RICE UNIVERSITY

Mapping the Structural Landscape of Protein Families with Geometric Feature Vectors

by

Drew Bryant

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Science

Approved, Thesis Committee:

Lydia E. Kavraki, Chair
Noah Harding Professor of Computer Science

Luay Nakhleh
Assistant Professor of Computer Science

Yousif Shamoo
Associate Professor of Biochemistry and Cell Biology

Houston, Texas

December, 2009

ABSTRACT

Mapping the Structural Landscape of Protein Families with Geometric Feature Vectors

by

Drew Bryant

This thesis describes two key results that can be used separately or in combination for protein function analysis. The Family-wise Analysis of SubStructural Templates (FASST) method uses all-against-all substructure comparison to determine family-wide sub-group organization by quantifying the substructural variation within a protein family. The results demonstrate examples of automatically determined sub-groups that can be linked to phylogenetic distance between family members, segregation by ligation state, and organization by ancestry among convergent protein lineages. The Motif Ensemble Statistical Hypothesis (MESH) framework constructs a representative template for each of the subgroups determined by FASST to build *motif ensembles* that are shown through a series of function prediction experiments to improve the function prediction power of existing templates. This work provides an unbiased, automated assessment of the structural variability of identified substructures among protein structure families and a technique for exploring the relation of substructural variation to protein function.

Contents

Illustrations

Tables

Chapter 1

Introduction

This thesis introduces a novel method to identify the variation of protein substructure geometry within a family of related proteins and a complementary method to construct generalized computational models of protein substructure motifs. The methods introduced here identify distinct clusters or sub-groups within a protein family by combining geometric comparisons among all protein structures within the family with unsupervised machine learning for cluster identification. The sub-group organization of a protein family is considered an intra-family "ontology" based upon substructure similarity. Using biological metadata, this thesis explains the significance of the identified sub-groups and then demonstrates how identified sub-groups can be used to construct sensitivity-improved substructure templates.

1.1 Protein Substructures

Understanding the link between protein structure and protein function is a fundamental problem that underlies diverse application areas including drug target identification, protein function prediction, and structure-based evolutionary analysis. The specific few amino acids that mediate the drug-binding affinity of targeted binding sites are an example of a *substructure* within a protein (see Fig. 1). The catalytic substructures of enzymatic proteins are intrinsically linked to enzyme function [9, 10, 11, 12], and establishing a mechanistic understanding of how specific structural features affect protein function is a central prob-

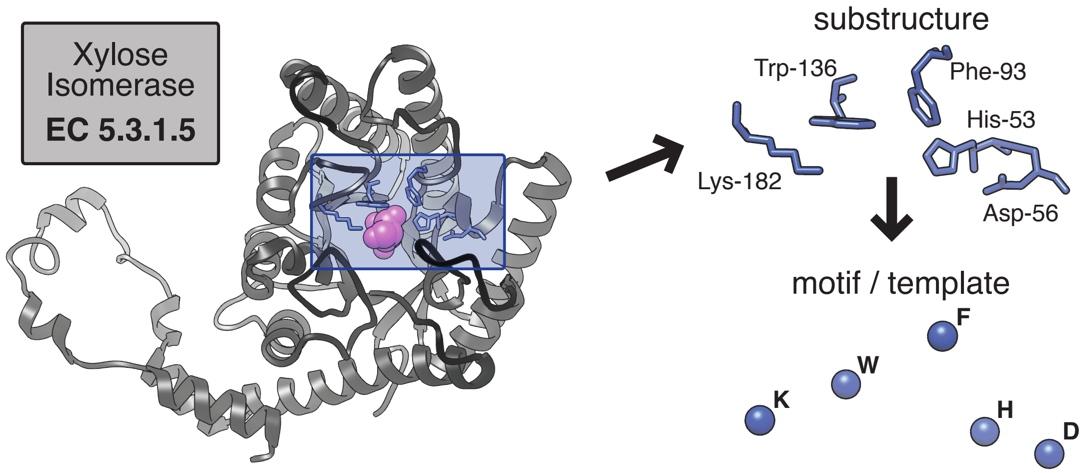


Figure 1.1 : **Protein substructures.** The xylose isomerase protein shown above has a sequentially non-contiguous, but spatially compact set of 5 functional residues that are shown above in stick representation. These 5 catalytically important residues constitute a *substructure* of the protein and can be represented by a motif/template. The C_α atom of each residue is modeled by the template along with one or more amino acid type labels. This template can then be compared to other protein structures to identify matching substructures which share chemical and geometric similarity to the template.

lem in structural genomics [13]. The analysis of the physico-chemical properties of the few amino acids constituting these substructures, common to families of functionally related proteins, can provide direct insight to the structural features that dictate a particular enzymatic function [10]. For example, the family of serine proteases is a well-established case of a common functional substructure, the HIS-ASP-SER catalytic triad, dictating a common function in the absence of sequence or fold similarity between chymotrypsins, subtilisins, and lipases [3, 14]. Conversely, in the case of TIM barrel proteins that share both sequence and fold similarity, differing functional substructures within the catalytic site imbue differing functions [15]. Therefore, because these functional substructures are essential determinants of protein function, computational approaches to analyze and compare substructures among proteins can provide fundamental insight to the molecular mechanisms that mediate protein function.

1.2 Significance of Substructure Analysis

Substructure analysis is of practical importance for identifying proteomic drug targets, finding potential drug side-effects, predicting protein function, and evolutionary analysis. Binding site substructures have been considered “receptor-based pharmacophores” [16], allowing a specific few amino acids to indicate likely interaction with a specific ligand-based pharmacophore. Substructural similarity at ligand-binding sites among proteins is indicative of similarity in ligand- and drug-binding properties [12, 11]. Exploitation of this property has been applied recently to identify new targets for existing drugs [17] and to computationally analyze potential drug side-effects [16]. Specifically, cross-species substructure analysis of binding sites among families of functionally homologous proteins can play an important role in lead evaluation [18, 16], and therefore computational approaches to analyze family-wise substructural variation are particularly relevant for modern drug development.

Furthermore, substructure comparison of catalytic sites among proteins has been shown to be a powerful technique for predicting the function of protein structures [14, 19, 20] and is an important component of structural genomics initiatives that seek to map and functionally annotate the space of protein structures [21, 13]. Finally, enzymes evolve under selective pressure to maintain biologically necessary functions [22], and functional substructure conservation in the absence of sequence of fold conservation has been established [23, 24]; substructure comparison may be the *only* way to establish homology between proteins that have significantly diverged in both sequence and fold [25]. Given the biological relevance of substructure analysis and the proliferation of available structures in the Protein Data Bank [26], computational approaches to substructure analysis can make meaningful contributions to our understanding of proteomics.

1.3 Overview

This thesis departs both from finding functionally significant substructures and from comparing substructures to identify biologically relevant matching proteins. Here, a meta-level approach to substructure analysis is presented that combines substructure comparison, unsupervised learning, dimensionality reduction and non-parametric statistical analysis to automatically identify intra-family "ontologies" by analyzing the structural diversity of a family of proteins. In this thesis, functionally homologous protein families are partitioned into sub-groups based upon substructural similarity, and the sub-group organization for a family is what this thesis refers to as an *intra-family ontology*.

The first method presented in this thesis, called the Family-wise Analysis of SubStructural Templates (FASST) method, can be used to identify the substructure-based intra-family ontology of a family of proteins, which automatically partitions a family into sub-groups based upon substructural similarity. The second method, the Motif Ensemble Statistical Hypothesis framework (MESH), exploits the substructure-based intra-family ontologies output by FASST to construct refined substructure representations that improve the function prediction power of existing templates.

1.3.1 FASST

FASST proceeds as follows. For a given enzyme family, a substructure template of the catalytic site is first defined from a literature reference or other source of substructure templates [27, 28, 29, 30, 4, 31, 32, 33]. Instances of the motif are then identified in each family member structure by a substructure search algorithm—LabelHash here [7]. Next, all-against-all pair-wise Least Root Mean Square Deviation (LRMSD) distance comparisons are computed between family members. The LRMSD of the catalytic site substructure from a given protein to the remainder of the family then encodes the family-wise relationship of the family members to one another as vectors of geometric features. Each geometric feature vector can then be interpreted as a point in a high-dimensional *geometric feature space*, where nearby points in this space indicate similar family-wise relationships for the corresponding substructures. The location of each protein substructure in geometric feature space is used to identify the place of each substructure in the overall substructure-based intra-family ontology output by FASST and to identify the amount of substructural variation present within a family. FASST then uses a Gaussian Mixture Model (GMM) clustering approach for unsupervised learning of the sub-groups within intra-family ontologies. A substructure-based intra-family ontology can then be compared to a biological ontology by mapping meta-data to each substructure for further analysis.

1.3.2 MESH

MESH utilizes the sub-groups identified by FASST to construct refined substructure templates that have improved *sensitivity*, and this procedure is demonstrated through a series of protein function prediction experiments. MESH constructs a representative motif for each identified sub-group. The collection of representative motifs, for the family, constitutes a single motif ensemble. To provide a statistically rigorous framework for calculating the statistical significance of substructure matches identified by motif ensembles, this thesis introduces a non-parametric model of substructural similarity for multi-structure templates.

1.3.3 Results

This thesis demonstrates with FASST that sub-groups within substructure-based intra-family ontologies can suggest biological sources of structural variation within a protein family. For the heme-dependent peroxidase family (EC 1.11.1.7) and the xylose isomerases (EC 5.3.1.5), this thesis shows that the observed intra-family ontology can be explained by the phylogenetic distance between members of the family. Structures of the thermolysin family of bacterial proteases are observed to have catalytic sites with both discrete and continuous modes of flexibility, and structures are known to transition between discrete structural conformation states upon ligation. Analysis of the family-wise structural variety of the serine protease catalytic triad, incorporating over 700 structures from 52 different species and 7 EC classes, demonstrates the ability of FASST to detect substructure variation among convergently related families where the template substructure resides in many configurations, including some spanning peptide chains.The substructural variation present within each family is automatically identified from the intra-family ontologies output by FASST.

MESH constructs sensitivity-improved motif ensembles from single structure motifs. The performance of the combined FASST-MESH methods is demonstrated in a series of protein function prediction experiments. When compared to single structure motifs, this thesis demonstrates that the FASST-MESH framework can significantly improve functional annotation sensitivity for structurally variable families of proteins, while maintaining annotation specificity, for the 15 protein families included in the study.

1.4 Contributions

Establishing a rational basis for structural variation, particularly at functional sites, has critical applications for drug target identification, side-effect prediction, protein function prediction, and molecular evolutionary analysis. In protein families that exhibit a common function, shared chemistry and geometry at catalytic site substructures provides a common, local point of comparison among proteins that may otherwise be highly differentiated at the sequence, fold, or domain topology levels.

The biological relevance of the functional substructures modeled by templates can be exploited for exploratory investigations of the role and structural conservation/variation of a substructure within a large protein family; the utility of this approach is demonstrated using FASST by comparing the structure-based intra-family ontologies output by FASST to biological ontologies such as phylogeny. Furthermore, selecting a single-structure template as a consensus model of a family-wide functional substructure can prove difficult [9] when functionally conserved protein families become large and species-diverse. The MESH framework transforms single-structure templates into *motif ensembles* to account for increasing family-wide substructural diversity and provides a robust procedure for identifying statistically significant matches to the motif ensemble as a whole. FASST and MESH directly contribute to substructure-based analysis by providing a template assessment and refinement procedure. FASST provides an additional avenue of exploratory investigation for selected substructures within a family of interest.

The popularity of incorporating only sequentially non-redundant structure subsets* in computational studies ignores a wealth of additionally available protein structures. As particularly demonstrated in Chapter 4.2, structures for sequentially identical proteins can harbor highly informative structural trends that can be connected with how the structure was crystallized and the ligation state of the protein when crystallized. Furthermore, including all available structures distinguishes outlier structures with more clarity and confidence than would be possible if only using a small subset of structures, and significant outlier structures are noted throughout Chapters 4.1, 4.2, and 4.3 in the analysis of intra-family ontologies output by FASST.

Pair-wise substructure comparison alone does not reveal all of the inter-connected struc-

*ASTRAL,nrPDB,etc.

tural relationships within a protein family. The family-wise substructure comparison approach implemented by FASST operates at a meta-level to pair-wise techniques in order to identify high-level trends latent to functional sites. By correlating high-level trends in substructure variation with biological metadata such as phylogeny, ligation state, and protein ancestry, FASST can be used as a tool for the exploratory analysis of structure-function relationships across large numbers of structures.

This thesis demonstrates an automated approach to augment existing substructure templates already available in repositories such as the Catalytic Site Atlas [29] by geometrically enriching motifs for families that exhibit high structural variability. As both the number and diversity of available structures for a given protein family continue to increase, the enhancement of substructure-based functional annotation methods to accommodate large families is necessary. The automated enrichment of available templates strengthens the function prediction power of these templates and facilitates the use of substructure-based analysis methods for large-scale, automated annotation of novel structures.

Chapter 2

Background

2.1 The Role of Substructures in Protein Evolution

The mechanisms of evolutionary conservation operate at multiple levels of resolution, from DNA and amino acid sequence to domain organization and fold topology to functional substructures, such as catalytic sites, and the rate of evolutionary divergence at these levels differs [23, 24]. In the absence of discernible sequence similarity, protein homology has been confirmed on the basis of topological similarity between structures, but over immense evolutionary time periods, even the fold topology of a protein may begin to drift[34]. In addition to fold topology, catalytic site substructures, including the chemistry of the amino acids involved in catalysis and the 3-dimensional orientation of these amino acids, are essential units of evolutionary conservation, because significant alterations to these substructures can result in loss of enzymatic function [22]. Protein substructures can therefore be considered one of the smallest proteomic elements of evolutionary conservation and should be considered in addition to sequence and topology when tracing the evolutionary history of protein families.

Protein substructures are capable of directly modulating differing enzymatic function among proteins sharing a common fold. The fold of triose-phosphate isomerase (TIM), termed "TIM-barrel", is an alternating α/β topology consisting of 8 units ($(\alpha\beta)_8$) and has been identified in proteins of widely varying enzymatic function. Examples of oxidoreductases, transferases, hydrolases, lyases, and isomerases that incorporate the TIM-barrel fold

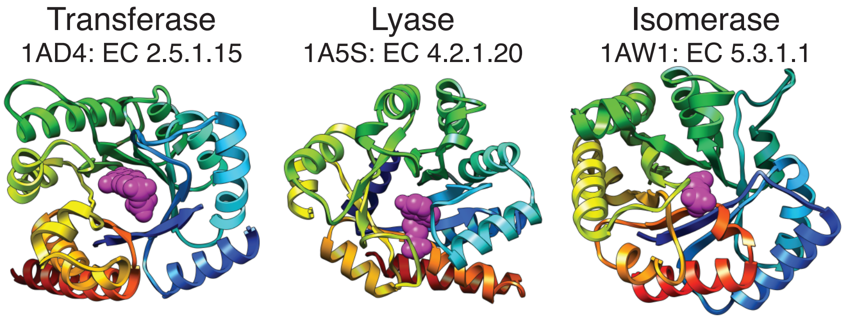


Figure 2.1 : **Same fold, different functions.** Dihydropteroate synthase (1AD4), tryptophan synthase (1A5S), and triose-phosphate isomerase (1AW1) are examples of proteins sharing the same TIM-barrel fold while catalyzing very different enzymatic reactions. Each structure is colored from blue (N-terminal) to red (C-terminal) with the bound ligand shown as pink spheres. Variable loop regions near the C-terminal end that link the main secondary structure elements are able to implement a variety of different functions by incorporating different substructural elements [1, 2].

have all been identified [2]. These TIM-barrel proteins are able to exhibit diverse functions while sharing a common topology by varying the residue composition of loop regions near the C-terminus as shown in Fig. 2.1 [1, 2]. The catalytic region of these enzymes are composed of residues separated in sequence but co-located with the C-terminal end in the folded peptide structure [1]. Therefore, because of the direct relationship between substructures and protein function, methods to analyze and compare substructures among proteins are important tools for understanding the link between structure and function.

Many striking examples of substructure conservation in the absence of higher-level topology and sequence conservation have been demonstrated [24] as well as many instances of convergent evolution to similar substructures from distinct ancestral sources [14, 19, 35]. The heme-dependent peroxidases, analyzed in Section 4.1, have been theorized to share an extremely ancient common ancestor [36] but bear little similarity to one another among modern species. As shown in Fig. 2.2, the mammalian and fungal versions of the enzyme are topologically distinct, and the sequence identity between the two proteins is only

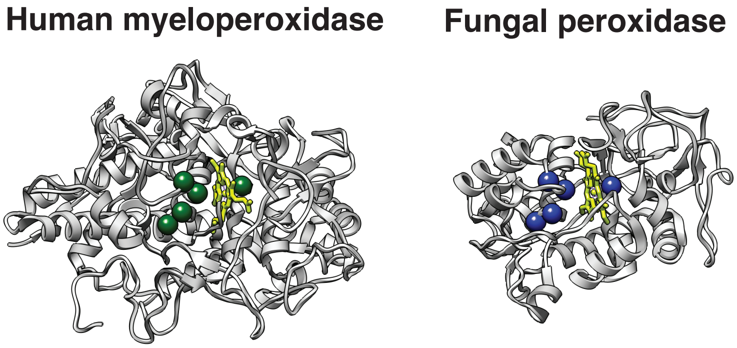


Figure 2.2 : **Substructural conservation in the heme-dependent peroxidases.** The catalytic site of the mammalian enzyme shares a common catalytic substructure with the sequentially and topologically distinct fungal version of the enzyme. The 5 catalytically necessary residues are shown with sphere representation at each C_α position above. The heme prosthetic group is shown in stick representation for reference in each structure above.

9%*. However, both the mammalian and fungal peroxidases share a common functional substructure in the catalytic site as shown in Fig. 2.2.

Convergent evolution to a common functional substructure–the "catalytic triad"–has been well-documented for the serine proteases [3]. The catalytic triad residues (HIS-ASP-SER) cleave ligand peptides at specific locations which depend upon the peptide's residue sequence. While these serine proteases share a common catalytic substructure, the overall fold between the chymotrypsin, subtilisin, and lipase superfamilies differ, as shown in Fig. 2.3. Therefore, substructure-based methods are capable of comparing very distantly related, or convergent components of protein structures where sequence- and topology-based methods may fail, because common substructures that can be linked to common functions may exist in the absence of fold or sequence similarity.

*Maximum possible sequence identity between all possible pairs of chains from structures 1CXP (mammalian) and 1ARU (fungal)

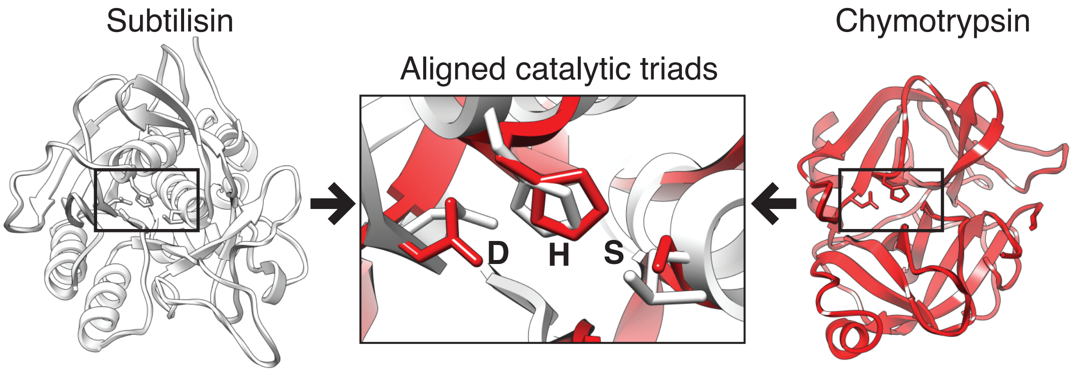


Figure 2.3 : **Different folds, same function.** Subtilisin and chymotrypsin have convergently evolved to contain a functionally equivalent set of 3 residues called the “catalytic triad” [3].

2.2 Substructure Identification Methods

Computational methods for finding functionally significant substructures and methods for comparing substructures to identify biologically relevant proteins with matching substructures are two complementary components of substructure analysis. As far as approaches capable of finding substructures are concerned, earlier work includes ligand-binding cavity identification (CavBase [27], CASTp [4]), structural pattern recognition (GASPS [30], FEATURE [37]), computational scanning mutagenesis (SNAP [33]), evolutionary analysis (ET [38], ConSurf [8]), expert knowledge (CSA [29]), and automatically curated databases (LigBase [28], SFLD [10], LigASite [32]). Substructures identified by these methods can be computationally represented, either in full or in part, by *templates* (also known as *motifs*) that model both the geometric and physico-chemical properties of a given substructure[†].

Substructure identification methods are a necessary component of large-scale automated pipelines and the FASST-MESH method introduced here is agnostic as to the source

[†] Other names used in the literature include local functional sub-domains, functional epitopes, amino acid constellations, receptor-based pharmacophores, and binding hot spots.

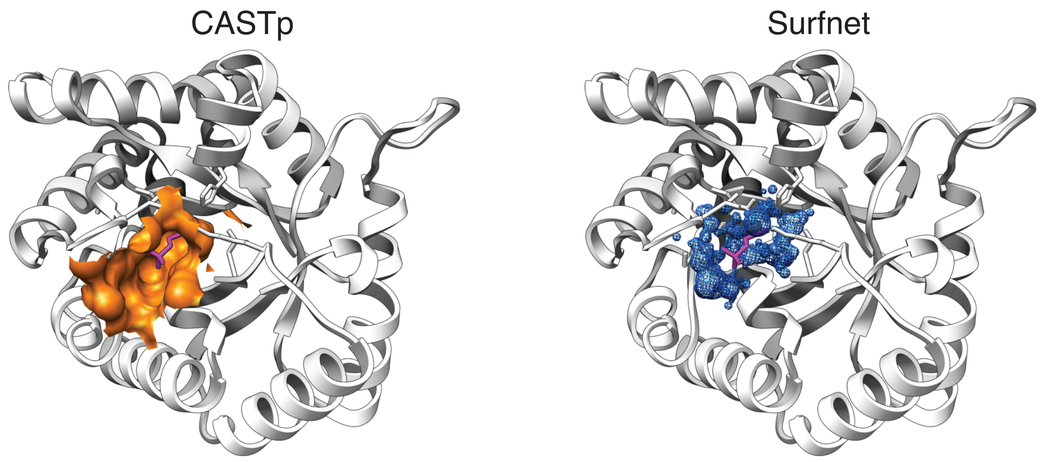
of substructure motifs. Methods for identifying cavities[‡] on protein surfaces have been investigated because enzymatic sites typically occur within the specific chemical microenvironments created by cavities. For example, the Computed Atlas of Surface Topography of proteins (CASTp) identifies cavities of varying size using alpha shapes [4]. An example of a cavity identified by CASTp is shown in Fig. 2.4. Directly translating all of the amino acids that compose a cavity to a motif is often suboptimal in terms of predictive power of the resulting motif, but this is addressed in Section 2.5. The Surfnet method can be used to identify interface regions between molecules and an example of a protein-ligand interface is shown in Fig. 2.4. Focusing the problem of identifying functional components of proteins towards relatively smaller regions of the protein structure, such as cavities, greatly reduces the complexity of automated motif selection and refinement.

2.3 Substructure Comparison Methods

Computationally identifying substructure matches in other proteins with statistically significant similarity to a template can indicate that a matched protein may share functional characteristics with the template [14]. Diverse approaches to template search and/or comparison have been developed and include: SPASM [39], ASSAM [40], PINTS [41], Jess [20], SiteEngine [42], Query3D [43], ProFunc [44, 45], ProKnow [46], SitesBase [31], GIRAF [47], MASH [48], LabelHash [7], SOIPPA [25], FEATURE [37], and pevoSOAR [49] to name a few.

While substructure comparison is a major component of FASST-MESH, the particular method selected for pair-wise substructure comparison need only be capable of generating LRMSD alignments between substructures. In this work, the LabelHash method [7] was

[‡] Also known as clefts or pockets



CASTp

Surfnet

Figure 2.4 : **Automated substructure selection methods.** The structure of triose-phosphate isomerase (TIM) (PDB 1AW1) is shown above with the ligand in stick representation and cavity regions as surface and mesh forms. CASTp [4] identifies a total of 40 different cavities (of widely varying size) for TIM; the cavity associated with the ligand binding site is shown alone. Surfnet [5] identifies interface regions between molecules, and above is shown the interface region between TIM and the bound ligand. For both CASTp and Surfnet, the residues associated with the identified cavities can serve as a source of substructure motifs to be used as input to the FASST-MESH method.

used for pair-wise substructure matching. Because FASST-MESH does not make heuristic assumptions tied to a specific substructure matching method or substructure model (such as C_α-only, C_α+C_β, pseudoatoms, etc.), different comparison methods or motif representations can be utilized. The ability of FASST-MESH to consume many different types of input motifs/methods is important for several reasons: many of the previously mentioned substructure comparison methods are targeted towards different scales of molecular comparison (ex. whole structure, domains, large cavities, 5-10 residues, 2-4 residues); biological queries should utilize the most sensical substructure model for the problem; providing a clean API for many different methods allows for cross-comparison/consensus among approaches to be used. The FASST-MESH method operates at a *meta-level* to these substructure comparison methods by synthesizing many pair-wise comparisons to identify high-level trends for large sets of protein structures (see Section 3 for further details).

2.4 All-against-all Comparison Methods

The FASST method presented here directly complements the *k*-partite [50], bipartite [51, 6] and product-graph-max-clique [52] approaches to all-against-all common substructure identification, because these methods can successfully identify common substructures between two [51, 52, 6] or more [50] binding sites. Several of the afore mentioned all-against-all methods have also been used to construct "similarity networks" of known ligand binding sites by using pair-wise similarity scores between binding sites in combination with linkage-based [51, 52, 6] clustering to build graphs of related sites. The GIRAF-based [47] approach to all-against-all comparison is highlighted here for detailed comparison to FASST-MESH.

For every structure in the PDB, GIRAF selects all residues that are $< 5\AA$ from a bound ligand and structures without a bound ligand are skipped; the residues selected are now "binding sites". Each binding site is then decomposed into tetrahedra (Delaunay tessellation) using all atoms (i.e. backbone atoms and side-chain atoms). Next, each tetrahedra is inserted into a relational database that is indexed by the constituent atom types, tetrahedra volume, and edge lengths. To identify a match to a given binding site, the database is queried for similar tetrahedra; the atoms representing the query and matched binding sites are then considered two graphs (i.e. all atoms of the binding site, not just the tetrahedra, are included). Iterative max-weight bipartite graph matching is then performed between the two binding site graphs; edge weights correspond to the Euclidean distance between paired atoms (i.e. one from each binding site) when optimally superimposed, and only atoms $< 2\AA$ apart are considered to be paired. For each iteration, the graphs are maximally matched (maximize number of paired atoms while minimizing RMSD) and then the

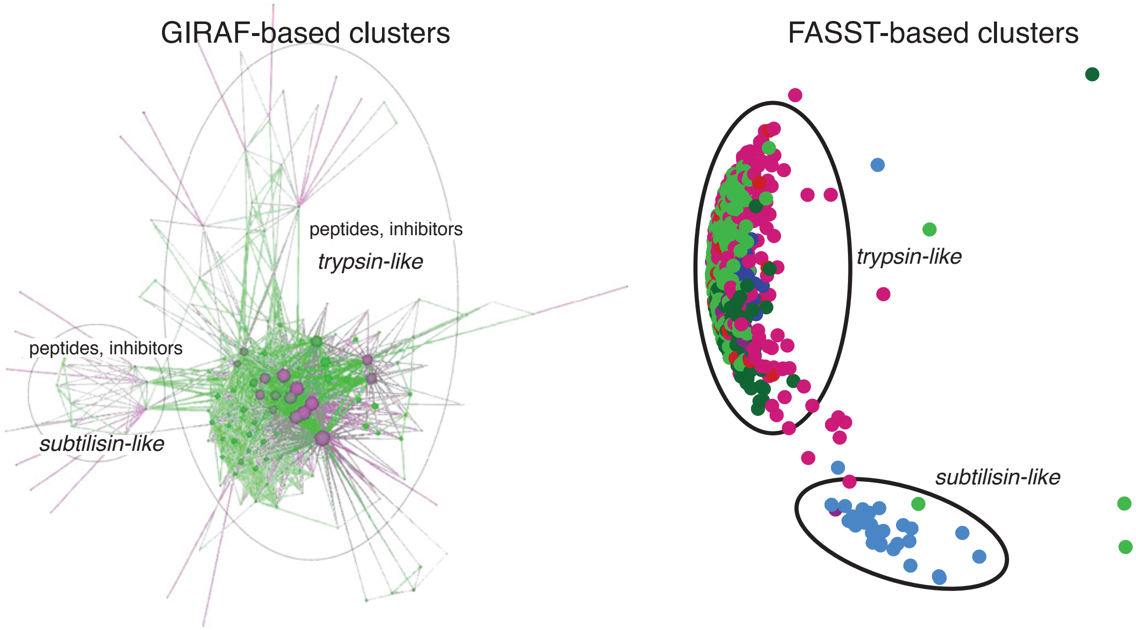


Figure 2.5 : **GIRAF and FASST clustering of the serine proteases.** The trypsin-like and subtilisin-like serine proteases are identified as separate clusters by both GIRAF and FASST, although the clustering approaches themselves differ. The trypsin-like cluster shown for GIRAF is actually an agglomeration of 19 individual clusters that Kinjo et al. identified manually. While FASST is only comparing the 3 C_α atoms of the catalytic triad, GIRAF compares the largest common substructure between every pair of bindings sites (nodes in the graph); the mean/std.dev. of the binding site atoms compared in the GIRAF-based network is 32/11, much larger than the 3 atoms used by FASST. While GIRAF and FASST differ in approach, the two methods are in agreement as to how the serine proteases should be structurally partitioned. Adapted from Kinjo et al. (2009) [6].

matching components are used to re-align the two binding sites into LRMSD alignment; this process iterates until the alignment converges. The final similarity score S for a pair of binding sites is then given by a combination of a modified Tanimoto coefficient[§] and LRMSD of the superimposed binding sites. The statistical significance of a match score S is obtained by calculating the p-value of the match given a random sample of binding sites in the database. See Kinjo et al. 2007 [47] and 2009 [6] for complete details.

The all-against-all, GIRAF-based approach from Kinjo et al. 2009 identifies clusters of "similar" binding sites using a similarity network approach that differs fundamentally from the FASST method introduced here. While FASST-MESH is presented fully in Chapter 3, important distinctions between the work of Kinjo et al. are outlined here. Given multiple binding sites, GIRAF may identify a different common substructure between each pair of sites; these common substructures can vary widely in size (number of atoms) and geometric similarity (LRMSD). FASST, however, compares a single substructure consistently between proteins; proteins either have the substructure or do not. Therefore, FASST is concerned with analyzing the geometric variability of a *particular* substructure among a set (i.e. family) of proteins while GIRAF may compare a *different* substructure between every pair of proteins. Furthermore, GIRAF employs a similarity measure that requires relative weighting of the match size and match LRMSD terms, while FASST uses the LRMSD distance metric alone.

Superficially, the "similarity network" of serine proteases (see Fig. 2.5) is very similar to the FASST clustering (as shown in Fig. 4.3) as far as cluster membership, but the method for determining sub-groups or clusters differs. An edge weight in the GIRAF-based network is the match p-value between a given pair of binding sites (nodes); the full graph is arrived at by selecting a global p-value threshold and only edges with p-value smaller than

[§] Tanimoto coefficient for sets A and B: $T(A,B) = \frac{|A \cap B|}{|A|+|B|-|A \cap B|}$

this threshold are retained. Clusters within the graph are identified using a hybrid technique that combines hierarchical single-linkage clustering followed by agglomerative complete-linkage clustering to arrive at the final clusters [6]. Therefore, the binding sites constituting a GIRAF cluster do not necessarily share any common substructure. The mean size of the common substructure between pairs for the GIRAF-based network in Fig. 2.5 is 32 atoms per binding site with a standard deviation of 11; FASST is only comparing the 3 C_α atoms of the catalytic triad among all of the serine protease structures (triad residues are shown in Fig. 2.3).

It is interesting to note that both GIRAF and FASST reach a similar conclusion given the fundamental differences in the methods, and it is likely attributed to the strong geometric distinctions between the subtilisins and trypsins at both catalytic triad residues and the surrounding binding cleft. The substructure-based all-against-all comparison implemented by FASST is most analogous to the seminal work of Holm and Sander [53] on mapping protein fold space via all-against-all Dali comparisons [54].

2.5 Motif Refinement and Optimization

The success of substructure comparison/search algorithms relies on high-quality motifs that accurately capture the geometric and chemical characteristics of a given functional site. Well-designed motifs should be both structurally similar to functionally analogous substructures and structurally different from functionally unrelated substructures in order to be used for accurate function prediction.

Previous work on Geometric Sieving (GS) has demonstrated utility for selecting a subset of residues from a larger functional site to build high-specificity motifs [55, 48]. GS estimates the *probability density function* (pdf) of all k-sized subsets from the original n-point motif ($\binom{n}{k}$ subsets) by computing matches from each subset motif to a random sam-

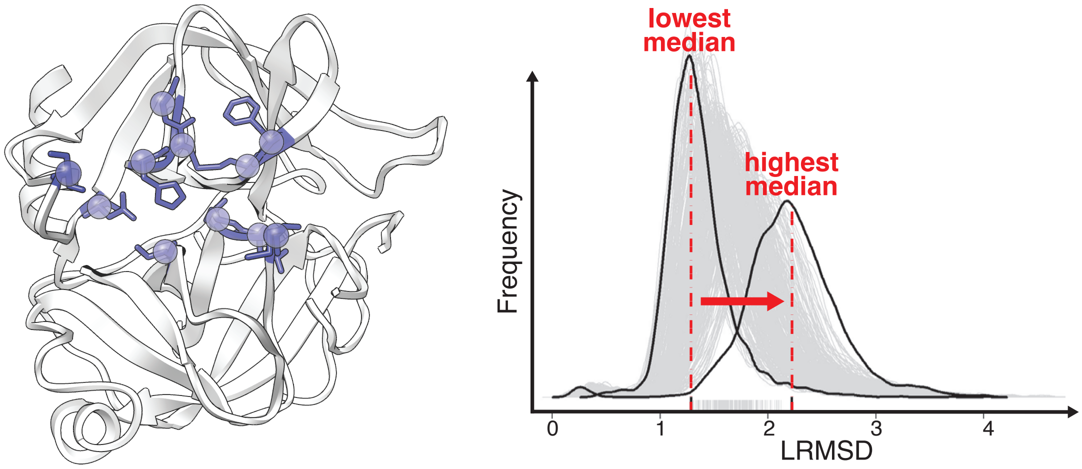


Figure 2.6 : **Geometric sieving.** A crude motif for chymotrypsin, containing both catalytic and nearby non-catalytic residues, is shown above with spheres indicating each of the 11 motif points. Every 7-point subset of the 11-point input motif is matched against a random sample of the PDB and these $\binom{11}{7} = 330$ smoothed pdfs are shown above on the right. The pdfs with highest and lowest medians are outlined in bold black while the other 328 profiles are in lighter gray. Hash marks along the x-axis denote the locations of each of the 330 medians. Geometric sieving selects the motif point subset with highest median because it is the most "geometrically unique" and least likely to match functionally unrelated targets at a given LRMSD.

ple of the Protein Data Bank (PDB) [56]. When selecting an optimized k-point subset of the n-point input motif, picking the subset motif with highest pdf median produces the most *specific* output motif, because this motif will match most unrelated protein structures with higher LRMSD (dissimilar to unrelated structures); subset motifs with low medians necessarily match more unrelated proteins with lower LRMSD (more similar to unrelated structures). As shown in Fig. 2.6, the medians of the bandwidth smoothed pdfs are used to differentiate the performance of each subset motif and the k-point subset that produced the highest median value is ultimately selected as the most "geometrically unique" subset motif.

Previous work on Cavity Scaling (CS) [57, 58] exploits the correlation between protein function and the presence of co-located ligand-binding cavities to improve the *specificity* of input motifs. The *cavity-aware* motif shown in Fig. 2.7(a) is a hybrid model that combines models for binding site residues and binding site cavity volumes, which are modeled by cavity spheres. Matches to a cavity-aware motif must geometrically match the motif points and have a cavity of similar volume to the motif. The additional cavity volume constraint is able to eliminate many spurious matches to substructure that share similar residues to the motif but lack a similar cavity volume. Different approaches for selecting and placing cavity spheres within a cavity-aware motif were investigated, because every additional cavity sphere further constrains the possible matches. Cavity spheres that eliminate very few matches are termed *low impact* while those that eliminate many matches are termed *high impact*. Fig. 2.7(c) shows how low and high impact cavity spheres can be distinguished by analyzing bandwidth smoothed pdfs, where each pdf represents an identical copy of the motif points but with a different sphere size. This process of identifying high impact cavity spheres is termed Cavity Scaling (CS).

Designing high-quality templates that accurately capture the functional essence of a substructure is critical and the performance of template-driven substructure comparison methods depends directly on the biological relevance of input templates. The FASST-MESH method introduced here contributes to both the identification and matching of templates.

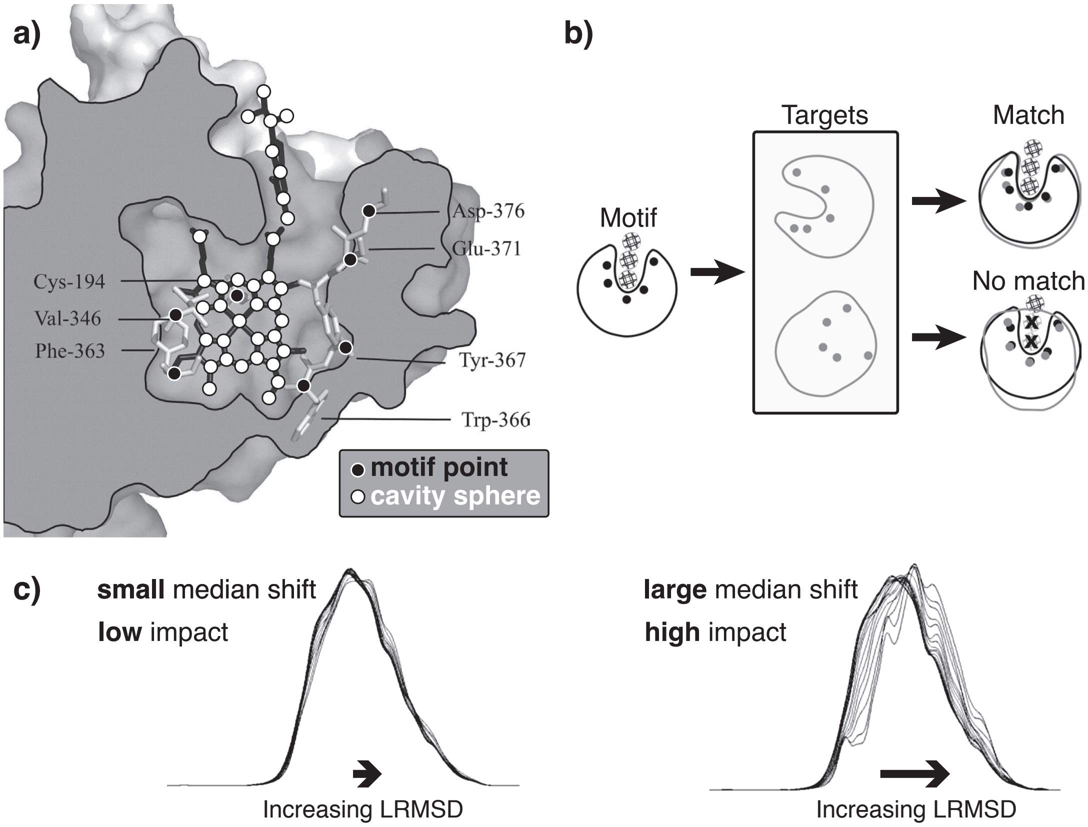
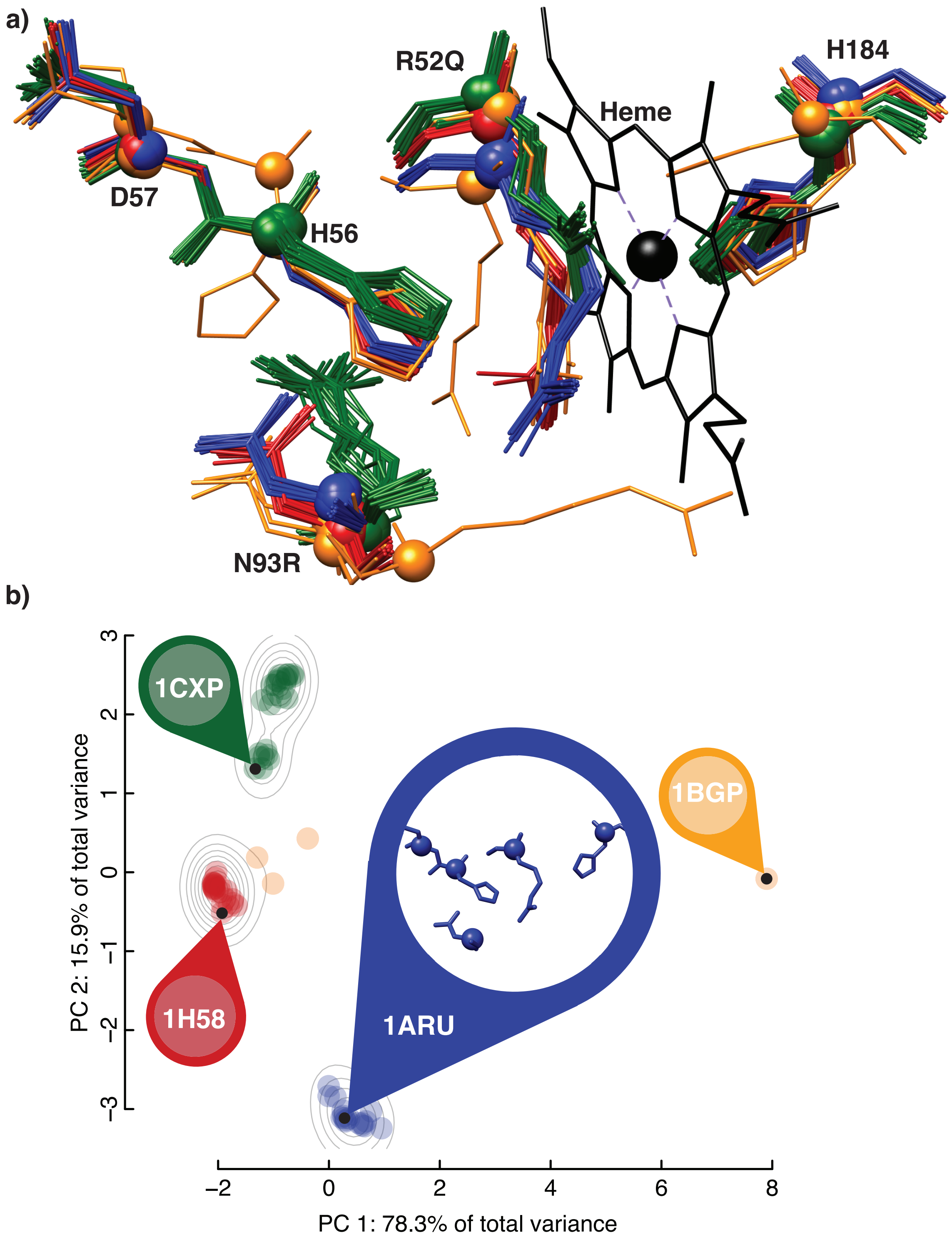


Figure 2.7 : **Cavity-aware motifs.** **(a)** The *cavity-aware* motif shown above is a hybrid model that combines both motif points (black circles) and cavity spheres (white circles). **(b)** Cavity-aware motifs eliminate erroneous matches where matched proteins lack similar cavity volumes. Both matches shown align with low LRMSD to the cavity-aware motif points, but the bottom, erroneous match is eliminated because it violates one more of the empty cavity sphere regions of the cavity-aware motif. **(c)** The number of matches eliminated by a cavity-aware motif depends on both the size and placement of cavity spheres. Cavity Scaling (CS) considers each cavity sphere individually at a range of radii. Cavity spheres that cause large shifts in pdf median are termed *high impact* while spheres that have little or no effect on pdf median are termed *low impact*.

Chapter 3

Methods

The *family-wise* substructure analysis method developed here (FASST) takes as input a user-defined substructure template called a *motif* and a *family* of protein structures, as defined by EC classification here, and outputs a substructure-based intra-family ontology that identifies one or more sub-groups of proteins within the larger family. Subsequent application of MESH to the sub-groups identified by FASST constructs a set of *consensus motifs*, collectively referred to as a *motif ensemble*, that can be used to represent the structural variety of the family for function prediction experiments. The combined FASST-MESH procedure is as follows: **(FASST: Step 1)** using LabelHash [7] (available online at *http://labelhash.kavrakilab.org*), or a another substructure search method (FASST is not tied to a particular search method), compute matches of the user-defined motif to identify analogous substructures in all family members, thereby creating one *propagated motif* per member; **(FASST: Step 2)** compute an all-against-all LRMSD alignment of each propagated motif, yielding a vector of substructure distances for each family member which we call a *geometric feature vector*; **(FASST: Step 3)** perform dimensionality reduction on the set of geometric feature vectors via principal components analysis (PCA) [59] and project each geometric feature vector onto the number of PCs necessary to preserve 90% of the original variance; **(FASST: Step 4)** cluster the dimensionality-reduced geometric feature vectors using a Gaussian Mixture Model (GMM) [60] to create a substructure-based intra-family ontology that identifies sub-groups within the family; **(MESH: Step 5)** build a set of consensus motifs to represent the sub-groups of the family by selecting an exemplar struc-


a)
D57
H56
R52Q
Heme
H184
N93R
b)
1CXP
1BGP
1H58
1ARU
3
2
1
0
−1
−2
−3
PC 2: 15.9% of total variance
−2
0
2
4
6
8
PC 1: 78.3% of total variance

ture from each sub-group or averaging substructures within a group; **(MESH: Step 6)** for functional annotation, match the consensus motifs against the Protein Data Bank (PDB) to search for proteins with substructural similarity to the original structure family and identify statistically significant matches using our non-parametric hypothesis testing framework for substructural similarity [48, 61], which is adapted and extended here to accommodate motif ensembles. Each of the steps is outlined in detail below.

3.1 Step 1: Motif Definition and Propagation

To quantify the geometric similarity between a pair of catalytic substructures, the LRMSD distance metric is commonly used, but to model the geometric similarity between a given catalytic site and a family of catalytic site substructures we introduce a simple extension to pair-wise LRMSD that will be referred to as geometric feature vectors.

The procedure for building geometric feature vectors begins with a single, user-defined motif, S^*, that represents the geometry and chemistry of a shared substructural element within the family. The S^* for each of the families included in this study were constructed from documented residues in the literature reference associated with each PDB structure

Figure 3.1 *(preceding page)*: **Clustering geometric feature vectors. (a)** Superposition of the propagated motifs for the animal and non-animal heme-dependent peroxidases of EC 1.11.1.7 demonstrates geometric variability. The color of each aligned substructure corresponds to its cluster assignment in (c), and it can be seen that closely aligned substructures in (a) correspond to co-located points in the intra-family ontology shown in (c). **(b)** Applying FASST to the family of peroxidases yields a family-wise geometric feature vector for each catalytic substructure in the family, reducing each substructure shown in (a) to a point in the intra-family ontology. Gaussian mixture model (GMM) clustering of the geometric feature vectors, projected onto a space of reduced dimension, identifies four clusters denoted by color. The gray isocontours show the smoothed density of substructures in each part of the geometric feature space. Each cluster identified constitutes a sub-group within the intra-family ontology.

listed in Table 1. For example, S^* for the heme-dependent peroxidases includes the C_α atom from each of the following residue numbers with the alternate amino acid labels shown in superscript: 52^{RQ},56^{H},57^{D},93^{NR},184^{H}; the 3-dimensional coordinates of each $C_\alpha \in S^*$ were taken from [PDB:1ARU] as noted in Table 1 and the residue numbers listed are according to [PDB:1ARU]. Care should be taken to define S^* with appropriate amino acid alternate labels (which allow for amino acid substitutions to be represented). While ConSurf [8] was used in this work, when available, an expert-curated multiple sequence alignment allows for the highest confidence in amino acid alternate selection.

First, the user-defined motif, S^*, is matched against a family of n protein structures, $F = \{f_1, ..., f_n\}$, as defined by Gene Ontology (GO) terms or Enzyme Classification (EC) levels, for example, to yield a set of matches $\mathbf{M}_{S^* \to F} = \{M_{S^* \to f_1}, ..., M_{S^* \to f_n}\}$. In this work, LabelHash [7], was used to identify substructure matches by searching each protein in F for similar substructures to the motif, S^*. Every match, $M_{S^* \to f_i} \in \mathbf{M}_{S^* \to F}$ is a bijection between S^* and a substructure of f_i, and defines a unique substructural element within f_i that will be referred to as a propagated motif, S^{f_i}. For algorithmic details of how LabelHash identifies substructure matches to templates/motifs see [7].

3.2 Step 2: Encoding Geometric Features

The pair-wise LRMSD between two propagated motifs will be denoted by $d(S^{f_i}, S^{f_j})$ and the geometric feature vector, $\mathbf{g}_i$, for a given f_i is defined as a vector of LRMSD values: $\mathbf{g}_i = \{d(S^{f_1}, S^{f_i}), ..., d(S^{f_n}, S^{f_i})\}$. The set of geometric feature vectors representing all structures in the family, F, will be denoted as $\mathbf{G} = \{\mathbf{g}_1, ..., \mathbf{g}_n\}$, and $\mathbf{G}$ constitutes an all-against-all alignment of the substructures that correspond to each respective protein in F. Each $\mathbf{g}_i \in \mathbf{G}$ defines a point in geometric feature space that represents the corresponding $f_i \in F$ and it is important to note that structures with similar family-wise distances will be nearby in the geometric feature space. By constructing the geometric feature space of a family, the structural variation present within an all-against-all substructure alignment (as shown in Fig. 1(a)) is preserved, but distilled into a much simpler representation that is more amenable to common machine learning techniques such as clustering.

3.3 Step 3: Dimensionality Reduction

Understanding the family-wise structural information encoded by **G** will lead to the motivation for the following step–dimensionality reduction. Let, for example, $n = 100$ and consider that the geometric feature vectors, $\mathbf{g}_i \in \mathbf{G}$, will be 100-dimensional, making analysis of the feature space difficult. It is often the case that many structures in a homologous family, as defined by EC or GO for example, will contain several crystallizations of the same protein, from the same species, causing some of the propagated motifs to be nearly identical in geometry. Because of these similar structures, a given $\mathbf{g}_i$ will have some very highly correlated features that increase the dimensionality of the feature vector representation, but do not each provide orthogonal information about the family-wise relationship of f_i to F. Removing similar structures via sequence-identity thresholds requires that a representative structure from the sequence-similar set to be selected. However, sequence-identity removal techniques do not consider the geometric diversity of available structures when selecting a representative structure. The method presented here allows all available structures for a family to be included without filtering for sequence-identity specifically because of the dimensionality reduction step. By including all available structures in the analysis, the method presented here does not make *a priori* assumptions about the sequential or structural diversity of a family of proteins.

While reducing the dimensionality of **G**, it is important to preserve the distances between substructures in feature space, since the purpose of geometric feature encoding is to find sub-groups of related substructures within F. We begin by finding the Principle Components (PCs) of $\mathbf{G}$ and then project $\mathbf{G}$ into a subspace of the PCs that captures at least 90% of the original variance in $\mathbf{G}$; we denote the lower-dimensional projection of $\mathbf{G}$ as $\mathbf{G}'$. The choice of a variance threshold directly impacts the dimensionality of $\mathbf{G}'$, but it is interesting to note that the conservative choice of 90% typically results in $\mathbf{G}'$ being 1- to 5-dimensional, even for large families of more than 1000 structures. PCA [59] was selected for simplicity, but many other dimensionality reduction methods, both linear and non-linear (for example SciMAP [62, 63]), could be substituted and would further improve the dimensionality reduction step. Fig. 1(c) shows the geometric feature vector encoded proteins for the 83-structure heme-dependent peroxidase family as points in the first and second principal components of $\mathbf{G}'$ which capture 94% of the original variance in $\mathbf{G}$; the total number of principal components to reach the minimum 90% variance threshold was 2-components for the peroxidases, so $\mathbf{G}'$ was 2-dimensional in this case. Thus, PCA is able to drastically reduce the dimensionality of the geometric feature space, which is vital to the performance of most clustering algorithms.

3.4 Step 4: Identifying Structural Sub-Groups

One approach to investigating the membership, types, and numbers of structurally related sub-groups within a larger family of proteins is to find clusters of geometrically related structures. Geometric feature vector encoding allows us to represent each protein in a family of structures as a point in feature space, and the process of finding groups or clusters of similar points in feature space can be delegated to an assortment of standard clustering methods.

To choose a clustering method, several key features were deemed important: the method should be able to identify the number of clusters, k, automatically; to avoid bias, no meta-data, such as species information, should be taken into account during clustering–unsupervised learning; the method should be able to identify instances where only a single cluster is sufficient to explain variation; the method should be robust to the presence of outliers; the method should be able to accommodate the presence of both very large, dense sub-groups and small, diffuse sub-groups. Methods that rely on a user-defined number of clusters, such as k-means, are difficult to apply to the problem of identifying significant clusters within F, because the number of clusters, k, is not known *a priori*.

To provide an automated, unbiased selection method for k, a Gaussian Mixture Model (GMM) approach using the MCLUST [60] package for the statistical language **R** was selected for use in this work. MCLUST incrementally adds multivariate Gaussians to the mixture model, fitting them through an iterative Expectation Maximization procedure, and assesses the Bayesian Information Criteria (BIC), while regularizing for model complexity to select a set of Gaussians that maximally explain the data, given the model complexity constraint. The GMM defines, for each data point, the probability that it belongs to the ith Gaussian mixture component and then a hard classification is performed to partition the data points into the mixture components from which the points were most likely to have been generated. The colors of the data points in Fig. 1(c) demonstrate the hard classification, into 4 sub-groups, made by the GMM for the peroxidase family of proteins (EC 1.11.1.7). The final organization of sub-groups based upon substructural similarity shown in Fig. 1(c) is the substructure-based intra-family ontology output by FASST.

3.5 Step 5: Constructing Consensus Motifs

As a family of protein structures grows both in numbers and structural diversity, building substructural templates for the family, as a whole becomes increasingly difficult, just as constructing HMM profiles [64] for a large set of diverse sequences is difficult. By representing each sub-group identified by GMM clustering with a distinct consensus motif, the entire family can then be represented as a collection of consensus motifs which we call a motif ensemble. To build a consensus motif for a given cluster, the propagated motifs belonging to proteins within that cluster were geometrically averaged to construct an artificial consensus structure by the method used in [65]. However, if a non-artificial consensus structure is desired, picking the structure nearest the cluster centroid would also be an effective strategy for finding a representative motif for the cluster. The consensus motif construction process is repeated for each of the *k* clusters identified during (**Step 4**), resulting in a motif ensemble that contains *k* consensus motifs. For example, four sub-groups were identified within the family of peroxidases (as shown in Fig. 1(c)), and therefore the motif ensemble for the family consisted of four consensus motifs, one for each sub-group.

3.6 Step 6: Formulating Hypothesis Tests for Structural Similarity

Comparing a template to target protein structures results in a set of substructure matches of varying quality. To distinguish erroneous matches that are likely to have occurred by chance alone and therefore not functionally related to the template from those matches which have *significant* similarity to the template requires a statistical model of substructure similarity. The non-parametric statistical framework for matching single-substructure motifs used in previous work [48, 7, 61] is extended in this work to address multiple-structure motif ensembles. A detailed discussion of the single-structure statistical model can be found in [48, 61] but is outlined briefly here to motivate the extension to motif ensemble statistical hypothesis testing.

3.6.1 Single-Structure Template Hypothesis Testing

The structural uniqueness of a match of motif S to a target structure T, $M_{S \to T}$ can only be evaluated with respect to a background structure reference set. A reference set should be selected such that is structurally diverse and contains protein structures functionally unrelated to the motif; a detailed analysis of the choice of reference sets can be found in [48] but in this work the 95% sequence identity non-redundant PDB (nrPDB$_{95}$) was used as a structural reference set. Given a background reference set, we can quantify whether the similarity between $M_{S \to T}$ and S is low, relative to the background, and could have occurred by chance, or that it is high, with respect to background, and is statistically significant.

The question of whether or not a match of motif S to a target structure T, $M_{S \to T}$ is significantly similar to S can be formulated as a hypothesis test: the null hypothesis (H_0) states that S and T are structurally dissimilar and that $M_{S \to T}$ occurred by chance; the alternative hypothesis (H_A) states that S and T are structurally similar and $M_{S \to T}$ defines a sub-structural element in T that is analogous to S. Given our definition of a background structural reference set, the p-value of $M_{S \to T}$, $p_{S \to T}$, is a measure of the structurally uniqueness of $M_{S \to T}$ with respect to the defined background reference set. By selecting a p-value threshold for statistical significance, α, we can reject H_0 for all $p_{S \to T} \leq \alpha$ and instead accept H_A and declare $M_{S \to T}$ to be statistically significant. Matching S versus all of the structures defined by the background reference set will yield a distribution of matches with varying levels of structural similarity to S, given by the RMSD of each match to S. By smoothing the RMSD distribution using the Sheather-Jones optimal bandwidth [66] we obtain a probability density function $\text{pdf}(r)$ over RMSD, r, for a given motif S; we denote this pdf as $\text{pdf}(r; S)$.

Given $\text{pdf}(r; S)$, the p-value measure of statistical significance of $M_{S \to T}$ can be found by calculating the probability of observing a match with RMSD, r, lower than the RMSD of $M_{S \to T}$, r_M, which can be written as $P(r \leq r_M; S)$ and defined to be: $\int_0^{r_M} \mathrm{pdf}(r;S)dr$. In summary, the p-value of a given match of a motif to a target protein structure is calculated by comparing the match RMSD to the population of match RMSDs that are expected to occur by chance alone. Using this technique, matches with statistically *unusual* amounts of geometric similarity to a motif can be readily identified without making assumptions about the structure of the match distribution.

3.6.2 Motif Ensemble Statistical Hypothesis Testing

The hypothesis testing framework used for quantitating the statistical significance of matches to a standard, single-structure motif, can be extended naturally to accommodate the notion of matching an ensemble of motifs. Given a motif ensemble with k consensus motifs $\mathbb{S} = \{S_1, S_2, ..., S_k\}$ we would like to know if the motif ensemble, $\mathbb{S}$, has statistically significant similarity to T. For each motif, $S_i \in \mathbb{S}$, we can calculate the p-value of matching S_i to T, $p_{S_i \to T}$, by matching S_i versus the background structure reference set and obtaining the probability density function over match RMSD, r, for motif S_i: $\mathrm{pdf}(r;S_i)$. This procedure produces a p-value for matching each S_i to T, $\mathbf{p}_{\mathbb{S} \to T} = \{p_{S_1 \to T}, p_{S_2 \to T}, ..., p_{S_k \to T}\}$ and, as for normal single structure motifs, an associated hypothesis test for each motif: the null hypothesis ($H_{0,i}$) states that S_i and T are structurally dissimilar and the match of S_i to T occurred by chance; the alternative hypothesis ($H_{A,i}$) states that S_i and T are structurally similar and the match of S_i to T defines a sub-structural element in T that is analogous to S_i. The overall null hypothesis for a match to the motif ensemble can now be stated in terms of the individual hypothesis corresponding to each consensus motif within the motif ensemble: $H_0 = \{H_{0,1}, ..., H_{0,k}\}$.

Because the overall null hypothesis, H_0, incorporates multiple hypothesis tests ($H_{0,1}$, ..., $H_{0,k}$), each of which can introduce new false positive matches, it is crucial to use a multiple testing correction procedure to account for the presence of multiple tests and control the *family-wise error rate*. The Hochberg p-value correction method [67] was selected to account for the presence of multiple tests for significance; Hochberg correction is applicable when the hypothesis tests are either independent or positively correlated [68]. After Hochberg multiple testing correction has been performed on the match p-value, $p_{S_i \to T}$, corresponding to each hypothesis test, $H_{0,i}$, each null hypothesis can then be independently evaluated: $p_{S_i \to T}^{\text{corrected}} < \alpha$. If any null hypothesis, $H_{0,i}$, is rejected, we then reject the overall null hypothesis, H, and consider the match between $\mathbb{S}$ and T to be statistically significant (a positive match).

MOTIF-ENSEMBLE-HYPOTHESIS-TEST($T, \mathbb{S}, \Omega, \alpha$)

```
for all $S_i \in \mathbb{S}$ do
    $\mathrm{pdf}(r; S_i, \Omega) \leftarrow$ MATCH($S_i, \Omega$)                ▷ probability density function
    $r_M \leftarrow$ MATCH($S_i, T$)                                  ▷ the LRMSD of the match
    $p_{S_i \to T} \leftarrow \int_0^{r_M} \mathrm{pdf}(r; S, \Omega)(r) dr$                 ▷ $p$-value of the match
end for
▷ Hochberg multiple testing correction of $p$-values
$\mathbf{p}'_{\mathbb{S} \to T} \leftarrow$ HOCHBERG($p_{S_1 \to T}, ..., p_{S_k \to T}$)

if MINIMUM($\mathbf{p}'_{\mathbb{S} \to T}$) $< \alpha$ then
    return statistically_significant
else
    return not_statistically_significant
end if
```

HOCHBERG($p_1, p_2, ..., p_n$)

```
$\mathbf{p}^{\textbf{sorted}} \leftarrow$ SORT-DESCENDING($p_1, p_2, ..., p_n$)
$\mathbf{p}^{\textbf{corrected}} \leftarrow \emptyset$
for $i = 1$ to n do
    $\mathbf{p}^{\textbf{corrected}} \leftarrow i * p_i^{\mathrm{sorted}}$
end for
return $\mathbf{p}^{\textbf{corrected}}$
```

Chapter 4

Family-wise Analysis of SubStructural Templates

The families of proteins included in this study were analyzed with FASST to construct intra-family ontologies that model the substructural diversity of each family. The underlying source of substructural variation could be clearly attributed to phylogenetic distance, ligation state, or protein ancestry in many cases. The families of proteins highlightd here have a source of substructural variation that can be concretely linked to a single biological factor, in order to better demonstrate the role of each variation source independently. Each structure family was defined by Enzyme Commission (EC) numbers and preference for inclusion into the data set was given to families with a large number of structures. A catalytic site template was defined for each family from a literature reference (see Table 1) using C_α positions. FASST then takes as input the family and template and outputs a substructure-based intra-family ontology for the family in order to identify the substructural variation within a family. The intra-family ontologies of highlighted families are examined in detail below.

4.1 Phylogenetic-based Intra-Family Ontologies

4.1.1 Heme-dependent Peroxidases

Heme-dependent peroxidases (EC 1.11.1.7) are ubiquitous enzymes responsible for moderating reactions with reactive oxygen species. The lactoperoxidases and myeloperoxidases found in animal leukocytes produce potent antibacterial agents and have been shown to

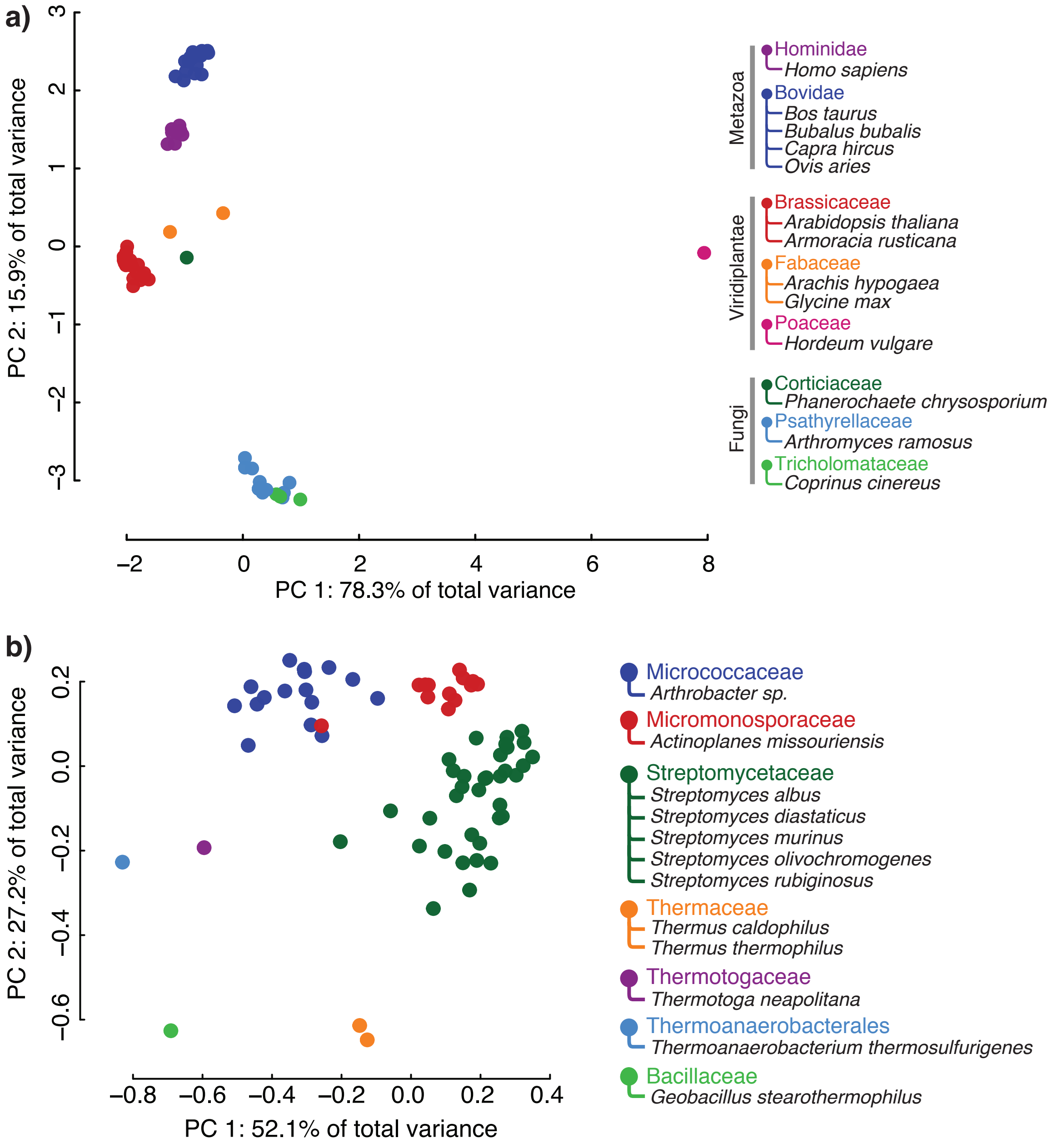


Figure 4.1 : **Phylogenetic-based intra-family ontologies.** **(a)** Substructures positions in the intra-family ontology colored by *Family*-level taxanomic classification reveal that phylogenetic distance between proteins is the main source of substructural diversity within the the family of heme-dependent peroxidases. **(b)** Xylose isomerase structures from 12 different species of bacteria and thermophilic archaea form sub-groups that can be mapped to the Family-level of taxonomic classification.

play a role in inflammatory responses [69]. The non-animal class II peroxidases, found in fungi, and class III peroxidases, found in plants, are both secreted enzymes that are thought to play multiple roles including organism development and pathogen defense [36].

The catalytic site region of the *Arthromyces ramosus* class II peroxidase enzyme [PDB:1ARU] includes the proximal (His-184) and distal (His-56) histidines coordinated to the heme group as well as the distal catalytic residues (Arg-52 and Asn-93) and the hydrogen-bonded Asp-57 [70]. Superposition of all of the heme-dependent peroxidase catalytic site structures, identified through motif propagation as outlined in Ch. 3.1, is shown in Fig. 3.1(a). Although the catalytic site motif can be identified within both animal and non-animal peroxidases, geometric variability of the catalytic residues is evident from the alignment.

The peroxidase intra-family ontology constructed by FASST (see Fig. 3.1(b)) reveals that the peroxidase structures segregate neatly into four main clusters that can be explained well by the phylogenetic ontology of the structures as shown in the corresponding Fig. 4.1(a) plot. The lactoperoxidase structures from *Capra hircus* (goat), *Bos taurus* (cow), *Ovis aries* (sheep), and *Bubalus bubalis* (water buffalo) form a single sub-group in the intra-family ontology nearby the distinct myeloperoxidase sub-group from *Homo sapiens*. The class III plant peroxidases from the *Brassicaceaa Family* form a tight sub-group in the intra-family ontology with the class III plant peroxidases of the *Fabaceae Family* near the perimeter, but outside the main sub-group. Finally, the class II fungal peroxidases form a fourth distinct sub-group most distant from the other three sub-groups.

The location of the peroxidase catalytic site substructures in the intra-family ontology appears to be highly correlated with the evolutionary history of the enzyme. The animal and non-animal peroxidases are theorized to have originated from two separate endosymbiotic events predating modern plant and animal cells [36]. The sequence identity between the human [PDB:1CXP] and fungal [PDB:1ARU] versions of the enzyme is 9% making a sequence-based approach to analyzing this family as a whole impossible.Pairwise sequence-identity between the labeled positions in Fig. 3.1(b) is consistently low as seen in the table below:

	1ARU	1BGP	1H58
1CXP	9%	7%	6%
1ARU	-	14%	7%
1BGP	-	-	40%

As shown in Fig. 2.2, the overall fold topology of the animal and non-animal peroxidases differ greatly and belong to separate fold classes within the CATH structural ontology [71]. However, the catalytic substructure represented by the motif provides a common point of comparison between these peroxidases and allows FASST to identify the significant family-wise catalytic site variation and underlying sub-groups within the larger protein family. By mapping the intra-family ontology to the *Family*-level phylogenetic ontology, FASST is able to propose a hypothetical explanation for the pattern of substructural conservation and variation within the family of peroxidases.

4.1.2 Xylose Isomerases

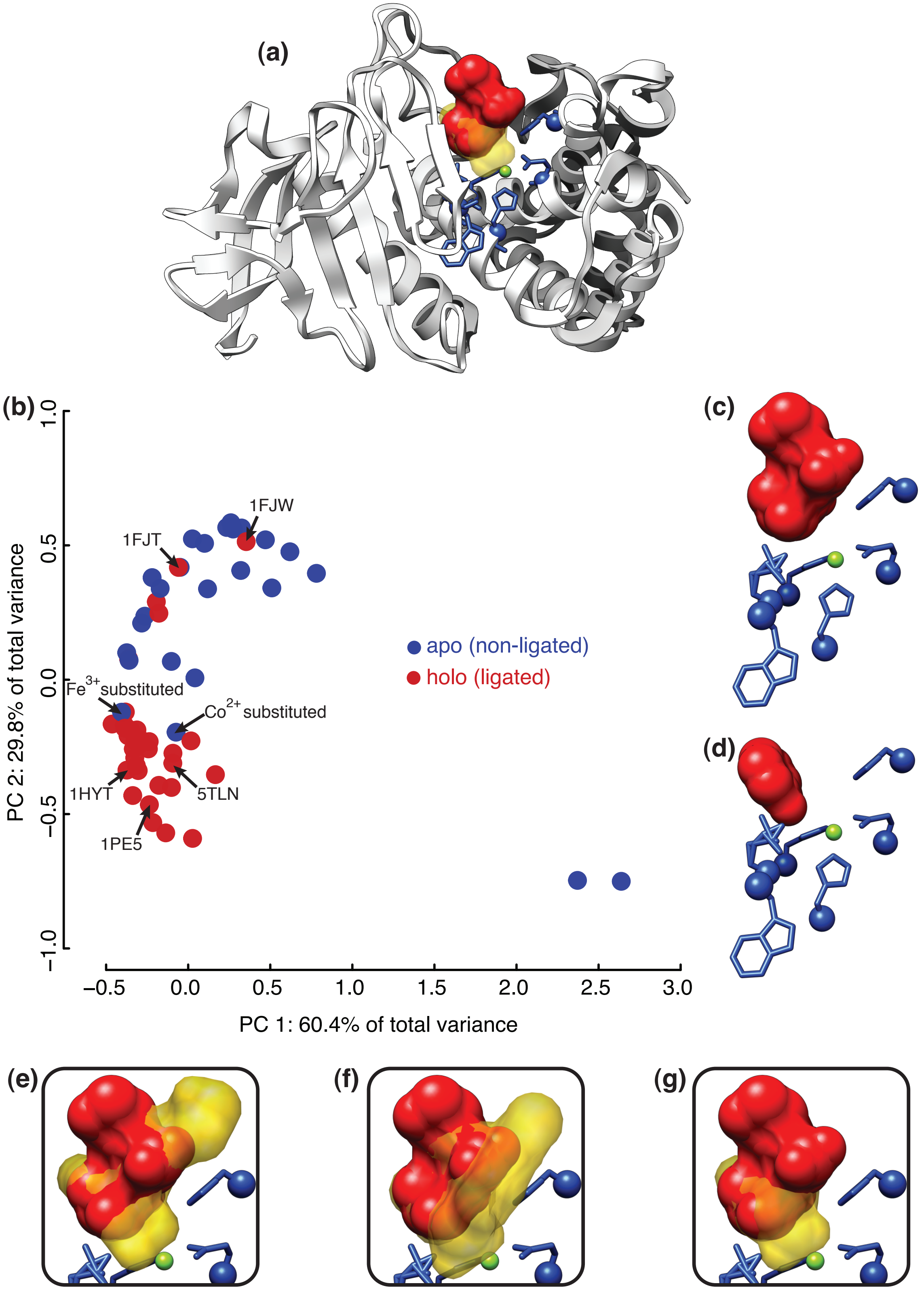
Metabolic engineering approaches to creating organisms capable of producing biofuels, such as ethanol, from previously unrecoverable plant biomass are being actively studied in the search for renewable energy sources [72]. Xylose isomerase is a key enzyme in many engineered biosynthetic pathways because of its ability to interconvert sugar isomers, allowing novel carbohydrate sources, such as plant biomass, to be utilized over more traditional sugar substrates such as glucose [73]. While members of the peroxidase family demonstrate topological diversity, the family of xylose isomerases (EC 5.3.1.5) are more topologically homogenous, and provide another clear example of sub-groups in a substructure-based intra-family ontology that can be linked to the corresponding phylogenetic ontology of the structures.

Applying FASST to the catalytic sites of 71 structures of xylose isomerase from 12 different species, including thermophilic archaea and several species of mesophilic bacteria, reveals that variation in catalytic site geometry within the family can be well-explained by the *Family*-level phylogenetic ontology of the family. As shown in Fig. 4.1(b), the closely-packed, but well-defined clusters of structures clearly map to the phylogenetic labeling at the *Family*-level of taxonomic classification. While the xylose isomerase family exhibits high structural conservation, understanding the substructural relationship between related members of enzymatic families, capable of catalyzing the same reaction under different environmental conditions, is an important step towards rational design of biosynthetic pathways.

4.2 Ligation-based Intra-Family Ontologies

Many proteins are known to undergo structural rearrangements and hinge-bending motions upon binding ligands or other proteins. Induced fit via amino acid rearrangements are a common feature of many catalytic sites, and the state of the catalytic site at a given time can often be partitioned into two states: *apo*, an open confirmation with no ligand, and *holo*, a closed confirmation with bound ligand. The thermolysins (EC 3.4.24.27) are a family of bacterial heat-stable metalloproteases that cleave peptide bonds at hydrophobic residue sites and have been shown to change confirmations upon ligand-binding [74].

The family of available thermolysins contains 59 structures of the protein from *Bacillus thermoproteolyticus* and a single structure from both *Staphylococcus aureus* and *Bacillus cereus*, all of which are gram-positive bacteria species (*Bacillales*). Because there are


(a)
(b)
1.0
0.5
0.0
−0.5
−1.0
PC 2: 29.8% of total variance
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
PC 1: 60.4% of total variance
1FJW
1FJT
Fe^{3+}substituted
Co^{2+}substituted
1HYT
5TLN
1PE5
● apo (non-ligated)
● holo (ligated)
(c)
(d)
(e)
(f)
(g)

roughly equal numbers of apo (non-ligated) and holo (ligated) structures within the family, and all but two of the structures are repetitions of the same protein from the same species, the effect of ligation state on the substructural variation of the catalytic site can be analyzed in isolation from other possible contributing factors such as phylogenetic distance. The substructure-based intra-family ontology for the thermolysins shown in Fig. 4.2(b) reveals that the structures partition very cleanly into two distinct sub-groups which can be clearly mapped to the ligation-state of the structure. However, as can be seen in Fig. 4.2(c), there are 5 holo structures in the apo region and 2 apo structures within the holo region.

Further investigation into the two apo outlier structures, shown to reside in the holo region of Fig. 4.2(b), reveals that these two proteins were artificially modified to coordinate

Figure 4.2 *(preceding page)*: **Ligation-state conformational changes in thermolysin.** **(a)** Backbone of thermolysin structure [PDB:1FJT] with coordinated valine-lysine dipeptide in red and template/motif residues shown in blue. Side-chains of the template residues are shown for reference, but only C_α coordinates are used by LabelHash in this paper. The yellow, semi-transparent volume corresponds to the superimposed benzylsuccinic acid ligand of [PDB:1HYT]. The coordinated Zn^{2+} ion is depicted as a small green sphere in the center of the template residues. The binding positions of the two ligands are superimposed to illustrate where the occupied regions of the thermolysin binding site differ between the two ligands. **(b)** Applying FASST to the family of thermolysin structures reveals that apo and holo structures segregate into different regions of the intra-family ontology. The segregation of structures seen indicates that the template residues undergo conformational change upon binding a ligand. The location of particular structures in the intra-family ontology are labeled for reference. **(c)** Holo outlier structure [PDB:1FJT] with bound valine-lysine dipeptide. The ligand sits in the side-chain recognition pocket of the thermolysin but does not interact with the Zn^{2+} ion and does not induce conformational change of the template residues. **(d)** Phenol ligand of holo outlier structure [PDB:1FJW] superimposed with the [PDB:1FJT] binding site for consistent reference. The phenol ligand also sits in the side-chain recognition pocket and does not induce conformation change of the template residues. **(e),(f),(g)** Ligated inhibitors from [PDB:5TLN], [PDB:1PE5], and [PDB:1HYT], respectively, in semi-transparent yellow superimposed with the [PDB:1FJT] binding site. These 3 inhibitors interact directly with the coordinated Zn^{2+} ion and induce conformational change in the binding site.

Co^{2+} and Fe^{3+} metals within their catalytic sites, instead of the normal Zn^{2+} metal found in nature. The substitution of Co^{2+} and Fe^{3+} for Zn^{2+} alters the geometry of the catalytic site, effectively converting thermolysin into the "closed," ligand-bound holo state [75]. This fact explains why these two artificially substituted apo outliers have higher substructural similarity to the holo structures and are co-located with the holo structures in the intra-family ontology shown in Fig. 4.2(b).

Closer examination of the five holo structures that reside in the apo region reveals that either lysine or phenol is bound to the structurally rigid side-chain recognition pocket of these structures in all five cases. In Fig. 4.2(c), the catalytic site of one of the five holo outliers [PDB:1FJT], where a valine-lysine dipeptide is bound near, but not within the catalytic site, is compared to a holo structure with a ligand bound for catalysis in Fig. 4.2(e,f,g). The ligand in Fig. 4.2(e,f,g) can be clearly seen to interact with the catalytic residues as well as the coordinated catalytic metal (Zn^{2+}) but the ligand of [PDB:1FJT] is bound just outside of the catalytic site. Binding of the valine-lysine/phenol ligands to the side-chain recognition pocket of thermolysin in the five holo outliers does not induce the catalytic site to alter its geometry, explaining the presence of these holo outliers in the apo region of the plot in Fig. 4.2(b). Therefore, FASST is able to accurately distinguish between the holo and apo structures of the family of thermolysins by modeling the family-wise substructural variation of the catalytic residues and constructing the substructure-based intra-family ontology.

4.3 Ancestry-based Intra-Family Ontologies

Some protein substructures have proven themselves, throughout the course of evolution, to be so well-suited at catalyzing particular reactions, that they have arisen independently in different kingdoms of life. One such example of convergent evolution in protein substructures is the HIS-ASP-SER catalytic triad which catalyzes the hydrolysis of peptide

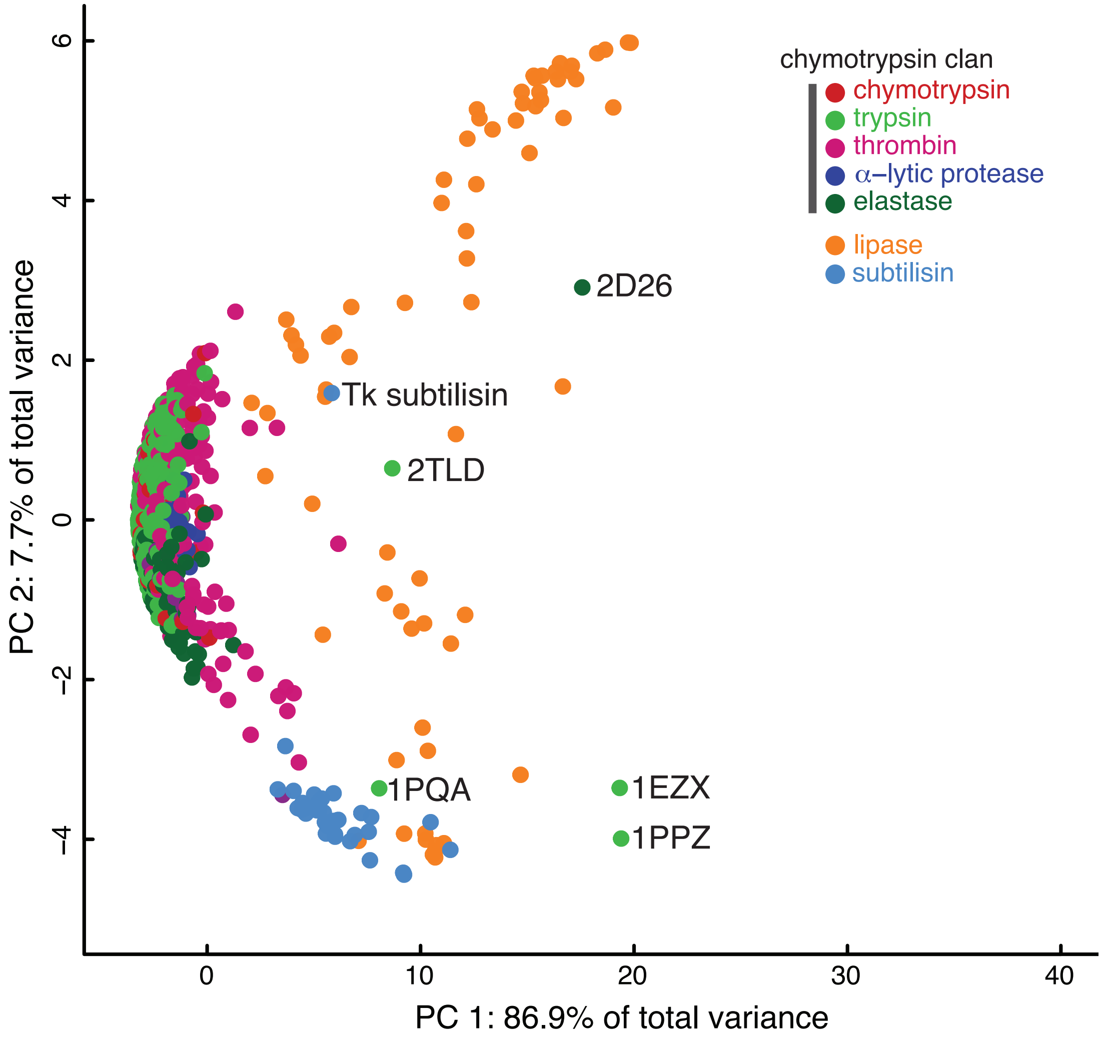


Figure 4.3 : **Catalytic triad diversity among serine protease families.** Comparing the geometry of the ubiquitous His-Asp-Ser catalytic triad across 730 structures, 52 species, and 7 EC families demonstrates the scalability of FASST to very large numbers of structures. All of the divergently-related families of the chymotrypsin clan form a single dense sub-group while the convergently-related subtilase family forms a separate sub-group in the intra-family ontology. The highly diverse family of lipases form several small sub-groups distinct from both the chymotrypsin-like and subtilisin-like structures. Several trypsin outlier structures are labeled and the references corresponding to each PDB entry document sources of catalytic site deviation.

bonds in many serine proteases [3]. The HIS-ASP-SER catalytic triad is a common substructure among many families of proteases and the geometry of the triad residues across protease families has been shown to be highly conserved [14]. To demonstrate the ability of FASST to detect substructure variation among convergently related families where the triad substructure resides in many configurations, including spanning peptide chains, we have considered all of the non-mutant protein structures from the following families in an analysis of the serine protease catalytic triad:

Family	**EC Class**	**# Structures**
Chymotrypsin	3.4.21.1	57
Trypsin	3.4.21.4	355
Thrombin	3.4.21.5	247
α-lytic protease	3.4.21.12	39
Elastase	3.4.21.36	90
Triacylglycerol lipase	3.1.1.3	107
Subtilisin	3.4.21.62	94

The mutant-filtered family of serine protease structures included 730 protein structures spanning 7 EC classifications and 52 species; the total number of structures in the table is 989 of which 259 are mutant structures. The input motif consisted of the C_α coordinates of the triad residues and was geometrically based upon the 1ACB chymotrypsin structure; this motif was able to accurately identify triad residues in all serine protease families, including cases where the triad residues span peptide chains. Correct identification of triad residues for all propagated motifs was subsequently confirmed prior to applying FASST.

The chymotrypsin, trypsin, elastase, thrombin, and α-lytic protease families are all divergently evolved proteases of the “chymotrypsin clan” (clan SA)[3] and share a common fold that differs from the convergently evolved subtilisin family of proteases. The triacylglycerol lipases have also convergently evolved the serine-based triad and form a third distinct evolutionary group [76]. Application of FASST to the families of serine proteases, as shown in Fig. 4.3, reveals that consistently, proteins of the chymotrypsin clan group together with high degrees of overlap in in the intra-family ontology and the subtilisin structures form a distinct sub-group outside of the chymotrypsin clan sub-group. Within the chymotrypsin clan, the different families of serine proteases show only subtle variations in triad geometry and are nearly inseparable from one another. It is evident from analysis of the intra-family ontology shown in Fig. 4.3 that the lipases exhibit much more catalytic triad geometric variability, overall, than either the subtilisins or chymotrypsins, as they can be seen in many different regions of the space.

Outlier structures within the intra-family ontologies output by FASST, labeled in Fig. 4.3, were further investigated. One of the most extreme outliers in Fig. 4.3 corresponds to a pancreatic elastase structure [PDB:2D26] complexed with α-1 antitrypsin, and this complex was documented to introduce extensive distortion to the catalytic site [77], well-explaining the distant position of this structure from other proteins in the intra-family ontology. Similarly, two trypsin outlier structures ([PDB:2TLD] and [PDB:1EZX]) denoted in Fig. 4.3 are complexed with a protein inhibitor that was documented to cause distortion of the catalytic site. Two trypsin structures ([PDB:1PQA] and [PDB:1PPZ]), crystallized with sub-atomic resolution, are also distant from the main chymotrypsin sub-group in the substructure-based intra-family ontology[78]. The single non-mutant Tk-subtilisin structure, from the archaeon *Pyrococcus kodakaraensis*, is found to be distant from both the chymotrypsin clan sub-group and main subtilisin sub-group, which suggests a mode of geometric variation different from that of prokaryotic subtilisins and chymotrypsin-like triads. Application of FASST to the serine proteases clearly demonstrates the extremely high degree of both chemical and structural conservation of the catalytic triad across very diverse species and proteins with diverse ligand specificities. Impressively, modeling only the triad C_α positions, as was done here, is sufficient to recover the super-family organization of the serine proteases.

Chapter 5

Protein Function Prediction

FASST provides a method to expose the underlying intra-family ontology of a protein family and the MESH framework utilizes the sub-groups within the ontology to enhance the function prediction power of substructure templates. Instead of representing an entire protein family with a single substructural motif, FASST-MESH uses an ensemble of motifs, where each motif within the ensemble is used to represent a sub-group within the intra-family ontology. MESH automatically constructs a representative consensus motif for each sub-group of geometrically related family members output by FASST (see Ch. 3.5). Collectively, the set of consensus motifs for all sub-groups within a intra-family ontology compose a motif ensemble. Earlier work investigated the performance of averaging all substructures within a family to identify a single family consensus template [79]. However, we found that for large geometrically diverse families, a single representative motif, based on any family member substructure or a substructure average of all members, could not sufficiently represent the entire family, just as building a single profile HMM for a large number of distantly related sequences can be difficult. Transitioning to the multiple-model motif ensemble, however, requires that the statistics employed by MESH to distinguish statistically significant matches take into account the presence of multiple tests for significance, one test for each consensus motif in the ensemble (see Ch. 3.6).

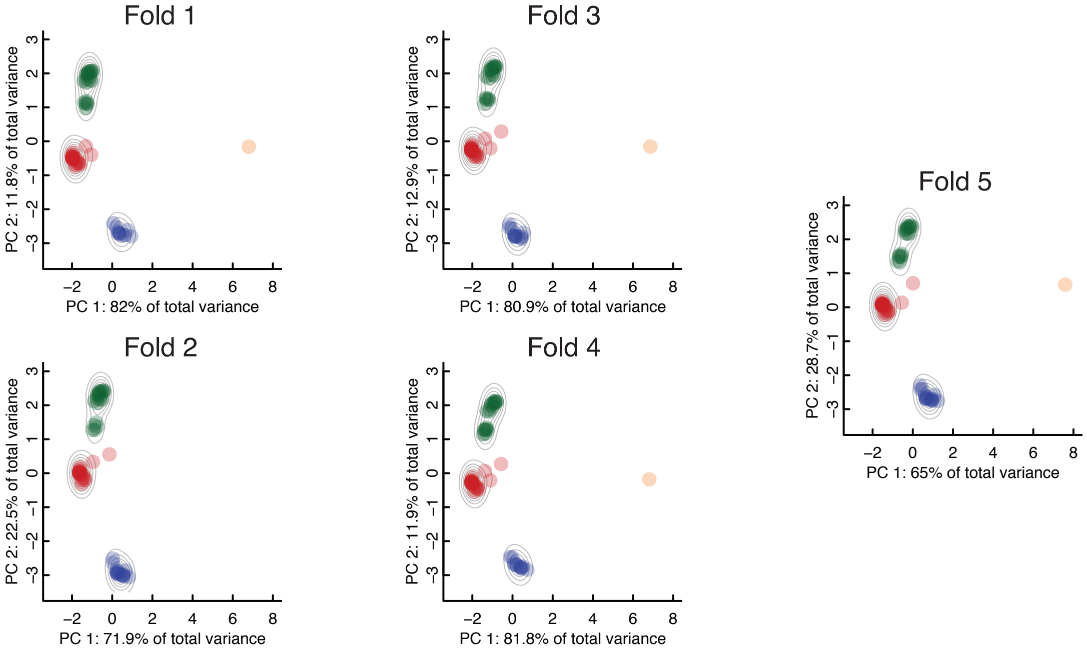


Figure 5.1 : **Robustness of clusters to data removal during 5-fold cross validation.** During each step of cross-validation, FASST-MESH is used to identify clusters and construct a motif ensemble for the family of peroxidases seen here.

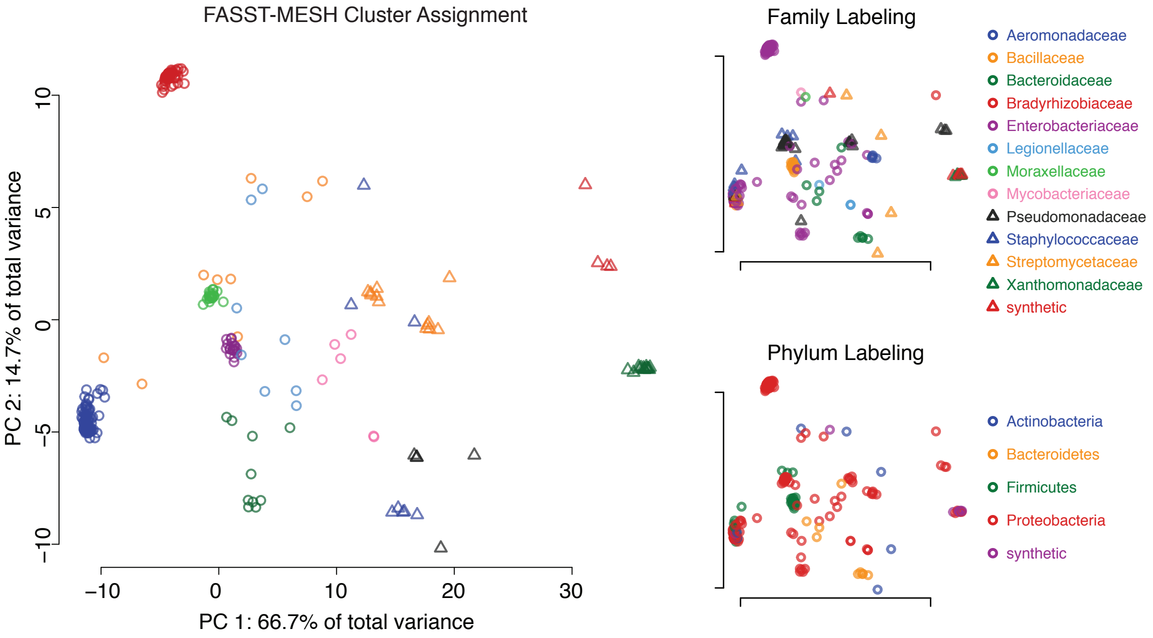


Figure 5.2 : **Sub-groups identified by FASST-MESH within the β-lactamases.** Applying FASST to expose the substructural diversity of a catalytic substructure among the β-lactamases reveals many distinct sub-groups within the family. The GMM clustering step of FASST identifies 13 sub-groups within the family and the colors/shapes of points in the intra-family ontology correspond to sub-group assignment. MESH then constructs one consensus motif for each sub-group identified, resulting in an ensemble of 13 motifs. Functional annotation sensitivity improves from 35.0% (single-structure motif) to 81.2% when using the motif ensemble constructed by FASST-MESH. For the highly diverse family of β-lactamases, the intra-family ontology output by FASST shows that many distinct sub-groups exist within the family. MESH takes advantage of this information to more completely model the geometric diversity present, thereby improving functional annotation coverage of the family. Mapping *Family*- and *Phylum*-level phylogenetic data to each of the substructures as shown in the corresponding plots on the right reveals that some, but not all, of the sub-groups identified are due to evolutionary distance between proteins. For example, the *Bacillaceae* proteins can be seen to form a single sub-group while *Enterobacteriaceae* proteins are distributed throughout the intra-family ontology in several sub-groups, indicating that another biological factor is working in concert with phylogenetic distance among the family of β-lactamases to produce the structural diversity uncovered by FASST.

5.1 Quantifying Function Prediction Performance

FASST-MESH was used to construct motif ensembles for 15 families of enzymes (see Table 1), as defined by Enzyme Commission (EC) number, and the performance of these motif ensembles was compared to single-structure motifs in a set of functional annotation experiments (see Table 2). Function prediction performance can be quantified by *sensitivity*, the percent of True Positives (TP) correctly identified (# TP / (# TP + # FN)), and *specificity*, the percent of True Negatives (TN) correctly identified (# TN / (# TN + # FP)). Because the process of constructing a motif ensemble can be considered *unsupervised learning* of the family substructure space, 5-fold cross-validation was implemented, where the motif ensemble was built from 4/5 of the data and then the last 1/5 was used for performance assessment. The robustness of the sub-groups identified in the intra-family ontology during cross-fold validation (as shown in Fig. 5.1) can be seen by the stability of the sub-groups during each of the 5 cross-fold validation steps. Two EC families included in the functional annotation experiments are discussed below, and each demonstrates a different extreme of sensitivity/specificity improvement after applying FASST-MESH.

5.2 β-lactamases

The diverse family of β-lactamases (EC 3.5.2.6) includes structures from 26 different bacterial species. Using the 13 sub-groups identified from the intra-family ontology output by FASST as shown in Fig. 5.2, MESH constructs a consensus motif for each sub-group, resulting in an ensemble of 13 consensus motifs. The β-lactamase motif ensemble, constructed by FASST-MESH, identified 81.2% of functionally homologous proteins (as defined by the EC class) with statistically significant substructure matches. The corresponding single-structure β-lactamase motif only identified 35.0% of functional homologs, and therefore FASST-MESH improved the functional annotation sensitivity of the single-structure motif by 2.3-fold while maintaining the high specificity of the single-structure motif.

5.3 Heme-dependent peroxidases

In the family of peroxidases (EC 1.11.1.7) analyzed in Fig. 3.1, a single-structure motif was capable of identifying a statistically significant match for 91.6% of the EC family, and therefore already showed high sensitivity. After applying FASST-MESH to the single-structure peroxidase motif, annotation sensitivity improved only slightly ($\sim$1% improvement) but the absolute number of false positive matches identified decreased from 131 to 78±8. The decrease in false positive matches, due to using a motif ensemble, occured because true positive matches tended to match multiple consensus motifs within the ensemble with low LRMSD, while many false positive matches have only marginally significant LRMSD to a single consensus motif, and applying multiple testing correction to the final set of matches for a given false positive often caused a single marginally significant match to move outside of the significance threshold.

As both the number and diversity of available structures for a given protein family continue to increase, the enhancement of substructure-based functional annotation methods to accommodate large families is necessary. This work demonstrates an automated approach (outlined in Ch. 3) to augment existing substructure templates already available in repositories such as the Catalytic Site Atlas (CSA) [29] by geometrically enriching motifs for families that exhibit high structural variability. The automated enrichment of available templates by FASST-MESH strengthens the function prediction power of these templates and facilitates the use of substructure-based analysis methods for large-scale, automated annotation of novel structures.

Table 5.1 : **Full protein family dataset used for functional annotation experiments.** For each EC class family, a single PDB structure was used to define an input motif (template). The list of amino acid numbers are documented functional residues found within the primary PDB (www.pdb.org) reference corresponding to each PDB structure. The superscript labels above each amino acid number are the possible amino acid types that can match at each motif point; further details of alternate amino acid label use can be found here [7]. Where multiple amino acid labels per motif point appear, they were determined using ConSurf [8].

EC class	**PDB ID (Chain)**	**Amino acid number**$^{\text{Labels}}$	**EC class size**
1.1.1.1	1HET (A)	46^{C},48^{S},67^{H},174^{C}	82
1.1.1.21	1US0 (A)	43^{D},48^{Y},76^{S},77^{K},110^{H}	89
1.11.1.7	1ARU (A)	52^{RQ},56^{H},57^{D},93^{NR},184^{H}	83
1.14.13.39	1DWW (A)	194^{C},346^{V},363^{F},366^{W},367^{Y}	126
2.5.1.18	2A2R (A)	7^{Y},13^{FLR},47^{ACFLM},108^{CFLY}	190
2.6.1.1	2QA3 (A)	32^{G},34^{G},183^{N},374^{R}	105
2.7.4.6	1NHK (R)	51^{Y},117^{H},119^{S},128^{K}	60
3.1.1.7	1H23 (A)	84^{W},117^{G},130^{Y},279^{W},330^{F}	110
3.1.3.1	1ANI (A)	51^{D},101^{D},102^{S},331^{H},412^{H},	44
3.1.3.48	2CM2 (A)	181^{DE},182^{FHMY},216^{S},221^{R},266^{Q}	248
3.2.1.1	1HT6 (A)	52^{G},178^{R},180^{D},205^{E},291^{D}	133
3.5.2.6	1YLJ (A)	70^{S},73^{K},130^{S},132^{N}	254
4.2.1.1	1HCB (A)	94^{H},96^{H},106^{E},119^{H},199^{T}	282
5.3.1.1	1YPI (A)	12^{K},95^{H},96^{S},165^{A}	95
5.3.1.5	1DID (A)	53^{H},56^{D},93^{F},136^{W},182^{K}	73

Table 5.2 : **Functional annotation performance of motif ensembles versus single-structure motifs at significance threshold of $\alpha = 0.01$.** For each single-structure motif, a motif ensemble was constructed using FASST-MESH. Next to each % sensitivity value is the total number of true positive (TP) matches; next to each % specificity value is the total number of false positive (FP) matches. The performance of motif ensembles was assessed using 5-fold cross validation and the sensitivity/specificity values correspond to mean ± standard deviation across the 5 folds. The x-fold improvement is calculated as: mean motif ensemble performance divided by single-structure performance.

	Single structure motif		**Motif ensemble (CV)**		**Improvement (x-fold)**	
EC class	%Sens. (#TP)	%Spec. (#FP)	%Sens. (#TP)	%Spec. (#FP)	Sens.	Spec.
1.1.1.1	52.4% (43)	99.2% (83)	74.3±7.0% (61)	99.4±0.0% (62±4)	1.4	1.0
1.1.1.21	93.3% (83)	99.1% (146)	93.2±4.8% (83)	99.2±0.1% (136±5)	1.0	1.0
1.11.1.7	91.6% (76)	99.1% (131)	92.7±10.0% (77)	99.5±0.0% (78±8)	1.0	1.0
1.14.13.39	90.5% (114)	99.3% (87)	96.1±2.7% (121)	99.4±0.0% (73±7)	1.1	1.0
2.5.1.18	25.3% (48)	99.1% (171)	46.3±5.1% (88)	99.2±0.0% (140±5)	1.8	1.0
2.6.1.1	66.7% (70)	99.1% (153)	82.9±5.4% (87)	99.3±0.0% (121±5)	1.2	1.0
2.7.4.6	81.7% (49)	99.2% (137)	88.3±2.6% (52)	99.4±0.1% (113±5)	1.1	1.0
3.1.1.7	98.2% (108)	99.2% (82)	99.0±2.0% (108)	99.4±0.0% (60±2)	1.0	1.0
3.1.3.1	84.1% (37)	99.1% (122)	100.0±0.0% (44)	99.3±0.0% (97±6)	1.2	1.0
3.1.3.48	28.6% (71)	99.1% (155)	56.1±3.6% (139)	99.4±0.1% (109±11)	2.0	1.0
3.2.1.1	83.5% (111)	99.1% (149)	88.7±7.9% (117)	99.4±0.1% (102±17)	1.1	1.0
3.5.2.6	35.0% (89)	99.2% (144)	81.2±6.3% (208)	99.4±0.0% (107±9)	2.3	1.0
4.2.1.1	87.9% (248)	99.1% (112)	95.3±3.5% (269)	99.6±0.0% (49±4)	1.1	1.0
5.3.1.1	78.9% (75)	99.1% (143)	82.1±10.9% (78)	99.4±0.1% (100±11)	1.0	1.0
5.3.1.5	97.3% (71)	99.1% (118)	98.5±2.3% (71)	99.4±0.1% (92±11)	1.0	1.0

Chapter 6

Discussion and Conclusions

Understanding the significant geometric variability among enzyme catalytic sites is an important component of structural analysis. As the number of solved protein structures grows, methods capable of summarizing and analyzing large amounts of structural data will become increasingly necessary. While whole structure alignment and protein fold analysis can be a valuable tool for assessing protein homology, in the absence of sequence similarity, extremely distantly related enzymes or enzymes which are examples of convergent evolution may be ill-suited to whole structure comparison techniques. However, when no detectable domain or fold homology exists, enzymes are still capable of exhibiting functional equivalence through chemically and geometrically synonymous functional substructures. Techniques capable of assessing the family-wise similarity of these conserved substructures can reveal new insights into the relationships among families of structures. FASST has the ability to recognize modes of family-wise geometric variation among substructures and knowledge of the substructural diversity of a family can guide hypotheses about the role of the substructure in different proteins.

6.1 Biological Significance of Intra-family Ontologies

In several families of proteins, possible biologial sources of geometric variation have been identified and linked with the the structural sub-groups automatically identified by FASST. In the peroxidase family, the geometric distance between catalytic sites appears to be correlated with phylogenetic distance. Organisms that are more closely related, such as the plant and fungal species, were shown to have more geometrically similar catalytic sites to one another than to more distantly related phyla, such as vertebrates. With the family of thermolysin structures, we demonstrated how FASST automatically captures modes of catalytic site flexibility, correctly segregating structures into sub-groups based upon ligation state. Using the families of serine proteases, we demonstrated how FASST extends naturally to very large numbers of structures and is still capable of identifying the major modes of geometric variation across vast numbers of species and triad configurations that include chain spanning and non-spanning instances. Finally, FASST is able to identify structural outliers within families, and these outliers were shown to have biochemical causes for substructural deviation from the remainder of the family, thereby guiding further inquiry to these anomalous structures.

FASST partitions a protein family into self-similar sub-groups of structures and in doing so, constructs an intra-family ontology that can then be linked with biological metadata to possibly explain the family-wise diversity. Here particular protein families have been highlighted whose substructural diversity can be clearly linked to a single biological ontology, such as phylogeny, ligation state, or ancestry. In several families included in the function prediction experiments, the sub-groups identified by FASST cannot be clearly attributed to a single biological factor. The β-lactamases are an example where some sub-groups clearly correspond to a single phylogenetic branch of bacteria, but other species of bacteria form multiple, distinct sub-groups as shown in Fig. 5.2. In the typical case, there are likely multiple biological factors working in concert to produce substructural variability. Future work will combine large-scale metadata analysis with FASST to automatically correlate likely biological factors, such as phylogeny, ligation state, and crystallization conditions, with FASST-identified sub-groups to unravel more complex relationships among functional substructures.

6.2 Differentiating Sequential and Structural Redundancy

Using FASST to analyze a catalytic site substructure of thermolysin among 61 sequence-similar proteins demonstrates how latent biological trends can be identified even within a sequentially-homogenous collection of structures. The thermolysin family examined here contained 59 different structures of the exact same enzyme from *B. thermoproteolyticus* and yet FASST was able to automatically uncover a structural trend where the catalytic substructure modified its position only upon binding ligands that interact directly with the coordinated zinc ion. If only sequentially non-redundant structures were used by FASST, this trend could not have been identified because of the miniscule number of sequentially-distinct crystallographic structures for thermolysin. This result demonstrates the additional information that can be garnered by researchers when all available structures are incorporated into a structural analysis. Similarly, the Multiple Solvent Crystal Structures (MSCS) technique utilizes repeated crystallizations of the same enzyme under different solvent conditions in order to probe for functional sites [80, 81]. Several of the available thermolysin structures incorporated in our study were produced as part of MSCS experiments [82, 83]. Techniques, such as FASST, that can detect subtle trends among sequentially-similar structure collections are important tools for analyzing and understanding structure-function relationships across large numbers of protein structures.

6.3 FASST-MESH Improves Single-Structure Templates

After identifying both the existence and membership of structurally defined sub-groups within a protein family via the automated FASST-MESH framework, this substructural information can be used to enhance existing substructural templates in order to more accurately represent large families with diverse catalytic site geometry. The function prediction experiments presented show that representing a structurally diverse family with a motif ensemble better captures the variety of substructures present within a given family and increases function prediction sensitivity while maintaining specificity. In cases where family-wide geometric diversity was found to be low, single structure motifs alone can have high sensitivity. However, even when geometric variability is low, motif ensembles created by FASST-MESH always maintain the function prediction performance of single structure motifs and demonstrate vast improvement in several cases among the families included in this study (see Tables 1 and 2). While LabelHash was used here as the underlying substructure comparison tool, this thesis is not attempting to compare the performance of LabelHash to other comparison tools. Rather, the purpose of the function prediction experiments presented here is to illustrate cases where a single-structure template insufficiently models a large class of functionally homologous, but structurally diverse proteins, and to demonstrate a method to improve the function prediction sensitivity of templates in general by using motif ensembles.

6.4 Automated Template Definition

In this thesis, the substructure templates given as input to FASST (see Table 1) were constructed only from residues that have been experimentally confirmed to play a role in enzyme function in order to separate the subproblem of template definition from template analysis. While the input single-structure templates used here were manually defined, a multitude of automated approaches to template definition are possible. Our previous work successfully used evolutionarily conserved residues, as determined by Evolutionary Trace[38], for automated template definition [48].

Because templates are an input parameter to FASST, different methods of identifying the residues constituting functional substructures can be used in conjunction with FASST, and by doing so, FASST provides an automated approach to further analyze and understand the role of these substructures. In future work, several substructure selection methods and databases, such as CASTp, ET [38], ConSurf [8], CSA [29], SNAP [33], and LigBase [28], will be used as sources for large numbers of templates. This thesis used only residues deemed to be functionally important by experimentalists, as defined by literature references, in order to isolate the performance of FASST-MESH from substructure selection methods.

6.5 Future Applications

FASST has been shown to be a powerful technique for assessing family-wise geometric variability among analogous protein substructures. Many proteins are known to have structurally conserved, but non-catalytic substructures, such as steric recognition sites, metal/ligand sequestering sites, phosphorylation sites, cofactor binding sites, or immunologically important substructural epitopes. Using the FASST-MESH approach for these non-catalytic substructures can be done without modification to the method because FASST-MESH makes no assumptions about the types of substructures modeled by templates nor underlying sources of structural variation. As the available number of protein structures continues to rapidly grow, methods for automated, large-scale analysis of structures such as FASST-MESH will be critical for identifying high-level structural trends among proteins and placing newly solved structures in the larger context of existing structural data.

Bibliography

[1] R. K. Wierenga, “The tim-barrel fold: a versatile framework for efficient enzymes.,” *FEBS Lett*, vol. 492, pp. 193–198, Mar 2001.

[2] N. Nagano, C. A. Orengo, and J. M. Thornton, “One fold with many functions: the evolutionary relationships between tim barrel families based on their sequences, structures and functions.,” *J Mol Biol*, vol. 321, pp. 741–765, Aug 2002.

[3] N. D. Rawlings and A. J. Barrett, “Families of serine proteases,” *Methods in Enzymology*, vol. 244, pp. 19–61, 1994.

[4] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang, “CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues.,” *Nucleic Acids Research*, vol. 34, pp. W116–8, Jul 2006.

[5] R. A. Laskowski, “Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.,” *J Mol Graph*, vol. 13, pp. 323–330, Oct 1995.

[6] A. R. Kinjo and H. Nakamura, “Comprehensive structural classification of ligand-binding motifs in proteins.,” *Structure*, vol. 17, no. 2, pp. 234–246, 2009 Feb 13.

[7] M. Moll and L. E. Kavraki, “Matching of structural motifs using hashing on residue labels and geometric filtering for protein function prediction,” in *Proc. of the Seventh Annual Intl. Conf. on Computational Systems Bioinformatics*, pp. 157–168, 2008.

[8] F. Glaser, T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal, “ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information,” *Bioinformatics*, vol. 19, no. 1, pp. 163–164, 2003.

[9] E. C. Meng, B. J. Polacco, and P. C. Babbitt, “Superfamily active site templates.,” *Proteins*, vol. 55, pp. 962–976, Jun 2004.

[10] S. C.-H. Pegg, S. D. Brown, S. Ojha, J. Seffernick, E. C. Meng, J. H. Morris, P. J. Chang, C. C. Huang, T. E. Ferrin, and P. C. Babbitt, “Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database.,” *Biochemistry*, vol. 45, pp. 2545–2555, Feb 2006.

[11] D. Rognan, “Chemogenomic approaches to rational drug design.,” *British Journal of Pharmacology*, vol. 152, pp. 38–52, Sep 2007.

[12] T. Klabunde, “Chemogenomic approaches to drug discovery: similar receptors bind similar ligands.,” *British Journal of Pharmacology*, vol. 152, pp. 5–7, Sep 2007.

[13] W. Hendrickson, “Impact of structures from the protein structure initiative,” *Structure*, vol. 15, no. 12, pp. 1528–1529, 2007.

[14] A. C. Wallace, R. A. Laskowski, and J. M. Thornton, “Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His-Asp catalytic triads in the serine proteinases and lipases,” *Protein Science*, vol. 5, no. 6, pp. 1001–1013, 1996.

[15] N. Nagano, C. A. Orengo, and J. M. Thornton, “One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions.,” *Journal of Molecular Biology*, vol. 321, pp. 741–765, Aug 2002.

[16] A. L. Bowman, M. G. Lerner, and H. A. Carlson, “Protein flexibility and species specificity in structure-based drug discovery: dihydrofolate reductase as a test system.,” *Journal of the American Chemical Society*, vol. 129, pp. 3634–3640, Mar 2007.

[17] A. Weber, A. Casini, A. Heine, D. Kuhn, C. T. Supuran, A. Scozzafava, and G. Klebe, “Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition.,” *Journal of Medicinal Chemistry*, vol. 47, pp. 550–557, Jan 2004.

[18] M. Hult, N. Shafqat, B. Elleby, D. Mitschke, S. Svensson, M. Forsgren, T. Barf, J. Vallgarda, L. Abrahmsen, and U. Oppermann, “Active site variability of type 1 11beta-hydroxysteroid dehydrogenase revealed by selective inhibitors and cross-species comparisons.,” *Molecular and Cellular Endocrinology*, vol. 248, pp. 26–33, Mar 2006.

[19] R. B. Russell, “Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution,” *Journal of Molecular Biology*, vol. 279, no. 5, pp. 1211–1227, 1998.

[20] J. A. Barker and J. M. Thornton, “An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis.,” *Bioinformatics*, vol. 19, no. 13, pp. 1644–1649, 2003.

[21] D. J. Rigden, “Understanding the cell in terms of structure and function: insights from structural genomics.,” *Current Opinion in Biotechnology*, vol. 17, pp. 457–464, Oct 2006.

[22] A. Andreeva and A. G. Murzin, “Evolution of protein fold in the presence of functional constraints.,” *Current Opinion in Structural Biology*, vol. 16, no. 3, pp. 399–

408, 2006.

[23] R. B. Russell, M. A. S. Saqi, R. A. Sayle, P. A. Bates, and M. J. E. Sternberg, "Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation," *Journal of Molecular Biology*, vol. 269, no. 3, pp. 423–439, 1997.

[24] N. V. Grishin, "Fold change in evolution of protein structures," *Journal of Structural Biology*, vol. 134, no. 2-3, pp. 167–185, 2001.

[25] L. Xie and P. Bourne, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments," *Proceedings of the National Academy of Sciences*, vol. 105, no. 14, p. 5441, 2008.

[26] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[27] S. Schmitt, D. Kuhn, and G. Klebe, "A new method to detect related function among proteins independent of sequence and fold homology.," *Journal of Molecular Biology*, vol. 323, pp. 387–406, Oct 2002.

[28] A. C. Stuart, V. A. Ilyin, and A. Sali, "LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures.," *Bioinformatics*, vol. 18, pp. 200–201, Jan 2002.

[29] C. T. Porter, G. J. Bartlett, and J. M. Thornton, "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.," *Nucleic Acids Research*, vol. 32, no. Database issue, pp. D129–33, 2004.

[30] B. J. Polacco and P. C. Babbitt, “Automated discovery of 3D motifs for protein function annotation.,” *Bioinformatics*, vol. 22, pp. 723–730, Mar 2006.

[31] N. D. Gold and R. M. Jackson, “Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships.,” *Journal of Molecular Biology*, vol. 355, pp. 1112–1124, Feb 2006.

[32] B. H. Dessailly, M. F. Lensink, C. A. Orengo, and S. J. Wodak, “LigASite–a database of biologically relevant binding sites in proteins with known apo-structures.,” *Nucleic Acids Research*, vol. 36, pp. D667–73, Jan 2008.

[33] Y. Bromberg and B. Rost, “Comprehensive in silico mutagenesis highlights functionally important residues in proteins.,” *Bioinformatics*, vol. 24, pp. i207–12, Aug 2008.

[34] R. Kolodny, D. Petrey, and B. Honig, “Protein structure comparison: implications for the nature of ‘fold space’, and structure and function prediction,” *Current Opinion in Structural Biology*, vol. 16, no. 3, pp. 393–398, 2006.

[35] P. F. Gherardini, M. N. Wass, M. Helmer-Citterich, and M. J. E. Sternberg, “Convergent evolution of enzyme active sites is not a rare phenomenon,” *Journal of Molecular Biology*, vol. 372, no. 3, pp. 817–845, 2007.

[36] F. Passardi, N. Bakalovic, F. K. Teixeira, M. Margis-Pinheiro, C. Penel, and C. Dunand, “Prokaryotic origins of the non-animal peroxidase superfamily and organelle-mediated transmission to eukaryotes,” *Genomics*, vol. 89, no. 5, pp. 567–579, 2007.

[37] I. Halperin, D. S. Glazer, S. Wu, and R. B. Altman, “The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications.,” *BMC Genomics*, vol. 9 Suppl 2, p. S2, 2008.

[38] O. Lichtarge, H. R. Bourne, and F. E. Cohen, “An evolutionary trace method defines binding surfaces common to protein families.,” *Journal of Molecular Biology*, vol. 257, pp. 342–358, Mar 1996.

[39] G. J. Kleywegt, “Recognition of spatial motifs in protein structures.,” *Journal of Molecular Biology*, vol. 285, pp. 1887–1897, Jan 1999.

[40] R. V. Spriggs, P. J. Artymiuk, and P. Willett, “Searching for patterns of amino acids in 3D protein structures.,” *Journal of Chemical Information and Computer Sciences*, vol. 43, pp. 412–421, Mar-Apr 2003.

[41] A. Stark and R. B. Russell, “Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures.,” *Nucleic Acids Research*, vol. 31, pp. 3341–3344, Jul 2003.

[42] A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson, “Recognition of functional sites in protein structures.,” *Journal of Molecular Biology*, vol. 339, pp. 607–633, Jun 2004.

[43] G. Ausiello, A. Via, and M. Helmer-Citterich, “Query3d: a new method for high-throughput analysis of functional residues in protein structures,” *BMC Bioinformatics*, vol. 6, no. 4, p. S5, 2005.

[44] R. Laskowski, J. Watson, and J. Thornton, “ProFunc: a server for predicting protein function from 3D structure.,” *Nucleic Acids Research*, vol. 33, p. W89, 2005.

[45] R. A. Laskowski, J. D. Watson, and J. M. Thornton, “Protein function prediction using local 3D templates.,” *Journal of Molecular Biology*, vol. 351, pp. 614–626, Aug 2005.

[46] D. Pal and D. Eisenberg, “Inference of protein function from protein structure.,” *Structure*, vol. 13, pp. 121–130, Jan 2005.

[47] A. R. Kinjo and H. Nakamura, "Similarity search for local protein structures at atomic resolution by exploiting a database management system," *Biophysics*, vol. 3, pp. 75–84, 2007.

[48] B. Y. Chen, V. Y. Fofanov, D. H. Bryant, B. D. Dodson, D. M. Kristensen, A. M. Lisewski, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs," *Journal of Computational Biology*, vol. 14, no. 6, pp. 791–816, 2007.

[49] Y. Y. Tseng, J. Dundas, and J. Liang, "Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns.," *Journal of Molecular Biology*, vol. 387, no. 2, pp. 451–464, 2009.

[50] M. Shatsky, A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson, "The multiple common point set problem and its application to molecule binding pattern detection.," *Journal of Computational Biology*, vol. 13, pp. 407–428, Mar 2006.

[51] A. Brakoulias and R. Jackson, "Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching," *Proteins: Structure, Function, and Bioinformatics*, vol. 56, no. 2, 2004.

[52] Z. Zhang and M. G. Grigorov, "Similarity networks of protein binding sites.," *Proteins*, vol. 62, no. 2, pp. 470–478, 2006 Feb 1.

[53] L. Holm and C. Sander, "Mapping the protein universe," *Science*, vol. 273, no. 5275, pp. 595–603, 1996.

[54] L. Holm and C. Sander, "Dali: a network tool for protein structure comparison.," *Trends in Biochemical Sciences*, vol. 20, pp. 478–480, Nov 1995.

[55] B. Chen, V. Fofanov, D. Bryant, B. Dodson, D. Kristensen, A. Lisewski, M. Kimmel, O. Lichtarge, and L. Kavraki, "Geometric Sieving: Automated Distributed Optimization of 3D Motifs for Protein Function Prediction," *Lecture Notes in Computer Science*, vol. 3909, p. 500, 2006.

[56] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[57] B. Chen, D. Bryant, V. Fofanov, D. Kristensen, A. Cruess, M. Kimmel, O. Lichtarge, and L. Kavraki, "Cavity-aware motifs reduce false positives in protein function prediction.," in *Proc. of the Fifth Annual Intl. Conf. on Computational Systems Bioinformatics (CSB2006)*, Imperial College Press, 2006.

[58] B. Chen, D. Bryant, V. Fofanov, D. Kristensen, A. Cruess, M. Kimmel, O. Lichtarge, and L. Kavraki, "Cavity scaling: automated refinement of cavity-aware motifs in protein function prediction.," *J Bioinform Comput Biol*, vol. 5, no. 2a, pp. 353–82, 2007.

[59] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, vol. 2, no. 11, pp. 559–572, 1901.

[60] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis and density estimation," *Journal of the American Statistical Association*, vol. 97, pp. 611–631, 2002.

[61] V. Y. Fofanov, B. Y. Chen, D. H. Bryant, M. Moll, O. Lichtarge, L. E. Kavraki, and M. Kimmel, "A statistical model to correct systematic bias introduced by algorithmic thresholds in protein structural comparison algorithms," in *IEEE International Conference on Bioinformatics and Biomedicine Workshop, 2008*, pp. 1–8, 2008.

[62] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, "Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction," *Proceedings of the National Academy of Sciences*, vol. 103, no. 26, p. 9885, 2006.

[63] E. Plaku, H. Stamati, C. Clementi, and L. E. Kavraki, "Fast and reliable analysis of molecular motion using proximity relations and dimensionality reduction," *Proteins*, vol. 67, no. 4, pp. 897–907, 2007.

[64] R. Finn, J. Tate, J. Mistry, P. Coggill, S. Sammut, *et al.*, "The Pfam protein family database," *Nucleic Acid Research*, vol. 36, no. Database issue, pp. D281–88, 2008.

[65] X. Wang and J. Snoeyink, "Multiple structure alignment by optimal RMSD implies that the average structure is a consensus," in *Proc. of the Fifth Annual Intl. Conf. on Computational Systems Bioinformatics*, Imperial College Press, 2006.

[66] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 53, no. 3, pp. 683–690, 1991.

[67] Y. Hochberg, "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.

[68] S. K. Sarkar and C. K. Chang, "The Simes method for multiple hypothesis testing with positively dependent test statistics," *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1601–1608, 1997.

[69] N. B. Loughran, B. O'Connor, C. Ó'Fágáin, and M. J. O'Connell, "The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions," *BMC Evolutionary Biology*, vol. 8, p. 101, 2008.

[70] K. Fukuyama, N. Kunishima, F. Amada, T. Kubota, and H. Matsubara, “Crystal structures of cyanide-and triiodide-bound forms of Arthromyces ramosus peroxidase at different pH values,” *Journal of Biological Chemistry*, vol. 270, no. 37, pp. 21884–21892, 1995.

[71] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, “CATH–a hierarchic classification of protein domain structures,” *Structure*, vol. 5, no. 8, pp. 1093–1108, 1997.

[72] K. Karhumaa, R. G. Sanchez, B. Hahn-Hägerdal, and M. F. Gorwa-Grauslund, “Comparison of the xylose reductase-xylitol dehydrogenase and the xylose isomerase pathways for xylose fermentation by recombinant Saccharomyces cerevisiae,” *Microbial Cell Factories*, vol. 6, no. 1, p. 5, 2007.

[73] A. J. van Maris, A. A. Winkler, M. Kuyper, W. T. de Laat, J. P. van Dijken, and J. T. Pronk, “Development of efficient xylose fermentation in Saccharomyces cerevisiae: xylose isomerase as a key component,” *Advances in Biochemical Engineering/Biotechnology*, vol. 108, pp. 179–204, 2007.

[74] H. M. Holden, D. E. Tronrud, A. F. Monzingo, L. H. Weaver, and B. W. Matthews, “Slow-and fast-binding inhibitors of thermolysin display different modes of binding: crystallographic analysis of extended phosphonamidate transition-state analogs,” *Biochemistry*, vol. 26, no. 26, pp. 8542–8553, 1987.

[75] D. R. Holland, A. C. Hausrath, D. Juers, and B. W. Matthews, “Structural analysis of zinc substitutions in the active site of thermolysin,” *Protein Science*, vol. 4, no. 10, pp. 1955–1965, 1995.

[76] D. Blow, “More of the catalytic triad,” *Nature*, vol. 343, no. 6260, pp. 694–695, 1990.

[77] A. Dementiev, J. Dobo, and P. G. W. Gettins, “Active site distortion is sufficient for proteinase inhibition by serpins: Structure of the covalent complex of α_1-proteinase inhibitor with porcine pancreatic elastase,” *Journal of Biological Chemistry*, vol. 281, no. 6, pp. 3452–3457, 2006.

[78] A. Schmidt, C. Jelsch, P. Ostergaard, W. Rypniewski, and V. S. Lamzin, “Trypsin revisited: crystallography at (sub) atomic resolution and quantum chemistry revealing details of catalysis,” *Journal of Biological Chemistry*, vol. 278, no. 44, pp. 43357–43362, 2003.

[79] B. Y. Chen, D. H. Bryant, A. E. Cruess, J. H. Bylund, V. Y. Fofanov, D. M. Kristensen, M. Kimmel, O. Lichtarge, and L. E. Kavraki, “Composite motifs integrating multiple protein structures increase sensitivity for function prediction,” in *Proc. of the Sixth Annual Intl. Conf. on Computational Systems Bioinformatics*, pp. 343–355, 2007.

[80] C. Mattos and D. Ringe, “Locating and characterizing binding sites on proteins.,” *Nat Biotechnol*, vol. 14, pp. 595–599, May 1996.

[81] C. Mattos, C. R. Bellamacina, E. Peisach, A. Pereira, D. Vitkup, G. A. Petsko, and D. Ringe, “Multiple solvent crystal structures: probing binding sites, plasticity and hydration.,” *J Mol Biol*, vol. 357, pp. 1471–1482, Apr 2006.

[82] A. C. English, S. H. Done, L. S. Caves, C. R. Groom, and R. E. Hubbard, “Locating interaction sites on proteins: the crystal structure of thermolysin soaked in 2% to 100% isopropanol.,” *Proteins*, vol. 37, pp. 628–640, Dec 1999.

[83] A. C. English, C. R. Groom, and R. E. Hubbard, “Experimental and computational mapping of the binding surface of a crystalline protein.,” *Protein Eng*, vol. 14, pp. 47–59, Jan 2001.