolve this problem by developing a binary representation of the entire PDB in a single file, compressing nearly seven gigabytes of data into approximately 1.4 gigabytes. We now read this file from local disk when initializing the distributed system, so all clients have low-latency access to all targets at all times. This resulted in an estimated 10 to 20 fold speedup on machines that could hold the entire binary PDB in memory, and vastly improved quality of service for other users on the development cluster. Further implementation details are available in Chapter 6.

5.2.2 Accelerating GS

Further modifications to GS can increase performance. In particular, let us now consider the optimization procedure ELIMINATE (Algorithm 3) which is called from GATHER. Note that when we call ELIMINATE during GATHER, all motif profiles are only partially computed. Eventually ANALYZE will identify the optimized motif by selecting the motif profile that has the highest median. A closer look at the computations happening during GATHER 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 ELIMINATE (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 GATHER. ELIMINATE need not be applied at every iteration of the outer loop of GATHER, as it will have a limited effect. Instead, we define a parameter called the *step size* and we call ELIMINATE after *step size* iterations of the outer loop of GATHER.

Algorithm 3 Eliminate

```
  Input: all motif profiles $S_\Omega$ from GATHER phase
  Calculate $r(S_\Omega)$ for all $S'_\Omega$
  Among all $r(S_\Omega)$, find $l$
  eliminate all $r(S'_\Omega)$ with $u < l$
  return to GATHER
```

As we pointed out above, when we call ELIMINATE during GATHER (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, ELIMINATE computes a 95% Confidence Interval $r(S''_\Omega)$ (see method of Efron and Tibshirani [91, 92, 93]), 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 ELIMINATE 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 GATHER, we do not calculate matches for eliminated subset motifs. If only one subset motif remains, or if GATHER 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.

5.2.3 Discussion

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 ELIMINATE to guide program logic, this can lead to accidental elimination of a subset motif. In order to reduce this possibility, ELIMINATE can be applied in a more adaptive manner, such as by running ELIMINATE less often when motif profiles have few samples.

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 [21]. 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.

5.3 Cavity Scaling

We have observed that the selection of C–sphere positions and radii can drastically affect the number of TP and FP matches eliminated, significantly influencing the effectiveness of some cavity-aware motifs. Some *high-impact* C–spheres have greater impact on FP match elimination than other *low-impact*

C–spheres. Without a method for identifying high-impact C–spheres, some cavity-aware motifs may not reduce as many FP matches as others, diluting the impact of adding volumetric representations to point-based motifs.

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

5.3.1 Markers of High-impact C–spheres

We have observed that motif profiles derived from cavity-aware motifs that include certain C–spheres have a tendency of shifting towards higher LRMSDs as C–sphere radius increases. Figure 5.1a demonstrates motif profiles computed with a motif that has exactly one C–sphere. Each motif profile corresponds to identical motif points with a C–sphere at an identical position, where the only difference is that radius changes evenly between zero and the C–sphere's maximum radius. As size increases, the motif profile changes very little. In comparison, in Figure 5.1b, for the same motif points and a C–sphere in a different position, as radius changes uniformly between zero and the C–sphere's maximum radius, many more matches shift towards higher LRMSDs, as mentioned in Section 3.2.3. High-impact C–spheres cause cavity-aware motifs to become more Geometrically Unique.

As matches shift towards higher LRMSDs, according to our statistical model in 3.2.4, statistically significant matches become statistically insignificant. This causes FP matches, which make up the dominating majority of matches computed in a motif profile, as mentioned in Section 3.2.4, to become TN matches. Therefore, C–spheres that cause more substantial shifts towards higher LRMSDs, as radius increases, cause more FP matches to become TN matches, relative to C–spheres that cause less substantial shifts in LRMSD. C–spheres that cause substantial shifts towards higher LRMSDs, therefore,

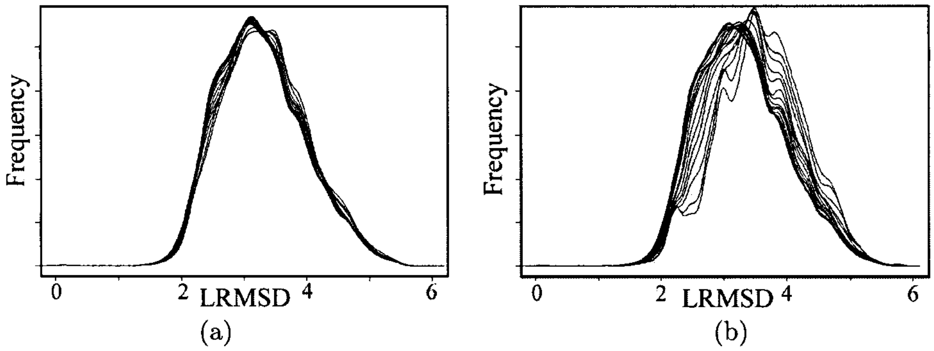


Figure 5.1 : Motif profiles for a low-impact C–sphere (a), and a high-impact C–sphere (b), as radius increases. For clarity, we provide 20 motif profiles for each C–sphere, showing how much the motif profile changes for a high-impact C–sphere. CS normally inspects only the motif profile with no C–spheres (the profile at the furthest left in both (a) and (b), and the motif profile corresponding to the C–sphere at maximum radius, at the furthest right in both (a) and (b).

are high-impact C–spheres. This is the primary principle which allows CS to distinguish high-impact C–spheres from low-impact C–spheres.

5.3.2 The Cavity Scaling Algorithm

As diagrammed in Figure 5.2, CS independently examines motif profiles for each C–sphere of the input, identifying which C–spheres are high-impact. We measure changes in motif profiles by comparing the median LRMSD, in order to distinguish shifts towards higher LRMSDs. Given an input motif S and one of its C–spheres, c_i, CS generates a variation of S which has no C–spheres, called S_p. Using S_p, CS applies CAMA to compute a motif profile against the PDB, which we call P_{S_p}. We then generate another variation of S, called S_{c_i}, that has only C–sphere c_i at its maximum radius, and compute a motif profile of S_{c_i}, called P_{c_i}, against the PDB. Comparison of the medians of P_{S_p} and P_{c_i}, $med(P_{S_p})$, and $med(P_{c_i})$, respectively, determines if c_i is a high-impact C–sphere. In order to determine if $med(P_{S_p})$, and $med(P_{c_i})$ vary substantially enough to identify c_i as a high-impact C–sphere, we used a simple empirical threshold of .5 LRMSD. An alternative threshold can be computed using confidence thresholds from a method of Efron and Tibshirani [91, 92, 93].

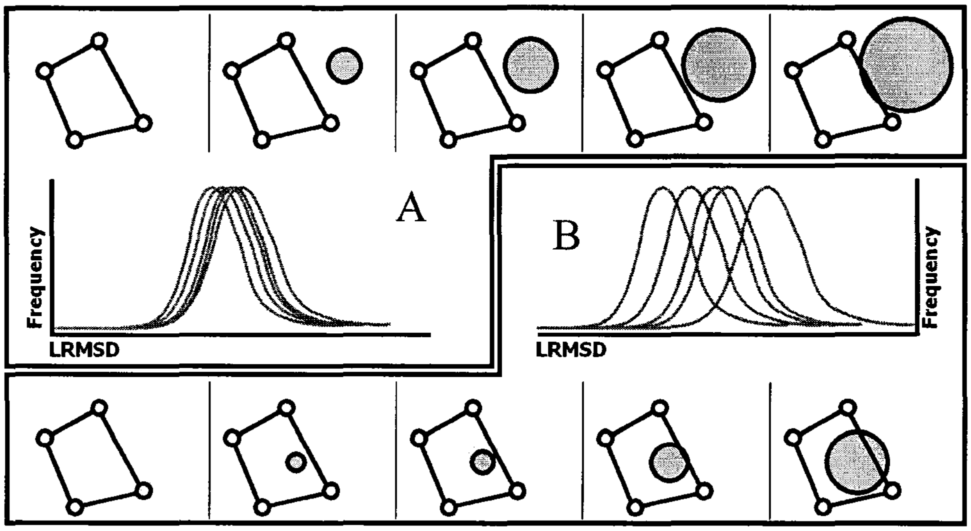


Figure 5.2 : How CS detects low-impact C–spheres (a) and high-impact C–spheres (b). Motif profiles corresponding to high-impact C–spheres vary significantly in their medians as C–sphere radius increases. Medians for low-impact C–spheres vary little.

As we will verify in our experiments, refined cavity-aware motifs eliminate most FP matches and maintain TP matches in comparison to manually defined cavity-aware motifs. In the future, this could be applied at a larger scale to explore more general representations of cavity-aware motifs, and provide feedback about C–sphere placements in motif design. CS only tests existing C-spheres to determine which are high-impact, and does not address the problem of finding high-impact C–sphere positions from the general set of all possible C–sphere positions. This is a subject of continuing investigation.

5.4 Discussion and Contributions

Geometric Uniqueness is a discriminating factor for identifying motifs that differ substantially from all known protein structures. Even though measuring Geometric Uniqueness is computationally expensive, the process is embarrassingly parallel and can be made efficient on clusters of networked machines.

More importantly, as demonstrated by CS, Geometric Uniqueness reflects criteria that affect the set of matches found without being customized for novel criteria. As additional biological information is used to identify other matching criteria, Geometric Uniqueness can continue to reflect the LRMSD of matches found.

MP is a technique that depends fundamentally on the existence of large amounts of data. As shown in Figure 6.11, biological matching criteria such as C–spheres can eventually eliminate so many matches that very few matches are found. Additional criteria could cause the measurement of the median to become ineffective, by destroying the monomodality of the motif profile, or by eliminating nearly all matches.

5.4.1 Contributions

MP is a novel and elegantly simple approach for refining motifs. While one method, MultiBind [17], is applicable to the refinement of point-based motifs, MP and the concept of Geometric Uniqueness are more general tools that can be applied in at least two settings, as we have demonstrated with GS and CS. MP is also the first fully automated algorithm for motif refinement, reducing dependence on expert knowledge. Finally, MP contributes a novel application of statistical median estimation to increase efficiency. In the next chapter, we will demonstrate experiments where MP identifies sensitive and specific refinements of several input motifs.

Chapter 6

Experimentation on Motif Profiling Methods

We have described GS and CS, two methods for refining motifs based on the abstract method of MP. In this chapter, we demonstrate that MP indeed identifies sensitive and specific motifs.

First, we show 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 that 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 GS, tested in the point-based 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.

Second, we perform a detailed analysis of each C–sphere in 18 cavity-aware motifs. CS identifies high-impact C–spheres, and high-impact C–spheres eliminate more FP matches than low-impact C–spheres. Using CS to refine our C–sphere selections, we produced refined motifs that we tested in the cavity-aware MASH pipeline. We observed that refined cavity-aware motifs preserve more TPs than our hand-designed motifs, while still eliminating many FPs.

6.1 GS Identifies Sensitive and Specific Motifs

In the previous section, we demonstrated that the point-based MASH pipeline was able to identify cognate active sites in functionally related proteins. Point-based MASH provides a foundation for the experimentation in this section, where we apply GS to refine motifs. In this section, we demonstrate that critical decisions in the design of GS permit GS to operate effectively and efficiently. We then demonstrate that optimized motifs generated by GS are in fact sensitive and specific, using the point-based MASH pipeline, and likely to contain evolutionarily significant and functionally documented amino acids.

We observed first that median LRMSD varies distinctly in the motif profiles of subset motifs derived from the same input set, and that changes in motif profiles correlate strongly with changes in median LRMSD. These differences in median LRMSD demonstrate that Geometric Uniqueness is adequately measured by the median LRMSD. We also observed that applying median estimation can substantially reduce the computation time necessary for identifying the most Geometrically Unique subset motif. These observations, on our small data set, demonstrate that GS is able effectively distinguish and isolate potential optimized motifs.

Then we test the sensitivity and specificity of refined motifs by comparing their performance to all possible refinements of the same input set, demonstrating that the subsets identified by GS are among the most sensitive and specific refinements possible. In addition, we will also study the importance of functionally documented, evolutionarily significant, and evolutionarily insignificant amino acids, and their impact on Geometric Uniqueness in motif refinements. These experiments demonstrate that GS identifies sensitive and specific optimized motifs, corroborating existing intuitions on motif design, which incorporate functionally documented and evolutionarily significant amino acids.

6.1.1 Primary Data

Input Sets The input sets chosen for this work were taken from ten well-studied proteins, listed in Figure 6.1. 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 6.2, 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 [94].

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

Figure 6.1 : Proteins used to test GS.

Selection Criteria Earlier work has produced examples of motifs designed with evolutionarily significant amino acids [21] and amino acids with documented function [19], which were sensitive and specific. Inspired by these approaches, we selected evolutionarily significant (^{E}, in Figure 6.2) and functionally documented (^{D}, in Figure 6.2) amino acids for each of our ten input sets, except Lysozyme (3lzt). Functionally documented amino acids are listed in Figure 6.5. We also included evolutionarily insignificant amino acids (^{I}, in Figure 6.2), 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,

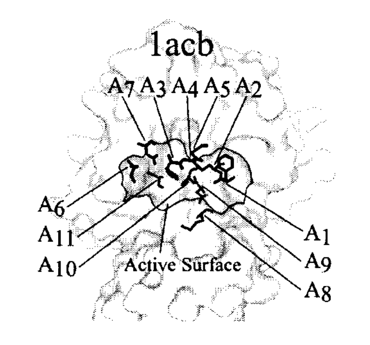
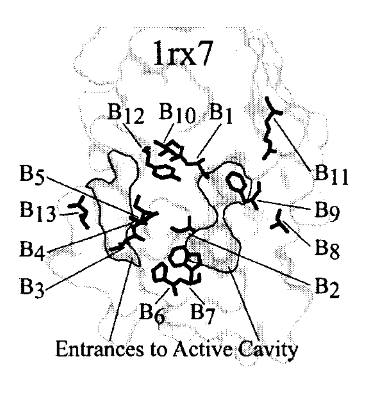
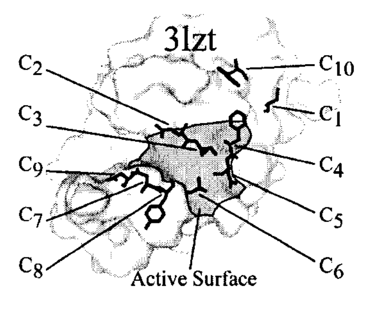
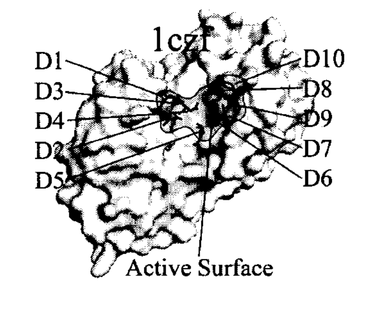
Diagram	tag	AA	#	Rank
 1acb A7 A3 A4 A5 A2 A6 A11 A10 Active Surface A1 A9 A8	A1	F^I	41	47.91
	A2	C^E	42	3.97
	A3	H^D	57	7.22
	A4	C^E	58	3.97
	A5	G^I	59	38.39
	A6	S^I	96	73.41
	A7	D^D	102	1.90
	A8	M^I	192	29.96
	A9	D^E	194	3.10
	A10	S^D	195	1.93
	A11	S^E	214	2.03
 1rx7 B12 B10 B1 B11 B5 B13 B9 B4 B8 B3 B2 B6 B7 Entrances to Active Cavity	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
 3lzt C2 C10 C3 C1 C9 C4 C7 C5 C8 C6 Active Surface	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
 1czf D1 D10 D3 D8 D4 D9 D2 D7 D5 D6 Active Surface	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

Figure 6.2 : Input sets used to test GS I

"AA": amino acid type; "#": PDB residue number; "Rank": ET rank.

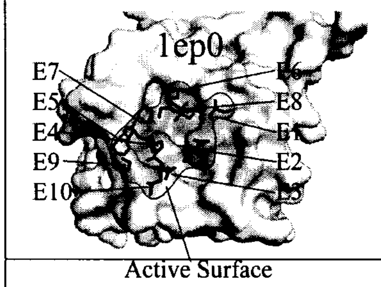
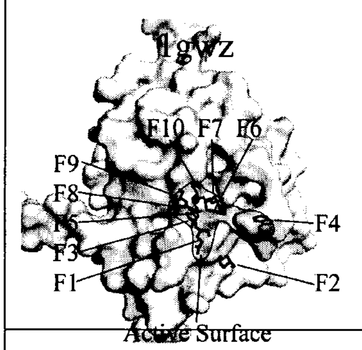
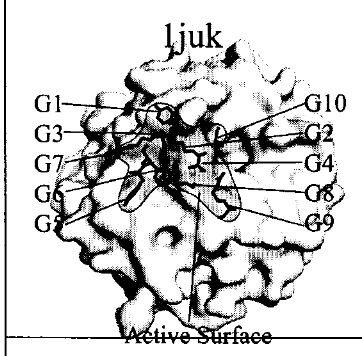
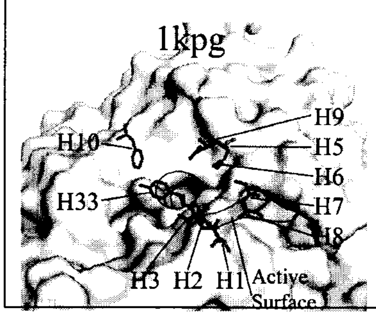
	tag	AA	#	Rank
Diagram  1ep0 E7 E5 E4 E9 E10 E6 E8 E1 E2 E3 Active Surface	E1	S^D	53	5.32
	E2	R^D	61	3.71
	E3	L^I	63	14.53
	E4	H^D	64	3.08
	E5	F^I	65	17.47
	E6	K^E	73	1.00
	E7	R^E	90	1.00
	E8	I^I	114	14.60
	E9	G^I	146	19.85
	E10	D^E	172	2.56
 1gwz F10 F7 F6 F9 F8 F5 F3 F1 F4 F2 Active Surface	F1	Q^E	327	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
 1juk G1 G3 G7 G6 G5 G10 G2 G4 G8 G9 Active Surface	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
 1kpg H10 H33 H9 H5 H6 H7 H8 H3 H2 H1 Active Surface	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

Figure 6.3 : Input sets used to test GS II

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

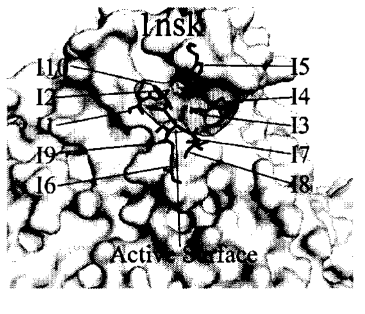
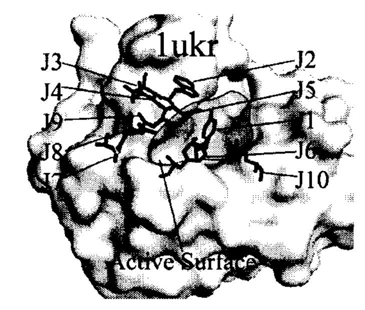
Diagram	tag	AA	#	Rank
 1nsk I1 I2 I3 I4 I5 I6 I7 I8 I9 Active Surface	I1	I^I	9	21.28
	I2	A^I	10	21.64
	I3	K^{DE}	12	2.51
	I4	P^E	13	4.16
	I5	Y^D	52	6.57
	I6	R^D	105	3.94
	I7	N^{DE}	115	3.39
	I8	I^I	116	22.74
	I9	I^I	117	19.26
	I10	W^D	118	4.80
 1ukr J3 J4 J9 J8 J7 J2 J5 J1 J6 J10 Active Surface	J1	Y^{DE}	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

Figure 6.4 : Input sets used to test GS III

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

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.

PDB ID	Amino Acids and Citations	EC class	Subset Size
1acb	Ser195 His57 Asp102 [95]	3.4.21.1	7
1rx7	Trp22 [96], and Gly15, Asp27, Phe31, His45, Ile50, Gly96 [97]	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 [98]	3.2.1.15	6
1ep0	Ser53, Arg61 and His64 [99]	5.1.3.13	6
1gwz	His454, Cys455, Ile459, [100]	3.1.3.48	6
1juk	Lys53, Lys110, Arg182, Gly233 [101]	4.1.1.48	6
1kpg	Gly72, Gly74, GLN99, Tyr33 [102]	2.1.1.79	6
1nsk	Lys12, Tyr52, Arg105, Asn115, His118 [103]	2.7.4.6	6
1ukr	Tyr70, Trp72, Glu79, Tyr81, Gln129, Glu170 [104]	3.2.1.8	6

Figure 6.5 : Functionally documented amino acids

Amino acids with documented function (and 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.

Using 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 6.2).

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 [89] (EC), which identifies families of functional homologs for each input set used (see Figure 6.5). 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 $H(S_i)$.

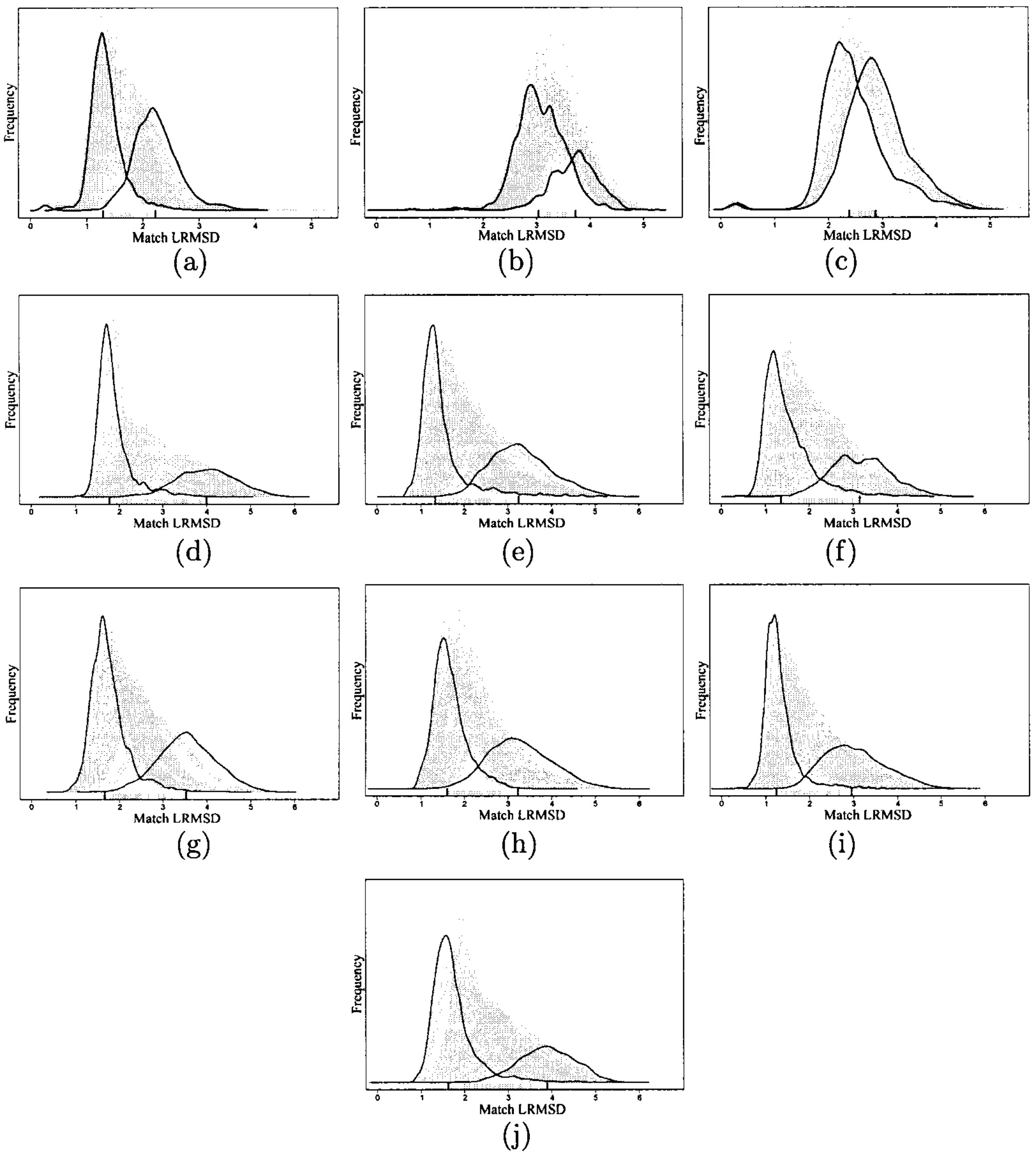
The Protein Data Bank In this paper, we use Ω_5, as mentioned in Section 3.1.4, which is sampled from the set of crystallographic protein structures in the PDB on Sept 1, 2005. PDB entries with multiple chains were divided into separate structures, producing 79322 structures. While this could prevent 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 [105] (MPI) protocol for interprocess communication, and prototyped on a cluster of 16 dual Athlon 1900MP machines with 1 gigabyte of RAM. Final data was run on the Rice Terascale Cluster (`http://www.rtc.rice.edu/`), a gigabit ethernet network of 140 dual Itanium2 machines, each running at 900Mhz, with 2 gigabytes of RAM per machine. Final data was also run on Ada, an experimental 28 chassis Cray XD1 with 672 2.2Ghz AMD Opteron cores. Each chassis on Ada is configured with six individual machines on a unified power source and an infiniband backbone, and each machine has two dual-core processors and 8 gigabytes of memory. The parameter ϵ, described in Section 2.2 was set to 7Å.

6.1.2 Median LRMSD Differentiates Motif Profiles

As mentioned in Section 6.1.1, 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.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 6.6 : Motif profile variation among different subset motifs

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.

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 that 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 6.6) 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 6.6, 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.

6.1.3 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 that need to be calculated before the optimized motif is identified. Using input sets from Section 6.1.2, we first generated matches without using the ELIMINATION optimization, mentioned in Section 5.2. Next, we repeated this calculation with the ELIMINATION optimization, with step sizes of 100 and 500, to stop sampling on motif profiles that 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, 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 Terascale Cluster, 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 that, with high probability, do not have the highest median. This is apparent in the number of matches necessary: 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

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

Figure 6.7 : 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.

less matches than GS without median estimation. Figure 6.7 describes the precise number of matches and time consumed.

Median estimation is very accurate. In every case described in Figure 6.7, 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 5.2.1) 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 6.1.4). 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.

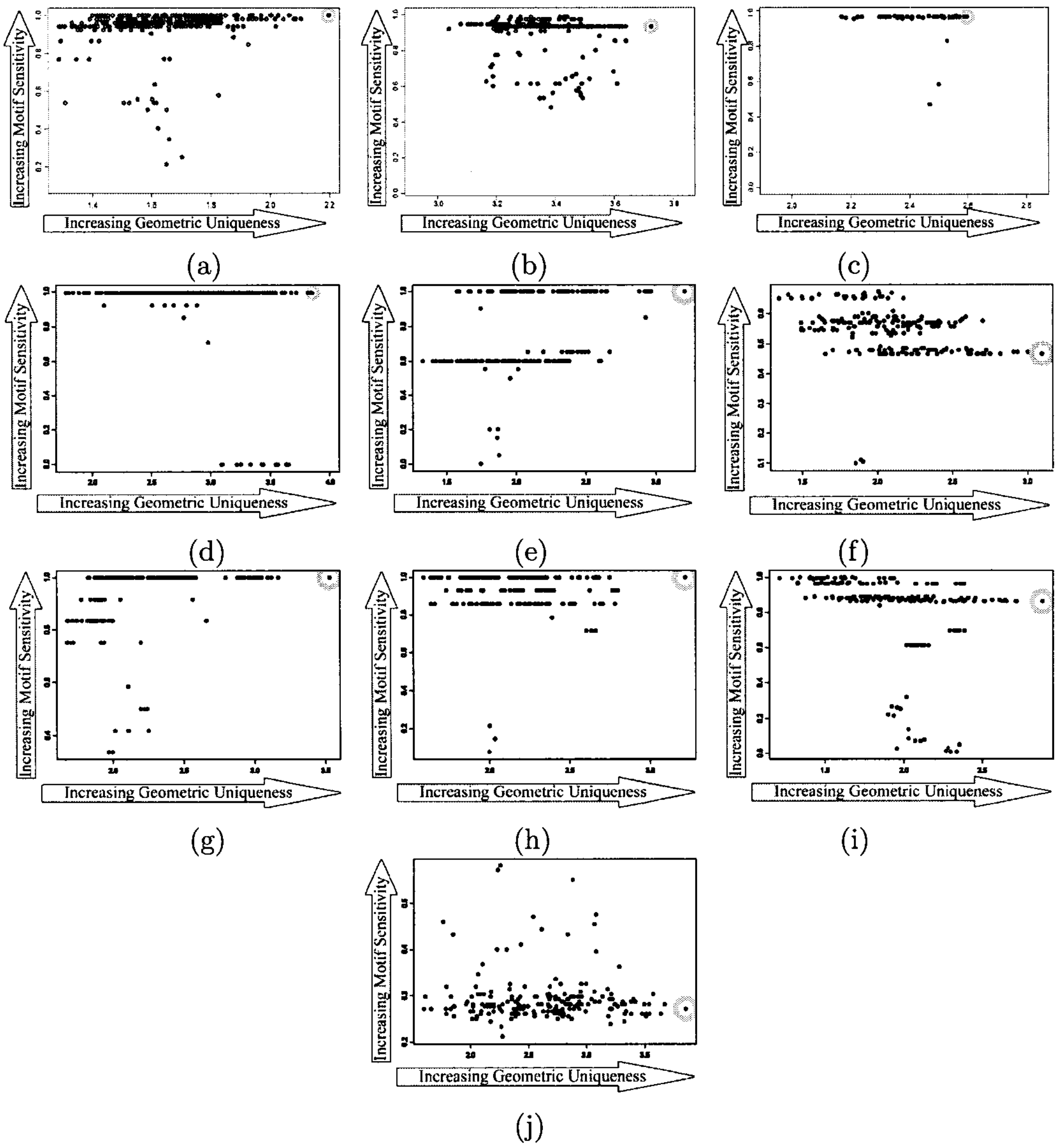
6.1.4 Optimizing Geometric Uniqueness Improves Motif Effectiveness

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 point-based 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 that, used with the point-based MASH pipeline, were quite sensitive at a high level of specificity (see Figure 6.8). 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 6.9 demonstrates the spectrum of sensitivity among the subset motifs observed. It is apparent that the sensitivity displayed by different subset motifs is radically affected by the selection of amino acids.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

Figure 6.8 : Comparison of subset motifs sensitivity I

Sensitivity (vertical axis) of (a) 1acb, (b) 1rx7, (c) 3lzt, (d) 1czf, (e) 1ep0, (f) 1gwz, (g) 1juk, (h) 1kpg, (i) 1nsk, (j) 1ukr, vs median LRMSD (horizontal axis). The most geometrically unique subset motifs, circled in grey, tended to be among the most sensitive, except in the case of beta-Xylanase (1ukr), where no subsets of the motif were very sensitive.

Sensitivity	1acb	1rx7	3lzt	1czf	1ep0	1gwz	1juk	1kpg	1nsk	1ukr
Max	100%	98.7%	96.7%	100%	100%	67.4%	100%	100%	100%	58.4%
Avg	94.2%	90.4%	93.4%	93.8%	75.5%	51.2%	93.9%	93.4%	81.7%	29.2%
GS	100%	93.3%	96.3%	100%	100%	46.6%	100%	100%	86.3%	27.0%

Figure 6.9 : Comparison of subset motifs sensitivity II

The table above 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%.

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 6.9. The two input sets that 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.

6.1.5 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 6.10. Median LRMSD was plotted on the horizontal axis.

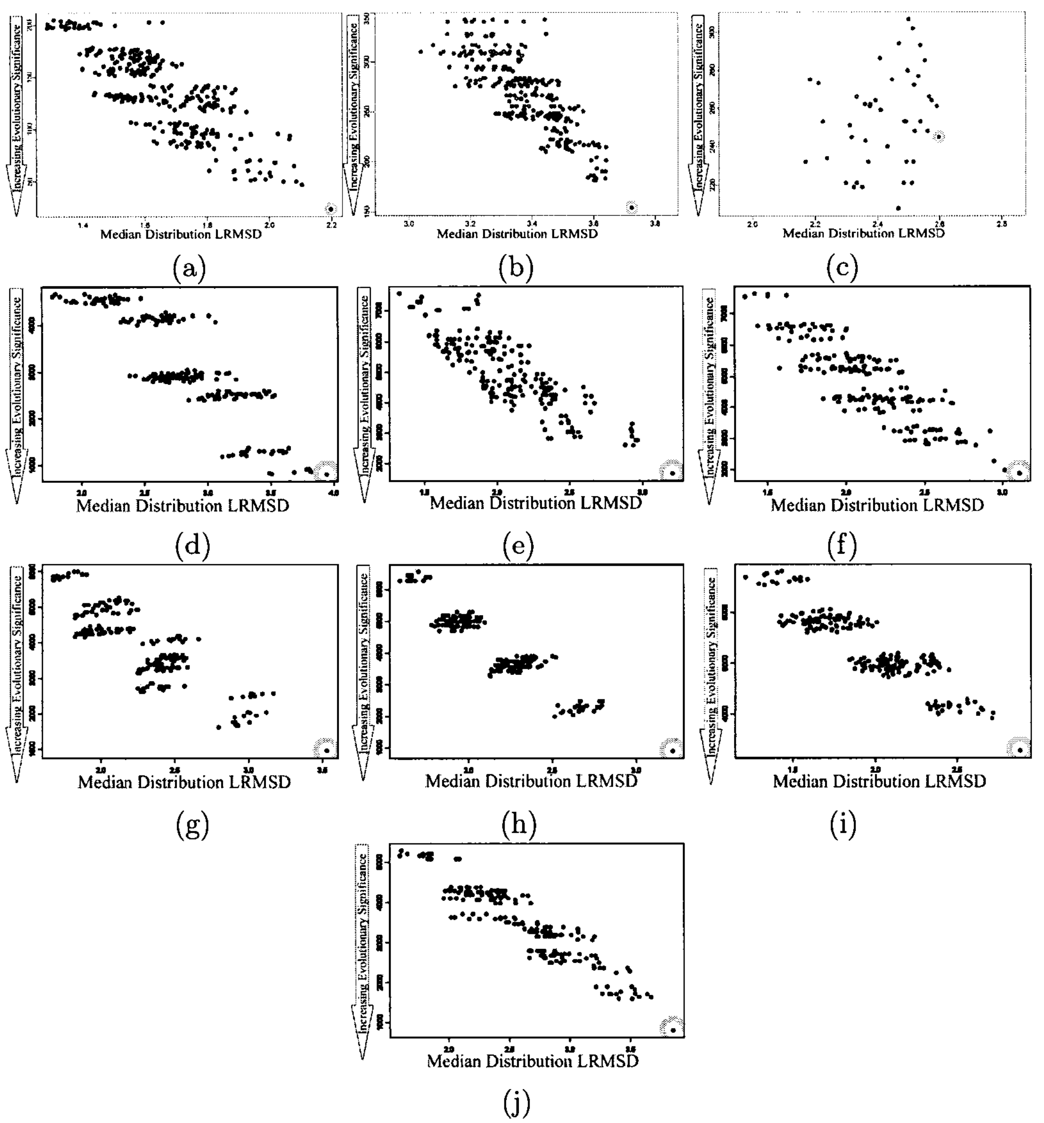


Figure 6.10 : Relationship between geometric uniqueness and evolutionary significance

Relationship of Geometric Uniqueness (horizontal axis) to Evolutionary Significance (vertical axis) in (a) 1acb, (b) 1rx7, (c) 3lzt, (d) 1czf, (e) 1ep0, (f) 1gwz, (g) 1juk, (h) 1kpg, (i) 1nsk, (j) 1ukr. Geometrically Unique subset motifs tended to be evolutionarily significant.

Observations Motif profiles with the highest median corresponded to the subset motif with the most evolutionarily significant amino acids (grey circles in Figure 6.10). 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.

6.1.6 Discussion

In this section, we presented experimentation using GS, a novel distributed algorithm for exhaustively refining input sets of candidate motif points into optimized motifs used in point-based MASH. We have implemented GS with techniques and optimizations 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 point-based MASH, 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 that 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 point-based MASH, and other pipelines using point-based motifs, in the form of a preprocessing step.

6.2 Cavity Scaling Identifies Effective Cavity-Aware Motifs

In the previous chapter, we described how CS uses MP to identify high-impact C–spheres. In this section, we provide verification for these claims by first demonstrating the distinct correlation between high-impact C–spheres and changes in the median LRMSD of motif profiles, as C–sphere radius increases. We then demonstrate that cavity-aware motifs that have been refined using CS preserve more TP matches and eliminate nearly as many FP matches, as C–sphere radius increases. Our results are computed on the same input data as in Section 4.2.1, so we do not repeat the description of this data set here.

6.2.1 Analysis of Individual C–spheres

Some C–spheres have a greater impact on FP match elimination than other C–spheres. We performed CS on each C–sphere in each of our 18 motifs, identifying which C–spheres were high–impact. 1ayl, used in Figure 6.11 is an excellent example, having several high- and low-impact C–spheres. All motifs had related behavior: Some motifs had many high-impact C–spheres, and others (1czf, 16pk, 8tln) had none, but significant increases in motif profile medians remained correlated to the elimination of FP matches in all examples.

Observations Motif profiles of some single-C-sphere motifs, computed over increasing radii, shift significantly in the median towards higher LRMSDs. These single-C-sphere motifs eliminate more FP matches as radii increase. Alternatively, motif profile medians of other single-C-sphere motifs that do not eliminate many FP matches also do not shift towards higher LRMSDs as radii increase. This is apparent in Figure 6.11, where we detail this effect for single C-sphere motifs based on 1ayl. In the inset graphs, identical copies of the 1ayl motif that contain only C-spheres 4 or 6 undergo significant changes in motif

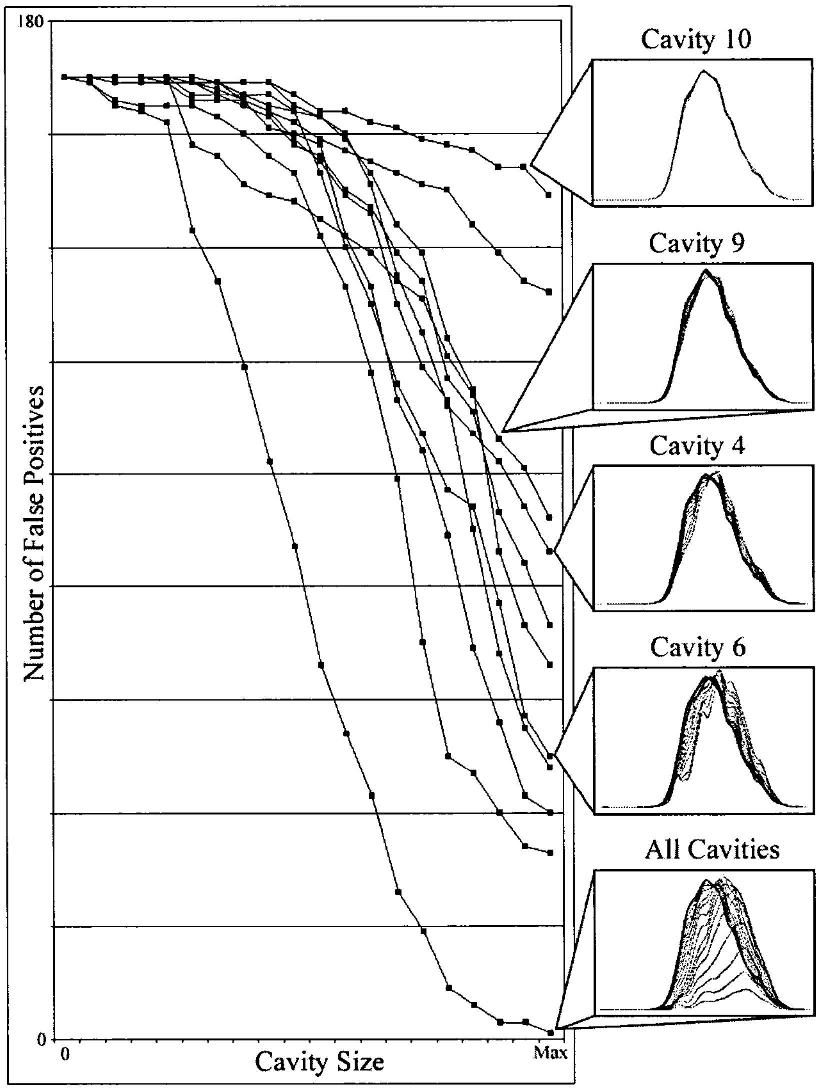


Figure 6.11 : Effect of individual C–spheres on motif specificity

As C–sphere size uniformly increases, as described in Section 4.2.2 (horizontal axis), some high-impact C–spheres, such as 4 and 6, eliminate more FP matches (vertical axis) than others, such as 10 and 9. Line plots show the number of remaining FP matches for a specific single-C-sphere motif, and for a motif containing all C-spheres. C-sphere positions relative to cavity shape are illustrated in the inset graphic. High-impact C–spheres, such as C–sphere 6, generate motif profiles whose medians shift towards higher LRMSDs as C–sphere radius increases. Other C–spheres, which do not eliminate as many FP matches, such as C–sphere 10, do not affect motif profiles as much. CS identifies C–spheres that eliminate more FP matches.

profile medians, towards higher LRMSDs, as radius increases. Simultaneously, as seen in the main graph, these single-C-sphere motifs, containing only C-sphere 4 or 6, rapidly eliminate FP matches. 1ayl motif copies with only C-spheres 9 or 10 experience insignificant changes in motif profile medians, and also eliminate FP matches more slowly, as radius increases. C-sphere positions relative to active site geometry are provided in the inset graphic in Figure 6.11. No correlation between high-impact C–spheres and cavity topography was apparent, emphasizing the difficulty of designing motifs with high-impact cavities.

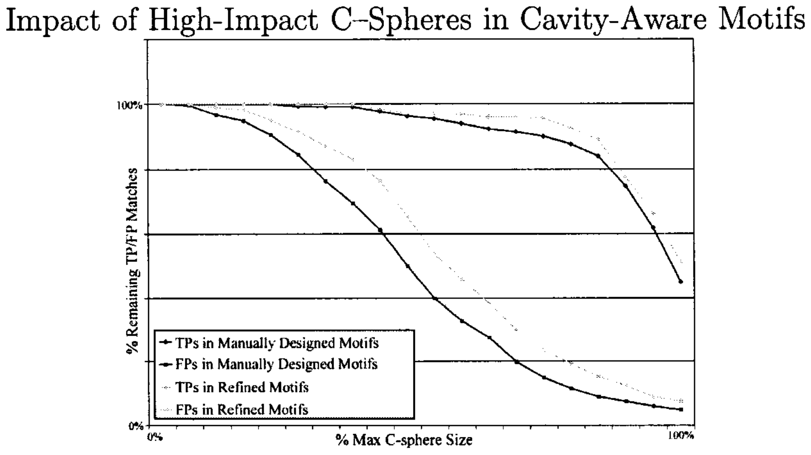
Motifs with only one C–sphere eliminate very few TP matches, but careful inspection indicates that individual cavities cause different TP matches to be rejected. This effect accumulates into the slow loss of TP matches observed in section 4.2.2.

6.2.2 Automatically Refined Cavity-aware Motifs

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

Experiment We applied CS on every C–sphere in every motif, and identified a set of high-impact C–spheres for all motifs except 1czf, 16pk and 8tln. We repeated the experiment described in Section 4.2.2 for the remaining motifs, using only high-impact C–spheres. We refer to these as automatically refined motifs. We compared our results to manually designed motifs used in Section 4.2.2, which contained all C–spheres.

Observations Like the axes of Figure 4.2.2, Figure 6.12 plots percent of maximum size (horizontal axis) versus the average percent of remaining TP and FP matches (vertical axis). Automatically refined cavity-aware motifs reject a large majority of FP matches, retaining a few more than manually



Impact of High-Impact C–Spheres in Cavity-Aware Motifs

Figure 6.12 : TP/FP matches preserved when using automatically refined cavity-aware motifs.

Axes here are identical those of Figure 4.6. Automatically refined motifs (gray) reject a large majority of FP matches, retaining slightly more than manually designed (black) motifs. Automatically refined motifs also preserve slightly more TP matches than manually designed motifs.

designed motifs. This is expected because low-impact cavities still eliminate some FP matches that are not eliminated in automatically refined motifs. Automatically refined motifs retained more TP matches on average than manually designed motifs, for the same reasons.

6.3 Discussion

We have tested two applications of the MP method, GS and CS. Together, these methods demonstrate that MP is capable of identifying sensitive and specific motif refinements in an automated way. In addition, using GS, we have demonstrated that efficient parallelization across many computers, and statistical median estimation can be used to mitigate the high computational costs of computing hundreds of motif profiles. These performance optimizations make the refined motifs computed by MP a practical and accessible form of motif refinement.

More importantly, however, MP represents an orthogonal direction in current the design of effective motifs. While many motifs are designed with expert knowledge about specific biological systems, and while the importance of expert knowledge is not diminished by MP, MP refines geometric aspects of motifs that human experts are incapable of perceiving. By selecting refinements of existing motifs, both point-based and cavity-aware, MP reduces the geometric and chemical similarity of existing motifs to the space of all known protein structures. MP demonstrates that computational refinement of existing motifs can compliment expert knowledge in motif design.

Chapter 7

Conclusions

Inspired by the need to determine protein functions on a large scale, this thesis presents one approach for identifying instances of geometric and chemical similarity to known active sites. We first designed MASH, a computational pipeline for identifying matches of geometric and chemical similarity between motifs and target proteins. We then used this pipeline to develop MP, a method for automated motif refinement that compliments expert knowledge in the design of motifs. MP is a unique and elegantly simple contribution to the study of motif refinement. While MULTIBIND [17] could also be applied to motif refinement, MP is the first to demonstrate measurably refined motifs. MP measures Geometric Uniqueness, which is a unique and generalizable concept that could be extended to refine other types of motifs, as we have demonstrated with point-based and cavity-aware motifs.

We designed point-based MASH to find matches for motifs that encode geometric information with chemical labels, priority rankings and alternate residue labels. MA, the algorithm we designed to identify matches, is the first algorithm that accepts ranking and alternate labels. We also developed a data-driven statistical model for measuring statistical significance. Testing these components together as point-based MASH, we observed that statistically significant matches could identify cognate active sites.

Next, we extended our input motifs with C–spheres representing active clefts essential for protein function. We use C–spheres to reject potential matches that do not identify similar cleft geometry. In addition, CAMA, our adaptation of MA for matching cavity-aware motifs, uses C–spheres for algorithmic optimization. Eliminating matches with C–spheres increases *p*-

values. In comparison to point-based motifs, cavity-aware motifs match many fewer FP matches while preserving most TP matches.

Cavity-aware motifs, which combine point-based and volumetric representations, represent a unique contribution to the study of active site representations. Cavity-aware MA is also the first algorithm that accepts cavity-aware motifs as input. Having demonstrated that these motifs can be successful in identifying matches to functionally related proteins, cavity-aware motifs can be useful starting points for other studies on protein structure representations. Cavity-aware motifs also suggest that integration of several types of related biological data can sometimes yield hybrid representations that can be effective identifiers of similar functional sites.

Point-based and cavity-aware MASH provided a platform to study the problem of motif refinement. We developed MP, a purely geometric analysis that uses Geometric Uniqueness as a criterion for motif refinement. Geometric Uniqueness estimates the relative geometric dissimilarity between a given motif and the set of all functionally unrelated proteins. As a motif refinement criterion, Geometric Uniqueness is orthogonal to existing expert knowledge, demonstrating that MP can compliment human experts in the design of effective motifs.

We applied MP in the refinement of point-based motifs, producing GS, an algorithm which refines selections of potential motif points into a subset motif with maximized Geometric Uniqueness. We implemented an efficient parallel version of GS that used statistical median estimation for further efficiency. In our experiments, we observed that GS identified optimized motifs which were among the most sensitive and specific of all possible refinements. We also applied MP in the refinement of cavity-aware motifs, producing CS, an algorithm which identifies high-impact C–spheres that eliminate many FP matches. Refined cavity-aware motifs, containing only high-impact C–spheres, tended to identify more TP matches while eliminating nearly as many FP matches as manually designed cavity-aware motifs.

Overall, our results demonstrate that large scale matching techniques can enable a data-driven statistical model that can identify matches to cognate active sites. Our results also demonstrate that large scale geometric comparison can be used for measuring Geometric Uniqueness, a novel geometric measurement applicable to motif refinement on several types of motifs. These methods provide one approach to the problem of Geometric and Chemical Matching problem, and one of the first approaches to the problem of Motif Refinement. Combined, we have completed and tested one comprehensive approach applicable to the identification of similar active sites.

In the near future, GS and CS could be combined to produce a new pipeline for designing motifs with optimal residue selection and high-impact C–spheres, potentially yielding additional sensitivity and specificity. Also, improvements to the design of C-spheres using different geometry, such as polyhedra and voxels, could also provide a higher resolution representation of active site cavities, which may eliminate more TP matches. Finally, additional analysis of multiple protein structures may yield further improvements in sensitivity and specificity.

In the distant future, our approach to the identification of enzymatic active sites could be extended or modified for the many upcoming challenges of building a more complete strategy for protein function prediction. In addition to the problem of predicting enzymatic active sites, which we have not solved, many challenges await in the prediction of different sites on protein surfaces, such as protein-protein and protein-DNA interaction sites. These sites have very different geometric properties, and often involve interactions between more amino acids, but may be identifiable using the same basic principles we applied for the design of MASH and MP. In particular, the use of MP could be applicable for the design of motifs that represent protein-protein interaction sites, since the large number of amino acids involved with protein-protein interactions undoubtedly contain many subset motifs that have little Geometric Uniqueness. Finally, one of the great advantages of protein-protein interaction surfaces is that their large geometric size encourages the design of geometrically large motifs, which naturally occur less often and have greater Geometric Uniqueness.

The application and re-application of geometric comparison was a recurring theme throughout our investigation. This suggests that algorithms like MA, used for computing aggregate measurements like Geometric Uniqueness, will continue to play a critical role in geometric observation systems and their applications to Functional Annotation.

Bibliography

[1] 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.

[2] Altschul S.F., Gish W., Miller W., Myers E.W, and Lipman D.J. Basic local alignment search tool. *J. Mol. Biol.*, 215:402–410, 1990.

[3] 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.

[4] Binkowski T.A., Naghibzadeh S., and Liang J. Castp: Computed atlas of surface topography of proteins. *Nucl. Acid. Res.*, 31(13):3352–55, 2003.

[5] Liang M.P., Banatao D.R., Klein T.E., Brutlag D.L., and Altman R.B. Webfeature: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucl. Acids Res.*, 31(13): 3324–7, 2003.

[6] Laskowski R.A., Luscombe N.M., Swindells M.B., and Thornton J.M. Protein clefts in molecular recognition and function. *Protein Science*, 5: 2438–2452, 1996.

[7] 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.

[8] Lichtarge O. and Sowa M.E. Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.*, 12(1):21–27, 2002.

[9] 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.

[10] Lichtarge O., Sowa M.E., and Philippi A. Evolutionary traces of functional surfaces along g protein signaling pathway. *Meth. Enzymol.*, 344: 536–556, 2002.

[11] 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.

[12] 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.

[13] 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*, 15(6):1530–6, Jun 2006.

[14] M.A. Huynen, B. Snel, C. von Mering, and P. Bork. Function prediction and protein networks. *Curr Opin Cell Biol*, 15(2):191–198, April 2003.

[15] E. Navieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21:i302–310, 2005.

[16] M. Lappe and L. Holm. Algorithms for protein interaction networks. *Biochem Soc Trans*, 33(3):530–534, June 2005.

[17] 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.

[18] Kinoshita K. and Nakamura H. Identification of protein biochemical functions by similarity search using the molecular surface database efsite. *Protein Science*, 12:15891595, 2003.

[19] 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.

[20] Stark A., Sunyaev S., and Russell RB. A model for statistical significance of local similarities in structure. *J. Mol. Biol.*, 326:1307–1316, 2003.

[21] 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.

[22] 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.

[23] J. Shapiro and D.L. Brutlag. Foldminer and lock 2: protein structure comparison and motif discovery on the web. *Nucleic Acids Res*, 32: W536–41, 2001.

[24] Laskowski R.A. SURFNET: A program for a program for visualizing molecular surfaces, cavities, and intramolecular interactions. *Journal Molecular Graphics*, 13:321–330, 1995.

[25] 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.

[26] 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.

[27] 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.

[28] C.A. Innis, J. Shi, and T.L. Blundell. Evolutionary trace analysis of tgf-beta and related growth factors: implications for site-directed mutagenesis. *Protein Eng.*, 13(12):839–47, 2000.

[29] O. Lichtarge, H. Yao, D.M. Kristensen, S. Madabushi, and I. Mihalek. Accurate and scalable identification of functional sites by evolutionary tracing. *J. Struct. Func. Gen.*, 4:159–66, 2003.

[30] 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.

[31] 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.

[32] Binkowski T.A., Joachimiak A., and Liang J. Protein surface analysis for function annotation in high-througput structural genomics pipeline. *Protein Science*, 14:2972–2981, 2005.

[33] 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.

[34] Liang J. Edelsbrunner H., Facello M. On the definition and the construction of pockets in macromolecules. *Discrete Applied Mathematics*, 88:83–102, 1998.

[35] 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)*, pages 311–23, August 2006.

[36] 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.

[37] 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.

[38] 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.

[39] Sheikh S.P., Zvyaga T.A., Lichtarge O., Sakmar T.P., and Bourne H.R. Rhodopsin activation blocked by metal-ion-binding sites linking transmembrane helices c and f. *Nat.*, 383:347–350, 1996.

[40] Verbitsky G., Nussinov R., and Wolfson H.J. Structural comparison allowing hinge bending. *Prot: Struct. Funct. Genet.*, 34(2):232–254, 1999.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] Connolly M.L. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221:709–713, 1983.

[48] 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.

[49] 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.

[50] Holm L. and Sander C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233:123–138, 1990.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] Smart O.S., Goodfellow J.M., and Wallace B.A. The pore dimensions of gramacidin A. *Biophysics Journal*, 65:2455–2460, 1993.

[56] Williams M.A., Goodfellow J.M., and Thornton J.M. Buried waters and internal cavities in monomeric proteins. *Protein Science*, 3:1224–35, 1994.

[57] Edelsbrunner H. and Mucke E.P. Three-dimensional alpha shapes. *ACM Trans. Graphics*, 13:43–72, 1994.

[58] 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.

[59] Wolfson H.J. and Rigoutsos I. Geometric Hashing: An overview. *IEEE Comp. Sci. Eng.*, 4(4):10–21, Oct 1997.

[60] 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.

[61] 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.

[62] Shatsky M., Nussinov R., and Wolfson H.J. A method for simultaneous alignment of multiple protein structures. *Proteins*, 56(1):143–56, 2004.

[63] Russell R.B. Detection of protein three-dimensional side chain patterns. new examples of convergent evolution. *J. Mol. Biol.*, 279:1211–27, 1998.

[64] Ullman J.R. An algorithm for subgraph isomorphism. *J. Assoc. Comp. Mach.*, 16:31–42, 1976.

[65] Alt H., Mehlhorn K., Wagener H., and Welzl E. Congruence, similarity, and symmetries of geometric objects. *Discrete Comput. Geom.*, 3:237–256, 1988.

[66] Akutsu T. On determining the congruity of point sets in higher dimensions. In *Proc. ISAAC: 5th Symp. Alg. Comp.*, 1994.

[67] Akutsu T., Tamaki H., and Tokuyama T. Distribution of distances and triangles in a point set and algorithms for computing the largest common point set. *Discrete Comput. Geom.*, 20:307–331, 1998.

[68] Daniel P. Huttenlocher, Klara Kedem, and Jon M. Kleinberg. On dynamic voronoi diagrams and the minimum hausdorff distance for point sets under euclidean motion in the plane. In *Symposium on Computational Geometry*, pages 110–119, 1992.

[69] L. Paul Chew, Michael T. Goodrich, Daniel P. Huttenlocher, Klara Kedem, Jon M. Kleinberg, and Dina Kravets. Geometric pattern matching under Euclidean motion. In *Proc. Fifth Canadian Conference on Computational Geometry*, pages 151–156, 1993.

[70] L. Paul Chew, Dorit Dor, Alon Efrat, and Klara Kedem. Geometric pattern matching in d-dimensional space. In *European Symposium on Algorithms*, pages 264–279, 1995.

[71] Jeff M. Phillips and Pankaj K. Agarwal. On bipartite matching under the rms distance. In *Proceedings of the 18th Canadian Conference on Computational Geometry (CCCG'06)*, pages 143–146, 2006.

[72] M.T. Goodrich, J.S.B. Mitchell, and M.W. Orletsky. Practical methods for approximate geometric pattern matching under rigid motions (preliminary version). In *Symposium on Computational Geometry*, pages 103–112, 1994.

[73] P. Indyk, R. Motwani, and S. Venkatasubramanian. Geometric matching under noise: Combinatorial bounds and algorithms. In *SODA: ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] Zdobnov E.M. and Apweiler R. Interproscan: an integration platform

for the signature-recognition methods in InterPro. *Bioinformatics*, 17: 847848, 2001.

[79] 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.

[80] Diestel R. *Graph Theory.* Springer, New York, USA, 1997.

[81] 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.

[82] de Berg M., van Kreveld M., and Overmars M.H. *Computational Geometry: Algorithms and Applications.* Springer, Berlin, Germany, 1997.

[83] Birnbaum Z.W. and Tingey F.H. One-sided confidence contours for probability distribution functions. *Ann. Math. Stat.*, 22(4):592–596, Dec 2003.

[84] Silverman B.W. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall: London, 1986.

[85] 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.

[86] 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.

[87] Crane B.R., Arvai A.S., Ghosh S., Getzoff E.D., Stuehr D.J., and Tainer J.A. Structures of the n^{ω}-hydroxy-l-arginine complex of inducible ni-

tric oxide synthase oxygenase dimer with active and inactive pterins. *Biochemistry*, 39:4608–4621, 2000.

[88] Adak S., Wang Q., and Stuehr D.J. Arginine conversion to nitroxide by tetrahydrobiopterin-free neuronal nitric-oxide synthase. *J. Biol. Chem.*, 275:33554–33561, 2000.

[89] International Union of Biochemistry. Nomenclature Committee. *Enzyme Nomenclature.* Academic Press: San Diego, California, 1992.

[90] Cassella G. and Berger R.L. *Statistical Inference.* Brooks/Cole Publishing Co., New York, USA, 1990.

[91] 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.

[92] Efron B. Better bootstrap confidence intervals (with discussion). *J. Amer. Stat. Assoc.*, 82:171, 1987.

[93] Efron B. and Tibshirani R.J. *An Introduction to the Bootstrap.* Chappman & Hall, London, 1993.

[94] Delano W.L. The PyMol molecular graphics system (2002), on world wide web: $http://www.pymol.org$, 2002.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] Snir M. and Gropp W. *MPI: The Complete Reference (2nd Edition).* The MIT Press, 1998.