STRUCTURAL PATTERN DETECTION AND DOMAIN RECOGNITION FOR PROTEIN FUNCTION PREDICTION

by
Süveyda Yeniterzi
2009

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
August 2009

STRUCTURAL PATTERN DETECTION AND DOMAIN RECOGNITION FOR PROTEIN FUNCTION PREDICTION

APPROVED BY:

to my parents
&
my sister Reyyan

Acknowledgements

This thesis owes its existence to the help, support, and inspiration of many people.

I especially would like to express my sincere appreciation and gratitude to Uğur Sezerman for his support and encouragement. He was always accessible and willing to help me with my research. As a result, research life became smooth and rewarding for me.

I would like to thank my thesis committee members Kemal Oflazer, Berrin Yanıkoğlu, Yücel Saygın and Devrim Gözüaçık for their valuable comments and suggestions.

The discussions and cooperations with Cenk Şahinalp, Bora Uyar and Fereydoun Hormozdiari have contributed substantially to this work. Thus, I would like to especially thank to them.

I would like to thank all the lab members of FENS 2014 especially to Özlem, İlknur, Burak, Ferhan and Hanife for their friendships and helps.

I would like to thank Erol Çöm for his help during the submission of this thesis.

I am indebted to Tübitak for providing financial support during my studies.

Last but not least, I owe special gratitude to my parents for their endless love, support and patience. My sister and best friend Reyyan deserves a special thank for her continuous and unconditional love.

STRUCTURAL PATTERN DETECTION AND DOMAIN RECOGNITION FOR PROTEIN FUNCTION PREDICTION

Süveyda Yeniterzi

MS Thesis, 2009

Thesis Supervisor: Assoc. Prof. Dr. Osman Uğur Sezerman

Keywords: Structural pattern detection, domain recognition, local structural alignment, graphlet mapping, function prediction, fold classification

ABSTRACT

Proteins are essential players of the cell that control and affect all functions. In proteins, structural patterns consist of a few amino acids which assemble in a specific arrangement. Due to their specific structures, they are recognized as the functionally important sites of the proteins, and conserved even in distantly related proteins. Moreover, several structural patterns merge and form domains which are also associated with the proteins function.

In this work, we introduced a method for finding structure patterns common to a protein pair by using graphlet mappings. We presented protein structures with graphs, and then generate graphlets. Local alignments are produced by mapping the generated graphlets from protein pairs. Moreover, by merging these local alignments, we tried to recognize functionally important domains.

These common domains are very useful in protein function prediction, fold classification and homology relationship detection. In this work, our algorithm was first applied to fold classification problem and 80% accuracy was observed. Furthermore, our algorithm was also used for protein function prediction and 97% accuracy was observed.

PROTEİN FONKSİYON TAYİNİ İÇİN
YAPISAL ÖRÜNTÜ TESPİTİ VE DOMEN TANINMASI

Süveyda Yeniterzi

MS Tezi, 2009

Tez Danışmanı: Doç. Dr. Osman Uğur Sezerman

Anahtar Kelimeler: Yapısal örüntü tespiti, domen tanınması, bölgesel yapı hizalaması, graf parçacıkları eşlemesi, protein fonksiyon tayini, işlevsel yapı ünitesi tayini

Özet

Proteinler hücrelerdeki fonksiyonları kontrol eden ve etkileyen önemli faktörlerdir. Proteinlerdeki birkaç aminoasitin belirli bir düzen içinde bir araya gelmesi ile yapısal örüntüler oluşur. Belirli düzenleri nedeniyle proteinlerin fonksiyon olarak önemli yerleri kabul edilen bu örüntüler, birbirlerine uzaktan akraba proteinlerde de korunurlar. Bunun yanında, bu tür birkaç yapısal örüntü bir araya gelerek protein fonksiyonunda önemli yeri olan domenleri oluşturur.

Bu çalışmada, iki proteindeki ortak yapısal örüntüleri graf parçacıkları eşlemesi kullanarak bulmaya çalışan bir metodu tanıtıyoruz. Protein yapıları, graf kullanılarak gösterildi daha sonra da graf parçacıkları yaratıldı. Her iki proteindeki graf parçacıkları birbirleriyle eşleştirilerek bölgesel yapı hizalamaları elde edildi. Ayrıca bu bölgesel yapı hizalamaları birleştirilerek fonksiyon olarak önemli domenler bulunmaya çalışıldı.

Bu ortak domenler, protein fonksiyon tayini ve işlevsel yapı ünitesi tayini ile homoloji ilişki tespitinde kullanılabilir. Çalışmada, algoritmamız öncelikle işlevsel yapı ünitesi tayin etme amacıyla kullanıldı ve %80 doğru sınıflandırma yapıldı. Ayrıca algoritmamız, fonksiyon tayin etme amacıyla da kullanıldı ve %97 doğru fonksiyon ataması gerçekleştirildi.

TABLE OF CONTENTS

List of Figures

List of Tables

TABLE OF ABBREVIATIONS

AFP	Aligned Fragment Pair
APGM	Approximate Graph Mining
BLOSUM	Blocks of Amino Acid Substitution Matrix
CATH	Class, Architecture, Topology, Homologous superfamily
CE	Combinatorial Extension
DALI	Distance Alignment Matrix Method
EC	Enzyme Commission
LG	Levitt-Gerstein
LGA	Local-Global Alignment
MSA	Multiple Sequence Alignment
PDB	Protein Data Bank
RMSD	Root Mean Square Deviation
SCOP	Structural Classification of Proteins
SSAP	Sequential Structure Alignment Program

Chapter 1

INTRODUCTION

1.1 Motivation

Proteins are essential players of the cell that control and affect all functions. Their role is mainly determined by their structure. Likewise, it is the amino acid sequence that determines the protein's structure. Therefore, there is a strong relationship among the sequence, structure and function of the proteins. This relationship is generally used to solve one of the most challenging problems in bioinformatics: the prediction of protein function. The classical method for protein function prediction is based on a pairwise sequential or an overall structural alignment of proteins. If similarities are detected in the alignment, the information about the function of well-known protein can be transferred to the successfully aligned unknown proteins [Saçan et al., 2007].

In spite of the relationship between sequence, structure and function of the proteins, the sequence similarity-based and the overall structure similarity-based approaches have limitations in function prediction. For instance, sequence alignments can provide insight into protein function; however, the sequence similarity-based approach fails when new proteins have a very low level sequence similarity with known proteins. Protein pairs that do not have high sequence similarity may still have similar functions due to the physicochemical properties conserved at the structural level [Liang et al., 2003]. Therefore, it has been concluded that the structure of the protein provide better sensitivity and predictive value for function prediction than does sequence similarity [Saçan et al., 2008].

Even though the knowledge of structural similarity has a great importance for function prediction, this approach fails to consider distantly related proteins that have only local similarities rather than global similarities. Since protein functions are determined from the specific structural regions, such as catalytic sites, binding sites, and protein-protein interaction sites, proteins are largely tolerant to mutations that happened in the non-functional regions of the structure. Thus, it is very common for two proteins with the same function to show only local similarities even though their structures are not globally similar. As a result, focusing only on the functionally important sites rather than the overall structure performs better in function prediction [Saçan et al., 2007, Ben-Hur and Brutlag, 2003]. The limitations of the above approaches led us to the development of this thesis.

We propose a method for finding structure patterns common to a protein pair by using a local alignment algorithm based on graphlet mappings. In our method, the proteins are represented with graphs and these graphs are used to detect graphlets of size 3 to 10. Topological similarities between two proteins are discovered by performing graphlet mapping. Moreover, our algorithm tries to assemble these aligned fragment pairs into a larger alignment for the purpose of recognizing structurally and functionally important domains shared between two proteins. Such domains are the most important factors in the identification of protein's function. Moreover, since structure patterns are better conserved than amino acidic sequences [Carugo, 2006], remote homology relationship between distantly related proteins can be recognized more reliably by using these local similarities.

1.2 Outline

The organization of the thesis as follows: Chapter 2 presents a brief biological background and an overview of the related works. In Chapter 3, we explain our approach in detail. Chapter 4 discusses the experiments and the results. Lastly, the conclusions and the future works are given in Chapter 5.

Chapter 2

BACKGROUND AND RELATED WORKS

2.1 Biological Background

2.1.1 Protein

A protein is composed of a chain of amino acids which are joined together by peptide bonds. There are a total of 20 amino acids and every amino acid has an amino group (NH_2), a carboxyl group ($COOH$), one carbon atom at the center which is also known as the alpha carbon (C_α), and a side chain attached to the C_α. These amino acids are listed in Table 2.1 [Kyte and Doolittle, 1982, Cooper and Hausman, 2004]. These amino acids have different biochemical properties such as hydrophilic or hydrophobic characters, resulted from their side chains. Since these properties affect the interactions of amino acid residues, they have a great influence on protein three-dimensional structure and as a result protein's main function. The distribution of hydrophobic and hydrophilic (polar and charged) amino acids determines the structure of the protein where the hydrophobic residues try to get a position in the protein core while the hydrophilic ones prefer to be outside.

When amino acids are strung together into a polypeptide chain, a water molecule is liberated from each joined amino acids. Therefore, rather than the original amino acids, the proteins composed of amino acid residues [Setubal and Meidanis, 1997]. These amino acid residues form the primary structure of the protein. When the sequence of

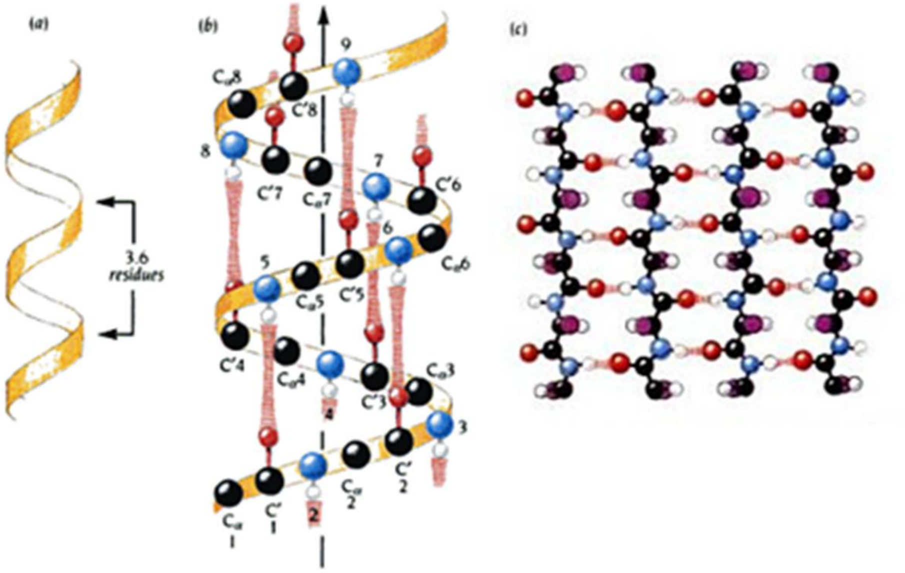


Figure 2.1: α-helices (a-b) and β-sheets (c) [Branden and Tooze, 1999]

amino acids are linked by hydrogen bonds, they form the secondary structures such as alpha(α) helices or beta(β) sheets. An α-helix on the average has 3.6 residues per turn and hydrogen bonds are formed between carboxyl and amino groups of the backbone atoms. An α-helix is one continuous sequence and its ends are generated by polar residues; therefore, they can be mostly observed on the surface of proteins. Similar to α-helices, in β-sheets hydrogen bonds are formed between backbone atoms of paralel strands. β-sheets occupy at least two continuous sequences each with approximately 5 to 10 residues long and either parallel and anti-parallel to each other [Branden and Tooze, 1999]. Examples to α-helices and β-sheets can be found in Figure 2.1.

α-helices and β-sheets form a spatial arrangement when certain attractions are present between them. This completely folded structure is called the tertiary structure. The folded structures of a protein can form an important functional site such as catalytic or binding sites [Branden and Tooze, 1999]. Therefore, structure of a protein is very important in function prediction.

Amino Acid	Abbreviations		Polarity	Charge	Hydropathy index
Alanine	Ala	A	nonpolar	neutral	1.8
Arginine	Arg	R	polar	positive	-4.5
Asparagine	Asn	N	polar	neutral	-3.5
Aspartic acid	Asp	D	polar	negative	-3.5
Cysteine	Cys	C	nonpolar	neutral	2.5
Glutamic acid	Glu	E	polar	negative	-3.5
Glutamine	Gln	Q	polar	neutral	-3.5
Glycine	Gly	G	nonpolar	neutral	-0.4
Histidine	His	H	polar	positive	-3.2
Isoleucine	Ile	I	nonpolar	neutral	4.5
Leucine	Leu	L	nonpolar	neutral	3.8
Lysine	Lys	K	polar	positive	-3.9
Methionine	Met	M	nonpolar	neutral	1.9
Phenylalanine	Phe	F	nonpolar	neutral	2.8
Proline	Pro	P	nonpolar	neutral	-1.6
Serine	Ser	S	polar	neutral	-0.8
Threonine	Thr	T	polar	neutral	-0.7
Tryptophan	Trp	W	nonpolar	neutral	-0.9
Tyrosine	Tyr	Y	polar	neutral	-1.3
Valine	Val	V	nonpolar	neutral	4.2

Table 2.1: List of amino acids

2.1.2 Structural Pattern

A sequence pattern is a biologically important nucleotide or amino acid sequence pattern that occurs frequently in many DNA strands or polypeptide chains. On the other hand, a structural pattern is a combination of few three-dimensional structural elements, which may not be adjacent. In proteins, structural patterns consist of several amino acids that form a specific geometric arrangement. These geometric arrangements can be associated with a particular function or a part of larger structural and functional unit [Branden and Tooze, 1999]. Although some structural patterns are regarded as an arrangement of secondary structures, such as the four-helix bundle motif; most patterns consist of several amino acids and they do not depend on any secondary structure. For instance, subtilisin, a bacterial serine protease, and chymotrpsin, a mammalian serine protease, have a common pattern called catalytic triad which consists of aspartic acid, histidine and serine. Even though, these two proteins share a structural pattern, their overall structures are quite different, and the elements of the catalytic triad are in different positions in the primary sequence [Petsko and Ringe, 2003].

2.1.3 Domain

A domain is a polypeptide chain or a part of a polypeptide chain which can fold independently into a stable tertiary structure. Domains are built from the different combinations of structural patterns [Branden and Tooze, 1999]. They are described as units of folding [Wetlaufer, 1973], compact structure [Richardson, 1981], function and evolution [Bork, 1991] which is not surprising since they are all related to each other. Therefore, domains are very important in finding protein's function, classifying protein's fold, and identifying homology relationships. Proteins may have either one domain or several domains which are called multi-domain. In multi-domain proteins, each domain can have a different function independent from the others, or they can work together in a concerted action. Domains form the functionally important sites of the proteins such as the catalytic sites of the enzymes or ligand binding sites. Moreover, since domains can fold independently, they play a significant role in protein folding by accelerating the folding process and reducing the potentially large combination of residue interactions.

2.2 Graph Representation of Protein Structures

Protein structure can be converted into a graph where the nodes represent the amino acids and the edges represent the contacts between residues. Contact map is one of the major graph representation techniques used in the literature [Vendruscolo and Domany, 1998, Zemla, 2003, Huan et al., 2004, Gupta et al., 2005, Bartoli et al., 2008, Küçükural et al., 2008]. In contact maps, the amino acids are represented with one of their atoms and the chosen atom's three dimensional coordinates are used in calculations. In order to decide which atoms represent the amino acids best, C_α, C_β and several other functional atoms were compared in [Torrance et al., 2005], and it is observed that C_α and C_β atoms have a better representation of the amino acids. Therefore, in this work, C_α atoms are used and it is assumed that two residues are in contact if three dimensional distances of their C_α atoms are smaller than a threshold. Several optimum distance thresholds were proposed in the literature such as $5.8A^o$ [Vendruscolo et al., 1997, Zaki, 2003], $6.8A^o$ [Miyazawa and Jernigan, 1985, Bahar and Jernigan, 1997, Shental-Bechor et al., 2005], and $8.6A^o$ [Zhao and Karypis, 2003, Atılgan et al., 2004, Taylor and Vaisman, 2006].

Besides the contact maps, another commonly used representation technique of protein structure is Delaunay tessellated graphs [Atılgan et al., 2004, Taylor and Vaisman, 2006, Küçükural et al., 2008] which have a different contact definition than contact maps. In a Delaunay tessellated graph, the edge lengths represents the physical distances between protein residues. On the other hand, in a contact map, all the edge lengths are equal to 1, which makes it a relational graph [Taylor and Vaisman, 2006]. In previous studies [Huan et al., 2004, Küçükural et al., 2008], it has been showed that Delaunay tessellated graph does not represents the structure of the proteins as good as the contact maps. Because of this, contact maps are employed in this work for the representation of protein structures.

2.3 Structural Alignment

Structural alignment is a method for discovering the similarities between proteins based on the proteins' shapes and three-dimensional conformations. During the evolution, protein structure is more conserved than the sequence; therefore, structural alignment is preferred in detecting evolutionary relationships between proteins with low sequence similarities. Moreover, structural alignment has been also a valuable tool in protein fold classification, protein structure modeling, and protein function prediction.

Many different overall structural alignment methods were developed. For instance, Combinatorial Extension (CE) [Shindyalov and Bourne, 1998] is a popular structural method which tries to assemble aligned fragment pairs (AFP) into a complete alignment. Similar to CE, distance alignment matrix method (DALI) [Holm and Sander, 1996] also breaks each structure into a series of fragments and brings together these fragments into a larger alignment using Monte Carlo simulation. Another widely used structure alignment method is sequential structure alignment program (SSAP) [Orengo and Taylor, 1996], which makes use of dynamic programming for detecting and combining local alignments. Finally, a recent method TM-align [Zhang and Skolnick, 2005] also uses dynamic programming with a novel method for weighting its distance matrix. TM-align uses inter structural residue distance vectors and an extended version of LG-scoring matrix TM-scoring. This algorithmic improvements accelerate the convergence of dynamic programming while overcoming the length difference problem of protein pairs. Therefore, TM-align performs better in both speed and accuracy over the existing methods. The quality of an alignment is measured with different methods such as root mean square deviation (RMSD), Levitt-Gerstein score (LG score) [Levitt and Gerstein, 1998], and local-global alignment (LGA) measure [Zemla, 2003].

2.4 Structural Pattern Detection

Many different methods have been developed in order to detect common structure patterns between proteins. Some algorithms rely on the structural alignments generated by superposition [Shapiro and Brutlag, 2004] while others apply geometric hashing to protein pairs [Nussinov and Wolfson, 1991, Barker and Thornton, 2003]. In this thesis, we focused on graph based approaches.

In one of these approaches, [Milik et al., 2003], the authors developed a search method for locating functionally and structurally common structures of protein pairs. Rather than using the backbone atoms, they chose specific atom types for each amino acid and found cliques of size four. Similar to our algorithm, discovered cliques from both proteins were compared and then merged to create a larger and continuous graph.

Graph theoretical representation and inexact subgraph matching approaches are also used in the determination of structural patterns. Similar to our method, in [Küçükural, 2008] the authors used contact maps for protein structure representation, and then used network properties such as connectivity, centrality, cliquishness to capture similar and conserved regions of proteins.

Another graph theoretic approach, [Wangikar et al., 2003], tried to detect structural patterns common in proteins from the same family. The method first generates all possible structural patterns in all proteins structures, and then detects the most observed pattern on the basis of content and geometric similarity.

Lastly, in [Jia et al., 2009], the authors developed a method called Approximate Graph Mining (APGM) which efficiently extracts and scores structure patterns from diverse proteins. Similar to our method, they represent the protein structures using graphs and take advantage of the substitution matrices in order to devise a novel graph data mining method to identify approximate matched frequent subgraphs. They applied their algorithm in protein fold classification problem where each discovered structure pattern was used as a feature in their classification scheme.

2.5 Domain Prediction

Protein domain prediction is significant for several reasons [Ingolfsson and Yona, 2008]:

Functional analysis of proteins:

Since domains are associated with protein function, finding domains is necessary for understanding the protein's function. Moreover, since domains are recurring patterns, determining the function of a domain will be useful in function prediction of many proteins which contain the same domain.

Structural analysis of proteins:

Since domains can fold independently into a stable tertiary structure, then protein structure determination is likely to be more successful if the protein can be divided into independent units such as domains.

Protein design:

Scientists make use of domain knowledge in protein engineering which is the design of new proteins and chimeras.

In the rest of this section, domain prediction methods will be explained briefly [Ingolfsson and Yona, 2008].

2.5.1 Experimental methods

In these experimental methods, a protein is chopped into its domains using proteases which are cellular enzymes that can cleave bonds between amino acids. By carefully manipulating experimental conditions, scientists make sure that the proteases can only access relatively unstructured regions of the protein, so that each fragment will contain a domain. Then with other experimental methods scientists try to understand the structure and function of these domains [Parrado et al., 1996].

2.5.2 Methods that use three dimensional structure

All methods in this category are based on the same general principle which assumes that domains are structurally compact and separate substructures. The differences in these methods are in the slightly different definitions of structurally compact substructures, and in the algorithms employed to search for these substructures. Some of these methods use various approaches to cluster residues into domains [Lesk and Rose, 1981], while others use top-down divisive approaches to split a protein into its domains [Xu et al., 2000, Alexandrov and Shindyalov, 2003].

2.5.3 Methods that are based on structure prediction

Since structure information is available for only a small number of proteins, several methods [Rigden, 2002, George and Heringa, 2002a] approach the domain prediction problem by employing structure prediction methods first. These algorithms can be quite effective in predicting domains; however, because of the structure prediction step, they are computationally intensive.

2.5.4 Methods based on similarity search

Methods that are based on similarity search use homologous sequences detected in a database search to predict domains. Most of these algorithms [Gracy and Argos, 1998, Heger and Holm, 2003, Portugaly et al., 2007] start with an all-vs.-all comparison of sequence databases, and then the similar sequences are clustered and split into domains.

2.5.5 Methods based on multiple sequence alignments

Another domain detection method is based on multiple sequence alignments (MSA). MSA-based approaches are the basis of several popular domain databases, such as Pfam [Bateman et al., 2004], and SMART [Schultz et al., 1998] which combine computational analysis and manual verification. Other MSA-based approaches [George and Heringa, 2002b] search their query sequences in a database to collect homologs and generate a MSA which is then processed to find domains. However, the quality of these methods depends on the number and composition of homologs used to construct the MSA.

2.5.6 Methods that use sequence-based features

Some methods try to utilize sequence-based features such as secondary structure information [Marsden et al., 2002], solvent accessibility, evolutionary profile, and amino

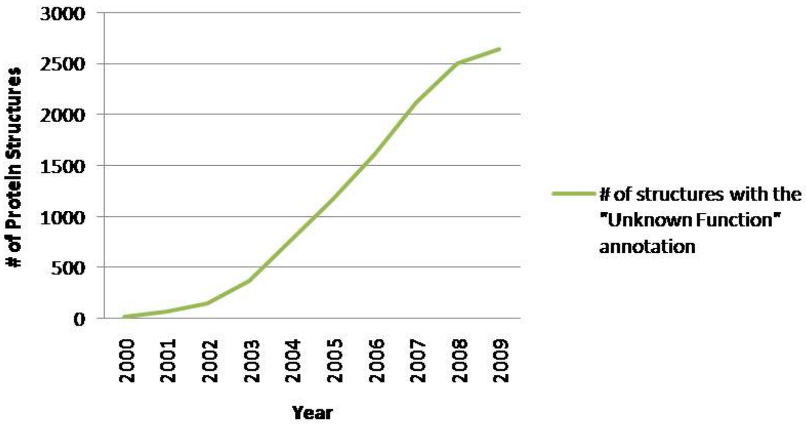


Figure 2.2: Yearly growth of the protein structures with the annotation ‘Unknown Function’ deposited in the PDB

acid entropy [Chen et al., 2006] for domain prediction.

2.6 Function Prediction

Protein function prediction is one of the most challenging problems of bioinformatics. Even though function of a protein can be determined from its’ structure, currently many proteins in the Protein Data Bank (PDB) are classified as ‘Unknown Function’ as can be seen in Figure 2.2. Besides these annotated proteins, many more proteins with unknown function are not even annotated. Therefore, what we see in Figure 2.2 is just the tip of the iceberg.

Many approaches were developed for predicting the protein’s function and these approaches are mostly based on detecting the similarities between a functionally annotated protein and the query protein, and then transferring the function information. During the evaluation of these approaches, three methods are generally used: prediction of Gene Ontology (GO) terms [Martin et al., 2004, Conesa et al., 2005], ligand binding site [Brylinski and Skolnick, 2008], and Enzyme Commission numbers [Dobson and Doig, 2005, Syed and Yona, 2009]. In this work, prediction of enzyme commission numbers is used for the evaluation purpose.

2.6.1 Prediction of Enzyme Commission Numbers

Enzymes are mostly protein based biomolecules that accelerate the rate of chemical reactions in a living organism. During these reactions, they convert a specific set of substrates into specific products. Since enzymes are selective for their substrates, they increase rates of only a few reactions which make the prediction of enzyme function an important problem.

The specific functions of enzymes are derived from their three dimensional structures, especially their active sites. Active site of an enzyme is the catalytic region that binds to the substrate and then carries out the reaction. Catalytic site structures are extremely conserved between distantly related enzymes. Since the catalytic site determines the activity of an enzyme, they can also be very similar in unrelated enzymes of similar function, such as the Ser-His-Asp catalytic triad [Torrance et al., 2005].

Many different methods were proposed for the prediction of enzyme function. The earlier researches focused on the sequence-based [Shah and Hunter, 1997] and the structure-based approaches [Rost, 2002]; however, lately different approaches based on alternative representation of proteins became popular. Features extracted from proteins such as secondary structure elements, contact energies, amino acid compositions, and physio-chemical properties are used for enzyme function prediction [desJardins et al., 1997, Cai and Chou, 2004, Han et al., 2004, Dobson and Doig, 2005, Borro et al., 2006, Syed and Yona, 2009]. Furthermore, information such as proteins' subcellular locations, tissue specificities and organism classifications are retrieved from databases for the same purpose [Lee et al., 2007]. Lastly, approaches that focused only on the functional regions such as catalytic sites were also proposed [Ben-hur and Brutlag, 2004, Torrance et al., 2005].

The International Union of Biochemistry and Molecular Biology have developed a nomenclature for enzymes, the Enzyme Commission number (EC number) [IUBMB, 1992], which is based on the function of an enzyme. In this numerical classification system, every enzyme consists of the letters 'EC' followed by four numbers seperated by periods such as EC.X.X.X.X. The first number indicates the general type of chemical reaction catalyzed by an enzyme. This top level classification divides enzymes into

6 categories: oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. The remaining three numbers represent a progressively finer classification of the enzyme and this classification is particular to each class. For instance, oxidoreductase class contains the enzymes that catalyze the transfer of electrons from one molecule (donor) to another (acceptor). In this class, second EC number represents the donor molecule, third number represents the acceptor molecule, and lastly fourth number represents the substrate [Ben-Hur and Brutlag, 2003].

It is important to note that EC numbers do not specify the enzymes; they classify the enzyme-catalyzed reactions. According to this classification scheme, different enzymes form different organisms have the same EC number if they catalyze the same reaction which is only possible if they share the same catalytic site structure. Therefore, in the EC number prediction systems, searching for similar catalytic site structure will perform better than using sequence or overall structure alignments for the following reasons [Torrance et al., 2005]:

- In order to carry out similar reactions, different proteins may independently evolve the same catalytic site structure. This phenomenon is known as convergent evolution and only the similar catalytic site structure can be used to predict the common function between these different proteins.

- In homologous enzymes of similar function, the catalytic site structure is conserved while the remaining protein structure has diverged to the degree that overall structure or sequence alignment cannot be used to predict the function.

- Although it is possible to identify distant homologues enzymes using the sequence methods, there may exist some ambiguities in the alignment, and a comparison of the catalytic site structures can be used as a disambiguation method.

- Moreover, similar catalytic sites that are spread over multiple protein chains can be identified easily by searching structurally similar catalytic sites rather than performing sequence or overall structural alignments.

- It is possible that two enzymes with different functions can be identified as homologues based on their sequence or overall structural alignments. In order to prevent the possibility of assigning these two enzymes to the same function class, their catalytic site structures have to be checked. Since the enzymes have different functions, their catalytic site structure will be dissimilar and this will prevent the misclassification.

2.7 Fold Classification

Proteins are made of polypeptide chains which are folded into a functional three dimensional structure. The folding process is the result of the interactions between the amino acids. These certain attractions form a spatial arrangement of the secondary structures. Therefore, finding the secondary structures of a protein is an important step in finding the three dimensional structure since it can reduce the search space.

A protein can be classified into fold classes according to its secondary structure components. Several databases have been developed for this purpose. Structural Classification of Proteins (SCOP) [Murzin et al., 1995] database is a manually created database for fold classification. SCOP database classifies proteins into structural domains based on their amino acid sequences and three dimensional structures. It has four hierarchical levels: class (general structure of the domain), fold (similar arrangements of secondary structures without evolutionary relation), superfamily (indicative of demonstrable evolutionary relationship without sequence homology), and family (some sequence similarity).

Besides SCOP, more automatic databases also exist such as CATH Protein Structure Classification [Orengo et al., 1997] database and Families of Structurally Similar Proteins (FSSP) [Taylor and Radzio-Andzelm, 1994] database. CATH is a semi-automatic classification system which also has four hierarchical levels: class (overall secondary-structure content of the domain), architecture (a large-scale grouping of topologies which share particular structural features), topology (high structural similarity without homology, equivalent to a fold in SCOP), and homologous superfamily (indicative of a demonstrable evolutionary relationship, equivalent to the superfamily level of SCOP). On the other hand, FSSP is purely automatically created database of

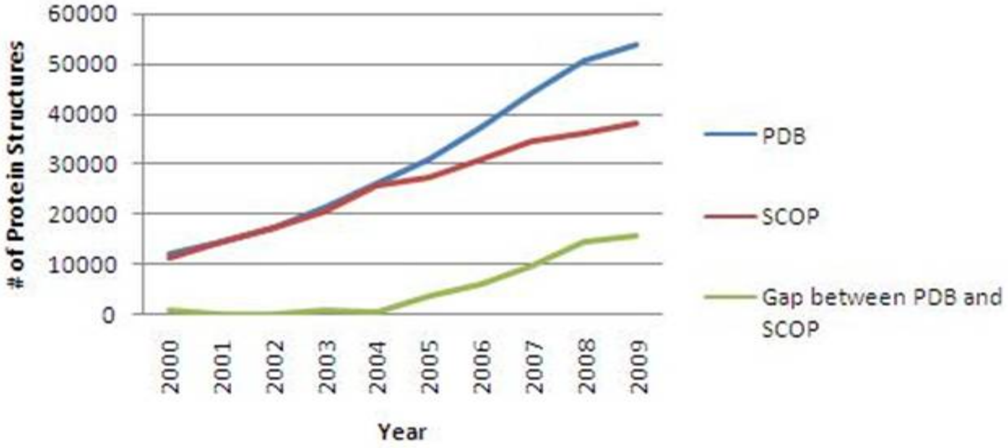


Figure 2.3: Yearly growth of protein structures in PDB and SCOP

structurally superimposed proteins generated using the DALI algorithm. This database does not classify the proteins. It compares the protein structures and allows the user to draw their own conclusion. Other automatic fold classification methods [Tan et al., 2003, Zerrin et al., 2004, Chen and Kurgan, 2007, Shamim et al., 2007] were also developed for fold classification.

Even though, important parts of the classification are performed manually in CATH, most of the work is done automatically. SCOP provides a better classification than CATH and all the other existing methods. Its' advantage over other systems is making use of human expertise which is needed to decide whether certain proteins are evolutionary related and therefore should be assigned to the same superfamily, or their similarity is a result of structural constraints and therefore should be assigned to the same fold. However, since SCOP is a manually generated database, it is incomplete and not up to date. If the yearly growth of protein structures in PDB and SCOP is compared, the gap between the number of PDB and SCOP structures grows in the last five years as can be seen in Figure 2.3. Therefore, there is a need for an automatic method that classifies proteins into different folds as accurate as SCOP does.

Chapter 3

METHODOLOGY

3.1 Introduction

Structural patterns consist of a few amino acids which assemble in a specific arrangement. Due to their specific structures, they are recognized as the functionally important sites of proteins, and even conserved in distantly related proteins. In our approach, we first represent the protein structures with graphs, and then generate the graphlets. In order to find the common structural patterns in protein pairs, local alignments are produced by mapping the generated graphlets from the same topologies. All the graphlet mappings are ranked with a scoring function which considers the residue distribution similarities of the mapped graphlets, connectivity, and evolutionary similarities of the mapped amino acids. Since our scoring function is based on structural arrangement and biochemical properties of amino acids, the graphlet mappings with high scores are treated as local structural alignments. In the rest of this thesis, graphlet mappings and local structural alignments are used interchangeably.

Domains are also associated with proteins function and they are built from structural patterns. Therefore, by merging the graplet mappings, we aim to construct functional domains. Moreover, many proteins have a multi domain structure and these different domains are associated with different functions. Our algorithm is designed to handle such situations by constructing all possible domains. A schematic illustration of our method is shown in Figure 3.1, and in the following sections, each step is explained in detail.

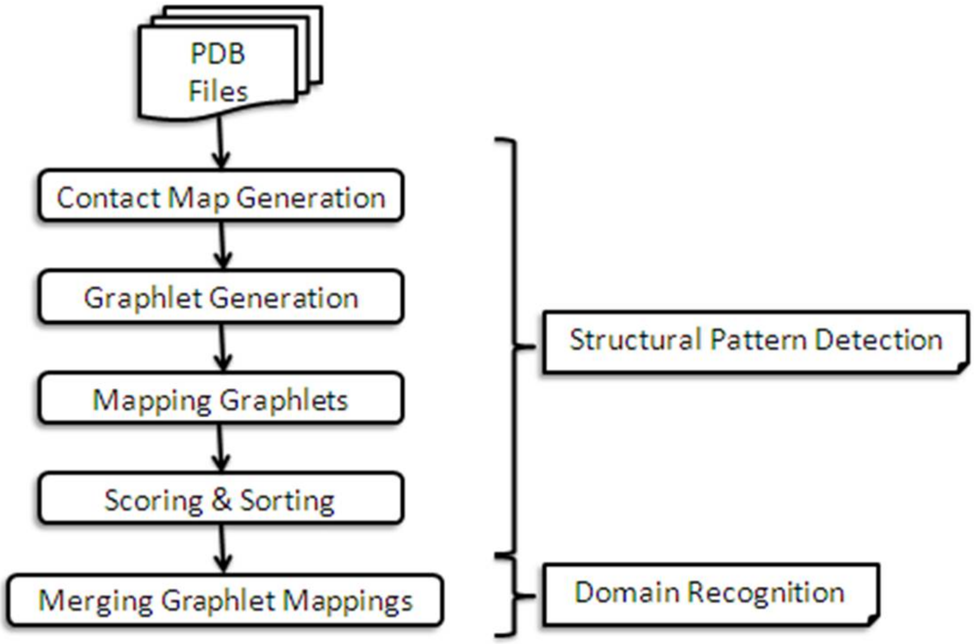


Figure 3.1: A schematic illustration of the methodology

3.2 Structural Pattern Detection

3.2.1 Contact Map Generation

The contact map is one of the major graph representation techniques for protein structures where the nodes represent the amino acids and the edges represent the contacts between residues. In this work, we assume that two residues are in contact if the three dimensional distances of their C_α atoms are smaller than a threshold. Several different optimum distance thresholds were proposed in the literature such as $5.8A^o$ [Vendruscolo et al., 1997, Zaki, 2003], $6.8A^o$ [Miyazawa and Jernigan, 1985, Bahar and Jernigan, 1997, Shental-Bechor et al., 2005], and $8.6A^o$ [Zhao and Karypis, 2003, Atılgan et al., 2004, Taylor and Vaisman, 2006]. All these thresholds were used in our experiments and the optimum distance threshold was decided according to the experimental results.

In the contact map generation step, three-dimensional atomic coordinates of all the residues are retrieved from the PDB files for each protein. These atomic coordinates are used to calculate the Euclidean distances between each residue pair. Two residues are assumed to be in contact if their distance is smaller than the threshold. During the implementation, the contact maps are represented with a binary two-dimensional matrix filled with 0. If two residues, i and j, are in contact, then the ij element of the matrix is changed to 1.

3.2.2 Graphlet Generation

After representing the structure of the proteins as graphs, the next step is to find the graphlets. A graphlet is a small connected induced subgraph of a graph. In this definition it is important to emphasize the definition of induced subgraph. A subgraph of G is a graph whose nodes and edges belong to G. On the other hand, an induced subgraph H of G is a subgraph of G, such that the edges of H consist of all edges of G that connect the nodes of H [Przulj et al., 2004, Hormozdiari et al., 2007].

Graphlets with 3, 4, 5 and 6 nodes have 141 possible graphlet topologies as shown in Appendix A. In this work, all these possible graphlet topologies are considered. Furthermore, we consider the cliques of sizes 7, 8, 9, and 10, which makes a total of 145 topologies. All these graphlets are generated by a program developed by Fereydoun Hormozdiari as the implementation of the paper [Hormozdiari et al., 2007]. The program takes contact maps as input and calculates the frequencies of all the graphlet topologies. If the frequency of a topology is below 1000, all the graphlets for that topology are generated.

For each graphlet topology, the algorithm starts by matching the topology's highest connected node to the nodes of the contact map, and considers each neighbor of that node as a possible neighbor of the node in the topology. Then the total number of counted graphlets are divided by the over counting factor of that topology. Over counting factor of a topology depends on the number of nodes with the highest connectivity value and the number of neighbors with similar contacts. For instance, two different topologies, topology 6 and 7 are shown in Figure 3.2. In topology 6, the node with the highest node connectivity is the 2^{nd} node. First, the algorithm tries to match this node to contact map nodes. When possible matches are discovered, the neighbors of the matched nodes are compared. As you can see, in topology 6, the 1^{st} and the

3^{rd} nodes have similar connections. Because of this similarity, this graphlet is counted twice during the comparison. Therefore, the over counting factor of topology 6 is 2. On the other hand, there are two nodes, the 1^{st} and the 4^{th}, with the highest connectivity value in topology 7. Moreover, these nodes' neighbors, the 2^{nd} and the 3^{rd} nodes have similar contacts. Therefore, this graphlet will be counted twice for the 1^{st} node, and again twice for the 4^{th} node, which makes the over counting factor equal to 4.

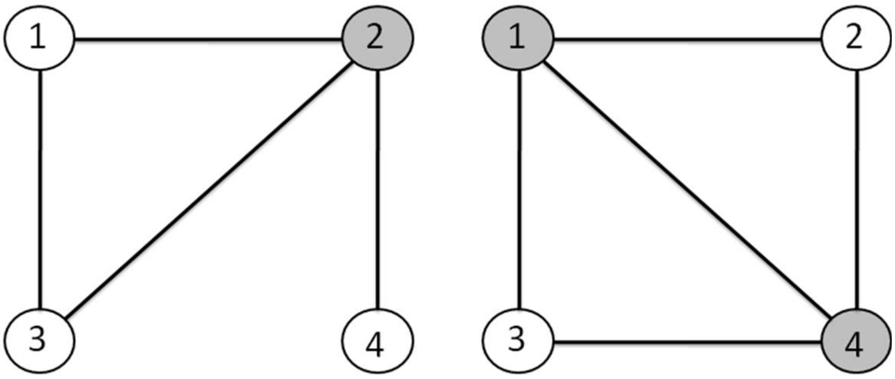


Figure 3.2: Topology 6 and 7

In the generated graphlets, the nodes are labeled with the residue numbers and their arrangement follows the graphlet topology. For instance, as seen in Figure 3.3, topology 11 consist of five nodes and there are only four edges which connects the 2^{nd} node to all the other nodes. Three example graphlets of this topology for two different proteins are shown in Figure 3.4. In this representation, the letters represent the one letter code of the amino acids, and the numbers in the parenthesis represent the residue numbers. As you can see, the residue numbers are not always in a sorted order. Their order is decided according to the topology. After the graphlet generation, the next step is finding isomorphic graphlets between two proteins.

3.2.3 Mapping Graphlets

At this step, we attempt to discover the topological similarities between protein pairs by mapping the generated graphlets. When graphlets of the same topology are detected for protein pairs, their isomorphism is checked. In this work, the isomorphism relation is defined as follows: given a labeled graphlet g_1 from $Protein$ 1 and a labeled graphlet

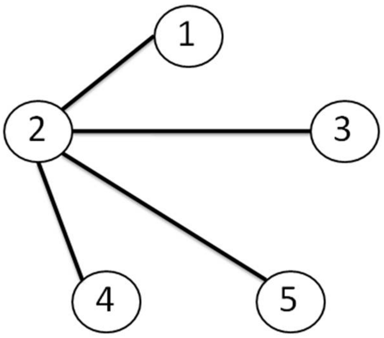


Figure 3.3: Topology 11

Protein 1 Topology 11	Node 1	Node 2	Node 3	Node 4	Node 5
Graphlet 1:	Y (13)	I (15)	M (34)	I (45)	E (47)
Graphlet 2:	N (87)	D (92)	K (90)	V (95)	L (119)
Graphlet 3:	W (89)	M (86)	A (93)	H (108)	G (120)

Protein 2 Topology 11	Node 1	Node 2	Node 3	Node 4	Node 5
Graphlet 1:	F (17)	V (21)	M (33)	I (35)	Q (36)
Graphlet 2:	Y (45)	L (32)	A (47)	H (61)	G (82)
Graphlet 3:	N (96)	D (119)	R (115)	I (121)	I (124)

Figure 3.4: Examples of graphlets

g_2 from *Protein* 2, the two graphlets are isomorphic when the bijection between the vertex sets of g_1 and g_2 will preserve the arrangements of the residues. According to this definition, our isomorphism detection is much simpler than the classical isomorphism definition. For instance, if we keep ordering of the g_1's node constants; then in classical isomorphism, the number of possible graphlets that needs to be checked is equal to the permutation of the number of g_2's nodes. However, according to our definition, the only permutation that enables a structural alignment is the one that has the same residue ordering as the graphlet it's aligned. For example, in Figure 3.5 and Figure 3.6, example graphlet mappings for topology 11 are shown. The first graphlet mapping is possible because at the end of the mapping, the aligned nodes of the proteins are in an ascending order without any disoriented mapping. On the other hand, the second graphlet mapping is not possible since the mappings become disoriented when the nodes are sorted. For the graphlets in Figure 3.4, all possible mappings are given in Figure

3.7.

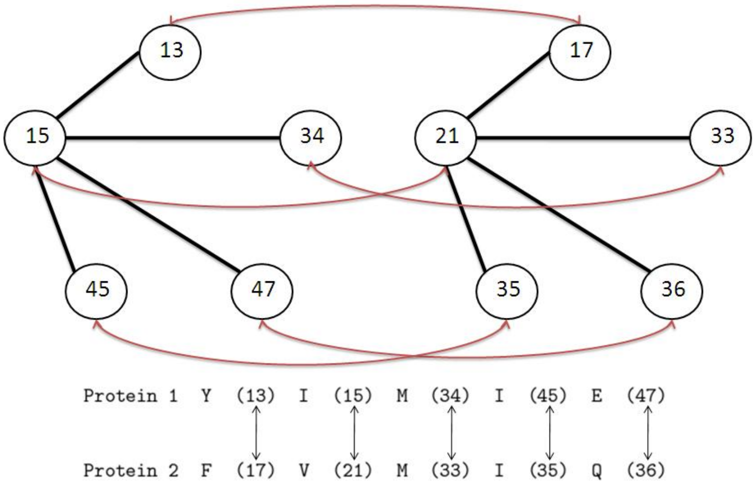

13
17
15
34
21
33
45
47
35
36
Protein 1 Y (13) I (15) M (34) I (45) E (47)
Protein 2 F (17) V (21) M (33) I (35) Q (36)

Figure 3.5: A possible graphlet mapping

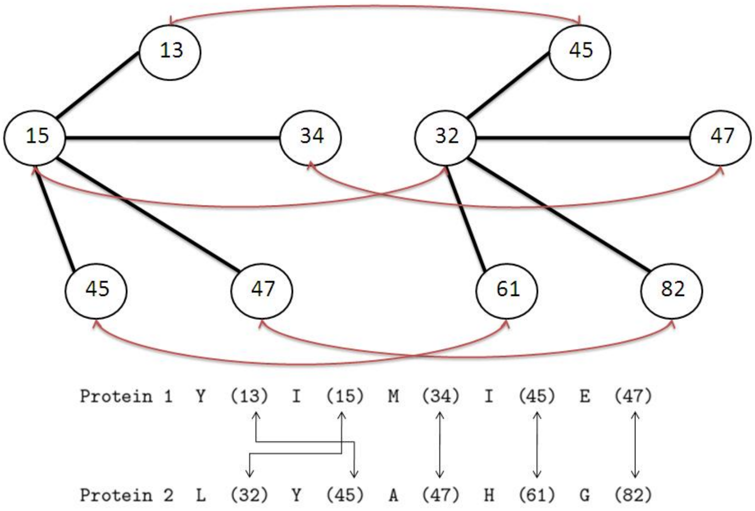

13
45
15
34
32
47
45
47
61
82
Protein 1 Y (13) I (15) M (34) I (45) E (47)
Protein 2 L (32) Y (45) A (47) H (61) G (82)

Figure 3.6: An impossible graphlet mapping

If an isomorphism exists between two graphlets according to our definition, then that local alignment is treated as a potential structural pattern shared by the two proteins. In order to decide whether a mapping is a definite structural pattern, it needs to be supported with the similar biochemical properties of the matched residues, or similar residual distributions of the graphlets. For this reason, in the next step, all possible mappings are ranked using a scoring function.

		Node 1	Node 2	Node 3	Node 4	Node 5
Protein 1	Graphlet 1:	Y (13)	I (15)	M (34)	I (45)	E (47)
Protein 2	Graphlet 1:	F (17)	V (21)	M (33)	I (35)	Q (36)
Protein 1	Graphlet 2:	N (87)	D (92)	K (90)	V (95)	L (119)
Protein 2	Graphlet 3:	N (96)	D (119)	R (115)	I (121)	I (124)
Protein 1	Graphlet 3:	W (89)	M (86)	A (93)	H (108)	G (120)
Protein 2	Graphlet 2:	Y (45)	L (32)	A (47)	H (61)	G (82)

Figure 3.7: Examples of graphlet mappings

3.2.4 Scoring

In this step, all the generated mappings are assigned a score based on their aligned amino acids' similarities, graphlets' residue distributions, and nodes' connectivity similarities. Therefore, our scoring function consists of three scores and the details of these scores are explained below.

3.2.4.1 Evolutionary Similarity Score

Amino acids have biochemical properties that influence their interchangeability in evolution. For instance, hydrophobic residues more likely get substituted for one another than do those of polar residues. Therefore, while calculating the similarity between two graphlets, it is important to use a scoring scheme that considers the evolutionary similarity and interchangeability of paired amino acids [Setubal and Meidanis, 1997]. For this reason, BLOSUM (BLOcks of Amino Acid SUbstitution Matrix) scores are used as one of the scoring function parameters.

BLOSUM matrices have been first proposed in [Henikoff and Henikoff, 1992] as a substitution matrix for protein sequence alignment. They are derived from aligned protein blocks and several sets were calculated from different blocks, each with a different sequence similarity percentage. For instance, BLOSUM62 matrix is constructed from sequence alignments with more than 62% identity. In this work, aligned residue pairs are scored using the BLOSUM62 matrix since it is specially designed for comparing moderately distant proteins. The BLOSUM62 matrix is given in Appendix B.

For two graphlets, g_1 and g_2, the evolutionary similarity score, $E\left(g_1, g_2\right)$, is calculated as follows: for each aligned amino acid pair, BLOSUM62 score is added and then the total BLOSUM62 score is divided to the number of amino acid pairs so that the average BLOSUM62 score can be obtained. This average residue BLOSUM62 score is used as the evolutionary similarity score of the mapping. The evolutionary similarity score values range from -4 to 11 due to the values of the BLOSUM62 matrix.

3.2.4.2 Residue Distribution Score

Besides evolutionary similarity, for a structurally consistent mapping, the distribution of the residues, i.e. the relative distances between the neighbor residues on the linear ordering of the protein, must be similar. In order to incorporate this property, the residue distribution score is defined as follows:

$$R\left(g_1, g_2\right) = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{min\left\{\left|g_1\left(i+1\right) - g_1\left(i\right)\right|, \left|g_2\left(i+1\right) - g_2\left(i\right)\right|\right\}}{max\left\{\left|g_1\left(i+1\right) - g_1\left(i\right)\right|, \left|g_2\left(i+1\right) - g_2\left(i\right)\right|\right\}} \tag{3.1}$$

where, g_1 and g_2 are the graphlets, n is the number of nodes, and $g()$ is the function that returns the residue number of a node. Equation 3.1 returns a value between 0 and 1. Therefore, graphlet mappings with similar residue distributions are rewarded with scores close to 1, whereas mappings with different residue distributions are penalized with scores close to 0.

3.2.4.3 Connectivity Score

Our last score is based on connectivity, a graph theoretical property that measures the number of neighbors of each residue in the protein [Küçükural et al., 2008]. Since we are looking for functionally shared motifs, it is important to have node alignments that have similar connectivity values. If in a mapping, the connectivity values of the aligned nodes are very different, then it is very unlikely that two graphlets share the same functionality. In order to reward residue alignments with similar connectivity values, the connectivity score is calculated as follows:

$$C\left(g_1, g_2\right) = \frac{1}{n} \sum_{i=1}^{n} \frac{min\left\{conn\left(g_1\left(i\right)\right), conn\left(g_2\left(i\right)\right)\right\}}{max\left\{conn\left(g_1\left(i\right)\right), conn\left(g_2\left(i\right)\right)\right\}} \tag{3.2}$$

where, g_1 and g_2 are the graphlets, n is the number of nodes, and $conn()$ is the function that returns the connectivity value of a residue. Similar to the residue distribution score, the connectivity score also assigns scores close to 0 if the graphlet mappings have a very different node connectivity.

After all three scores are calculated, the total score is calculated as follows:

$$TotalScore\,(g_1, g_2) = coef_1 * E\,(g_1, g_2) + coef_2 * R\,(g_1, g_2) + coef_3 * C\,(g_1, g_2) \tag{3.3}$$

where the $coef$'s represents the coefficients that the scores are multiplied with.

3.2.4.4 Parameter Optimization

As shown in Equation 3.3, our scoring function is the linear sum of R, C, and E scores. These score values differ greatly from each other since the E score can be any value between -4 and 11 while the R and the C scores are restricted to the interval between 0 and 1. Because of this, each score has a different weight on the total score. In order to prevent this, we decided to add coefficients to our scoring function. Moreover, with optimum coefficients, we can also achieve a better ranking of our mappings. For finding the optimum parameters, multidimensional linear regression was performed where the TM-align similarity score which will be explained in Section 4.1.2 was used as the dependent variable and the R, C, and E scores were used as the predictor variables. The results of the regression analysis will be given in Section 4.1.4.

When all the graphlet mappings are scored, they are sorted in preparation for the merging step.

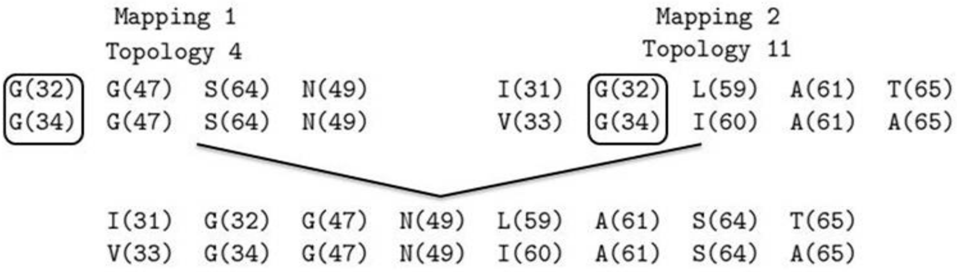
3.3 Domain Recognition

3.3.1 Merging Graphlet Mappings

Local structural alignments are obtained in the previous step with the graphlet mappings. It is possible to detect domains by extending these local alignments into longer alignments. Therefore, after the mappings are scored and ranked, we merge these mappings in order to obtain longer alignments or trees as we have called them. Our merging process is based on three conditions:

Condition I

Two mappings can be merged if they have at least one common residue pair. For instance, the mappings in Figure 3.8 can be merged due to their common amino acid alignment `G(32)-G(34)`.



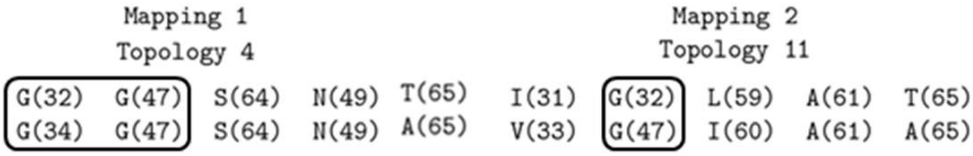
```
       Mapping 1                               Mapping 2
       Topology 4                              Topology 11
G(32)  G(47)  S(64)  N(49)          I(31)  G(32)  L(59)  A(61)  T(65)
G(34)  G(47)  S(64)  N(49)          V(33)  G(34)  I(60)  A(61)  A(65)

       I(31)  G(32)  G(47)  N(49)  L(59)  A(61)  S(64)  T(65)
       V(33)  G(34)  G(47)  N(49)  I(60)  A(61)  S(64)  A(65)
```

Figure 3.8: Example for condition I

Condition II

Two mappings can be merged if their residue pairings do not conflict. For example, the mappings in Figure 3.9 cannot be merged since the first mapping's `G(32)-G(34)` and `G(47)-G(47)` alignments are in conflict with the second mapping's `G(32)-G(47)` alignment. As seen in this example, `G(32)` and `G(47)`are aligned with each other in Mapping 2, while they are aligned with different residues in Mapping 1.



```
         Mapping 1                                  Mapping 2
         Topology 4                                 Topology 11
G(32)  G(47)  S(64)  N(49)  T(65)   I(31)  G(32)  L(59)  A(61)  T(65)
G(34)  G(47)  S(64)  N(49)  A(65)   V(33)  G(47)  I(60)  A(61)  A(65)
```

Figure 3.9: Example for condition II

Condition III

Two mappings can be merged if there is no conflict in their residue orderings. For instance, the mappings in Figure 3.10 cannot be merged because the second mapping's

L(45)-I(60) alignment disrupts the ascending residue order of second sequence in the merged mapping G(34) I(60) G(47).

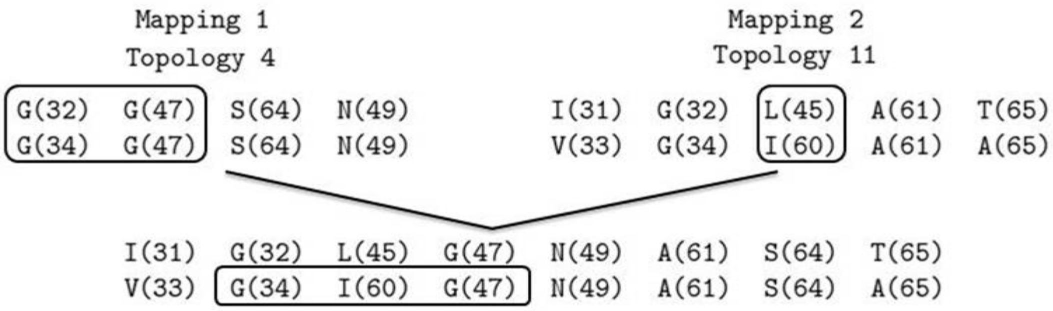

Mapping 1
Topology 4
G(32) G(47) S(64) N(49)
G(34) G(47) S(64) N(49)
Mapping 2
Topology 11
I(31) G(32) L(45) A(61) T(65)
V(33) G(34) I(60) A(61) A(65)
I(31) G(32) L(45) G(47) N(49) A(61) S(64) T(65)
V(33) G(34) I(60) G(47) N(49) A(61) S(64) A(65)

Figure 3.10: Example for condition III

If two alignments satisfy all the three conditions, then they can merge and form a longer alignment. In our algorithm, a large tree can be constructred by merging mappings starting from the highest scoring one. However, this algorithm ignores two probable circumstances. The first one is if we only focus on the best scoring mapping, then a mapping with a slightly smaller score can be ignored if it contains residues from a completely different portion of the proteins. Since they do not have a common residue pair, the second mapping will be lost in the merging process even though it is a correct alignments. This is a very common case for multi-domain proteins.

The second probable case is although we tried to perfect our scoring scheme with coefficients obtained from the regression analysis which will be explained in detail in Section 4.1.4, it is not definite that the highest scoring mapping is always the best alignment. Sometimes, a mapping which conflicts with the best scoring mapping can be a better alignment. In that case, again this mapping will be lost in the merging process.

In order to prevent the above two situations, all possible trees are generated in the merging process. The merging process starts from the highest scoring mapping, and continues with the next higher score mapping. All new mappings are first compared with the existing trees to check whether they have at least one common residue pairing. If they have no common residue pairing, than a new tree is formed for the mapping; however, if they have a common pair, then these trees and the mapping are checked for the situations below:

- If the mapping is in conflict with these trees; then a new tree is created for that mapping.
- If the mapping satisfies conditions II and III with one of the existing trees, then the mapping is added to that tree.
- If the mapping satisfies conditions II and III with more than one of the existing trees, then these trees are checked with each other in order to detect whether they are in conflict or not:
    - If all these trees are in conflict, then the mapping is added to the one with the highest average node score.
    - If some of the trees are not in conflict, then these trees are merged and the mapping is added to this new merged tree. With this condition, trees that cover different portions of the proteins may be merged and one global alignment can be formed.

The flow diagram of the merging algorithm is given in Figure 3.11.

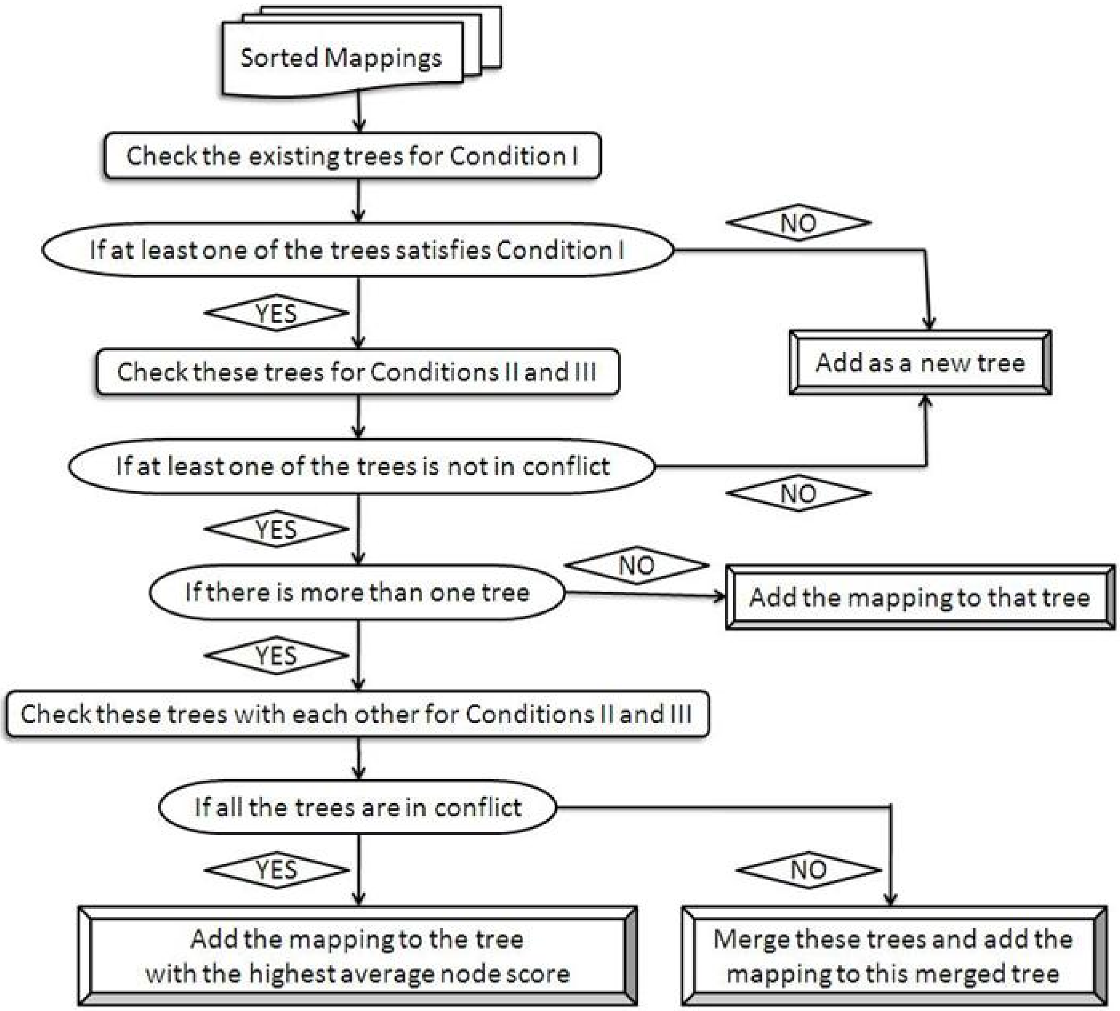

Sorted Mappings
Check the existing trees for Condition I
If at least one of the trees satisfies Condition I
NO
YES
Add as a new tree
Check these trees for Conditions II and III
If at least one of the trees is not in conflict
NO
YES
NO
If there is more than one tree
Add the mapping to that tree
YES
Check these trees with each other for Conditions II and III
If all the trees are in conflict
YES
NO
Add the mapping to the tree with the highest average node score
Merge these trees and add the mapping to this merged tree

Figure 3.11: Flow diagram of the merging algorithm

Chapter 4

EXPERIMENTS AND RESULTS

4.1 System Improvement

After the algorithm design, we tried to improve the efficiency of our scoring function and determine an optimum contact map threshold by performing several experiments on a small data set. In this section, the details of these experiments are explained along with the data set and the evaluation criteria.

4.1.1 Data Set

We performed all our experiments on a set of protein pairs. These proteins were chosen from ASTRAL 40 database [Chandonia et al., 2004] which contains protein pairs with sequence identity less than 40%. This database was created according to SCOP classification; therefore, the protein pairs are remote homologous and from the same sub-family. Random 10 protein pairs were chosen. These protein pairs, their lengths, SCOP families, and the sequence similarity percentages can be found in Table C.1 in Appendix C.

4.1.2 Evaluation

Since we focused on local structural alignments, it is not possible to evaluate our alignment results with measuring techniques such as RMSD value, LG score, or LGA measure. Therefore, we decided to evaluate our local alignments by comparing them with the results of a protein overall structural alignment method. We used TM-align [Zhang and Skolnick, 2005] because it has better accuracy than other structural alignment methods. For each mapping generated by our method, its' similarity to the alignment resulted from TM-align is detected and a TM-align similarity score is assigned. The TM-align similarity score is calculated as follows: our aligned residue pairs are compared with the pairs aligned by TM-align. The number of the same residue alignments is divided to the alignment length, which gives us the similarity percentage of the two alignments. An example comparison is given in Figure 4.1 where all residue alignments are same except the last one which therefore returns an 83.33% similarity score.

```
Our alignment:
A(16) G(32) T(84) C(34) G(47) L(86)
A(17) G(34) I(83) S(36) G(47) V(113)
TM-align alignment:
A(16) G(32) T(84) C(34) G(47) L(86)
A(17) G(34) I(83) S(36) G(47) V(92)
Alignment accuracy according to TM-align :  83,33%
```

Figure 4.1: Example TM-align comparison

Observing a graphlet with clique of size 7, 8, 9 or 10 in a contact map is a very low probability. For this reason, in all our experiments, the detected mappings always consist of 4, 5 or 6 node alignments. Possible TM-align similarity scores are 0%, 25%, 50%, 75%, 100% for a mapping with 4 residue alignments; 0%, 20%, 40%, 60%, 80%, 100% for a mapping with 5 residue alignments; and 0%, 16.66%, 33.33%, 50%, 66.66%, 83.33%, 100% for a mapping with 6 residue alignments.

4.1.3 Determining the Score Thresholds

As explained previously in Section 3.2.1, in the literature, different cutoff distances such as 5.8A^o, 6.8A^o, and 8.6A^o have been proposed for contact map generation. All three cutoff distances were evaluated during our system development. Three different contact maps were generated for all the twenty proteins. When these contact maps were compared with each other, it was observed that the contact maps produced with thresholds $5.8A^o$ and $6.8A^o$ are same for all the proteins except one. Therefore, only contact maps with $6.8A^o$ and $8.6A^o$ thresholds were used in the rest of the experiments.

In the next step, the created contact maps were used to generate graphlets for all the 145 different topologies shown in Appendix B. The graphlets of the two proteins were mapped for each protein pair. During this mapping process, only the graphlets from the same topologies, and the similar residue orderings were aligned. As stated before, the resulted mappings were compared with the alignments produced by TM-align in order to calculate the TM-align similarity score. Moreover, as explained in Section 3.2.4, the mappings were scored according to the evolutionary and connectivity similarities of the aligned amino acids and the residue distribution similarity of the aligned graphlets. We compared this score with the TM-align similarity score in order to determine the thresholds for our scoring function components which are R score for the residue distribution similarity, E score for the evolutionary similarity, and C score for the connectivity similarity. The frequencies of the scoring function components were determined for each TM-align similarity score. For instance, all the mappings' R scores were divided into intervals of size 0.1, and the number of mappings was counted within these intervals for each TM-align similarity score. In order to clarify the methodology, an example frequency table for R score is given in Table 4.1 for $8.6A^o$ contact map threshold. Similar to the R score, intervals of size 0.1 were used for C score which also takes values between 0 and 1. On the other hand, since E score can take values ranging from -4 to 11, we used intervals of size 0.5 for E score.

					TM-align	Similarity	Score						
R score	**0%**	**16.66%**	**20%**	**25%**	**33.33%**	**40%**	**50%**	**60%**	**66.66%**	**75%**	**80%**	**83.33%**	**100%**
0 - 0.1	42376	380	329	800	46	23	24	0	0	0	0	0	0
0.1 - 0.2	167987	3837	2484	1641	585	284	147	31	3	0	0	0	0
0.2 - 0.3	348152	10102	5841	2482	2343	692	623	97	42	1	0	0	0
0.3 - 0.4	588994	17919	9253	2724	5200	1637	1774	278	171	20	19	1	0
0.4 - 0.5	715400	21354	10182	2031	8413	3502	4148	1063	1055	47	157	38	0
0.5 - 0.6	695437	18768	8549	1571	9025	3336	5926	1808	3014	93	424	488	6
0.6 - 0.7	563537	12512	6958	1106	7402	3138	6052	2201	4455	183	766	2224	98
0.7 - 0.8	339320	6904	4302	621	4763	2698	4962	2365	5534	193	1475	3631	1075
0.8 - 0.9	190812	2729	1223	401	2128	983	2706	1214	3774	279	1223	4984	2233
0.9 - 1.0	77642	919	269	101	779	254	942	355	1281	169	656	2454	3822

Table 4.1: R score frequency table for graphlet mappings generated with $8.6A^o$ contact map threshold

When the frequencies were determined for all the score intervals and TM-align similarity scores, we observed that in all ten protein pairs, we have many mappings with 100% TM-align similarity score. For this reason, we decided to focus only on the TM-align similarity scores of 100% for determining our score thresholds. The number of mappings with 100% TM-align similarity score was determined for each interval of the three scoring function components. These values are represented in the graphs below in Figure 4.2, Figure 4.3, and Figure 4.4 which is a detailed version of Figure 4.3.

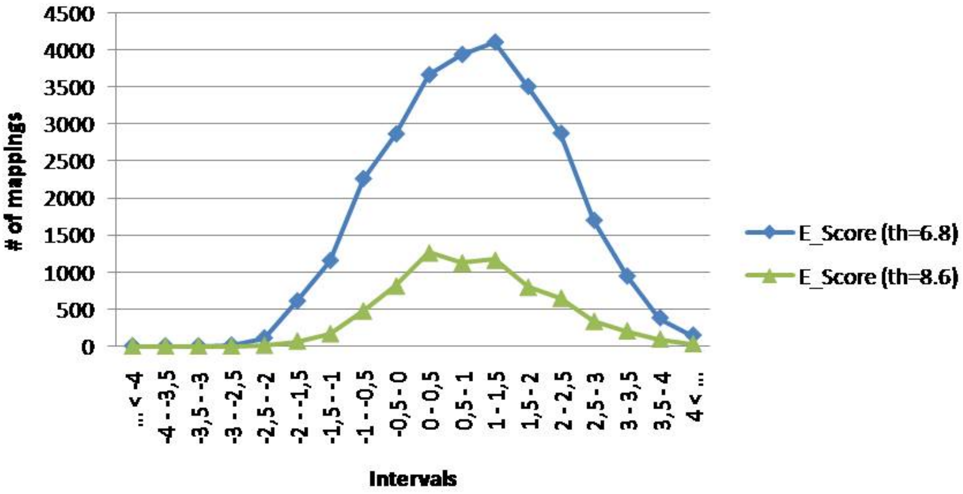


Figure 4.2: The distribution of E score for the number of mappings with 100% TM-align similarity score

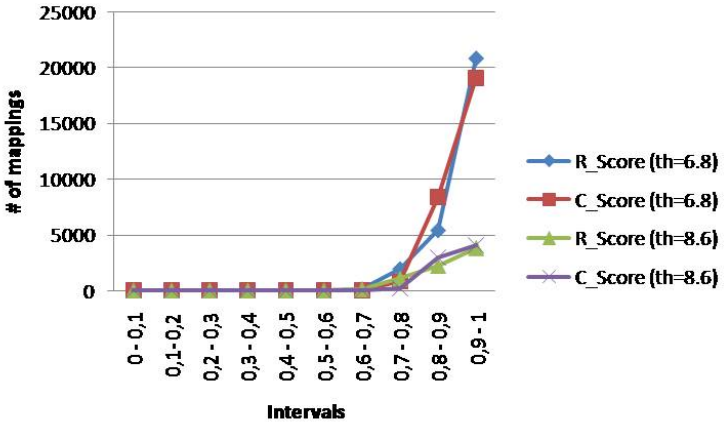


Figure 4.3: The distribution of R and C scores for the number of mappings with 100% TM-align similarity score

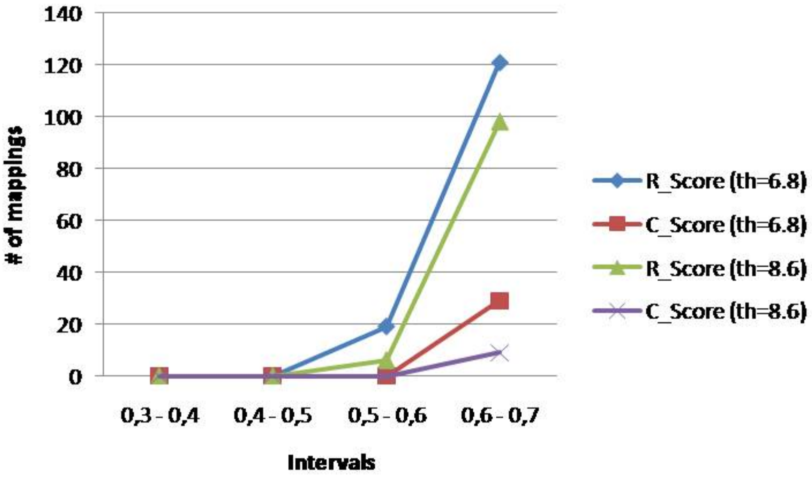


Figure 4.4: Detailed distribution of the R and C scores for the number of mappings with 100% TM-align similarity score

As seen in Figure 4.2, for 6.8A^o contact map threshold, none of the mappings with 100% TM-align similarity score has E score less than -3, and for 8.6A^o contact map threshold, none of the mappings have E scores less than -2.5. Furthermore, as seen in Figure 4.4, for both contact map thresholds, no mapping exists with an R score less than 0.5 or C score less than 0.6. Therefore, we used these values as score thresholds in our scoring function. As a result, graphlet mappings that have evolutionary less similar amino acid alignments, or mappings with different residue distributions were eliminated in the scoring section. When these eliminations were performed for the graphlet mappings of ten protein pairs, a minimum 9.97% and a maximum 38.39% decrease was observed in the number of mappings generated with 6.8A^o contact map threshold. These percentages are even more drastic for mappings generated with 8.6A^o contact map threshold, with a minimum 40.14% and maximum 68.11% decrease. The numbers of eliminated and remained mappings for each contact map cutoff distance are represented in Figure 4.5.

4.1.4 Determining the Coefficients of the Scoring Function

As mentioned in Section 3.2.4.4, multidimensional linear regression was performed with the purpose of determining the scoring function coefficients. The coefficients obtained

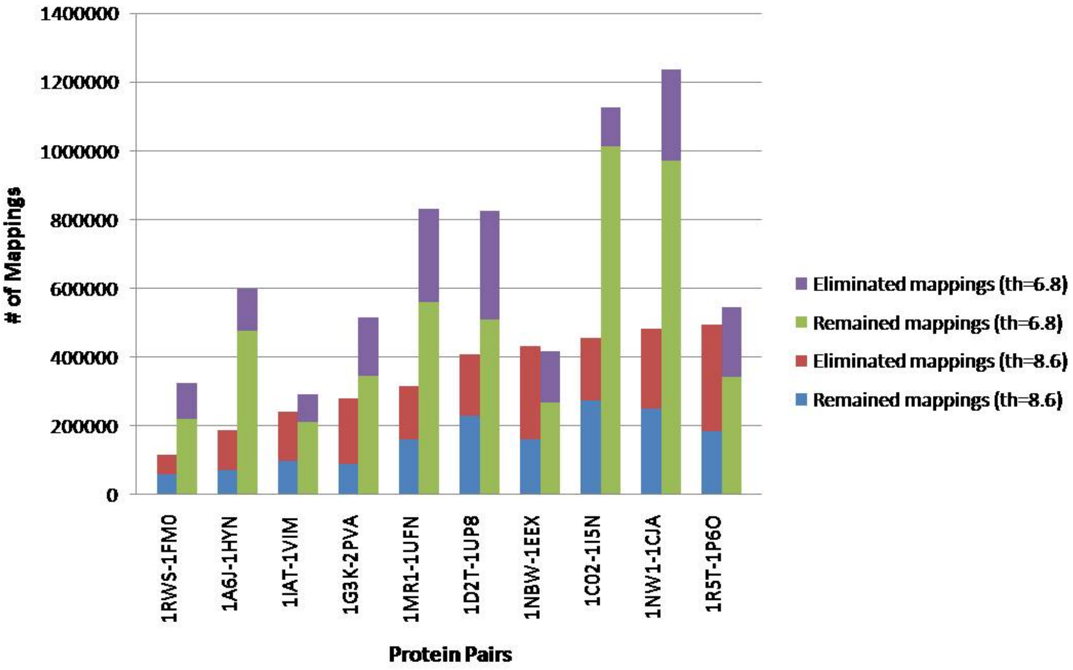


Figure 4.5: The numbers of eliminated and remained mappings

from the regression are given in Table 4.2 and Table 4.3. In addition, other regression statistics can be found in Appendix D.

In the multidimensional linear regression, several approaches for selecting the subset of predictor variables were proposed. In statistical methods, the order of the predictive variables entering into the model is determined according to the strength of their correlation with the dependent variable. In our regression analysis, we used the stepwise regression which tests the regression model at each stage for predictive variables to be included or excluded. The best model in our experiment was the case that includes all three predictive variables to the model. This result indicates that all the three scores in our scoring function are significant and necessary for a good alignment.

Furthermore, when the determined coefficients are compared between each other, for both regression models, E score's coefficient is the minimum one. This result was expected because while the intervals for the E score values are very wide, between -4 and 11; the intervals for the R and C score values are very close, between 0 and 1. Moreover, it has been observed that C score has the biggest coefficient in both regression models

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-124.783	.394		-316.810	.000
C Score	160.159	.458	.361	349.600	.000
E Score	9.656	.028	.350	343.708	.000
R Score	29.262	.237	.124	123.711	.000

Table 4.2: Coefficients of the scoring function for graphlet mappings generated with $6.8A^o$ contact map threshold

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-123.157	.460		-267.655	.000
C Score	138.215	.520	.367	265.811	.000
E Score	7.612	.034	.306	221.188	.000
R Score	51.582	.307	.225	167.852	.000

Table 4.3: Coefficients of the scoring function for graphlet mappings generated with $8.6A^o$ contact map threshold

which proves that connectivity of the residues are important in structural alignment.

4.1.5 Determining the Optimum Contact Map Threshold

As mentioned before, the protein pairs used in these experiments were chosen from the same superfamilies as can be seen in Table C.1. For this reason, we decided to determine our contact map threshold by performing fold classification in our data set. All-vs.-all pairwise search was performed on the data set where the first proteins were compared with all the second proteins. Since our aim in this process was to determine the optimum contact map threshold, all these experiments were performed for two different cutoff distances, $6.8A^o$ and $8.6A^o$. After graphlets of all the proteins were generated for two contact map thresholds, they were mapped to each other. A mapping was eliminated if it does not satisfy one of the below conditions:

- R score must be bigger than or equal to 0.5.

- C score must be bigger than or equal to 0.6.

- If contact map threshold is $6.8A^o$, than its' E score must be bigger than or equal to -3.
- If contact map threshold is $8.6A^o$, than its' E score must be bigger than or equal to -2.5.

Mappings that satisfy all the above conditions were ranked using the scoring function with the corresponding coefficients. The mappings were merged starting from the top score graphlet mapping. In all of the merging processes, only the highest scoring 500 mappings were used. At the end of the merging process, the highest score was used in the classification process. After all the comparisons were finished, the scores were ranked for each protein. The fold of a protein was determined using the fold of the hit protein. If the hit protein is in the same superfamily with the searched protein, a correct prediction has been achieved.

When fold classification was performed for two different contact map thresholds, 70% accuracy was obtained with graphlets generated from the contact map with cutoff distance $6.8A^o$ while 80% accuracy was observed with threshold $8.6A^o$. Since graphlets generated from $8.6A^o$ contact map thresholds have better classification accuracy, we decided to choose $8.6A^o$. Moreover, cutoff distance $8.6A^o$ seems to be more efficient than $6.8A^o$. In Figure 4.5, we showed the number of mappings obtained from contact maps with different threshold. A substantial difference is observed in the number of mappings when thresholds $6.8A^o$ and $8.6A^o$ are compared. Even though, the number of mappings obtained from the contact maps with $8.6A^o$ is very much smaller than the number of mappings obtained from the contact maps with $6.8A^o$, its classification accuracy is higher. Therefore, $8.6A^o$ is chosen as the contact map threshold of our system.

4.2 Function Prediction

4.2.1 Data Set

In addition to the system improvements, some experiments were also performed to evaluate the performance of the system in function prediction. As explained in Section 2.6.1, EC (Enzyme Commission) number prediction is one of the most common techniques used for evaluating function predictions. Our experiments were performed on a set of enzyme pairs. This data set was obtained from [Küçükural, 2008] where it had been used for the same purpose. The data set contains 44 protein pairs and they were all specially chosen from remote homologues and the lengths of the protein sequences are at least three times longer than its corresponding pair. These protein pairs and their EC numbers can be found in Table C.2 in Appendix C.

4.2.2 Results

The accuracy of the function prediction is calculated using the EC number prediction as mentioned before. Similar to the fold classification, an all-vs.-all search was performed in the data set. In this search, our algorithm first found the local structural alignments and recognized the common domains shared between two proteins, and then the calculated scores were ranked for each protein. The function of the protein was determined using the function of the hit protein. If the hit protein has the same EC number of the searched protein, a correct prediction has been achieved. If the correct prediction has not been reached in the top hit, then for the evaluation purposes, top 5 and 10 hits are considered whether a protein with the same function can be found in those hits.

When only the top hits are considered for function prediction, our accuracy rate is 97.05%. This accuracy rate is much higher than the accuracy rates reported in [Küçükural, 2008] for the same data set. The results are shown in Table 4.4.

Moreover, when the score of the top hits are compared with the following hits' scores, big fold differences are observed. In our data set, for the correctly classified enzymes, the minimum observed fold difference between the score of the correct protein function assignment and the score of the highest scoring wrong protein function is 2.28

	Our method	[Küçükural, 2008]
Top score	97.05%	55.66%
Top 5 score	97.05%	77.94%
Top 10 score	97.05%	88.24%

Table 4.4: Function prediction results

and these fold differences increase until 13.7. These high fold differences prove that the high accuracy of our system is not by chance.

Furthermore, in the second column of Table 4.5, for each protein the similarity between the recognized domains and alignments obtained from TM-align is given. Moreover, domains' coverage percentages are also given in the last column. These high accuracies with low coverage percentages indicate that without performing an overall alignment, our algorithm is able to recognize local domain regions successfully, and its residue alignments accuracies are very similar to TM-align results. However, our TM-align similarity accuracies are 0% in several proteins which contain Receptor L domains. We observed that in these proteins, TM-align can align only the half of the amino acids, and the remaining amino acids are aligned with gaps. This was an unexpected observation because in the rest of the proteins TM-align aligns the 95-100% of the amino acids. Since even for the proteins with receptor L domain our algorithm predict the functions of the proteins correctly, we believe that the low accuracy values of protein with receptor L domain are resulted from TM-align's weak performance.

Protein	Domain	Accuracy (%)	Coverage(%)
1A81	SH2	95	41.23
1B90	Glyco_hydro_14	100	96.96
1EMS	HIT	98.96	75.78
1FD9	FKBP_C	93.67	69.91
1GAX	Anticodon_1	99.2	88.11
1GPM	GMP_synt_C and GATase	54.92	37.96
1GWE	Catalase	35.71	31.39
1ITO	Peptidase_C1	100	56.25
1KI0	Pacifastin_I	100	47.5
1LAR	Y_phosphatase	100	26.66
1LCK	SH2 and SH3_1	96.72	62.88
1M6B	Recep_L_domain	0	0
1M8P	APS_kinase	100	74.43
1MIR	Peptidase_C1	100	50
1N8Y	Recep_L_domain	0	0
1N8Z	Recep_L_domain	0	0
1NYQ	tRNA-synt_2b and tRNA_SAD	98.34	54.01
1O6K	Pkinase	97.33	61
1PBH	Peptidase_C1	100	45.83
1QCF	SH2 and SH3_1	100	26.22
1SY7	Catalase	100	36.32
1WAA	I-set	96.66	67.41
1YGU	Y_phosphatase	97.72	30.87
2A91	Recep_L_domain	0	0
2AHX	Recep_L_domain	0	0
2B3O	SH2	96.9	37.45
2ESM	Pkinase	96.25	50.5
2F2U	Pkinase	97.5	50.31
2FH7	Y_phosphatase	0	0
2J0J	Pkinase_Tyr	97.97	40.57
2NLK	Y_phosphatase	98.75	27.77
2NP0	Toxin_R_bind_N and Toxin_R_bind_C	96.36	12.82
2RD0	PI3K_C2	100	78.4
2Z6B	Phage_lysozyme and Gp5_OB	96.62	54.26

Table 4.5: Domain prediction results

Chapter 5

CONCLUSIONS

In this thesis, we presented a method for finding structure patterns common to a protein pair by using graphlet mappings. Identifying these common structures patterns from diverse protein structures is one of the most challenging problems in bioinformatics due to several difficulties which we tried to overcome with our algorithm.

One of the difficulties is that proteins contain hundreds of amino acids with thousands of atoms and chemical bonds; therefore, they are large and complex geometric structures. In order to simplify the protein structures; we represented them with contact maps and during the construction of contact maps, C_α atoms were used since they represent an amino acid better than other atoms. Moreover, different contact map thresholds were tried with the purpose of finding the best cutoff distance. At the end of these experiments, it is observed that structurally important domains can be recognized better from graphlets generated with the contact map threshold of $8.6A^o$ which is also more efficient than $6.8A^o$.

Furthermore, in these large protein structures, we do not have any knowledge about the possible location or geometric shape of the structural patterns. Therefore, during our graphlet generation step, our algorithm searches all parts of the protein with the purpose of identifying all possible structural patterns. Our 145 different graphlet topologies cover all possible structural patterns during this step.

Last but not least, because of the evolutionary mutations, the common structure patterns between two proteins may show small variations such as different amino acids or compositions. In order to tolerate such differences, we allow different residue distributions and amino acid mismatches in our alignments. However, in order to prevent the alignment of very diverse structures, we incorporated the evolutionary similarity score and the residue distribution score into our scoring scheme. Moreover, we also included the connectivity score into our scoring function to find functionally similar structures, and at the end of the regression analysis, this score proves to be an important factor in a good alignment.

We found structural pattern with our graphlet mapping algorithm, and then by merging these local alignments, we tried to recognize domains that are common between protein pairs. These common domains are very useful in finding a protein's function, classifying a protein's fold, and identifying homology relationships. In this thesis, our algorithm was first applied to a fold classification problem on a small data set and 80% accuracy rate was observed. Then, in a larger data set, we tried to predict the proteins' functions using the domains that are discovered with our algorithm. The accuracy rate of predicting the correct function for our data set was 97.05% which is better than the previously published results on the same data set.

Currently, our algorithm can perform local alignments between only two proteins; however, with small improvements in our graphlet mapping step, multi-structural alignments can be obtained. A multi-structural alignment between proteins with the same function can be very useful in finding the functionally important sites. Besides the multi-structural alignment, our algorithm can be also used to develop a global structural alignment method. Our algorithm already assembles short local alignments into a longer alignment in the merging step. Using these longer alignments, an overall alignment can be obtained by matching the unaligned amino acids from the protein pair. A global alignment obtained this way can be more accurate than previously developed global alignment methods since this alignment conserves the structural patterns that are common between protein pair. Lastly, our algorithm currently performs function prediction according to the top scoring domain. In multi-domain proteins, such an assignment will be misleading since different domains of the protein may have different functions. During our merging process, we produce all possible domains; therefore, with small modifications, multi-label classification can be performed.

Bibliography

Nickolai Alexandrov and Ilya Shindyalov. PDP: protein domain parser. *Bioinformatics*, 19(3):429–430, 2003.

Ali Rana Atılgan, Pelin Akan, and Canan Baysal. Small-world communication of residues and significance for protein dynamics. *Biophysical Journal*, 86:85, 2004.

Ivet Bahar and Robert L. Jernigan. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *Journal of Molecular Biology*, 266(1):195–214, 1997.

Jonathan A. Barker and Janet M. Thornton. An algorithm for constraint-based structural template matching: application to 3d templates with statistical analysis. *Bioinformatics*, 19(13):1644–1649, 2003.

Lisa Bartoli, Emidio Capriotti, Piero Fariselli, Pier L. Martelli, and Rita Casadio. The pros and cons of predicting protein contact maps. *Protein Structure Prediction*, 413: 199–217, 2008.

Alex Bateman, Ewan Birney, Lorenzo Cerruti, Richard Durbin, Laurence Etwiller, Sean R. Eddy, Sam Griffiths-jones, Kevin L. Howe, Mhairi Marshall, and Erik L. L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Research*, 32 (Database-Issue):138–141, 2004.

Asa Ben-hur and Douglas L. Brutlag. Sequence motifs: highly predictive features of protein function. In *Feature extraction and foundations*. Springer Verlag, 2004.

Asa Ben-Hur and Douglas L. Brutlag. Remote homology detection: a motif based approach. In *ISMB (Supplement of Bioinformatics)*, pages 26–33, 2003.

P. Bork. Shuffled domains in extracellular proteins. *FEBS Lett*, 286:47–54, 1991.

Luiz C. Borro, Stanley R.M. Oliveira, Michel E.B. Yamagishi, Adaulto L. Mancini, Jose G. Jardine, Ivan Mazoni, Edgard H. dos Santos, Roberto H. Higa, Paula R. Kuser, and Goran Neshich. Predicting enzyme class from protein structure using bayesian classification. *Genet Mol Res*, 5:193–202, 2006.

Carl-Ivar Branden and John Tooze. *Introduction to Protein Structure*. Garland Publishing, 2nd edition, 1999.

Michal Brylinski and Jeffrey Skolnick. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of Sciences*, 105:129–134, 2008.

Yu-Dong Cai and Kuo-Chen Chou. Using functional domain composition to predict enzyme family classes. *J Proteome Res*, 4:109–111, 2004.

Oliviero Carugo. Rapid methods for comparing protein structures and scanning structure databases. *Current Bioinformatics*, 1:75–83(9), 2006.

John-Marc Chandonia, Gary Hon, Nigel S. Walker, Loredana Lo Conte, Patrice Koehl, Michael Levitt, and Steven E. Brenner. The ASTRAL compendium in 2004. *Nucleic Acids Res*, 32(Database issue), 2004.

Ke Chen and Lukasz Kurgan. PFRES: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics*, 23(21):2843–2850, 2007.

Lusheng Chen, Wei Wang, Shaoping Ling, Caiyan Jia, and Fei Wang. KemaDom: a web server for domain prediction using kernel machine with local context. *Nucleic Acids Research*, 34(Web Server-Issue):158–163, 2006.

Ana Conesa, Stefan Gotz, Juan M. Garcia-Gomez, Javier Terol, Manuel Talon, and Montserrat Robles. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.

Geoffrey M. Cooper and Robert E. Hausman. *The cell: a molecular approach(3rd ed.)*. Sinauer, 2004.

Marie desJardins, Peter D. Karp, Markus Krummenacker, Thomas J. Lee, and Christos A. Ouzounis. Prediction of enzyme classification from protein sequence without the use of sequence similarity. In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*, pages 92–99. AAAI Press, 1997.

Paul D. Dobson and Andrew J. Doig. Predicting enzyme class from protein structure without alignments. *J Mol Biol*, 345(1):187–199, 2005.

Richard A. George and Jaap Heringa. SnapDRAGON: a method to delineate protein structural domains from sequence data. *J Mol Biol.*, 316(3):839–851, 2002a.

Richard A. George and Jaap Heringa. Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins*, 48(4):672–681, 2002b.

Jérôme Gracy and P. Argos. Automated protein sequence database classification. i. integration of compositional similarity search, local similarity search, and multiple sequence alignment. *Bioinformatics*, 14(2):164–173, 1998.

Nitin Gupta, Nitin Mangal, and Somenath Biswas. Evolution and similarity evaluation of protein structures in contact map space. *Proteins*, 59(2):196–204, 2005.

LY Han, CZ Cai, Zhi Liang Ji, Zhi Wei Cao, J Cui, and Yu Zong Chen. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.*, 32:6437–6444, 2004.

Andreas Heger and Liisa Holm. Exhaustive enumeration of protein domain families. *J Mol Biol.*, 328(3):749–767, 2003.

Steven Henikoff and Jorja G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919, 1992.

Liisa Holm and Chris Sander. Mapping the protein universe. *Science*, 273(5275):595–603, 1996.

Fereydoun Hormozdiari, Petra Berenbrink, Nataa Przulj, and S. Cenk Sahinalp. Not all scale-free networks are born equal: The role of the seed graph in PPI network evolution. *PLoS Comput Biol*, 3(7):e118, 2007.

Jun Huan, Wei Wang, Deepak Bandyopadhyay, Jack Snoeyink, Jan Prins, and Alexander Tropsha. Mining protein family specific residue packing patterns from protein structure graphs. In *RECOMB '04: Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, pages 308–315. ACM Press, 2004.

Helgi Ingolfsson and Golan Yona. Protein domain prediction. In Mitchell Guss Bostjan Kobe and Thomas Huber, editors, *Structural Proteomics High-Throughput Methods*, volume 426, pages 117–143. Humana Press, 2008.

IUBMB. *Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992.* Academic Press, 1992.

Yi Jia, Jun Huan, Vincent Buhr, Jintao Zhang, and Leonidas N. Carayannopoulos. Towards comprehensive structural motif mining for better fold annotation in the "twilight zone" of sequence dissimilarity. *BMC Bioinformatics*, 10, 2009.

Alper Küçükural. Novel techniques for protein structure characterization using graph representation of proteins, 2008.

Alper Küçükural, O. Uğur Sezerman, and Aytül Erçil. Discrimination of native folds using network properties of protein structures. In *APBC*, pages 59–68, 2008.

Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.

Bum Ju Lee, Heon Gyu Lee, Dae-sung Kim, and Keun Ho Ryu. Feature extraction in spatially-conserved regions and protein functional classification. In *FBIT '07: Proceedings of the 2007 Frontiers in the Convergence of Bioscience and Information Technologies*, pages 165–170. IEEE Computer Society, 2007.

Arthur M. Lesk and George D. Rose. Folding units in globular proteins. *Proc. Natl. Acad. Sci. USA.*, 78:4304–4308, 1981.

Michael Levitt and Mark Gerstein. A unified statistical framework for sequence comparison and structure comparison. *PNAS*, 95(11):5913–5920, 1998.

Mike P. Liang, Douglas L. Brutlag, and Russ B. Altman. Automated construction of structural motifs for predicting functional sites on protein structures. In *Pacific Symposium on Biocomputing*, pages 204–215, 2003.

Russell L. Marsden, Liam J. McGuffin, and David T. Jones. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.*, 11(12):2814–2824, 2002.

David M.A. Martin, Matthew Berriman, and Geoffrey J. Barton. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, 5, 2004.

Mariusz Milik, Sandor Szalma, and Krzysztof A. Olszewski. Common structural cliques: a tool for protein structure and function analysis. *Protein Eng*, 16(8):543–552, 2003.

Sanzo Miyazawa and Robert L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. page 534552, 1985.

Alexey G. Murzin, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, 1995.

Ruth Nussinov and Haim J. Wolfson. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci U S A*, 88(23):10495–10499, 1991.

Christine A. Orengo and William R. Taylor. SSAP: Sequential structure alignment program for protein structure comparison. *Computer Methods for Macromolecular Sequence Analysis*, 266:617–635, 1996.

Christine A. Orengo, A. D. Michie, Susan Jones, David T. Jones, Mark B. Swindells, and Janet M. Thornton. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

Juan Parrado, Francisco Conejero-Lara, Richard A.G. Smith, Julian M. Marshall, Christopher P. Ponting, and Christopher M. Dobson. The domain organization of streptokinase: nuclear magnetic resonance, circular dichroism, and functional characterization of proteolytic fragments. *Protein Sci.*, 5(4):693–704, 1996.

Gregory A. Petsko and Dagmar Ringe. *Protein structure and function: From sequence to consequence.* New Science Press, 2003.

Elon Portugaly, Nathan Linial, and Michal Linial. EVEREST: a collection of evolutionary conserved protein domains. *Nucleic Acids Research*, 35(Database-Issue):241–246, 2007.

Natasa Przulj, Derek G. Corneil, and Igor Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

Jane S. Richardson. The anatomy and taxonomy of protein structure. *Advances in protein chemistry*, 34:167–339, 1981.

Daniel J. Rigden. Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments. *Protein Engineering*, 15(2): 65–77, 2002.

Burkhard Rost. Enzyme function less conserved than anticipated. *J Mol Biol*, 318(2): 595–608, 2002.

Ahmet Saçan, Özgür Öztürk, Hakan Ferhatosmanoğlu, and Yusu Wang. LFM-Pro: a tool for detecting significant local structural sites in proteins. *Bioinformatics*, 23(6): 709–716, 2007. ISSN 1367-4803.

Ahmet Saçan, Hakkı I. Toroslu, and Hakan Ferhatosmanoğlu. Integrated search and alignment of protein structures. *Bioinformatics*, 24(24):2872–2879, 2008.

Jorg Schultz, Frank Milpetz, Peer Bork, and Chris P. Ponting. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences*, 95(11):5857–5864, 1998.

Carlos Setubal and Joao Meidanis. *Introduction to Computational Molecular Biology.* 1997.

Imran Shah and Lawrence Hunter. Predicting enzyme function from sequence: A systematic appraisal. In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*, pages 276–283. AAAI Press, 1997.

Mohammad Tabrez Anwar Shamim, Mohammad Anwaruddin, and Hampapathalu A. Nagarajaram. Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. *Bioinformatics*, 23(24):3320–3327, 2007.

Jessica Shapiro and Douglas Brutlag. FoldMiner: Structural motif discovery using an improved superposition algorithm. *Protein Science*, 13(1):278–294, 2004.

Dalit Shental-Bechor, Safak Kırca, Nir Ben-Tal, and Türkan Haliloğlu. Monte carlo studies of folding, dynamics, and stability in alpha-helices. *Biophysical Journal*, 88 (4):23912402, 2005.

Ilya N. Shindyalov and Philip E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11(9):739–747, 1998.

Umar Syed and Golan Yona. *Enzyme function prediction with interpretable models*, volume 541, chapter 17, pages 373–420. Humana Press, 2009.

Aik Choon Tan, David Gilbert, Tan David, and Yves Deville. Multi-class protein fold classification using a new ensemble machine learning approach. *Genome Informatics*, 14:206–217, 2003.

Susan S. Taylor and Elzbieta Radzio-Andzelm. Three protein kinase structures define a common motif. 2(5):345–355, 1994.

Todd J. Taylor and Iosif I. Vaisman. Graph theoretic properties of networks formed by the delaunay tessellation of protein structures. *Physical Review E*, 73:041925, 2006.

James W. Torrance, Gail J. Bartlett, Craig T. Porter, and Janet M. Thornton. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J Mol Biol*, 347:565–581, 2005.

Michele Vendruscolo and Eytan Domany. Efficient dynamics in the space of contact maps. *Fold Des*, 3(5):329–336, 1998.

Michele Vendruscolo, Edo Kussell, and Eytan Domany. Recovery of protein structure from contact maps. *Fold Des*, 2(5):295–306, 1997.

Pramod P. Wangikar, Ashish V. Tendulkar, S. Ramya, Deepali N. Mali, and Sunita Sarawagi. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *Journal of molecular biology*, 326(3):955–978, 2003.

Donald B. Wetlaufer. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A*, 70:697–701, 1973.

Ying Xu, Dong Xu, and Harold N. Gabow. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, 16(12):1091–1104, 2000.

Mohammed J Zaki. Mining protein contact maps. In *The 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD*, 2003.

Adam Zemla. LGA: A method for finding 3d similarities in protein structures. *Nucleic Acids Res*, 31(13):3370–3374, 2003.

Işık Zerrin, Berrin Yanıkoğlu, and O. Uğur Sezerman. Protein structural class determination using support vector machines. In *Lecture Notes in Computer Science*, pages 82–89. Springer, 2004.

Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, 33(7):2302–2309, 2005.

Ying Zhao and George Karypis. Prediction of contact maps using support vector machines. In *Proc. of the IEEE Symposium on BioInformatics and BioEngineering*, pages 26–36. IEEE Computer Society, 2003.

Chapter A

Graplet Topologies

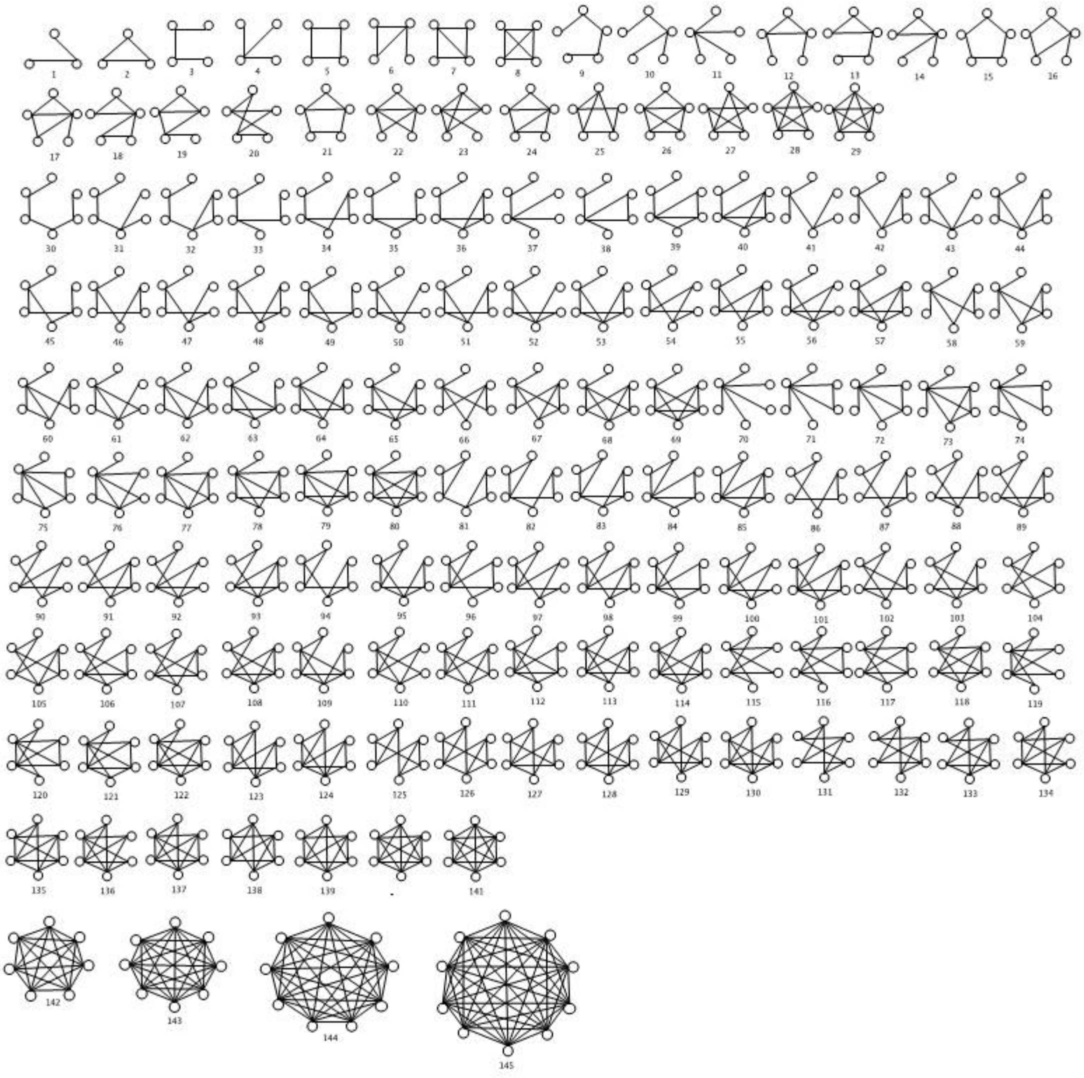


Figure A.1: Graplet topologies used

Chapter B

BLOSUM62 Matrix

A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

Table B.1: BLOSUM62 matrix

Chapter C

Data Sets

Protein 1	Length	SCOP Class	Protein 2	Length	SCOP Class	Sequence Similarity (%)
1R5T	141	c.97.1.1	1P6O	156	c.97.1.2	15.4
1D2T	222	a.111.1.1	1UP8	597	a.111.1.2	16.5
1IAT	556	c.80.1.2	1VIM	192	c.80.1.3	17.4
1G3K	173	d.153.1.4	2PVA	332	d.153.1.3	18.0
1NW1	365	d.144.1.8	1CJA	327	d.144.1.3	18.8
1A6J	150	d.112.1.1	1HYN	293	d.112.1.2	24.6
1NBW	113	c.51.3.2	1EEX	178	c.51.3.1	24.7
1RWS	68	d.15.3.2	1FM0	81	d.15.3.1	25.9
1C02	166	a.24.10.2	1I5N	128	a.24.10.3	26.2
1MR1	97	d.217.1.2	1UFN	94	d.217.1.1	27.0

Table C.1: Protein data set for system improvement

Protein 1	Chain	EC Number	Protein 2	Chain	EC Number
1A81	A	2.7.1.112	1JWO	A	2.7.1.112
1B90	A	3.2.1.2	1CQY	A	3.2.1.2
1EMS	A	3.6.1.29	2FIT	A	3.6.1.29
1FD9	A	5.2.1.8	1YAT	A	5.2.1.8
1GAX	A	6.1.1.9	1WK9	A	6.1.1.9
1GPM	A	6.3.5.2	2VPI	A	6.3.5.2
1GWE	A	1.11.1.6	1YE9	A	1.11.1.6
1ITO	A	3.4.22.1	1SP4	A	3.4.22.1
1KI0	A	3.4.21.7	2PK4	A	3.4.21.7
1KI0	A	3.4.21.7	5HPG	A	3.4.21.7
1LAR	A	3.1.3.48	2B49	A	3.1.3.48

Protein 1	Chain	EC Number	Protein 2	Chain	EC Number
1LAR	A	3.1.3.48	2GJT	A	3.1.3.48
1LAR	A	3.1.3.48	2I75	A	3.1.3.48
1LAR	A	3.1.3.48	2PA5	A	3.1.3.48
1LCK	A	2.7.1.112	3CQT	A	2.7.10.2
1M6B	A	2.7.1.112	3C09	A	2.7.10.1
1M8P	A	2.7.7.4	2PEY	A	2.7.1.25
1MIR	A	3.4.22.1	1SP4	A	3.4.22.1
1N8Y	C	2.7.10.1	3C09	A	2.7.10.1
1N8Z	C	2.7.1.112	3C09	A	2.7.10.1
1NYQ	A	6.1.1.3	1TJE	A	6.1.1.3
1O6K	A	2.7.11.1	2NP8	A	2.7.11.1
1PBH	A	3.4.22.1	1SP4	A	3.4.22.1
1QCF	A	2.7.10.2	3CQT	A	2.7.10.2
1SY7	A	1.11.1.6	1YE9	A	1.11.1.6
1WAA	A	2.7.11.1	2YZ8	A	2.7.11.1
1YGU	B	3.1.3.48	2B49	A	3.1.3.48
1YGU	B	3.1.3.48	2I4G	A	3.1.3.48
1YGU	B	3.1.3.48	2I75	A	3.1.3.48
1YGU	B	3.1.3.48	2PBN	A	3.1.3.48
2A91	A	2.7.1.112	3C09	A	2.7.10.1
2AHX	A	2.7.1.112	3C09	A	2.7.10.1
2B3O	A	3.1.3.48	2B49	A	3.1.3.48
2ESM	A	2.7.1.37	2NP8	A	2.7.11.1
2F2U	A	2.7.1.37	2NP8	A	2.7.11.1
2FH7	A	3.1.3.48	2GJT	A	3.1.3.48
2FH7	A	3.1.3.48	2I4G	A	3.1.3.48
2FH7	A	3.1.3.48	2I75	A	3.1.3.48
2J0J	A	2.7.10.2	2OFV	A	2.7.10.2
2NLK	A	3.1.3.48	2OC3	A	3.1.3.48
2NLK	A	3.1.3.48	2QEP	A	3.1.3.48
2NP0	A	3.2.1.52	2QN0	A	3.4.24.69
2RD0	A	2.7.1.153	2V1Y	A	2.7.1.153
2Z6B	A	3.2.1.17	3LZM	A	3.2.1.17

Table C.2: Enzyme data set for function prediction

Chapter D

Regression Statistics

Contact map threshold	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
$6.8A^o$	1	.613	.376	.376	25.62049
$8.6A^o$	1	.624	.390	.390	21.66936

Table D.1: Model summaries

Contact map threshold	Model	Sum of Squares	df	Mean Square	F
$6.8A^o$	Regression	2.625E8	3	8.752E7	1.333E5
	Residual	4.366E8	665172	656.410	
	Total	6.992E8	665175		
$8.6A^o$	Regression	1.057E8	3	3.522E7	7.501E4
	Residual	1.656E8	352572	469.561	
	Total	2.712E8	352575		

Table D.2: ANOVA tables