Fall 2008

Integrated mining of feature spaces for bioinformatics domain discovery

Pradeep Chowriappa

INTEGRATED MINING OF FEATURE SPACES FOR BIOINFORMATICS DOMAIN DISCOVERY

by

Pradeep Chowriappa, B.S., M.C.A.

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

COLLEGE OF ENGINEERING AND SCIENCE
LOUISIANA TECH UNIVERSITY

November 2008

LOUISIANA TECH UNIVERSITY

THE GRADUATE SCHOOL

November 19th, 2008
Date

We hereby recommend that the dissertation prepared under our supervision by Pradeep Chowriappa

entitled Integrated Mining of Feature Spaces for Bioinformatics Domain Discovery

be accepted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Computational Analysis and Modeling

Supervisor of Dissertation Research

Head of Department
Computational Analysis and Modeling
Department

Recommendation concurred in:

Advisory Committee

Approved:
Director of Graduate Studies

Approved:
Dean of the Graduate School

Dean of the College

GS Form 13a
(6/07)

ABSTRACT

One of the major challenges in the field of bioinformatics is the elucidation of protein folding for the functional annotation of proteins. The factors that govern protein folding include the chemical, physical, and environmental conditions of the protein's surroundings, which can be measured and exploited for computational discovery purposes. These conditions enable the protein to transform from a sequence of amino acids to a globular three-dimensional structure. Information concerning the folded state of a protein has significant potential to explain biochemical pathways and their involvement in disorders and diseases. This information impacts the ways in which genetic diseases are characterized and cured and in which designer drugs are created. With the exponential growth of protein databases and the limitations of experimental protein structure determination, sophisticated computational methods have been developed and applied to search for, detect, and compare protein homology. Most computational tools developed for protein structure prediction are primarily based on sequence similarity searches. These approaches have improved the prediction accuracy of high sequence similarity proteins but have failed to perform well with proteins of low sequence similarity. Data mining offers unique algorithmic computational approaches that have been used widely in the development of automatic protein structure classification and prediction.

In this dissertation, we present a novel approach for the integration of physico-chemical properties and effective feature extraction techniques for the classification of

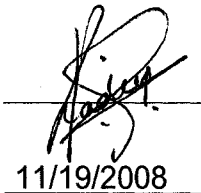
proteins. Our approaches overcome one of the major obstacles of data mining in protein databases, the encapsulation of different hydrophobicity residue properties into a much reduced feature space that possess high degrees of specificity and sensitivity in protein structure classification. We have developed three unique computational algorithms for coherent feature extraction on selected scale properties of the protein sequence. When plagued by the problem of the unequal cardinality of proteins, our proposed integration scheme effectively handles the varied sizes of proteins and scales well with increasing dimensionality of these sequences. We also detail a two-fold methodology for protein functional annotation. First, we exhibit our success in creating an algorithm that provides a means to integrate multiple physico-chemical properties in the form of a multi-layered abstract feature space, with each layer corresponding to a physico-chemical property. Second, we discuss a wavelet-based segmentation approach that efficiently detects regions of property conservation across all layers of the created feature space.

Finally, we present a unique graph-theory based algorithmic framework for the identification of conserved hydrophobic residue interaction patterns using identified scales of hydrophobicity. We report that these discriminatory features are specific to a family of proteins, which consist of conserved hydrophobic residues that are then used for structural classification. We also present our rigorously tested validation schemes, which report significant degrees of accuracy to show that homologous proteins exhibit the conservation of physico-chemical properties along the protein backbone. We conclude our discussion by summarizing our results and contributions and by listing our goals for future research.

APPROVAL FOR SCHOLARLY DISSEMINATION

The author grants to the Prescott Memorial Library of Louisiana Tech University the right to reproduce, by appropriate methods, upon request, any or all portions of this Dissertation. It is understood that "proper request" consists of the agreement, on the part of the requesting party, that said reproduction is for his personal use and that subsequent reproduction will not occur without written approval of the author of this Dissertation. Further, any portions of the Dissertation used in books, papers, and other works must be appropriately referenced to this Dissertation.

Finally, the author of this Dissertation reserves the right to publish freely, in the literature, at any time, any or all portions of this Dissertation.

Author 

Date 11/19/2008

DEDICATION

Dedicated to the four pillars of my life

"Matha, Pitha, Guru, Dievam."

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to the people of Louisiana Tech University who provided me with the kind and generous support that made this dissertation possible.

I am extremely indebted to my advisor, Dr. Sumeet Dua, for his constant and meticulous supervision of my research. His insatiable desire for creative solutions to problems in the field of bioinformatics has left an everlasting mark on me. He has always been available whenever I needed him, and most of all, I would like to thank him for being an inspiration. I would also like to thank Dr. Hilary Thompson and Dr. Jinko Kanno with whom I have had the privilege to work closely. Their valuable suggestions and encouragement have brought richness to this body of work. I would also like extend my gratitude to my other committee members Dr. Weizhong Dai and Dr. Katie Evans for their patience and kind cooperation.

I would also like to thank my "buddies" at the Data Mining Research Laboratory (DMRL) for tolerating me and making me feel at home. I enjoyed many fruitful discussions with them which resulted in valuable insights. Only Louisiana Tech could bring together these minds, and it has been my privilege to know them. Thank you Shraddha Pathak, Ramakrishnan Rajagopalan, Prithi Srinivasan, Kameshwari Palepu, and Naveen Kandiraju - pleasure learning with you. I would also like to give special thanks to

Harpreet Singh, Sheetal Saini, Alan Alex, and Sathya Alagiriswami for their support and help.

I would like to express my sincere appreciation to Brandy McKnight, Krithika Bhuvaneshwar, and Raelyn Williams for their technical support and insight into my doctoral dissertation presentation. And last but not the least I would like to thank my parents and friends for their constant encouragement and support.

CHAPTER 1

INTRODUCTION

Biology has been transformed greatly in the last century, gradually growing into a data rich field inviting scientific interests from a variety of researchers from disciplines including, but not limited to computer and computational sciences, mathematics, statistics, and engineering. The challenges imposed by biological problems have provided a much needed impetus for advancements, both in theory and in application, and have led to the development of unique multi-disciplinary fields including bioinformatics and biomedical computing. This chapter provides a brief overview of bioinformatics, emphasizing its growth and impact, as well as the scientific need for its advancement.

1.1 Biology and Bioinformatics

Biological research has witnessed a paradigm shift from in vivo and in vitro to in silico experimentation [1]. This development has been attributed to the development of bioinformatics, which has broadened the field of biology into new and otherwise unknown directions. Like other natural sciences, biology is fostered by human curiosity about natural phenomena [2]. As such, it explores the very existence of life. However, biology is still an immature science in which we cannot make predictions based on general principles [3]. Modern biotechnology began in the 1970s with the cloning and isolation of genes. Techniques offered by molecular biology and the completion of

human chromosome sequencing have brought bioinformatics to the forefront of the biological sciences. The new techniques developed in the automation of protein and DNA sequencing have made bioinformatics irreplaceable to both general and molecular biology.

Specifically, technological advances in computer science have made it possible to optimize the storage and use of data collected through years of experimental trials. This blend of technology and legacy techniques has bridged the gap between wet lab experiments and engineering computer simulations, hence opening a new area of research. Thus, bioinformatics emphasizes the management and analysis of biological information stored in large databases. In short, it is a science that consists of the amalgamation of biology, computer science, and mathematics. The ultimate goal of researchers in bioinformatics is to abstract knowledge and principles from large-scale data to represent and predict computational systems of higher complexity for cells and organisms.

1.2 The Growth of Bioinformatics

Technological advances in computer science have positively impacted bioinformatics, and with hardware and software becoming more economical, the scope of bioinformatics continues to grow. The Human Genome Project (HGP)[1], a project designed to map and sequence the complete human chromosome, as well as other important organisms, started in the mid-1990s, and to date has sequenced 100,000 nucleotide sequences (National Human Genome Research Institute[2]). A decade of research has resulted in a vast amount of data, with many sequence analysis problems for

[1] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
[2] http://www.genome.gov/

which conventional algorithms prove inadequate. This inadequacy is attributed to the inherent complexity of biological systems and the lack of knowledge of molecular organization.

Apart from the above issue, scientists must take into account that biological sequences are inherently noisy, due to variability arising from random events amplified by evolution. Machine learning processes play a vital role in removing noise from the sequences. They are suited for characterizing large amounts of data and noisy patterns in the absence of general theories. The idea behind these approaches is to learn the theory automatically from the data through a process of model fitting or learning from examples. This process is also called training.

1.3 The Impact of Bioinformatics

Bioinformatics has provided the basis for the future of large-scale biology: relative data rich science with inexpensive resources. Research and development can be done with modest equipment and public resources [4]. The advances in high performance computing are synonymous with advances made in biotechnology. With the internet providing a means to distribute data and software, researchers are able to perform sophisticated analyses on remote high performance servers.

The effects of data mining in bioinformatics can be enumerated as follows:

1. **Economical Impact**: Data mining provides an affordable solution to traditional (wet lab) techniques and may potentially replace them. This replacement could be brought about by using existing data to identify and remove those data that have no potential use, thereby speeding up the process and reducing the cost.

2. **Better Understanding**: Data mining has played a vital role in pointing out trends that would generally go unnoticed [4]. For instance, computational techniques, have provided scientists a better understanding of disease pathways, and have provided a better comprehension when identifying potential medication through the use of data mining and modeling, and various other visualization and simulation tools.
3. **Bioinformatics Tools**: Bioinformatics tools are designed to accurately identify and analyze gene and protein expressions with respect to healthy and diseased tissue at different stages of disease. This identification function means that bioinformatics technologies can be used to identify markers for cancer diagnosis, to monitor disease progression, and to identify therapeutic drug targets.

1.4 Computational Challenges in Bioinformatics

The rate at which data is being generated from high throughput biological projects continues to out distance the ability to interpret them, even when researchers use the fastest computers available today. This exponential growth of biological data has fuelled an overarching need for knowledge discovery efforts, which derive information from a growing body of invalidated data. For example, freely available protein databases provide new opportunities for the discovery and research. The ability to determine the structure of a protein without relying on sequence similarity is an important impetus for researchers in bioinformatics and has recently generated a great deal of scientific interest.

For computational scientists, these newly found and constantly improving abilities to determine protein structure without sequence similarity provide various opportunities

to participate in data-driven biological knowledge discoveries in which they derive biological hypotheses from hidden patterns discovered in this large volume of data.

The data and information discovery focus of our research is analyzing and interpreting patterns, trends, and anomalies from high dimensional protein databases. Figure 1.1 shows the yearly growth of the Protein Data Bank; the number of proteins in the bank reached 50,000 on April 22nd, 2008 [8]. Figure 1.2 shows the number of curated sequences of the SwissProt Database, which has reached 362,782 protein and nucleotide sequences [9].

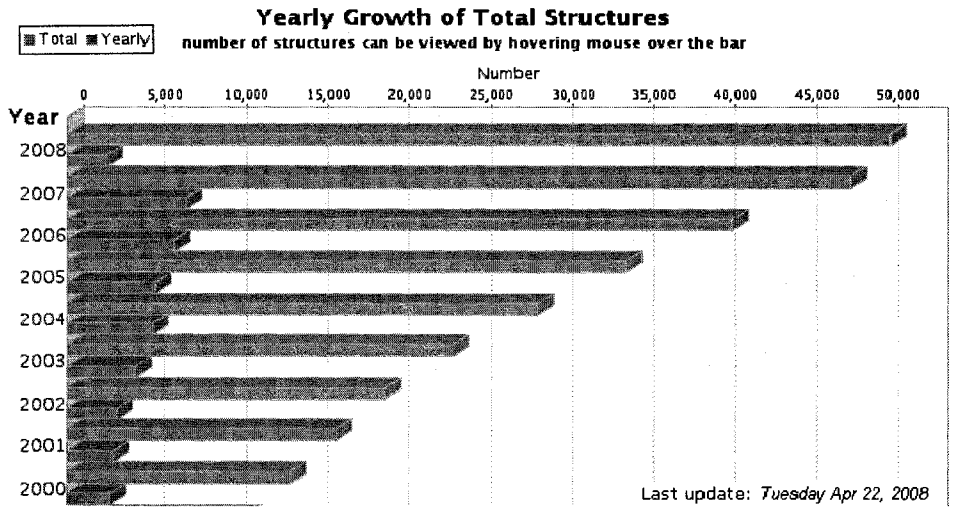


Figure 1.1 Yearly growth of the Protein Data Bank.

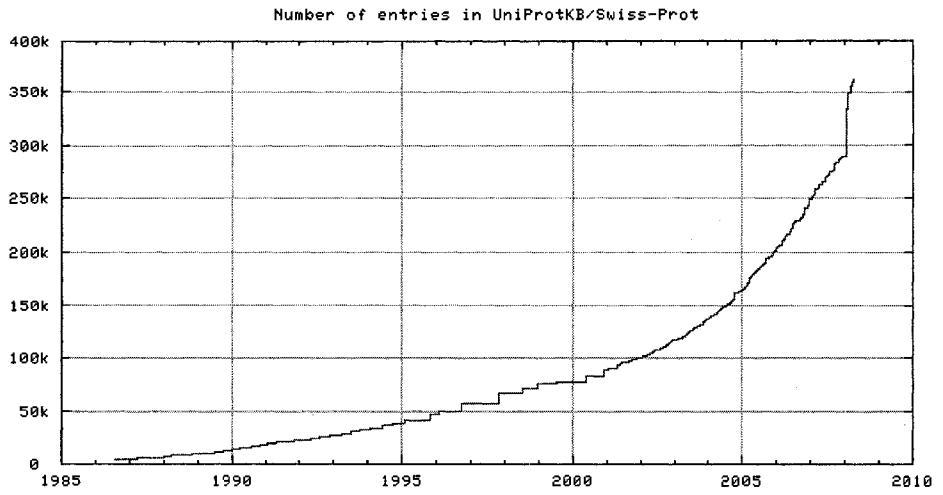


Figure 1.2 Number of curated sequences of the SwissProt database.

1.5 Knowledge Discovery in Databases

The new biological discovery opportunities can be divided into six specific data mining challenges in bioinformatics that are enumerated in the following section 1.5.1. First, data mining, by definition, involves the use of sophisticated data analysis tools for the discovery of previously unknown, valid patterns and relationships in large datasets in general [5]. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms, such as neural networks or decision trees that improve performance automatically through experience). Consequently, data mining consists of more than collecting and managing data; it also includes analysis and prediction. Data mining, also known as Knowledge Discovery in Databases (KDD), has been defined as "[t]he nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [17]. Data mining uses machine learning, statistical techniques, and visualization techniques to discover and present knowledge in an easily comprehensible form. Data mining algorithms include classification, clustering, and prediction [18].

1.5.1 Steps in Knowledge Discovery in Databases

Data mining is an iterative, data-driven, knowledge-discovery process that includes the following steps, each of which poses challenges for researchers in bioinformatics. In this section 1.5.1, we provide a brief overview of the knowledge discovery process in protein databases, which is illustrated in Figure 1.3.

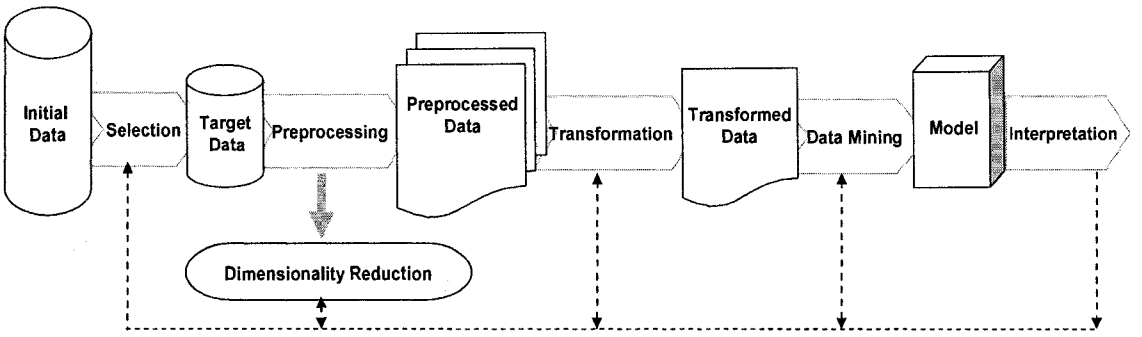


Figure 1.3 Data mining as a step in KDD.

1. **Data Selection**: The overwhelming size of the protein database calls for researchers to develop techniques that are applied both to reduce the volume of information and to maintain the integrity of the original dataset, in order to obtain a reduced representation of the dataset.
2. **Data Transformation**: Many high-throughput protein data-capturing devices and methods are still in early developmental stages; therefore, data from the databases are plagued with noise. In this step, the data are altered or consolidated into forms appropriate for mining.
3. **Data Mining**: Patterns and relationships are found in the data. Methods such as association rule mining, classification, and clustering form the core methods of this process.

4. **Interpretation or Pattern Evaluation**: Evaluation of the results obtained in the previous method is performed. These results are interpreted in order to extract interesting patterns that represent knowledge and are based on measures of interest.
5. **Knowledge Representation**: The results are correlated with supporting evidence extracted from existing biological literature. Visualization and knowledge representation techniques are then each used to illustrate the extracted knowledge to the user.

1.5.2 Feature Extraction

When dealing with high dimensional data, the two main challenges are (1) the algorithm's ability (or inability) to scale large datasets and (2) the Curse of Dimensionality. First, high dimensionality leads to inefficient space and time complexities as the dataset's dimensionality increases. Second, the Curse of Dimensionality is caused by the exponential increase in resources associated with adding extra dimensions to the data.

Dimensionality reduction is a technique in the data preprocessing step of KDD, which reduces the dataset's size by removing the attributes that are irrelevant to the particular task of data mining. Feature extraction and feature selection are two broad categorizations of techniques that fall under the umbrella of dimensionality reduction. The dimensionality reduction challenges inherent to the handled data are enumerated below.

1. **Multidimensional Mapping:** A dimensionality reduction technique is required to map the high dimensional data to a low dimensional space.

2. **Estimating Information Loss (Gain):** A good dimensionality reduction technique can be identified by the amount of information it retains in the reduced dataset.
3. **"Small n Large P Problem":** The imbalance of many genes relative to fewer samples creates a high likelihood of finding "false positives" due to chance – both in finding differentially expressed genes, and in building predictive models.
4. **Unbalanced Datasets:** This challenge is relevant to classification problems that arise when the data is constricted by classes that do not have equal representations. Unbalanced representation causes misrepresentation of classes, and learning tends to be biased.
5. **Validation:** We need robust methods to validate the models and assess their accuracy and likelihood.

This work is aimed toward the creation of better dimensionality techniques that can be extracted from both the sequential and structural properties of proteins, keeping in mind the challenges of handling proteomic data.

1.6 Proteins: Sequence, Structure, and Function

A protein is defined by a chain of amino acids. On average, the length of a protein ranges from 200 to 5000 amino acids. The twenty known amino acids are each represented by a letter. Thus a protein sequence is viewed as a long combination of 20 letters. The protein folding problem has been one of the greatest challenges to researchers in bioinformatics. The problem is that predicting the native three dimensional (3-D) structure of a protein from its sequence can be difficult due to the folding and misrepresentation of the protein.

According to the central dogma of protein folding, the protein sequence (the primary sequence) dictates how the protein folds in three dimensions. It is the protein's specific 3-D structures that enable it to function, by dictating the function of the protein and the way it interacts with other proteins. In this section 1.6, we provide an overview of the structure of a protein as shown in Figure 1.4. Typically the structure of a protein starts with its primary sequence. The resultant degrees of structural conformation are governed by the interaction with the environment and adjacent residues.

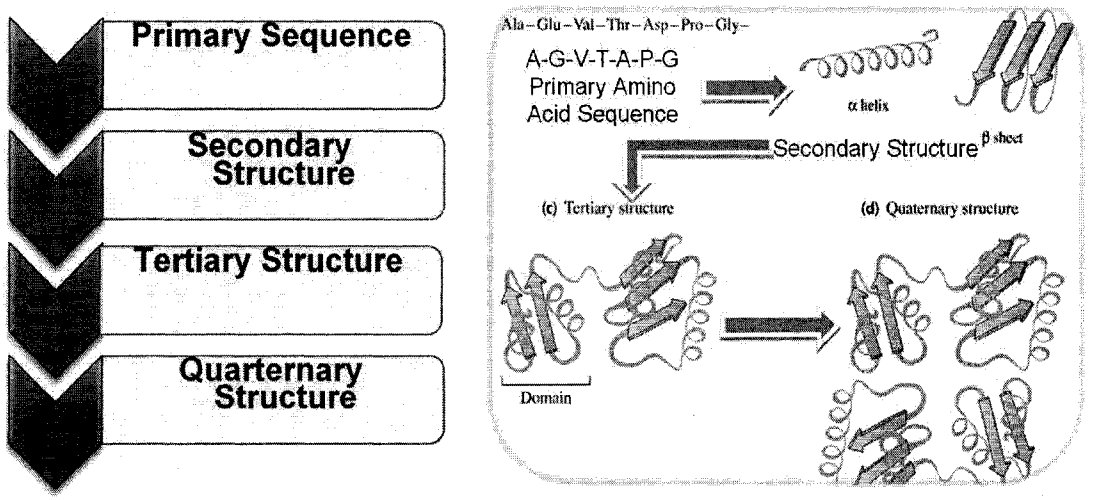


Figure 1.4 Hierarchy of a protein structure.

1.6.1 Primary Sequence

The sequence of amino acids in each protein is determined by the gene that encodes it. The initial process involves the transcription of the gene into a messenger RNA (mRNA), which in turn is translated into a protein by a ribosome. The primary structure is often called the "covalent structure" of a protein, since the covalent bonding mainly defines the primary structure of a protein. However, the other levels of protein structure involve many non-covalent interactions.

1.6.2 Secondary Structure

The secondary structure defines the local spatial arrangement of the main-chain amino acids and is the result of local hydrogen bonds being created along the peptide backbone. The three most common folding patterns found in the secondary structure are the alpha helices, beta sheets, and turns that have well determined and distinct shapes.

1.6.3 Tertiary Structure

The packing of secondary structures results in a third level 3-D tertiary structure. The assembly and interactions of helices and sheets form the tertiary structure. This structural level denotes the "global folding" of a single polypeptide chain. The tertiary structure is determined by a phenomenon called the hydrophobic effect [12]. The folding of the polypeptide chain results in the exposure of the polar residues on the outer surface, while the non-polar amino acids are hidden within the structure. In our work, we are particularly interested in the tertiary structure, especially since the function of a protein depends on it.

1.6.4 Quaternary Structure

As illustrated in Figure 1.4 (on page 21), the quaternary structure involves the stable association of multiple polypeptide chains resulting in an active multi-subunit structure. All proteins do not exhibit this type of structure. Typically, each polypeptide present within a multi-subunit protein folds independently into a stable tertiary structure, and the folded subunits then associate with each other to form the final structure. Not every possible combination of amino acids can form a stable protein sequence that folds and functions properly. Evolution selects only those sequences that can fold into a stable functional structure.

The terms "motifs" and "domains" are commonly used to describe protein structure and function. A motif is a simple combination of a few conservative secondary structure elements. Some, but not all motives are associated with a specific biological function. A domain is the fundamental unit of structure. A domain combines several secondary structure elements and motifs, not necessarily contiguous that are packed in compact globular structures. A domain can fold independently into a stable 3-D structure and has a specific function. A protein may consist of a single domain or of several different domains, or of several copies of the same domain.

1.7 Conclusion

The immediate goal of protein sequence or structure data analysis and visualization is to gain insight into novel protein functions, anonymous protein complexes, and uncharacterized biological processes. With the high-throughput protein data generation projects, only a small percentage of the data can reach the final protein interaction database due to either unavoidable errors or quality issues. Thus scientists need to assess the biases in each data generation method and develop sophisticated data mining algorithms to make use of all available protein data sources [1]. Any knowledge representation scheme should be expressive enough to capture current knowledge details and flexible enough to keep up with future technological advancements and shifting biological interests. We have adopted these principles in our research.

CHAPTER 2

LAYOUT OF RESEARCH

Researchers have been working on the previously mentioned computational challenges for protein mining for decades. Addressing these challenges, we have just started to realize the factors involved in protein folding and structure determination, the steps of data mining, the process of data selection and transformation, the application of realistic evaluation criteria, and the representation of data. As the size of protein data grows at an exponential rate, the significance of using this extracted knowledge is exemplified for system biology and drug development. The presented body of research is aimed at alleviating these challenges. In this chapter, we present our novel research contribution that investigates various dimensionality integration schemes for the properties of proteins using the data mining framework and put forth our research layout.

2.1 Classifications of Protein

The rapid growth in the number of protein sequences and in 3-D structures has made it practical and advantageous to classify proteins into families and more elaborate hierarchical systems. Proteins are grouped together on the basis of structural similarities in the following classification schemes namely, FSSP (Families of Structurally Similar Proteins), CATH (Class (C), Architecture (A), Topology (T), and Homologous superfamily (H)), and SCOP (Structural Classification of Proteins) databases. SCOP is

based on human expert intervention, the FSSP on automatic methods, and CATH on a mixture of both human intervention and automatic methods. These three databases are described in detail below. Other databases, which we do not mention, collect proteins on the basis of sequence similarities to one another, e.g. PROSITE, SBASE, PFAM, BLOCKS, PRINTS, and PRODOM. Several collections contain information about proteins and their structural similarities.

2.1.1 The Protein Data Bank

The Protein Data Bank[3] (PDB) [13] is a database of crystallographic protein structures, a repository of the 3-D Cartesian co-ordinate information of atoms in the amino acid molecules of the protein chain, which are either experimentally determined using x-ray, electron or neutron diffraction, or nuclear magnetic resonance, or are computationally determined by homology or comparative modeling. The bank holdings are increasing at a rapid rate and currently include more than 34,000 determined protein structures.

2.1.2 The SCOP Database

The SCOP[4] (Structural Classification of Proteins) [3] is a manually maintained database that provides a detailed and comprehensive description of the evolutionary and structural relationships of all known protein structures. The extent of the evolutionary relationships of proteins is described at the lower two levels of protein clustering, the family and the superfamily. In this case, the geometrical relationships are described at the fold level. The evolutionary classification incorporated by SCOP is produced by human experts, because to date, automatic classification techniques can only measure a few

[3] PDB http://www.rcsb.org/pdb/
[4] SCOP http://scop.berkeley.edu/

evolutionary changes, and cannot provide insight into the full extent of these changes. This problem makes such techniques less accurate and less efficient.

The fundamental unit of SCOP is the protein domain. A domain is defined as an evolutionary unit observed in nature, either in isolation or in more than one context in multi-domain proteins. The protein domains are classified hierarchically into families, super families, folds, and classes. The major classes are all α, all β, α+β, α/β, and miscellaneous small proteins,' which often have little secondary structure. The July 2005 SCOP release contained 25,973 PDB entries, in 70,859 domains.

2.1.3 The FSSP Database

FSSP[5] [14] is known as fold classification based on the structure-structure alignment of proteins and families of structurally similar proteins. FSSP is based on a fully automated structure comparison algorithm, DALI[6] [15] that calculates a structural similarity measure between pairs of protein chain structures taken from the PDB. This measure is represented in terms of z-score values. First, FSSP chooses a subset of representative protein structures from the PDB and employs the DALI algorithm for the z-scores for all pairs of selected representatives. Next, the z-scores between each representative and the corresponding PDB structures are calculated. For each query structure there is a subset of structural neighbors from the set of representatives and a list of sequence homologs from the PDB. The database entry for this protein structure contains structure-structure alignments with its neighbors along with the list of sequence homologs. Alignments are based purely on the 3-D co-ordinates of the proteins and are derived by the comparison algorithm DALI. A fold tree is generated by applying an

[5] FSSP http://www.sander.ebi.ac.uk/dali/fssp/[6] DALI www.ebi.ac.uk/dali/
[6] DALI www.ebi.ac.uk/dali/

average-linkage hierarchical clustering algorithm to this all-against-all z-score matrix. FSSP is a fully automated structural comparison scheme. Hence, frequent updates of new proteins to the database by the DALI search engine are feasible. FSSP was recently extended by a new database, called DALI, which consists of all-against-all z-scores between chains and domains of a larger representative protein set. This set is built so that no two protein chains exhibit more than 90% sequence similarity.

2.1.4 The CATH Classification

CATH7 [2] is a hierarchical classification of protein domain structures, which groups proteins at four major levels, class (C), architecture (A), topology (T), and homologous superfamily (H). A consensus approach is used to assign domains to proteins using various algorithms. The hierarchal class level describes secondary structures found in the domain and is created automatically. There are four class types: mainly-α, mainly-β, α-β, and proteins with few secondary structures. The topology level clusters together all similar structures with similar sequential connectivity between their secondary structure elements. The homologous superfamily, which is the fourth-level family in the hierarchy, contains structures that exhibit high structural and functional similarity. The similarities among these structures are calculated by a measure called SSAP at both the topology and homologous superfamily levels. The CATH database is connected to the dictionary of homologous superfamilies (DHS) database [37], which permits further analysis of structural and functional features of evolutionary related proteins.

---

[7] CATH http://www.cathdb.info/latest/index.html

2.2 Motivation and Contribution

In Section 2.1, we see the different structural classification schemes available and the features on which they are based. In the following discussion, we enumerate our objectives and contributions. Protein structure and residue conservation can provide information about protein function and protein functional context not apparent from protein sequence analysis. Thus, by studying protein structures, we can understand the functional roles of previously uncharacterized proteins in different environmental conditions.

It has long been recognized that the regular, organized structure of a protein embedded in a non-isotropic environment will be reflected in the sequence of chemical properties of the residues in the protein. The physico-chemical properties of the less conserved residues still encode the information necessary for folding. Hydrophobicity is to a high degree conserved in structurally equivalent positions among evolutionary related proteins, even when the individual amino acid residues are different [6]. Several qualitative, quantitative, and algorithmic techniques have been introduced to model and detect the periodic variation in chemical properties along the protein sequence that are characteristic of secondary structural features [7]. Hydrophobicity and hydrophilicity are incontrovertibly physico-chemical properties in characterizing protein structures. Hydrophobicity scales [8] are intended to be representative of natural phenomenon and to be the predictable result of differences in the inter-molecular forces between water and the amino acid and of those between the amino acid and some other medium. Because of these measurements, hydrophobicity allows a better understanding of how amino acids interact within proteins and provides a way to predict structural properties [9] and [10]. It

is also known that most hydrophobic sequences in a protein are found in the interior of the native structure, and the most hydrophilic sequences are found on the exterior [9]. The structure of a protein can be associated with its hydropathy, and its synthesis can consequently be employed as a viable descriptor for structural classification and prediction. However, to allow such exploitation of the predictive power of hydrophobicity, the most accurate evaluations and representations of the hydrophobicity and hydrophilicity of amino acids should be formulated [9].

In this research, we present novel methods for the encapsulation of different hydrophobicity scales into a coherent feature expression, which is then employed for classification. The feature vector is further refined in our experimentation to include other stereo-chemical properties so that we can study the effects and contributions of those properties to the structural state. We ultimately aim to classify proteins to their respective secondary structural classes using sequence based properties (physico-chemical properties). With the majority of the algorithms claiming appreciable degrees of accuracies of classification with high sequence similarity while failing to reproduce the same with proteins of low sequence similarity, we have an impetus to develop algorithms that encapsulate the following objectives.

2.2.1 Objectives

Multiple scales are available for the measurement of the hydropathic character of a protein. Each scale depicts a different aspect of the intermolecular forces involved and the properties of the proteins itself; there are 40 such scales [8]. Though there are conflicting rudiments among these scales, embedded associations exist ([7] provides an excellent discussion on 37 of the published scales and the subjective correlations between

them), and the correlations between the scales can be mined for classification. However, there is no known work that combines these properties in a coherent fashion for a synthesized feature set.

Figure 2.1 provides a tree representation of the existing amino acid indices (scales), where each node in the tree corresponds to a scale. These nodes are categorized based on the properties they represent [11].

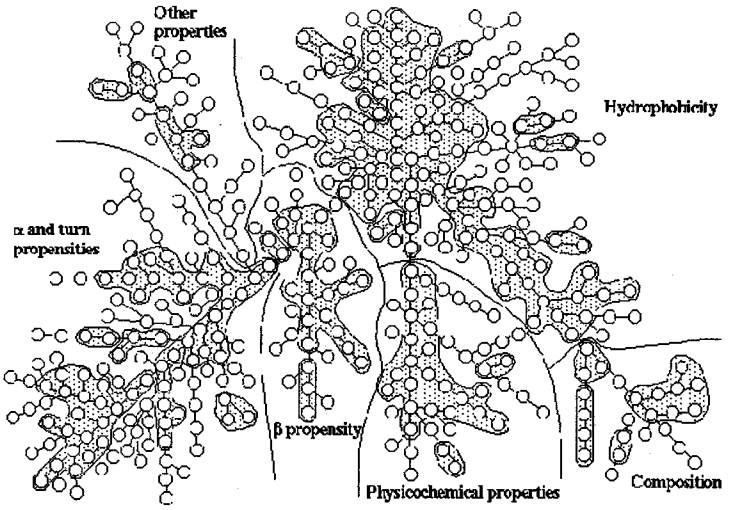


Figure 2.1 Available amino acid indices [11].

Our objectives are enumerated below and described through section 2.2.1:

1. To test the efficacy of the physico-chemical properties of proteins as effective structural descriptors,
2. To extract features by merging of physico-chemical properties,
3. To enhance the detection of structurally conserved regions among homologous proteins, and
4. To annotate the functionality of proteins using the conserved regions of proteins.

To address the fold classification problem we look to answer the following key questions. Due to the numerous scales of hydrophobicity available, can a coherent measure of similar features between different hydrophobicity scales be discovered? Can these measures be exploited for efficient and accurate structural classification of these proteins? Can a mechanism be developed to discover the candidate pairs of such scales for coherence measurement? We aim to develop a computing schema which will both discover an equal-sized feature vector for proteins of unequal sizes involved in the study of fold classification and significantly reduce the dimensionality of the search space. We will evaluate the efficacy of the feature space by the use of different supervised classification algorithms.

We approach the problem in a unique way for the following reasons. Hydrophobicity is a key element contributing to the folding state of the protein; we believe that its scales need to be better signified in constructing a stereo chemical property-based feature vector. Different scales of hydrophobicity represent unique protein behavior and should be constructively aggregated for superior feature representation. The presence of an unbalanced number of proteins in different fold classes and the unequal length of these proteins is not an exception, but a norm, and a consistent cardinality of feature vector for such proteins needs to be discovered. This discovery will enable uniformity in feature treatment by the classification schema. The technique should allow for the merging of other stereo-chemical properties to hydrophobicity for performance enhancement.

Consequently, we will define a distinctive data mining profile generation schema for proteins to enhance the detection of structurally conserved regions among

homologous proteins. As mentioned previously, the expressions of the hydrophobic effect are palpable in many facades of protein sequence-structure-function dependencies. These effects include the stabilization of the folded conformation of globular proteins in solutions, the subsistence of amphipathic structures in peptides or of membrane proteins at lipid boundaries, and protein-protein interactions associated with protein subunit assembly, protein-receptor binding, and other intermolecular bio-recognition processes [12].

Our objective is to identify conserved hydrophobic residues among structurally related proteins, using hydrophobicity scales for classification. By doing so, we reduce our feature space and show that the reported conserved hydrophobic residues are sufficient to differentiate between native and non-native proteins at both the class and fold levels of the structural classification of proteins (SCOP) hierarchy[8]. We focus on five well-known scales of hydrophobicity: the Kyte and Doolittle scale, the Hopp Woods scale, the Janin scale, the Rose et al. scale, and the Eisenberg et al. scale [6]. Employing the principles of graph theory and incorporating the metric of mutual information to identify compact structural units, we aim to extract frequently occurring patterns using a discriminative weighing function.

For the functional annotation of proteins using the conserved regions of proteins, the contribution of different protein regions towards the bio-chemical function is determined by the interactions formed with substrates, cofactors, and other residues. Traditional sequence-based techniques of homology transfer are sensitive and unreliable, forcing researchers to venture into structure alignment and structure pattern matching

[8] http://scop.mrc-lmb.cam.ac.uk/scop/

techniques. Though more effective, the dependence of traditional sequence-based techniques on 3-D coordinate information makes them computationally expensive on larger datasets. Our objective is to create a unique representation scheme known as "Protein Maps" for a given protein, so that we can capture structural makers across the different scales, which are functionally significant. Using spectral base analysis, we aim to report regions of functional significance in the protein, through the protein map. We further wish to extend our validations and to develop a framework that can be extended to the entire Protein Data Bank (PDB) [13].

2.3 Research Layout

This dissertation is divided into three major research contributions. These contributions are detailed in Chapters 3, 4, and 5. With the numerous physico-chemical properties for a given protein, we provide, in Chapter 3, an in-depth look at different hydrophobicity scales and the vital role they play in protein folding. A methodology for coherent feature extraction based on protein sequence information from selected hydrophobicity scales is provided in the chapter. The detailed experimentation discussed in this work demonstrates results with enhanced specificity and sensitivity of protein structural classification using new feature sets, and the results are compared to previous results in this area.

The insights obtained from Chapter 3 provide us the impetus to develop an algorithm that could integrate multiple physico-chemical properties at one time. Hence, in Chapter 4, we report a unique representation scheme known as the "protein maps," aimed at capturing structural markers across a myriad of physico-chemical properties, for a given protein. Conserved protein sequence residues help determine the bio-chemical

function, which is obtained by interactions formed with substrates, cofactors, and other residues [14]. Traditional sequence-based techniques of homology transfer are sensitive and unreliable, forcing researchers to venture into structure alignment and structure pattern matching techniques. Though more effective, their dependence on 3-D coordinate information makes them computationally expensive to apply to larger datasets. Thus, we hypothesize that correlated mutations of physico-chemical interactions between residues reveal residue conservation patterns that are unique to homologous proteins. Integration is traditionally inhibited by a two physico-chemical properties at one time limit. In our study, we use wavelet-based analysis which reports regions of functional significance in the protein. We have validated our study and reported its significance.

Finally, in Chapter 5, we experiment with more accurate ways to identify protein cores. The interactions among residue clusters serve as potential nucleation sites in the folding process. Evidence postulates that residue interactions are governed by the hydrophobic propensities that the residues possess [15]. An array of hydrophobicity scales have been developed to determine the hydrophobic propensities of residues under different environmental conditions. Thus, in Chapter 5, we propose a graph theory-based data mining framework to extract and isolate protein structural features that sustain invariance in evolutionary related proteins. This isolation has been done through the integrated analysis of five well-known hydrophobicity scales over the 3-D structure of proteins. We conjecture that proteins of the same homology contain conserved hydrophobic residues and exhibit analogous residue interaction patterns in the folded state. The results shown in Chapter 5 demonstrate that discriminatory residue interaction

patterns shared among proteins of the same family can be employed for both the structural and the functional annotation of proteins.

In our results for the methods proposed in Chapters 3, 4, and 5, we obtained an average accuracy of 90% in protein classification with a significantly small feature vector compared to previous results.

2.4 Datasets Used

Our dataset consists of proteins initially used in the studies conducted by [16], [17], and [18]. The original dataset consists of independent training and testing sets proteins. The training set, extracted from the PDB-select, consists of 408 proteins distributed across 25 fold classes. The testing set, also extracted from the PDB-select, consists of 174 randomly chosen proteins, resulting in a dataset of 582 proteins from 25 fold classes and 5 structural classes of variable sizes. For training, we have adopted the PDB dataset to directly compare our results with previous work in the area. The proteins used in the dataset have been randomly selected from the SCOP 1.61[9] and ASTRAL 1.61[10] databases with a sequence similarity of less than 40%. To reduce the selection bias, we use 10-fold validation of the split between training, test, and averaged results. Figure 2.2 provides a graphical representation of the dataset with the classes, the sub-classes, and the respective percentages of proteins used for training and testing.

[9] http://scop.mrc-lmb.cam.ac.uk/scop/
[10] http://astral.berkeley.edu/

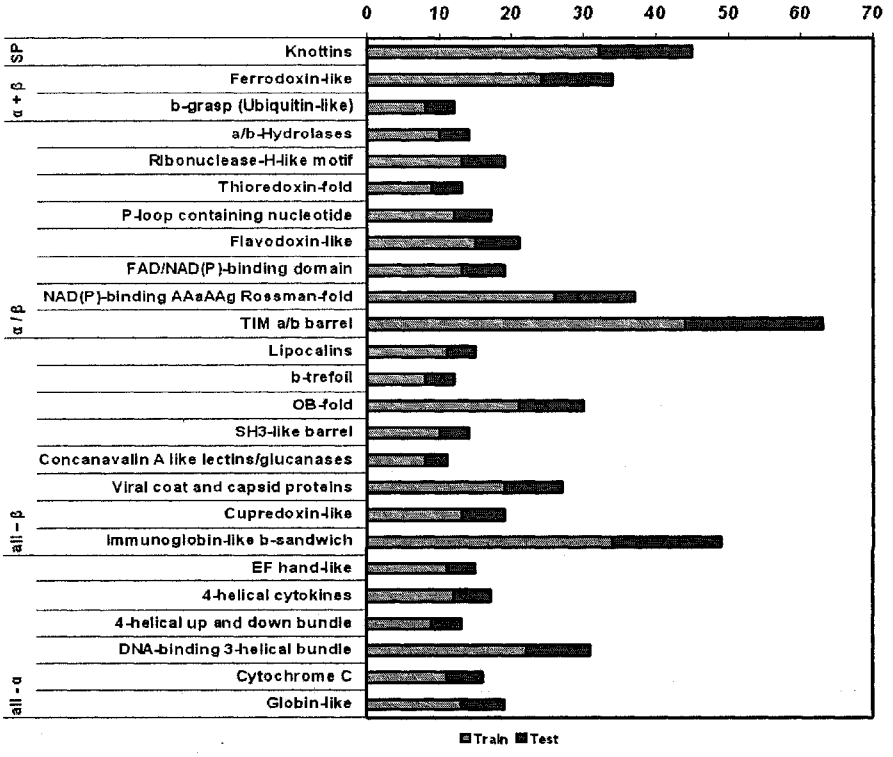


Figure 2.2 Dataset used with classes.

To test the feature vector on a dataset containing two classes (binary classification), we choose two well-known datasets. The first dataset C1 obtained from [19] is unbalanced, consisting of distinctly related proteins from the all-α class- nuclear receptor ligand-binding domain proteins (NB, 16 proteins of typical length ranging between 210 to 260 residues each) against the prokaryotic serine proteases family (PSP, 10 proteins each of length averaging between 190 to 250 residues long) from the all-β classes of proteins. The second dataset, C2, is balanced, consisting of proteins from the eukaryotic serine proteases family (ESP, 19 proteins of length between 200 to 260 residues on average) and from the PSP family, belonging to the same class of all-β proteins. Both datasets (C1 and C2) contain proteins filtered under 60% pair-wise

sequence similarity to remove highly homologous proteins, with a resolution of <=3 and an R factor of <= 1.0. The datasets can be obtained from the "culled PDB list[11]."

To test the performance of the feature vector in a multi-class classification, we choose a dataset consisting of 106 proteins, from three structural classes: all-β, α/β, and α+β of the ASTRAL SCOP 1.71 database with less than 40% pair-wise identity. We consider two important fold classes of all-β proteins. The first fold class consists of 38 proteins of the immunoglobulin-like beta sandwich class of proteins (IgFF). Each protein is 260 to 300 residues long. These proteins exhibit heterogeneity of tissue and species distribution/ functional implications. The domains of these proteins are more conserved than their sequences. The second fold class of the all-β family consists of 35 trypsin-like serine proteases proteins. The trypsin-like serine proteases fold (TSP) has smaller than average surface areas, smaller radii of gyration, and higher Cα atom densities (approximately 238 residues in length on an average). These findings imply that proteases are, as a group, more tightly packed than other proteins, as also evidenced in [14]. There are also notable differences in secondary structure content between the folds of these proteins.

Next, we introduce the third random class of proteins for classification, taking into account the local bias caused by the binary class dataset. This third class consists of proteins chosen at random from an unrelated structural class of proteins. In order to reduce the effect of this class on classification results, we ensure that no structural uniformity exists among these proteins. This lack of uniformity results in a class of 33 proteins, each an average of 160 residues long, belonging to both the α/β and α+β

[11] http://dunbrack.fccc.edu/Guoli/pisces_download.php

structural classes. All the proteins of the dataset satisfy the criteria of < 40% of identity.

2.5 Conclusion

The chapters in this dissertation are a compilation of three published contributions and one contribution which is currently under review:

1. Sumeet Dua, Pradeep Chowriappa, Ramakrishnan Rajagopalan (2006) Computational Prediction of Protein Structure Using Self-Similarity Based Classification, *International Symposium on Computational Biology and Bioinformatics*
2. Sumeet Dua, Pradeep Chowriappa, Ramakrishnan Rajagopalan (2007) Spectral Coherence Feature Extraction from Stereochemical Scales for Protein Classification, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Under Review
3. Pradeep Chowriappa, Sumeet Dua, Jinko Kanno, Hilary Thompson (2008) Protein Structure Classification Based on Conserved Hydrophobic Residues, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, In Press
4. Sumeet Dua, Pradeep Chowriappa (2008) Protein Maps: Physico-chemical Properties Integration for Functional Annotation of Proteins, *7th Asia-Pacific Bioinformatics Conference*, The Asia Pacific Bioinformatics Conference (APBC), Submitted.

Chapters 3, 4, and 5 are based on the publications listed above (listed in order of appearance in this dissertation). Publications 3 and 4 refer to the same issue. Each chapter is divided into four sections, the introduction, which is not explicitly enumerated and related literature, results, and discussion, which are explicitly enumerated.

CHAPTER 3

DISCOVERY OF COHERENCE BETWEEN HYDROPHOBICITY SCALES

It has long been recognized that the regular, organized structure of a protein embedded in a non-isotropic environment is reflected in the sequence of chemical properties in protein residues. Several qualitative, quantitative, and algorithmic techniques have been introduced to model and detect the periodic variation in chemical properties along the protein sequence that are characteristic of secondary structural features [7]. Hydrophobicity and hydrophilicity are incontrovertibly such physico-chemical properties in characterizing protein structures. Hydrophobicity scales [8] are intended represent natural phenomenon, the predictable result of differences in the inter-molecular forces between water and amino acid, and the predictable result of differences in the intermolecular forces between the amino acid and some other medium. Hydrophobicity both allows us to better understand how amino acids interact within proteins and provides us a way to predict structural properties [9] and [10]. Most hydrophobic sequences in a protein are found in the interior of the native structure, and the most hydrophilic sequences are found on the exterior [9]. The structure of a protein can be associated with the hydropathy, and its synthesis can consequently be employed as a viable descriptor for structural classification and prediction. However, to allow such exploitation of the predictive power of hydrophobicity, the most accurate evaluations and

representations of the hydrophobicity and hydrophilicity of amino acids should be formulated [9].

In this research, we present a novel method for the encapsulation of different hydrophobicity scales into a coherent feature expression, which is then employed for classification. The feature vector is further refined in our experimentation to include other stereo-chemical properties which will help us to study the effects and contributions of those properties to the structural state. In the past, researchers have relied on physico-chemical properties to extract relevant structural information given the sequence information [12], [13], [14], and [17]. These properties, namely amino acid composition, predicted secondary structure, hydrophobicity, normalized van der waals volume, polarity, and polarizability [20], were extracted using three global descriptors. However, the commonality that the majority of the related literature [16], [17], [18] share is the prioritization of the machine learning process of classification. Suitable signature profiles of various proteins belonging to different fold classes are usually constructed based on the selected properties. For example, Dubchak et al. [20] have calculated descriptor parameters such as composition, transition, and distribution, laying the foundational work for [12], [13], and [14] to try new classifiers on the dataset.

We believe that these previous works suffer from three key limitations:

1. Lacking discussion as to why these few (six) specific physico-chemical properties were chosen,
2. Lacking implicit mechanism to infer similarities between scales, or to employ those to diminish redundancy in feature representation and increase precision in classification, and

3. Lacking quality measurement for descriptor accuracies and reproducibility.

In this chapter, we address the fold classification problem and attempt to answer the following key questions:

1. Can a coherent measure of similar features between different hydrophobicity scales be discovered?
2. Can these features be exploited for efficient and accurate structural classification of these proteins?

And lastly,

3. Can a mechanism be developed to discover the candidate pairs of such scales for coherence measurement?

In pursuit of these aims, we develop a computing schema to discover an equal-sized feature vector for proteins of unequal sizes involved in the study. In doing so, we significantly reduce the dimensionality of the search space. The efficacy of the feature space is evaluated by the use of different supervised classification algorithms, and detailed experimental results are presented and discussed.

Multiple scales are available for the measurement of the hydropathic character of a protein. Each scale depicts different aspects of the intermolecular forces involved, along with the properties of the proteins. We have examined thirty-seven such scales [8]. Though there are conflicting rudiments between these scales, embedded associations do exist ([7] provides an excellent discussion on 37 of the published scales and the subjective correlations among them), and the correlations among the scales can be mined for classification. However, there is no known work that combines these properties in a coherent fashion for a synthesized feature set.

In our work, correlations among hydrophobicity scales are interpreted as magnitude squared coherences in frequency domain. Frequency domain measures tend to be more encompassing and provide a more complete description of all common oscillatory inputs. This procedure facilitates analysis of the distribution of coherence across multiple frequencies and can lead to a better understanding of the nature of the common inputs involved. Consequently, magnitude squared coherence yields the enhanced information that both is synergic to and complementary among the scales, and that produces a comprehensive measure of the property.

Some other interesting elements of the problem should also be noted. The presence of an unequal number of proteins in different structural classes (unbalanced data) is standard, not the exception. To avoid over-fitting classifiers to certain classes, we compare the performance of different multi-class classification algorithms. We perform our analysis using the Random Forest classification algorithm and variants of the multi-class Support Vector Machine (SVM) algorithm. Further, we provide an in-depth analysis of class level accuracies in addition to the overall specificity and sensitivity. Evaluation and analysis of the physico-chemical property impact on classifier efficacy are provided and make it possible to examine the effectiveness of classifiers in capturing the structural similarities of proteins.

The rest of the chapter is organized as follows. Section 3.2 presents related literature in this area. Section 3.3 describes the training and testing dataset used in our study. Section 3.4 describes the proposed methodology, including feature vector estimation, classification, and scale choice. We present our results in Section 3.5 and conclude with a discussion and our conclusions in Sections 3.6 and 3.7.

3.1 Related Literature

The exponential growth of proteomic data has fuelled an overarching need for machine-learning algorithms that use protein sequence property information for classification to known fold or structural classes. The ability to determine the structure of a protein without relying on sequence similarity is an important impetus for bioinformatics researchers and has recently generated a great deal of scientific interest.

Researchers often rely on physico-chemical properties to extract relevant structural sequence information. Dubchak et al. [20] investigated a machine learning approach to process six physico-chemical properties for structural prediction and yielded significant results. These properties, amino acid composition, predicted secondary structure, hydrophobicity, normalized van der waals volume, polarity, and polarizability, were extracted using three global descriptors. The descriptor composition was used to describe the global composition of a given amino acid's properties in a protein. In the remaining properties, parameter transition was used to compute the frequencies with the property changes along the length of the protein, and the descriptor distribution was used to describe the distribution pattern along the sequence [20]. This work paved the way for researchers to examine better classification models for vast and constantly evolving data [16], [17], and [18]. However, the new models still relied on the same or similar datasets and features extracted from the older data [20]. The work pursued by Tan et al. [17], proposed an ensemble machine learning method aimed at improving the coverage of classifiers under the multi-class imbalanced datasets by integrating knowledge from different base classifiers and utilizing the feature space described by [20]. They applied frequency-based discretization, and concatenation of the six features to introduce a

Bayesian classification schema, and their contributions mainly addressed the imbalanced nature of data.

The work of Venkatarajan et al. [6] is aimed at identifying qualitative descriptors, which use multidimensional scaling, a classification approach that reconstructs synthesized qualitative descriptors based on the geometrical configuration of a large point set into lower dimensions. Five synthesized descriptors based on 237 physico-chemical properties for all 20 amino acids were reported.

In our research, an adroit utilization of physico-chemical properties can be attributed to the design of coherence-based feature profiles. The coherence-based profiles then overcome inherent problems of large dimensionality and unequal cardinality of search space in such domains. Results demonstrate that these descriptors effectively capture the structural behavior of proteins based on classification accuracies. Table 3.1 shows the individual contribution of each property as per [16], compared to our approach for using coherence among hydrophobicity scales as a feature descriptor.

Table 3.1 Prediction accuracy for different parameters [20].

Parameter	**SVM Ind-test**
Composition	32.7%
Secondary Structure	29.5%
Hydrophobicity	23.5%
Volume	21.8%
Polarity	20.9%
Polarizability	20.2%
Proposed Coherence Features of Hydrophobicity	***62.64%***

Note that our proposed feature space for hydrophobicity alone can achieve up to two-and-a-half times better classification accuracy than the individual properties reported in previous work. As shown later in our experimental results, we have also boosted accuracy by appending other physico-chemical properties.

We approach the problem in a unique way with the following four factors for our motivation:

1. With hydrophobicity being a key contributing element to the folded state of the protein; we believe that its scales need to be better signified in constructing a physico-chemical property-based feature vector;
2. Different scales of hydrophobicity represent the unique behavior of proteins and should be constructively aggregated for superior feature representation;
3. The presence of an unbalanced number of proteins in different fold classes and the unequal length of these proteins is not an exception, but a norm, and a consistent cardinality of a feature vector for such proteins should be discovered to enable uniformity in feature treatment by the classification schema; and
4. The technique should allow for the merging of other physico-chemical properties to hydrophobicity for performance enhancement. Consequently, we define a distinctive data mining profile generation schema for proteins. Our choice of scales is based on existing correlations between scales, and we use coherence between the selected scales to encapsulate structural discriminators that can be used for the classification of proteins into their respective structural classes.

3.2 Dataset

Our dataset consists of publicly available proteins[12] initially used in the study conducted by [16], [17], and [18]. The original dataset consists of independent sets of training and testing proteins. The training set, consisting of 408 proteins distributed across 25 fold classes, was extracted from PDB-select. These proteins were randomly selected from SCOP 1.61[13] and ASTRAL 1.61[14] databases with a sequence similarity of less than 40%. The testing set consisted of 174 randomly chosen proteins, resulting in a dataset of 582 proteins from 25 different fold classes and five structural classes of variable sizes. To reduce the selection bias, we used ten-fold validation of the split between training and test and averaged results.

[12] http://www.nersc.gov/~cding/protein/
[13] http://scop.mrc-lmb.cam.ac.uk/scop/
[14] http://astral.berkeley.edu/

3.3 Methodology

Our method consists of three main components: feature extraction, supervised classification, and schema for the hydrophobicity scales as a candidate for coherence based analysis. The overall methodology is presented in Figure 3.1.

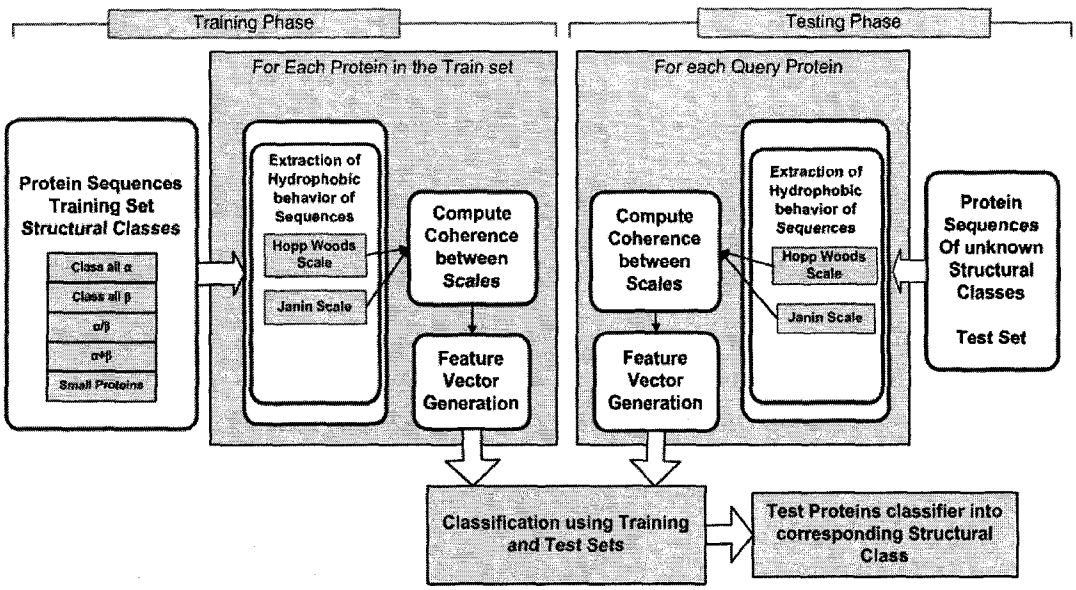


Figure 3.1 Coherence-based feature extraction and classification.

3.3.1 Feature Extraction

Efficient and accurate classifier design depends on choosing discriminatory features intuitively derived from data. Statistical properties such as mean, variance, covariance, and correlation are used as potential descriptors to discriminate between classes of data and have been effective with small datasets. However, intrinsic inconsistencies such as redundancies and outliers can compromise the effectiveness of these properties to capture discriminatory patterns, especially for data mining applications.

We use a method that imbibes the intrinsic statistical properties, such as mean and standard deviation, along with mathematical principles to provide the necessary levels of abstraction to deal with multi-feature datasets. In this pursuit, we incorporate spectral coherence as a feature vector design tool for multi-class classification.

Let $C_1, C_2, ..., C_n$ be different classes. Each class C_i, where $i \in 1..n$ in itself has variable number m samples of proteins such that

$$P_{i1}, P_{i2}, .., P_{im} \in C_i, \text{ where } m \geq 1 \quad \forall C_i . \tag{3.1}$$

As in Eq. 3.1, for every protein (P_i), a sequence of amino acids can be expressed as a sequence of C_α atoms (backbone) of the individual amino acid.

Let each hydrophobicity scale be V_a where $a \in \{1,2\}$ and $|V_a|$=20 referring to hydrophobic propensities corresponding to the 20 known amino acids. Thus as in Eq. 3.2, let G_a be the corresponding representation of the protein sequence P_i given the hydrophobicity scale V_a, such that

$$G_a(j) = \{P_i, V_a : P_i(j) \rightarrow V_a(P_i(j))\}, \text{ where } j = 1,..,N-1 . \tag{3.2}$$

Spectral coherence between scales is computed using the following steps

1. Segmentation of hydrophobic representation of sequence

   Each G_a is subject to segmentation of length L, with overlap of length D.

   Let $G_{a1}(j)$ *where* $j=0,..,\ L-1$, be the first segment,

   then $G_{a1}(j)=G_a(j)$ *where* $j=0,\ ..,\ L-1$.

   Similarly, $G_{a2}(j)=G_a(j+D)$ *where* $j=0,..,L-1$, and finally $G_{ak}(j)=G_a(j+(k-1)D)$ *where* $j=0,\ ..,\ L-1$.

2. Let us suppose we have k segments such that

   $G_{a1}(j),.., G_{ak}(j)$ are the resultant segments that cover the entire sequence; i.e.

   $(k- 1) D+ L= N,$ the length of the protein sequence.

3. The computation of the power spectra PG_a for individual G_a given its segment from step i is computed using the Fast Fourier Transformation, given the window *size* (ω).

4. Given the PG_a*'s* of protein P_i, the cross spectra is computed as described by Welch in [21].

5. Magnitude squared coherence between G_1 and G_2 is calculated as follows

$$MSC(P_i) = \left| \frac{(P_{G_1G_2})^2}{(P_{G_1G_1} * P_{G_2G_2})} \right|.$$

By definition coherence is the vector property that quantifies the degree of interference. By interference, we imply that if at least two vector-like entities are combined, and if the relative phase between them is positive, then they can add constructively or subtract destructively. The Welch's averaged, modified periodogram method for computing coherence uses the Fast Fourier Transformation (FFT) to estimate the power spectra of a vector [21]. This computation involves segmentation with overlaps of the feature vector into windows of fixed length, taking the modified periodograms of these segments and finally finding the average of these modified periodograms. The coherence-based features offer better discriminatory properties, as shown in Figure 3.2, than selecting the top 20% of FFT coefficients that contain the most energy, and using them as features for classification [22].

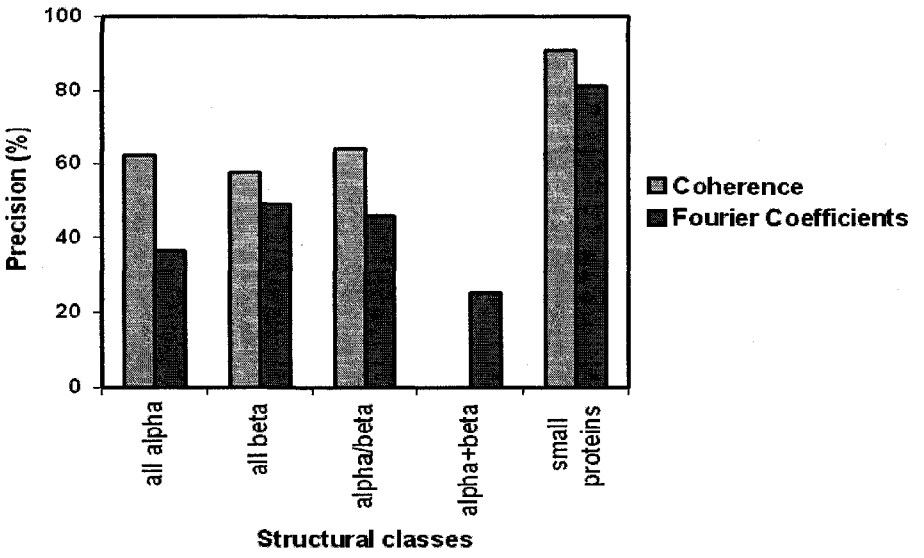


Figure 3.2 Feature vector performance.

The hydrophobic representation of protein P_i belonging to C_n, with scale V_1, results in $G_1^{Cn} = \{ax_1, ax_2, ..., ax_N\}$, when subjected to the computation of coherence and is segmented into segments of length L with overlapping regions of length D. We then determine the power spectral estimate of the vector G_1^{Cn} as the average of the periodograms, which is computed using FFT for each window of G_1^{Cn}. We denote these coefficients as PG_1. Similarly, we determine the spectral estimate for $G_2^{Cn} = \{bx_1, bx_2, ..., bx_N\}$ and denote these coefficients as PG_2. Both PG_1 and PG_2 are power spectral representations of the same protein PG_1 in two different scales V_a where $a \in \{1,2\}$. This representation can be extended in a straightforward manner to the estimation of the cross spectrum. Modified cross periodograms are computed for each pair of segments, and the average of these modified cross periodograms constitutes P_{12} (C_n). The mean squared coherence $MSC(Pi^{Cn})$ is the estimate of the two vectors G_1^{Cn} and G_2^{Cn} of protein P_i

belonging to class C_n using Welch's averaged, modified periodogram method, which is given by the relation in Step 5 above.

For training purposes, we represent each protein in the training set as its computed MSC of the scales as Feature Vector FV, represented as in Eq. 3.3

$$FV(P_i) = MSC(P_i) \quad \forall P_i \text{ , where } i \in \{1,..,m\} \text{ and } P_i \in C_n. \tag{3.3}$$

The resultant is a single vector of attributes for each protein P_i belonging to class C_n. In our experiments, coherences computed from the hydrophobicity scales of Hopp Woods and Janin give us better results than other combinations. A window size of seven is chosen because of its superior performance in our method. The coherence between these two vectors will result in a vector of consistent length 32, which is used as the feature descriptor for each protein. These vectors are then subjected to multi-class classification in the subsequent steps.

3.3.2 Classification

To classify the feature vectors, we employ Random Forest Classification and multi-class Support Vector Machines.

3.3.2.1 Random Forest

We use Random Forest [23] to determine the similarity of proteins within a family. Random Forest Classification uses a collection of independent decision trees, instead of one tree. Each tree is grown using a subset of the possible attributes. In order to accurately classify the protein, we use each tree as "votes" for one class. We then assign the most popular class to the tree. Interested readers are referred to [23] and [24] for more details on Random Forest Classification.

3.3.2.2 Support Vector Machines (SVM)

We also use SVM with different kernel functions for classification. SVMs view classification as a quadratic optimization problem. This method is chosen because of its superior generalization in high dimensional data and fast convergence in training [25]. In general, SVMs plot the feature vector for each sample in the training set resulting in a high-dimensional feature space. Each vector is labeled with its class identifier referred to as training IDs. A hyperplane drawn between the training IDs maximizes the distance between the different classes. The following kernel functions are explored in our study: linear, polynomial, and radial basis. The shape of the hyperplane is generated by the kernel function, though many experiments select the polynomial kernel as optimal.

We have applied "one-against-one" classification [25] for each of the n classes. In this case, $n(n-1)/2$ classifiers are generated to train the data, where each training vector is compared with two different classes, and the error (between the separating hyperplane margins) is minimized. The classification of the testing data is accomplished by a voting strategy [26] where the winner of each binary comparison registers on a counter. The winners are the classes with the highest counter value after all classes have been compared and the results reported as in Section 3.4.

3.4 Results

The following section 3.4 of this chapter enumerates the results of our experiments. We have divided our results into categories: choice of scales and multi-class classification. Assignment into these two categories is based on the two contributions made. First, we briefly describe the method which involves how two scales of physico-

chemical properties are chosen from existing scales. Second, we describe a set of experiments that are carried out using the coherence between the two chosen scales.

3.4.1 Choice of Scales

More than 37 scales have been used to estimate the hydrophobicity of amino acids. These scales have proven useful in providing insight into the measurement of the hydropathic character of a protein. Each of the 37 scales depicts a different aspect of the intermolecular forces within the protein and the properties of the protein [7]. In our technique, spectral coherence is calculated for a pair of hydrophobic scales, and the choice of scales to be included is contingent upon the relative affinity of these scales to the proteins to the training classes. These vectors are clustered using a hierarchical clustering approach, and we hypothesize that the scales that exhibit low affinity in discovered clusters should be chosen for spectral-coherence analysis. Also, any methodology that is applied to such scales should account for inequality in protein sizes. Proteins are represented in a 3-D domain, such as the one shown in Figure 3.3, where one dimension (x-axis in Figure 3.3) refers to the protein index; a second represents relative amino acid composition (y-axis in Figure 3.3); and the third (z-axis in Figure 3.3) represents the hydrophobicity scale under consideration. Relative amino acid composition of protein *(P)* is computed based on the frequency *(F)* of an individual amino acid *(aa)* in *P*, as in Eq. 3.4

$$aa_i(P) = F(aa_i)/length(P), \text{ where } i = 1,..,20. \tag{3.4}$$

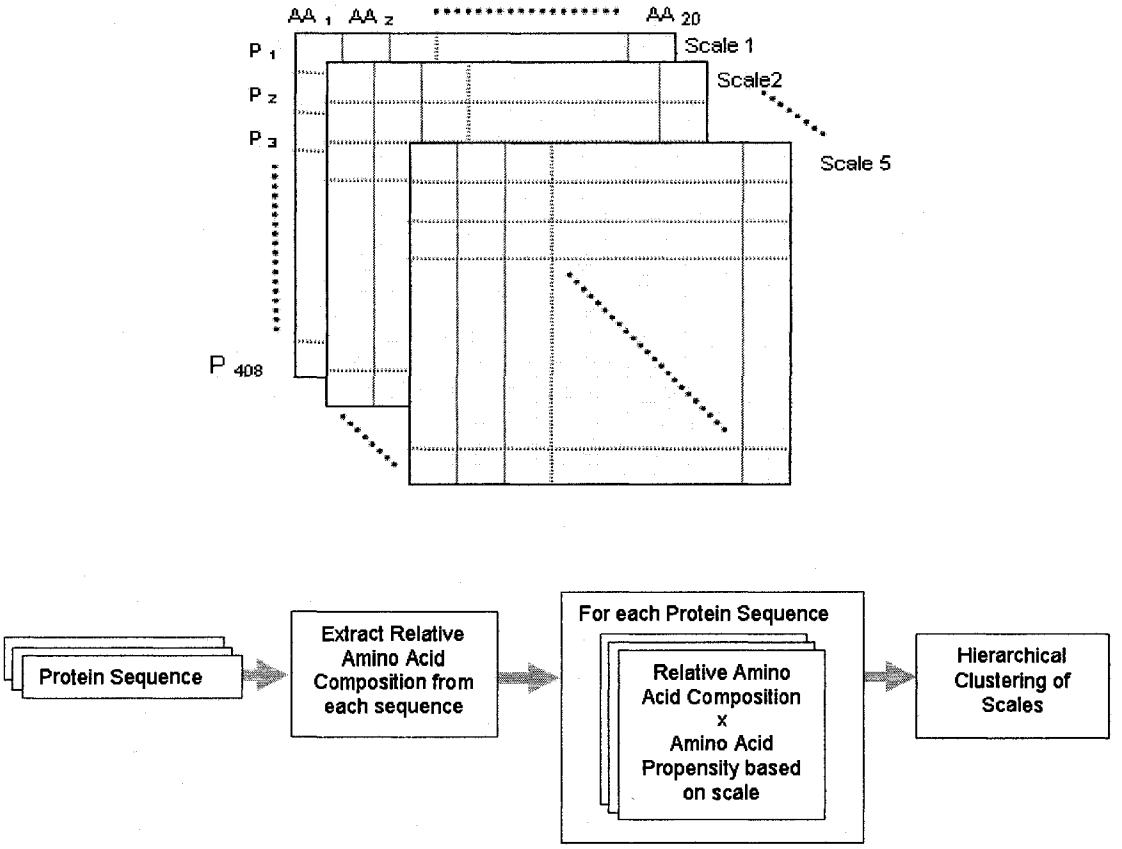


Figure 3.3 Hydrophobicity data of (training) proteins and methodology.

The resultant descriptor of protein P, with respect to an individual amino acid, is defined as a 20 dimensional tuple that contains the frequency of occurrence of each individual amino acid in the protein sequence. Thus P_n is represented as in Eq. 3.5

$$P_n = \{aa(1), aa(2), .., aa(20)\}. \tag{3.5}$$

Thus, the dataset is represented by the corresponding profiles of proteins. For comparison, we normalize each hydrophobicity scale *(S)* by its mean (μ) and standard deviation (σ) by the relation shown in Eq. 3.6

$$S'(i) = (S(i) - \mu)/\sigma . \tag{3.6}$$

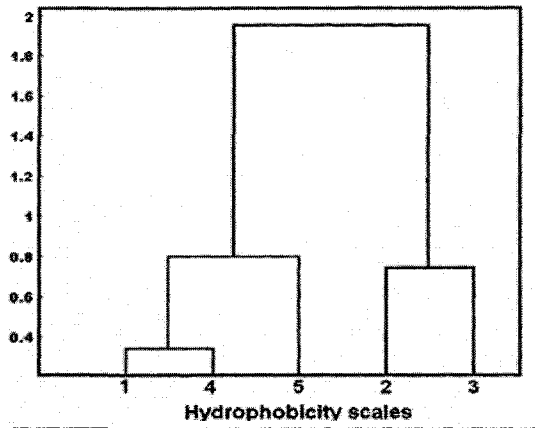
We generate the weighted-relative amino acid composition of each protein. The relative amino acid composition for each amino acid $P_n(aa(i))$ of the protein is multiplied by the weight assigned to it by the corresponding hydrophobicity scale $S(aa(i))$, defined by

$$WP_n = S'(aa(i))xP_n(aa(i)) . \tag{3.7}$$

In our dataset, the resultant is a multi-dimensional problem that involves 408 proteins of the training set, each represented by the corresponding profiles of 20 amino acid compositions for a given scale (Figure 3.3). With 37 known hydrophobicity scales, the clustering of scales takes place in a *20 * 408* dimensional space to choose those scales that exhibit the least degree of correlation. The datasets of the hydrophobicity scales are available on our project website.

We then extract the Eigenvector, which processes the highest Eigen value with respect to weighted amino acid composition. This Eigenvector acts as a weighted representation of amino acids for each scale. We then perform hierarchical clustering of Eigenvectors that represent respective scales to identify those scales that exhibit the least correlation in *20*408* dimensional spaces. Complete linkage distance is used to identify correlations between scales when clustering. As shown in Figure 3.4, scales Hoop Woods, Rose, and Eisenburg cluster together, as do scales Kyte and Doolittle, and Janin. These clusters exhibit maximum inter-cluster correlation and minimum intra-cluster correlation defined by the Euclidean distance in the complete linkage distance calculation. To narrow the choice in deciding which two scales generate the best accuracy, all possible combinations of the scales between the two clusters are carried out

using the suggested framework. The clustering results and the classification accuracies of the hydrophobicity pairs are reported in Figure 3.4.



Legend	
ID	Scales
1	Hoop Woods
2	Janin
3	Kyte & Doolittle
4	Rose
5	Eisenberg

Combination of Scales	SVM (C-SVC) (%)	Random Forest (%)	SMO (%)
(1) and (2)	**58.62**	**62.64**	**55.17**
(1) and (3)	56.90	54.02	52.87
(1) and (4)	50.00	55.74	48.27
(1) and (5)	52.87	51.15	46.55
(2) and (3)	52.87	55.17	48.85
(2) and (4)	52.30	54.59	48.27
(2) and (5)	54.60	51.15	45.40
(3) and (4)	50.57	58.05	52.30
(3) and (5)	50.00	54.02	49.43
(4) and (5)	52.3	56.87	48.85

Figure 3.4 Hierarchical clustering of scales of hydrophobicity.

3.4.2 Multi-Class Classification

One of the objectives of our study is to demonstrate that the hydrophobic behavior of proteins of the same family is similar and that the coherent pattern of hydrophobicity is useful to classify proteins into five structural classes. The five classes we use are the all α, the all β, the α/β, the α+β, and the small proteins. Additionally, we measure the performance of two algorithms: multi-class SVM (C-SVC) and Random Forest. Weka

tools[15] are used to implement [12] Random Forest analysis. For SVM analyses, the LIBSVM package[16] developed by Chang et al. [15], including parameterized kernel functions formulations and multi-class classification, as shown in Table 3.2 is used.

Table 3.2 Different parameter settings used in C-SVC classification using SVM.

	Experiment 1 - Feature Vector Size: 64	Experiment 2 - Feature Vector Size: 32
Deg of Kernel	1	5
Gamma	0.35	0.3
Penalty cost (complexity)	2	4

The meta-classifier is used for multi-class datasets with two class classifiers. This classifier is also capable of applying error correcting output codes for increased accuracy. We later complete a one-against-all transformation to convert the single multi-class problem into several two class problems. We set the number of trees to be generated to 20 for both the 64 and 32 feature vector length based experiments. Due to time constraints, the results for SVM, shown in Table 3.3, are reported only for radial basis function with different experimental conditions.

[15] http://www.cs.waikato.ac.nz/ml/weka/
[16] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Table 3.3 Comparison of results obtained using two feature vector lengths compared with previous results [17].

		Feature Vector size: 64		Feature Vector size: 32	
Classes	Tan et al [7] (%)	Random Forest (%)	SVM (C-SVC) (%)	Random Forest (%)	SVM (C-SVC) (%)
All Alpha	76.40	54.50	42.40	48.5	54.55
All Beta	86.50	67.90	71.72	64.2	64.15
Alpha/ Beta	53.10	73.80	70.49	70.5	62.30
Alpha+Beta	55.00	0.00	0.00	0.00	7.14
Small Proteins	100.00	76.90	53.85	69.2	76.92
Overall Accuracy	**74.2**	**62.64**	**58.62**	**58.62**	**58.05**

To determine the accuracy of our methods, we perform our experiments in two phases. We begin our first experiment by determining the contribution of hydrophobicity with coherence computed at a frequency of 128, resulting in a feature vector of 64 points. In our second experiment, we reduce our frequency to 64, and the resultant feature vector consists of 32 features. We perform classification using the two classifiers and then perform a one-to-one comparison on the individual structural class accuracies of [17], the last known work in the area. Tan et al. [17] have reported better accuracy than [20] and [16]. These results are presented in Table 3.3 above. We achieve classification accuracy comparable to [21] when using hydrophobicity alone in our feature extraction approach. In [21], six different properties have been used.

3.4.3 Appending Other Properties to the Feature Vector

In order to improve classification accuracy, other physico-chemical property parameters are appended to the 32-size feature vector of our experiment, as shown in Table 3.4. These parameters (adopted from [16] for comparison purposes) are predicted secondary structure and percentage amino acid composition. These extended parameters

result in a total feature vector size of 72. We choose hydrophobicity with a vector of size 32 so that our parameters relatively match in size, allowing us to maintain a balanced cardinality of different properties within the feature vector. This feature vector is then subject to classification using multi-class SVM (C-SVC) and multi-class Random Forest with the same experimental setting. These results (for Experiment-3) are presented in Table 3.5.

Table 3.4 Property features added to boost the classification accuracy.

Parameter	*Vector Size*
Hydrophobicity	32
Secondary Structure	20
Amino Acid Composition	20

Table 3.5 Comparison of results obtained from Experiment-3 with the results of [17].

		Experiment 3 Feature Vector size 72	
Classes	**Tan et al [7] (%)**	**Random Forest (%)**	**SVM (C-SVC) (%)**
All Alpha	76.40	78.80	75.76
All Beta	86.50	90.6	83.02
Alpha/ Beta	53.10	93.4	88.52
Alpha+Beta	55.00	14.30	28.51
Small Proteins	100.00	92.30	84.62
Overall Accuracy	**74.2%**	**83.33%**	**79.31%**

The Random Forest Classifier outperforms SVM (C-SVC). To demonstrate the actual number of true alarms, we have provided the confusion matrices in Figure 3.5. Various conclusions and discussions are addressed in the following sections 3.4.4 and 3.5.

Confusion Matrix (Random Forest)					
	a	b	c	d	g
a	**26**	2	5	0	0
b	1	**48**	4	0	0
c	1	3	**57**	0	0
d	0	6	6	**2**	0
g	0	1	0	0	**12**

(a)

Confusion Matrix (C-SVC)					
	a	b	c	d	g
a	**25**	4	3	1	0
b	0	**44**	7	2	0
c	1	6	**54**	0	0
d	1	3	6	**4**	0
g	0	2	0	0	**11**

(b)

Figure 3.5 Confusion matrices.

3.4.4 Testing the Efficiency of Feature Vector

We also test the efficacy of our feature vector by subjecting it to various classification algorithms. As shown in Figure 3.6, the precision obtained by the Random Forest and SMO classifiers overshadow the performance of the Linear SVM and the RBF-SVM algorithm, in the class of small proteins. However, in the all-α, all-β, and α/β protein classes, all the algorithms perform relatively equally. In the α+β protein class, the Random Forest and SMO classifiers obtain negligible degrees of accuracy, and the Linear SVM and RBF SVM classifiers perform at 100% accuracy. This observation indicates the α+β class reduces the overall accuracy obtained from the Random Forest and SMO classifiers and behaves like an outlier class.

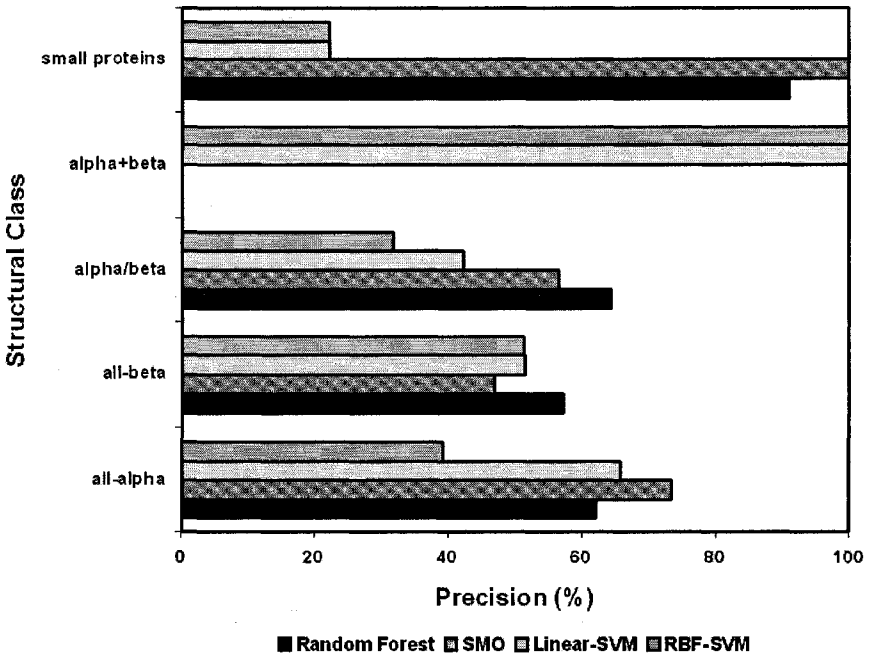


Figure 3.6 Overall classification accuracy achieved by different classifiers.

3.5 Discussion

While a great deal of evidence suggests that hydrophobicity is a key physico-chemical property that is related to the structural behavior of proteins, the quantification of this fact has been attempted by few researchers. According to Ding et al. [16], hydrophobicity contributes an average of 23 percent toward the effective classification of proteins [16]. We have shown that contribution can be far larger (62%) with the applicability of an improved feature vector and an adaptation of more than one scale. The study reinforces our theory that hydrophobicity is a key contributor to protein classification into known families. We further elucidate the class-level accuracies of our classification to better understand the results and interpretations. For the clarity of space,

we have only included representative results from our extensive experimentation, and readers are referred to our project website for more details[17].

To provide a quick visual indication of whether a detailed analysis of mining results will uncover any nuggets, we produce the lift charts [27]. In the process of uncovering the effect of the hydrophobicity descriptors in the feature vector, we analyze lift charts. Figure 3.7 demonstrates the results obtained using the C-SVC classifier and using hydrophobicity alone for class all β.

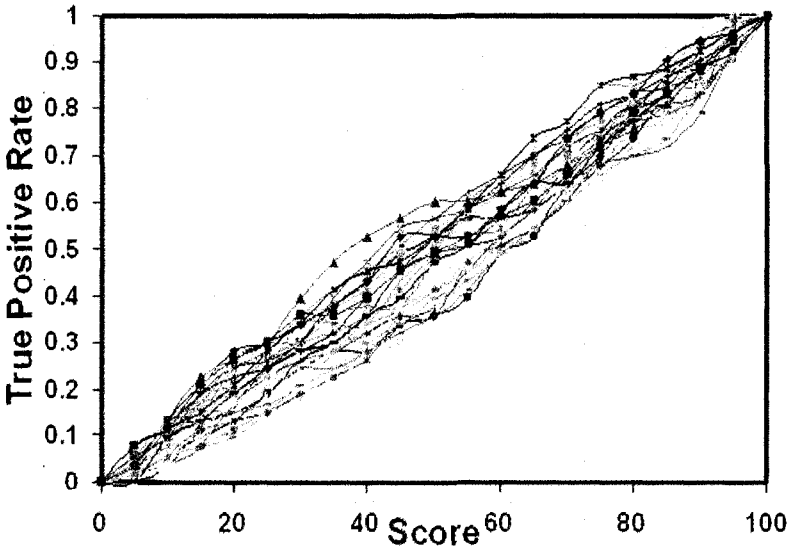


Figure 3.7 Lift Curves for class all β.

The curves display true positive rates for the parameters in the feature vector, individually plotted for each of the five classes. The plots of classes all α, all β, and α/β demonstrate that their feature vectors are good discriminators in the identification of proteins that belong to their respective classes. Correspondingly, plots of class α+β and small proteins have curves that are separated, implying that the feature vector is not

[17] www.latech.edu/~pch008/spectral_prot07

comparatively effective in protein classification for those classes. We also obtain the lift curves (results omitted here for clarity) for the feature vectors used in [16] and [17]. Those curves using our coherent scales are more compact and less scattered than those using other scales.

Our second objective is to evaluate the performance of classifiers that best suit the nature of our study. We choose Neural Networks and SVMs (and its modifications) from existing research such as [16]. However, these techniques are not effective in handling multi-class classification, especially in an imbalanced dataset. For analysis, we plot the ROC curves for individual classes (from results obtained in experiment-3). Only the curves for Random Forest Classifiers are shown here. The slopes of the curves arch toward the top left corner of the plot for the all-α, all-β, α/β, and small proteins classes. Several of these plots are presented in Figure 3.8. The curve location indicates that the Random Forest Classifier is effective in classifying proteins into their respective classes with a higher degree of accuracy.

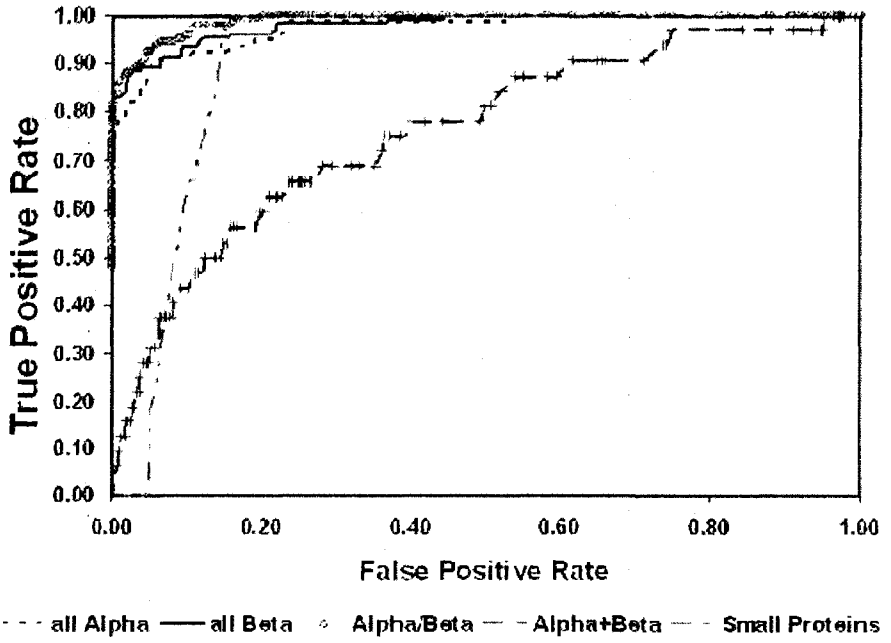


Figure 3.8 ROC plots.

3.5.1 Detailed Lift Curve Analysis

The lift charts provide a quick visual indication of whether a detailed analysis of mining results will uncover any new information [27]. In the process of uncovering the effect the hydrophobicity descriptors have on the feature vector, we perform the analysis of lift charts. In this chapter, for the purpose of comparison, we generate lift curves using the features of coherence along with the standard features of Composition, Secondary Structure, Hydrophobicity, Volume, Polarity and Polarizability as reported by Dubchak et. al. in [16]. Since the original results are based on SVM, we use a common multi-class SVM (C-SVC) of the LIBSVM package [26] to generate results from the extracted features using coherence and the provided features of Dubchak et al. [20] on the test datasets.

The individual curve in the plots represents specific parameters in the feature vector. We create individual plots for each of the five unbalanced structural classes in the dataset. The relationship among the curves is important for analysis. If the lift chart indicates little or no difference among classes, then we can assume that the parameters are good discriminators. However, if the curves are distributed or scattered, then the parameters in the feature vector are poor discriminators.

As reported earlier in the chapter, the feature vector of coherence of hydrophobicity scales depicts varied degrees of performance towards the different structural classes of proteins. For comparison, we perform the lift plot analysis using the feature vectors of hydrophobicity of [20]. Figure 3.9 contains the plots obtained with respect to each structural class based on the hydrophobicity features extracted from the feature vector of [20], forming a feature vector of size 20.

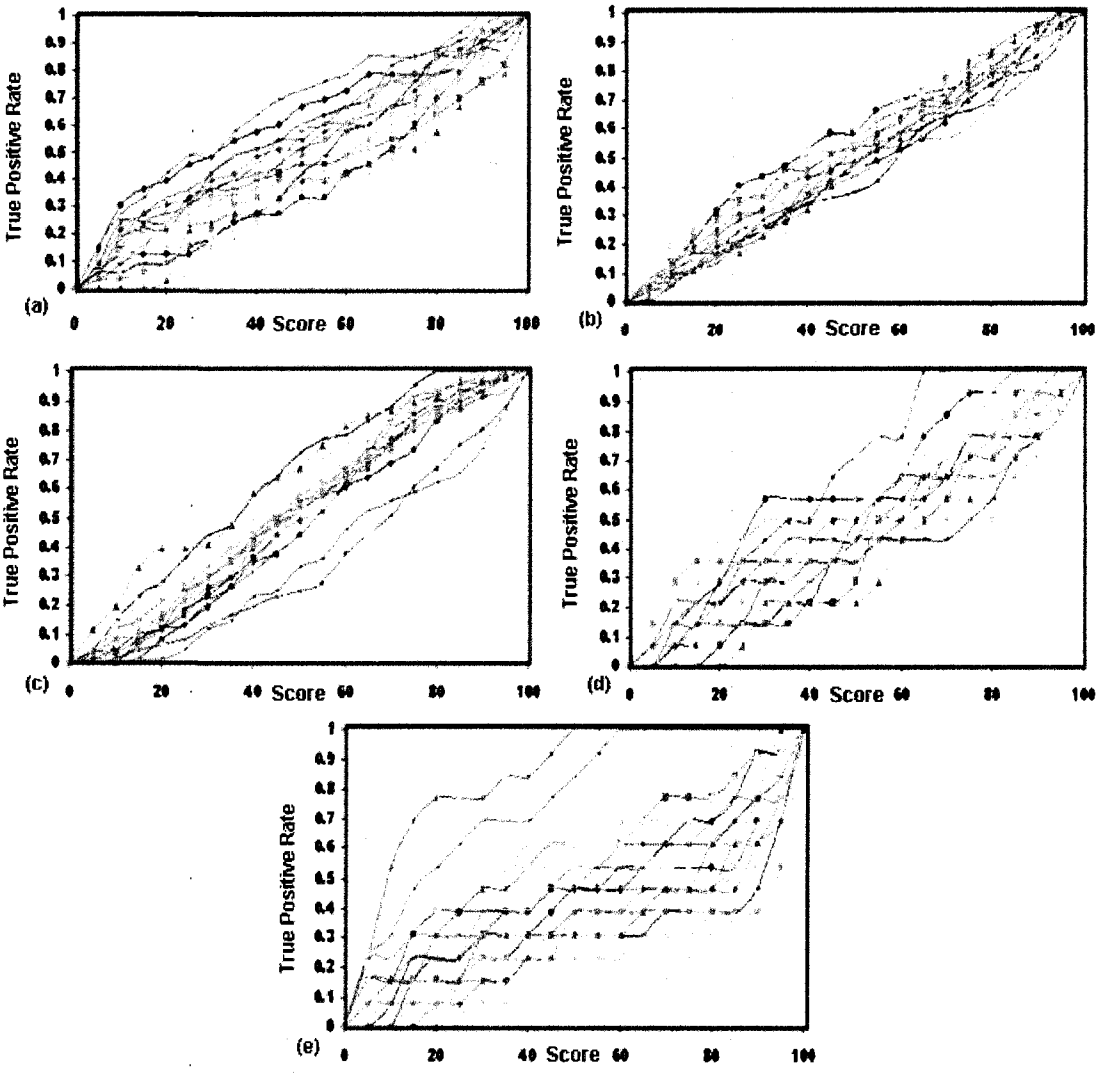


Figure 3.9 Lift Curve analysis of the hydrophobic property attributes.

The curves in Figure 3.9 are more distributed for four of the five structural classes than the curves of Figure 3.10. However, for Figure 3.10, the plots of structural classes α+β and small proteins exhibit a higher degree of scatter, similar to those obtained in Figure 3.9. This scatter reinforces our theory that the weakness of distinctiveness between points is dependant on class representation. However, the lift curves are more scattered than in Figure 3.10, indicating that the feature vector using coherence of hydrophobicity scales is superior in distinguishing proteins of the test set. We can therefore conclude that

the proposed hydrophobicity vector outperforms the vector described in Ding and Dubchak [16].

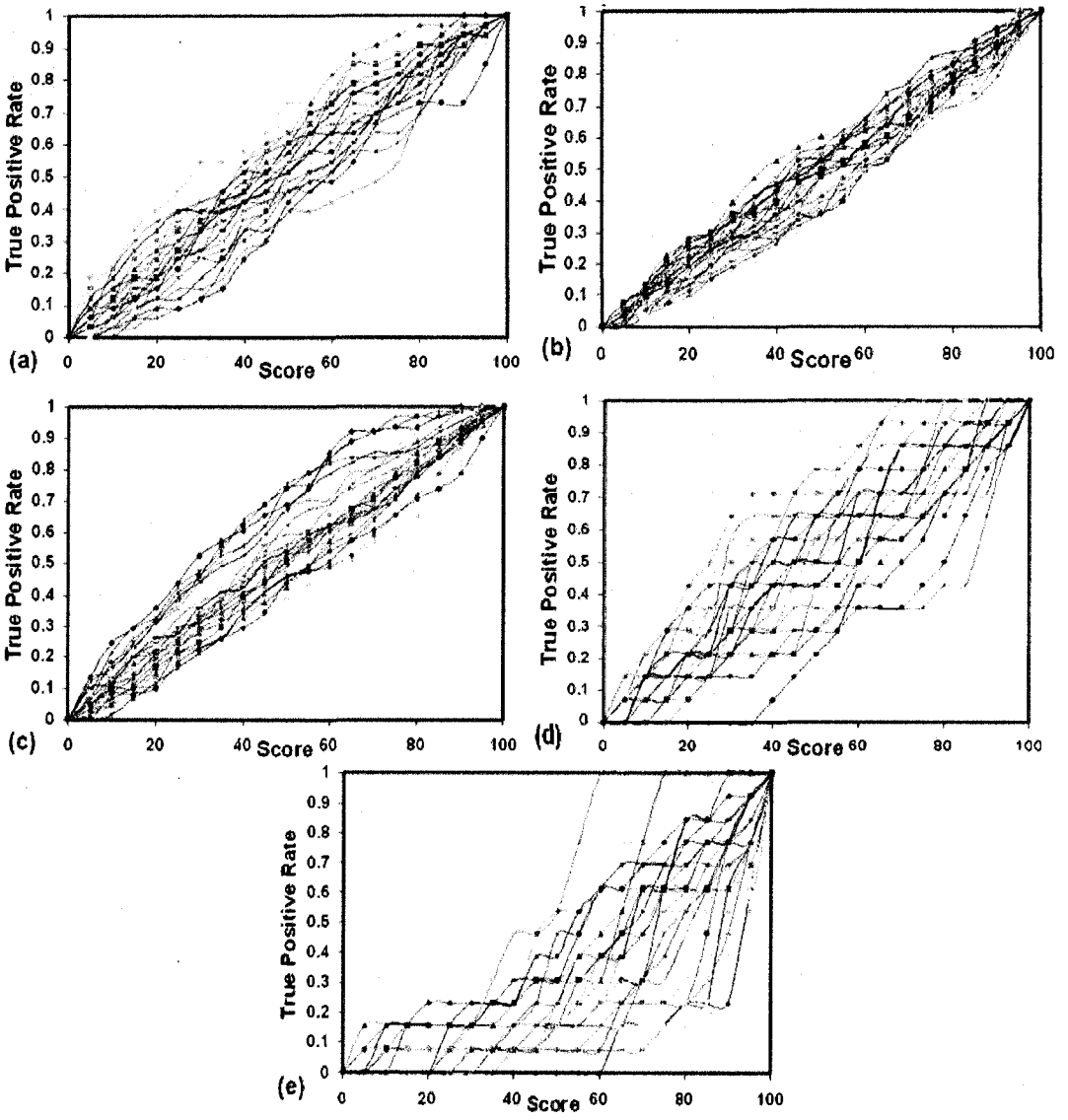


Figure 3.10 Lift Curve analysis of feature vector.

3.5.2 ROC Analysis

In addition to the above, we also want to compare the class level performance of the feature vector using Random Forest with a different classifier – namely the C-SVC of the LIBSVM, as a justification for the choice of classifier used in our study. neural networks and SVM and its modifications are used [12] and [17]. However, these techniques do not effectively handle multi-class classification, especially in an unbalanced dataset. Modifications to SVM address these issues, and new and more efficient algorithms have been developed (see [17]). Each newly proposed algorithm has outperformed the other in accuracy. Along the same lines, we have used the multi-class Random Forest algorithm and a multi-class SVM (C-SVC) algorithm from the LIBSVM package in our study. Using the results obtained from Experiment-3, we plot the ROC curves for individual classes as shown in Figure 3.11.

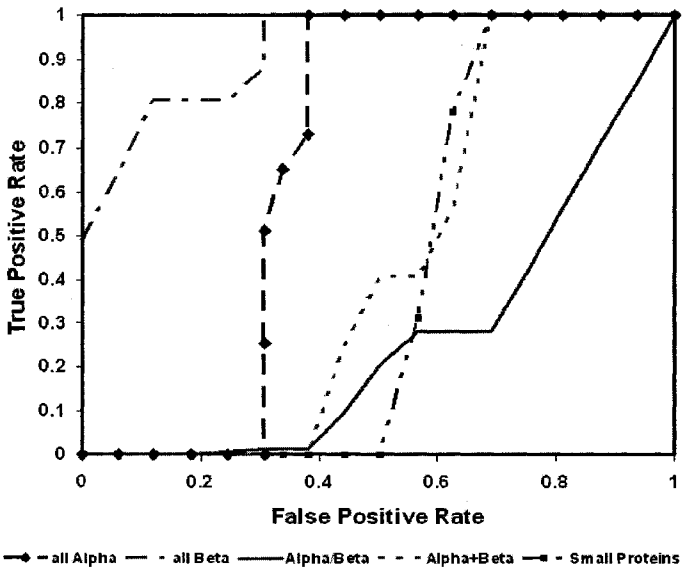


Figure 3.11 ROC plot of the performance of feature vector.

The multi-class SVM (C-SVC) algorithm performs effectively only in the all α and all β structural classes. Its performance in the remaining classes is below par. This indicates that the SVM (C-SVC) classifier may not be built to handle the nature of the dataset, and Random Forest may be more effective in handling datasets of this nature.

3.5.3 Order of Combination of Parameters

In order to boost the classification accuracy, we append the parameters to the existing feature vector of size 32. We observe an improvement in accuracy. However, there seems to be a fluctuating effect on the accuracy based on the order in which parameters are appended to the feature vector. Three parameters contribute to making the feature vector in Experiment-3; Table 3.6 contains the different combinations of these parameters and the corresponding overall accuracies obtained with both classifiers. Figure 3.12 shows significant variations in the overall accuracies of the Random Forest Classifier, based on the order of the parameters, while the C-SVC classifier is unaffected. Despite this difference in performance, both classifiers perform best when coherence of hydrophobicity scale features are placed first and followed by secondary structure and amino acid composition.

Table 3.6 Effect of combination of parameters on overall accuracy.

Order of Combination of parameters	Overall Accuracy	
	Random forest (%)	SVM (CSVC)(%)
Hydrophobicity + Secondary Structure+ Amino Acid Composition	**83.33**	**79.31**
Hydrophobicity + Amino Acid Composition+ Secondary Structure	81.03	78.17
Amino Acid Composition+ Hydrophobicity + Secondary Structure	80.45	78.16
Amino Acid Composition+ Secondary Structure+ Hydrophobicity	81.60	78.74
Secondary Structure+ Amino Acid Composition + Hydrophobicity	80.46	79.31
Secondary Structure+ Hydrophobicity + Amino Acid Composition	81.03	79.31

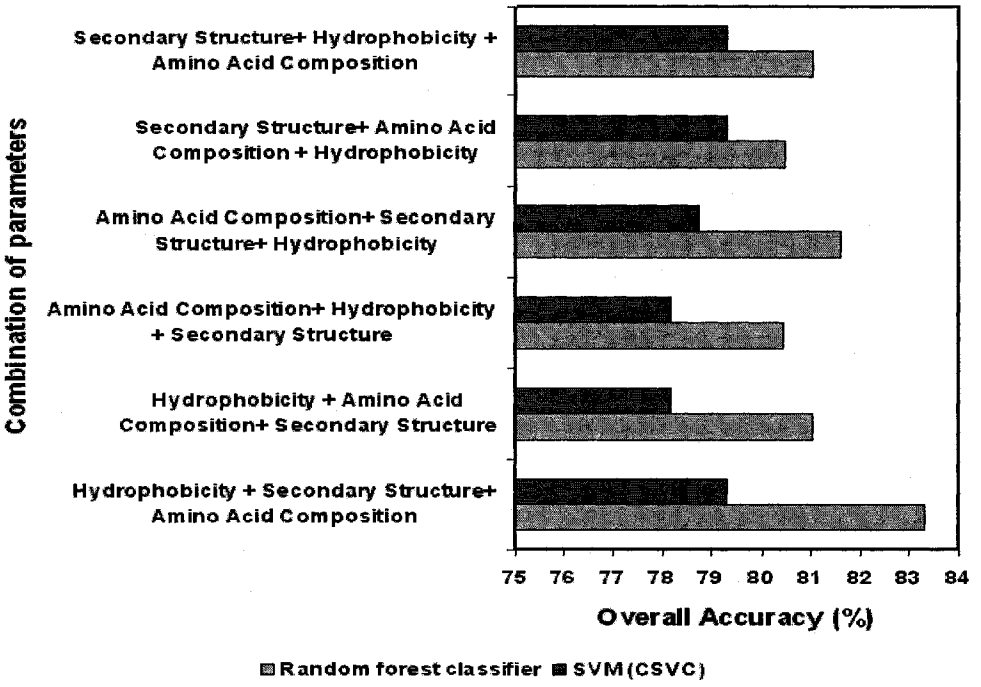


Figure 3.12 Effect of combination of parameters on the overall accuracy.

3.6 Conclusion

The quest to decipher the structure and function of a protein from its amino acid sequence has provided an interesting challenge. Due to the sheer quantity of existing protein data, this challenge naturally presents itself as a complex computational problem requiring the deployment of novel data mining techniques. Existing research in the area uses feature vectors generated from these property values to predict the secondary class by constructing a machine-learning classifier. It has long been recognized that the regular, organized structure of a protein embedded in a non-isotropic environment is reflected in the sequence and hydrophobic physico-chemical properties of the residues in the protein. The usefulness of hydrophobicity can provide a clearer understanding of how amino acids interact within proteins, as well as providing a basis upon which one can predict the structural properties of proteins from sequence information.

In this work, we have discovered points of spectral similarity among the hydrophobicity scales which produce the resultant coherence vector used for classification. These features are subjected to a random tree classifier for multi-class classification. Similar classification is performed using support vector machines, and the performance of each method is compared to the performances of the other methods to evaluate the strength of the proposed feature vectors and algorithms in terms of true-positives and false-negatives for individual structural classes. In another set of experiments, we append the feature vectors for other physico-chemical properties to the computed hydrophobicity features and study the change (boost) in specificity and sensitivity of classification on an incremental basis. We discover that the treatment of hydrophobicity in previous literature has suffered from an inadequate feature

representation of hydrophobicity scales. Discovering and representing novel feature vectors that exploit embedded similarities in these properties can significantly enhance the accuracy of structural prediction.

Although the elucidation of the contribution of individual physico-chemical properties of protein sequences is far from complete, computing methods such as the one proposed can assist in a better understanding of the contributions of physico-chemical sequence properties to the intricate world of protein folding.

CHAPTER 4

PROTEIN MAPS: INTEGRATION OF PHYSICO-CHEMICAL PROPERTIES FOR FUNCTIONAL ANNOTATION OF PROTEINS

In Chapter 3, we attempted to address the fold classification problem and provided insights to the integration of different hydrophobicity scales. Our aim for this chapter is to utilize the integration of physico-chemical properties for the identification of domains. We also propose a characterization scheme to enable the classification of functionally related proteins.

We discussed experiments in which we discovered points of spectral similarity among hydrophobicity scales to produce coherence vectors used for the classification of evolutionary related proteins. In Chapter 4, we concentrate on experimentation with the identification of conserved regions unique to a family of proteins. It has long been recognized that the regular, organized structure of a protein embedded in a non-isotropic environment is reflected in the sequence and hydrophobic or physico-chemical properties of the residues in the protein. The usefulness of hydrophobicity can provide a clear understanding of how amino acids interact within proteins and can provide a basis upon which one can predict the structural properties of proteins from sequence information.

The idea of protein structural domains goes back at least to Wetlaufer [28], who defined a domain as a small number of continuous regions of a protein chain that can be

enclosed in a single compact volume. Additionally, [28] made the first clear distinction between continuous domains, those formed from a single chain segment, and discontinuous domains, those formed from multiple chain segments. The work of Lijjas and Rossman [29], with the creation of contact maps (adjacency matrices), enabled domain assignment. They observed a large number of inter-residue contacts within a domain and relatively few between domains. This observation forms the basis of many modern automated structure-based domain assignment methods.

Since domains are evolutionary conserved units among proteins of the same family [30] and [31], a majority of proteins consist of multiple domains. It is therefore necessary to develop techniques that aid in the delineation of regions over the sequence that belong to a domain and those that do not. This development is vital, as it is believed that domains determine the function and evolutionary relationships of proteins.

In this chapter we investigate the role physico-chemical properties play in domain identification. Studies have shown that changes in protein properties are brought about by the cumulative effects of several small adjustments, many of which are propagated over significant distances in the 3-D structure. Trace evidence of such coordinated mutations brought about by evolution are present in the protein sequence data within members of any protein family [30]. Researchers have historically relied on computational techniques that depend on sequence homology or structural homology, or sometimes both for domain identification.

It is well known that sequence homology techniques are currently unable to keep up with the newly generated protein sequences. Assigning incorrect functions that are linked to the true ones, therefore, requires new automatic strategies addressing domain

identification, that is important for helping to identify and assign specific protein functions [32]. This insensitivity to sequences of low similarity has researchers investigating more reliable techniques. In fact, studies have revealed that residues distant in sequence but near in 3-D space undergo simultaneous compensatory variation to conserve their overall physico-chemical properties [32]. However, these studies have met with only limited success. There are currently no reliable techniques for the identification of conserved residues that affect functionality across homologous proteins. The estimated accuracy of statistical contact predictions has been 15-20% at best [30]. Our impetus being to improve this accuracy, we propose to develop a technique that utilizes physico-chemical properties derived from sequences to aid in functional annotation.

4.1 Related Literature

Thus detecting the domain structure of a protein is a challenging problem. Given the protein sequence, there are no clear signals that indicate when one domain ends and another begins. To quantify the likelihood that a sequence position is either part of a domain or the boundary of a domain, several measures, based on the multiple sequence alignment reflecting the structural properties of proteins, can be informative of the protein domain structure. Previous traditional domain prediction techniques can be roughly placed into the following categories.

4.1.1 Sequence Homology Based Domain Prediction Methods

These methods, which work on the principle of multiple sequence alignment (MSA), are straightforward and widely used. Because they work on the principle of MSA, the sequences are aligned to other sequences that have known domain information,

or seed proteins. The following are examples of such techniques that include ADDA [33], Biozon [34], Dopro [35], Matao [36] and Ginzu [37].

4.1.2 Structure Based Threading Techniques [38]

Structural information can help detect the domain structure of a protein. Domain delineation based on structure is currently best done manually by experts [34]; the SCOP domain classification [39], which is based on extensive expert knowledge, is one such example. These techniques use no form of sequence similarity to determine the domain of the protein non-sequence homology based methods and are useful in the absence of homologous sequences (or seed proteins). In such cases, a target protein may be structurally similar to a protein of known 3-D structure, even if there is no significant sequence similarity. In such a case, domains can be predicted using fold recognition or threading techniques where the target sequence is aligned into a given structure or fold. Here, domains can be predicted using fold recognition or threading techniques, where the target sequence is aligned into a given structure or fold as in Dompred [40], SSEP-domain[18] , DOMPRO[19], and GLOBPLOT [41].

Well-known domain databases for protein domains can again be divided roughly into two categories as described above. The PFAM [42] and SMART[43] databases are useful for several reasons. They are based on multiple sequence alignment, and are considered to be the largest existing databases, and rely on expert knowledge. Additionally, they are driven by methods such as Hidden Markov Models (HMMs) and Artificial Neural Networks (ANNs). Similarly, the PRODOM [44] and TIGRFAMS[20]

[18] http://www.bio.ifi.lmu.de/cafasp/
[19] http://www.ics.uci.edu/baldig/dompro.html
[20] http://www.tigr.org/TIGRFAMs

databases identify domains based on evolutionary relationships using advance machine learning techniques. These databases are considered accurate in their predictions but have been restricted to a few well studied families of proteins.

In Chapter 3, we attempted to address the fold classification problem and provided insights to the integration of different hydrophobicity scales. Our aim in this chapter is to utilize the integration of physico-chemical properties for the identification of domains. We also propose a characterization scheme to enable the classification of functionally related proteins. We hypothesize that evolutionary related proteins exhibit correlated behavior across regions, along their backbones, and over a myriad of interacting physico-chemical property residues in unison, revealing a pattern that is unique to different functional families of proteins. We propose protein maps to help capture the co-evolution of residues through spectral analysis of independent physico-chemical properties.

Chapter 4 is organized as follows. We first describe the physico-chemical properties used in this study. We then briefly describe how we divide a sequence into subsequences and extract the features by transforming the sequence into the frequency domain. We then describe the steps that are taken to create the protein map for a given physico-chemical property. The protein map is subjected to wavelet-based segmentation which clusters regions of the protein map and to identify regions that exhibit similarity over the entire sequence. The most coherent cluster is chosen and reported as a domain for further validations described in Section 4.3.

4.2 Methodology

The methodology as shown in Figure 4.1, involves the creation of a protein map from a given protein using physico-chemical properties as descriptors. To discuss the methodology in detail, we first define a protein '*P*' as a sequence of amino acids of finite length '*n*'. The following are some of the key concepts followed in this work.

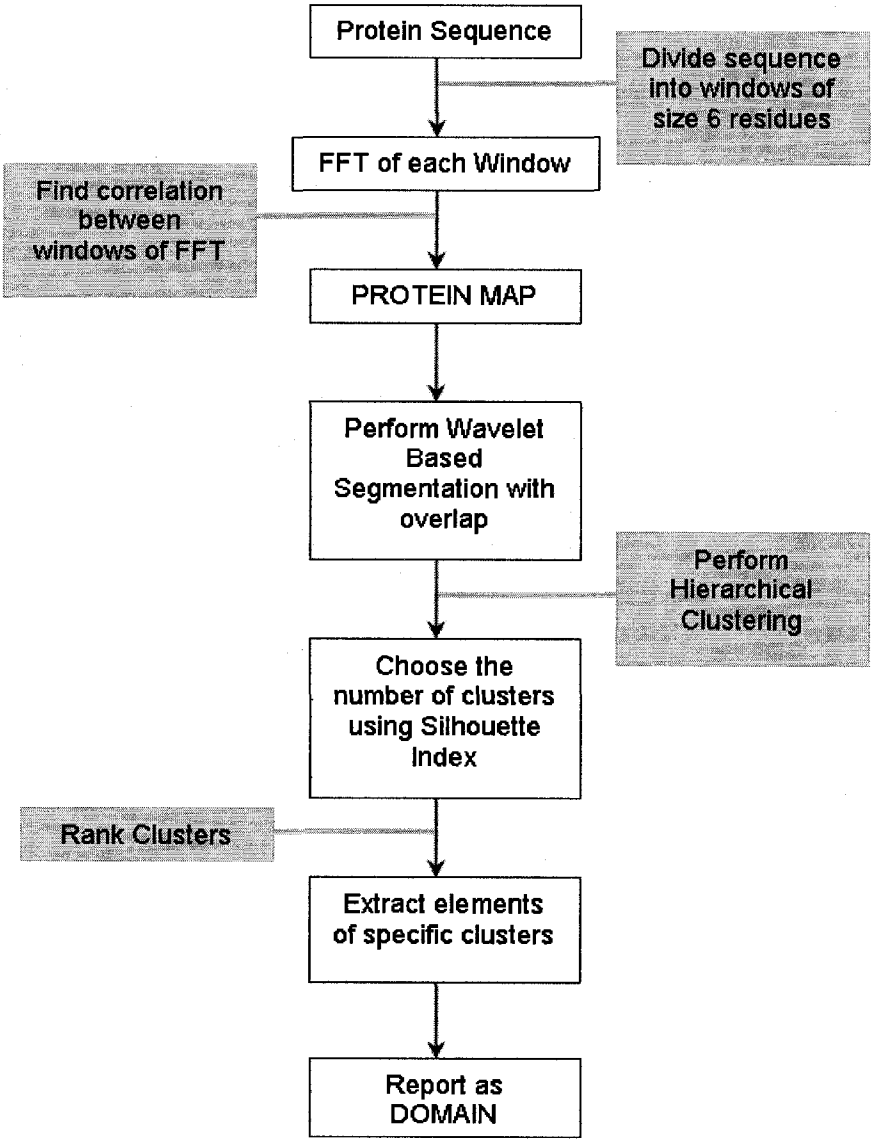


Figure 4.1 Proposed methodology for the discovery of domains.

4.2.1 Amino Acid Descriptors

In Chapter 4, we have used the quantitative descriptors for the 20 amino acids as proposed by Venkatarajan and Braun [6]. Using the method of multi-dimensional scaling, Venkatarajan and Braun summarized information from 237 known physico-chemical properties aimed at providing useful information for the identification of protein homologues on the basis of property-based motifs. They provided five-dimensional numerical descriptors for each amino acid, from the first five Eigenvectors as seen in Table 4.1., referred to as E1 through E5.

Table 4.1 Venkatarajan and Braun components.

Amino Acid	Components				
	E1	E2	E3	E4	E5
A	0.008	0.134	-0.475	-0.039	0.181
R	0.171	-0.361	0.107	-0.258	-0.364
N	0.255	0.038	0.117	0.118	-0.055
D	0.303	-0.057	-0.014	0.225	0.156
C	-0.132	0.174	0.070	0.565	-0.374
Q	0.149	-0.184	-0.030	0.035	-0.112
E	0.221	-0.280	-0.315	0.157	0.303
G	0.218	0.562	-0.024	0.018	0.106
H	0.023	-0.177	0.041	0.280	-0.021
I	-0.353	0.071	-0.088	-0.195	-0.107
L	-0.267	0.018	-0.265	-0.274	0.206
K	0.243	-0.339	-0.044	-0.325	-0.027
M	-0.239	-0.141	-0.155	0.321	0.077
F	-0.329	-0.023	0.072	-0.002	0.208
P	0.173	0.286	0.407	-0.215	0.384
S	0.199	0.238	-0.015	-0.068	-0.196
T	0.068	0.147	-0.015	-0.132	-0.274
W	-0.296	-0.186	0.389	0.083	0.297
Y	-0.141	-0.057	0.425	-0.096	-0.091
V	-0.274	0.136	-0.187	-0.196	-0.299

As per [4], the components (scales) E1 to E3 are useful in describing the hydrophobicity, size, and helical propensity of a protein sequence. E4, on the other hand, is a useful descriptor for partial specific volumes, relative abundance of amino acids, and the number of codons. The β strand forming propensity seems to be the dominant factor for E5. We propose to use these five components for the creation of a protein map for a given protein.

4.2.2 Creation of Protein Maps

The correlated compensation of properties is balanced over the entire sequence, making it vital to capture this characteristic across the entire length of the protein. We thus propose dividing the protein into subsequences, using the concept of sliding window with overlap, defined in Eq. 4.1

$$N = ((\xi - \mu) + s), \tag{4.1}$$

where 'ξ' is the arbitrary length of the protein sequence (the number of residues) and 'μ', the size of the sliding window set at six residues and 's' set at one less than the size of the sliding window. Figure 4.2 provides a pictorial representation of sliding window over the physico-chemical profile of a protein sequence.

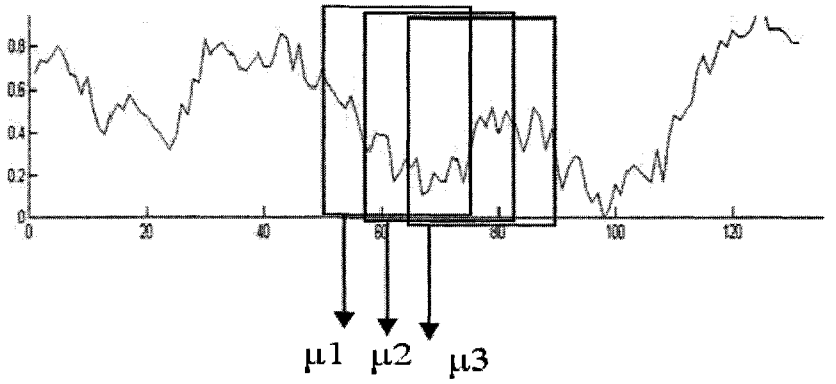


Figure 4.2 Creation of protein maps.

This process creates N subsequences for a given protein sequence. Since not all proteins are of equal length, and since we must keep the length of a window constant for any given protein, the number of windows, *'N'*, varies from protein to protein. Each subsequence (window) is subject to the Fourier transform, defined by the relation shown in Eq. 4.2

$$X(k) = \sum_{j=1}^{N} \mu(j) \omega_N^{(j-1)(k-1)}, \tag{4.2}$$

where $\omega_N = e^{(-2\pi i)/N}$.

Our aim is to capture the localized changes in physico-chemical properties. By extracting the Fourier coefficient of the window, we obtain a profile for each window in the frequency domain that enables us to capture the transient behavior of physico-chemical property over the sequence.

4.2.2.1 Correlated mutations scores

An important source of information about the structural flexibility of a position can be found in the profile of a protein. Traditionally, a count of the number of pair-wise contacts between residues on opposite sides of that position is necessary for each sequence position [45]. Minima in the profile correspond to regions where fewer interactions occur across these sequence positions, implying relatively higher structural flexibility and suggesting a domain boundary. Contacts between residues in a protein are usually predicted based on correlated mutations.

We believe that the correlations that exist between the physico-chemical behaviors of localized regions over the entire sequence of the protein provide a better

understanding of physico-chemical interactions between the residues of a neighborhood and help identify compact structural domains. This correlation lends valuable insight into the structure of bio-chemical property conservation over homologous proteins.

The correlated mutation score between frequency coefficients of two windows k and l is defined as in Eq. 4.3

$$Corr\ (k,l) = \frac{\sum_{i=1}^{n}(k_i - \bar{k})(l_i - \bar{l})}{\sqrt{\sum_{i=1}^{n}(k_i - \bar{k})^2 \sum_{i=1}^{n}(l_i - \bar{l})^2}}, \tag{4.3}$$

where $k,l \in 1..N$ subsequences of a given protein.

Here k_i is the amino acid propensity in position i, of subsequence k of the protein sequence (similarly for the window l). The resultant correlation matrix consists of the correlation coefficients of all possible pair-wise combinations of Fourier coefficients of windows 'μ' for a given protein. The textured representation matrix, capturing the correlated behavior of residues, is known as a 'layer' of the protein map using a given physico-chemical property. The algorithm of this process is described in Figure 4.3. Since we are using the five components or scales as described in Section 4.3.2.1, a myriad of five layers, constitute the resultant protein map of protein 'P'.

Algorithm 1 *Creation of Protein Map*
Input: Protein Sequence PS_i
Output: Protein Map $PM(PS_i)$

1. For the given physico-chemical property, convert PS_i to its corresponding signal
2. Divide PSi (signal) into subsequence of length six using overlapping windows of *step size=1*
    a. Extract the Fourier coefficients of each window using FFT
    b. Select the first *50%* of coefficients of each window
3. Compute the correlated mutation score between every possible combination of subsequences
    a. $Mat\_PS(k,l)=correl\_coeff(window_k(P_i), window_l(Pi))$ *where* $0<k,l<=j$

Figure 4.3 Algorithm1 for the creation of a layer in protein map.

4.2.2.1.1 Wavelet-based segmentation

Once we have created a Protein Map for protein 1AAQ, shown in Figures 4.4 and Figure 4.5, our next objective is to predict those proteins that significantly contribute to domains. We plan to use the existing correlated mutations between localized regions to make this prediction. We propose a novel wavelet-based segmentation approach for the identification of conserved correlated segments for a given protein map. To assess the significance of correlated mutation scores, we subject each layer to z-score normalization, thus the normalized and correlated mutation score r is defined as in Eq. 4.4

$$z - score(r) = (r - \mu)/\sigma, \tag{4.4}$$

where μ and σ correspond to the mean and standard deviation of the correlated mutation sores of a given layer of the protein map.

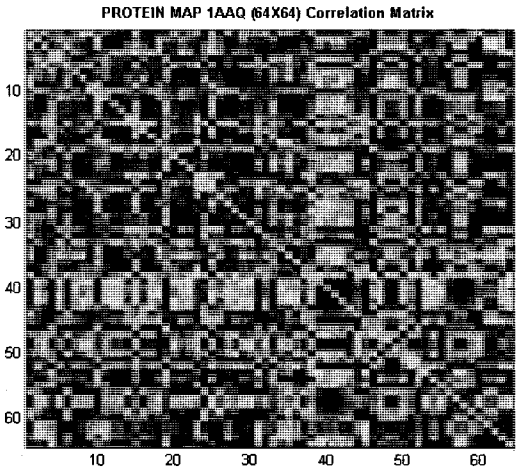


Figure 4.4 Layer of Protein Map for Protein 1AAQ.

Figure 4.5 Structure of protein 1AAQ.

4.2.2.1.2 Wavelet transform

Since each frequency component can be analyzed with a different resolution and scale, wavelet functions are capable of the multi-resolution representation of a signal. This multi-resolution representation allows the wavelet transform to represent discontinuities in the signal by using "short" functions, and, at the same time, emphasizing low frequency components using "wide" functions [46].

The *Continuous WT* decomposes a signal into a set of scaling functions by using a wavelet functions basis, as in Eq. 4.5

$$(W_a f)(b) = \int f(x)\psi^*_{a,b}(x)dx\,. \tag{4.5}$$

With the basis of wavelet functions obtained by scaling and shifting a single mother wavelet function $\psi(x)$, given as follows in Eq. 4.6

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right). \tag{4.6}$$

The general norm states that the mother wavelet should only satisfy the zero-average condition as in Eq. 4.7

$$\int\psi(x)dx = 0\,. \tag{4.7}$$

The Discrete wavelet transform, on the other hand is obtained by taking $a=2^n$ and $b \in Z$.

4.2.2.1.3 Segmentation of Protein Map

The protein map is an aggregate representation of the transient behavior of different physico-chemical properties. It provides a means for conserved residues to analyze a protein under a myriad of properties. We propose a method to identify these regions for a given layer of a protein map where the layer is broken down into segments consisting of correlation coefficients that correspond to specific localized regions over the sequence of the protein. The steps are as follows:

1. **Segmentation**: The layer is segmented into non-overlapping segments of uniform dimensions.

2. **Application of DWT to Individual Segment:** The approximation coefficients are extracted from each segment.
3. **Clustering of Segments:** The approximate coefficients of each segment are hierarchically clustered, keeping the maximum number of clusters extracted at twenty; we call each cluster an *fA* (frequency aggregate) based on the similarity of wavelet coefficients.

4.2.3 Generation of Frequency Aggregates

We adopt a hierarchical clustering-based approach to identify clusters of protein map segments that exhibit similar characteristics. As mentioned, the approximate coefficients of each segment are applied as time-frequency descriptors to group the segments of a layer of the Protein Map. We adopt the 'Euclidean distance' approach to measure the similarity between the approximate coefficients of segments. As seen in Figure 4.6 each frequency aggregate is a collection of segments.

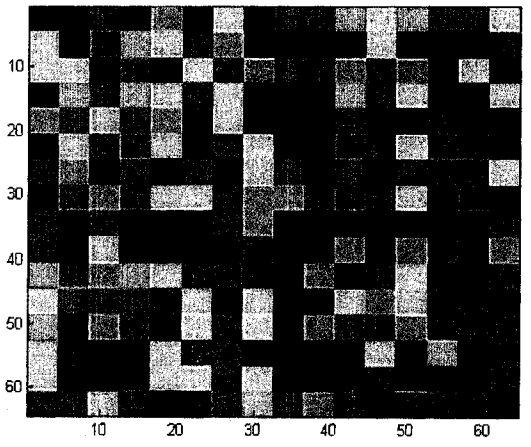


Figure 4.6 Segmented protein map for protein 1AAQ after DWT.

We rank the silhouette scores of each cluster in the hierarchy and choose those segments that constitute the cluster of highest rank. Each segment of the *fA* corresponds to the correlated mutation scores of the windows of the sequence. It is thus simple to back track to those regions for the given protein. Figure 4.7 provides an overview of the resultant hierarchical clustering of segments and the resulting frequency aggregates of a single layer of the protein map of protein 1AAQ.

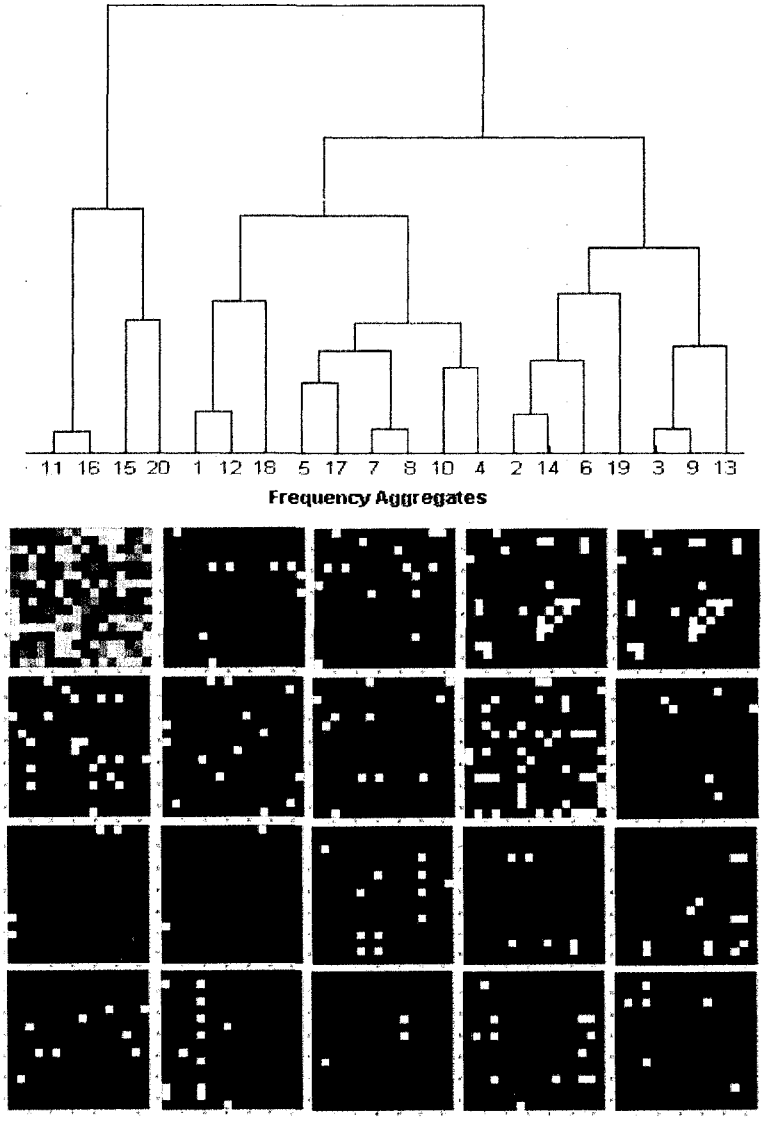


Figure 4.7 Clustered segments of a layer of a protein map.

4.2.3.1 Conservation measures

The hierarchical clustering of segments of a given protein is carried out for each physico-chemical property. We follow the above process of generating *fA* for each layer of the protein map. The generation of *fA*s facilitates back tracking to specific sequence positions on the protein that could constitute conserved domains for each property. To quantify the likelihood that a sequence position is part of a domain or is at the boundary of a domain across the five physico-chemical properties, we define a simple weighing scheme to measure the likelihood that a given position on the sequence constitutes a domain.

As in the MSA of proteins, key positions along the backbone which are crucial to stabilize the protein structure or which play an important functional role (as in the active site of an interaction site), are revealed. These positions tend to be more conserved than others and strongly favor amino acids with similar and very specific physico-chemical properties because of structural and functional constraints.

Based on this concept, we align the generated *fA* for each physico-chemical property and weigh the probability of occurrence of a residue at the given location as conserved across the properties. This probability E_i as in Eq. 4.8, acts as an indicator of those residues that strongly constitute domains.

$$E_i(P_i) = \frac{\#of \quad occurances \quad at \quad P(i) \quad as \quad conserved}{\# \quad of \quad physico-chemcial \quad properties}, \tag{4.8}$$

where E_i is the estimated probability of conservation for a residue at location i. This results in a weighted representation of a domain of protein P_i.

4.2.4 Analysis of the Structural Environment of Conserved Residues

We analyze all the conserved residues and compare the structural environment to amino acids in the naturally occurring proteins in the dataset, using packing density, hydrogen bonding, and solvent accessibility. The following is a brief description of the methods used to determine the parameters. The values computed are presented later in Section 4.4.2.

1. **Packing Density (Ooi Number):** A contact number with other residues within an 8 Å radius is computed using the method of [47]. Because the longest distance from C^{i}_{α} *to* C^{i+1}_{α} is approximately 4 $\mathring{A}$, the nearest neighbor residues on either side of the dipeptide are omitted.
2. **Hydrogen Bond Information:** Hydrogen bond information is defined using a donor-acceptor distance of $\leq$ 3.5 $\mathring{A}$. Angular criteria are not considered because side-chain atoms are not equally positioned by crystallography, and not all hydrogen atom positions are fixed by the positions of the heavier atoms. Hydrogen bonding is examined from a side chain at positions *i* to the residues other than those at positions *i-1, i* and *i+1*, the average number of hydrogen bonds (dipole interactions) that can be formed by the residue in a given position.
3. **Solvent Accessibility:** The solvent accessible contact area of amino acids is calculated using the method of [48], with a probe radius of 1.4 $\mathring{A}$. The percentage of accessible contact area of the total atoms is used.

4.3 Results and Discussion

This section 4.3, enumerates the results obtained for each validation proposed herein. Largely automated sequence comparison protocols are responsible for databases of aligned protein domains such as PFAM, and SMART. The assignment of domain boundaries for entries in these databases sometimes originates in a manually-curated 'seed' alignment, as is the case for PFAM. Alternatively, computer analysis is applied based either on the recurrence of similar sequence segments in different proteins at different distances from the *N*- and *C*-termini, or on duplicated segments observed in protein sequences.

4.3.1 Accurate Domain Assignment

Accurate domain assignment requires, ideally, structural information, or otherwise the repeated occurrence of a domain in different contexts. Domain identification is observed in protein families that lack relevant structural information and whose structures comprise several domains. If these domains are only observed in a single order, or if sequence comparisons fail to reveal their presence elsewhere, then the current protein domain databases will erroneously assign a single domain to the whole protein.

For our experiment, we have randomly chosen proteins from the Swiss_Prot and SMART databases that belong to the Trypsin and Eukaryotic families. Table 4.2 shows these databases and results. Figures 4.8, 4.9, 4.10, and 4.11 show the results of comparisons made with the domain assignments of Swiss-Prot.

Table 4.2 Domain validation of Trypsin and Eukaryotic proteins.

	ID	AC	Domain in SMART	Domain Identified (cut off 0.4 and above)
Trypsin	1433Z_BOVIN	P63103	3-242	1-235
	3BHS_VACCV	P26670	2-144, 43-155, 57-91, 87-198, 167-197, 190-239, 190-267	1-345
	3HAO-PSEFL	Q83V26	13-126, 51-121, 51-86, 84-184, 104-166	10-180
	ACT10-DICDI	Q54GX7	6-376; 3-376 (LS)	1-375
	ACT12-ARATH	P53497	7-377; 4-377, 230-241	2-370
	ACT17-DICDI	Q554S6	6-374; 3-374	1-370
	ACT18-DICDI	P07828	6-374; 3-376, 208-216	1-375
Eukaryotic	2SS1-ARATH	P15457	1-21, 59-153; 5-27, 5-18, 59-153, 85-96	1-155
	2SS2-ARATH	P15458	1-21, 45-158; 5-27, 45-158, 62-73, 76-85, 89-101	1-160
	2SS2-BRANA	P01090	5-24, 60-167; 2-168, 42-65, 47-74, 92-104, 102-153, 104-153, 105-159, 129-143	1-165
	2SS2-CAPMA	P30233	5-27, 40-149; 2-75, 2-20, 36-59, 68-81, 88-145, 92-141, 106-130	3-140
	2SS3-ARATH	P15459	1-21, 58-151; 5-27, 58-151	2-140
	2SS4-ARATH	P15460	1-21, 58-155; 5-27, 5-18, 58-155	1-155
	2SS4-BRANA	P17333	5-27, 60-169; 1-116, 2-170, 29-86, 42-65, 47-74, 87-148, 92-104, 111-146, 130-146	1-145

SWISSPROT ID: 1433Z_BOVIN
AC NUMBER: P63103

CHAIN	1-245
MOD_RES	1
MOD_RES	184
MOD_RES	232
CONFLICT	25
HELIX	3-14
TURN	15-17
HELIX	19-30
TURN	31-33
HELIX	38-66
HELIX	74-104
HELIX	106-108
HELIX	112-131
HELIX	136-157
HELIX	167-179
TURN	180-182
HELIX	185-201
HELIX	212-225

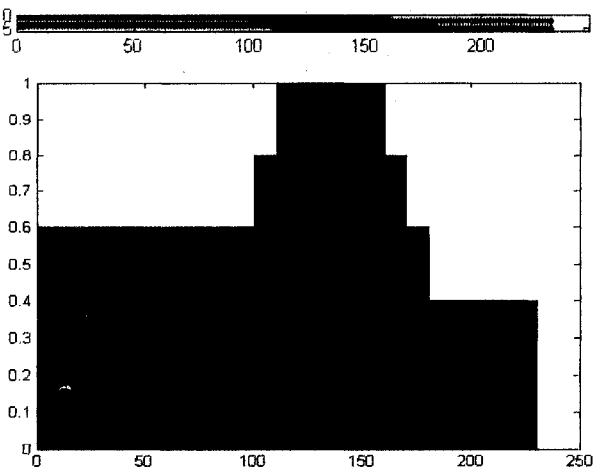


Figure 4.8 Degree of conservation of protein 1433Z_BOVIN.

SWISSPROT ID: 3BHS_VACCV
AC NUMBER: P26670

CHAIN	1-346

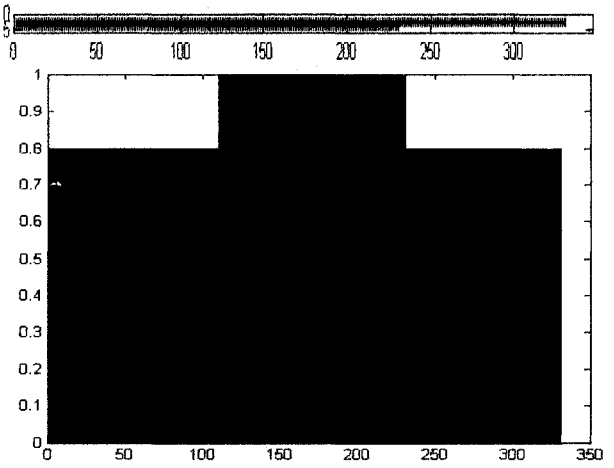


Figure 4.9 Degree of conservation of protein 3BHS_VACCV.

SWISSPROT ID: 2SS1_ARATH
AC NUMBER: P15457

SIGNAL	1-21
PROPEP	22-37
CHAIN	38-73
PROPEP	74-83
CHAIN	84-162
PROPEP	163-164

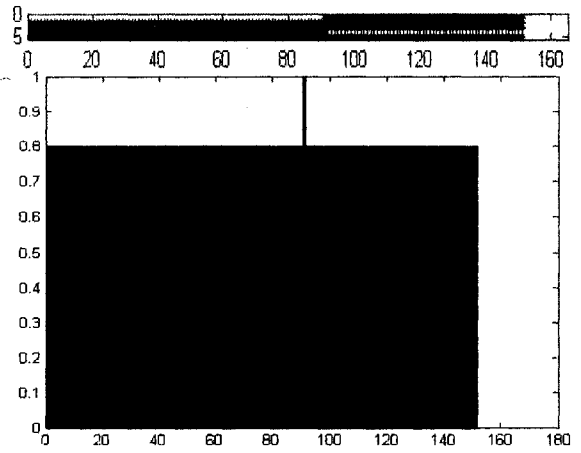


Figure 4. 10 Degree of conservation of protein 2SS1_ARATH.

SWISS PROT ID: 2SS2_ARATH
AC NUMBER : P15458

SIGNAL	1-21
PROPEP	22-37
CHAIN	38-72
PROPEP	73-88
CHAIN	89-170

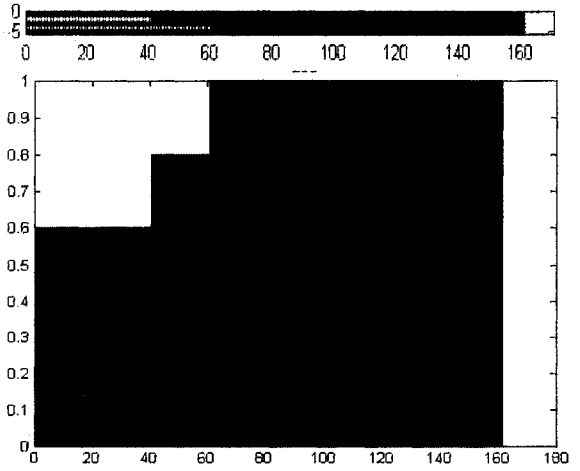


Figure 4.11 Degree of conservation of protein 2SS1_ARATH.

4.3.2 Residue Type Based Measures

Physico-chemical properties of proteins may also help predict domain boundaries, since they tend to have different characteristics around domain transition points than in domain core positions. For example, hydrophobic residues tend to cluster inside domain cores with hydrophilic residues occupying more exposed locations in a protein structure,

and, therefore are more likely to be in inter-domain regions. Similarly, certain amino acids such as cystines and prolines are crucial in defining protein structure, and therefore tend to occur in different frequencies in core domain and inter-domain regions of a protein. The value of considering residue composition in detecting domain boundaries is also demonstrated in the work done by Miyazaki et al. [49]. In order to exploit these sources of information, we must first define several measures: those for hydrophobicity; those for molecular weight; and those for the amino acids cystine, valine, proline, and glycine, all believed to be instrumental in defining protein structure. In addition, RasMol[21] classification of amino acids must be completed to create and measure a set of non-redundant classes (acyclic [ARNDCEQGILKMSTV], aliphatic [AGILV], aromatic [HFWY], buried [ACILMFWV], hydrophobic [AGILMFPWYV], large [REQHILKMFWY], negative [DE], positive [RHK], and small [AGS]). For each measure, the score of an alignment column is defined as the average of all residue scores, where residue scores are defined in the range of 0-1. Hydrophobicity and molecular weight residue scores are adopted from Black and Mould [50], and class scores are simply defined by the presence (score 1) and absence (score 0) of the residue in the class.

4.3.3 Structural Environment of Conserved Residues

To score residue presence, we first conduct a comparative study to verify the validity criteria which will test the structural environment of the reported conserved residues. We used a dataset consisting of the protein sequences reported by the Munich Information Center for Protein Sequences (MIPS[22]) yeast protein-protein interaction dataset of family (3.1.1), were reported in the PARTSLIST [51] database. The listed

[21] http://www.umass.edu/microbio/rasmol/
[22] http://mips.gsf.de/

proteins were also cross-ranked with representatives from two other well-known, functional classifications, namely the Julia classification by Wilson et.al., JMB 297(1)[23], and the GenProtEC[24] classification for E. coli. Three families, namely (3.1.1), (3.4.21) and (3.2.1), were considered, and a total of 64 proteins were used for analysis.

The first test of validation, we compared the relative composition of amino acids of the conserved regions to the entire proteins. Ideally, for a good dimensionality reduction, the conserved amino acid composition should exhibit a trend similar to that of its natural occurrence. From our test, it can be seen, as in Figure 4.12, that the behavior tends to hold true for all three families of proteins, as well as with the entire dataset.

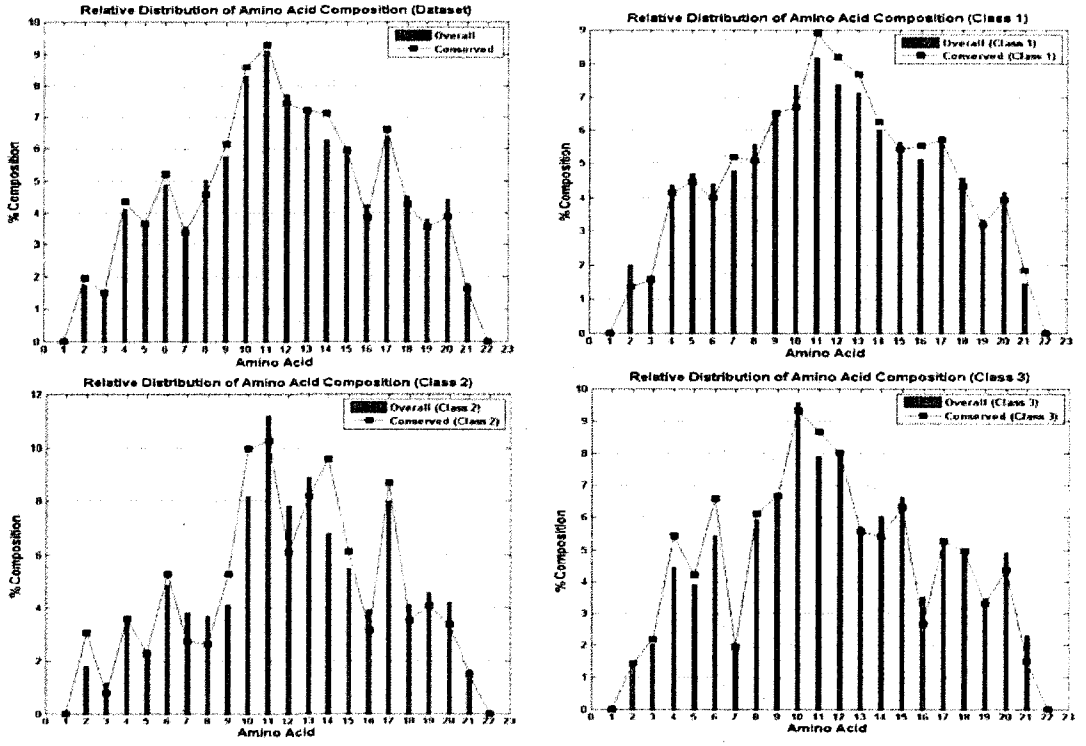


Figure 4.12 Comparison of reported relative amino acid composition.

[23] http://bioinfo.mbb.yale.edu/align/scop/tables/
[24] http://genprotec.mbl.edu/

We further reinforce our observation by subjecting the proteins of family (3.1.1) (Class 1) to the described validation criteria. As illustrated in Figure 4.13, it is clear that the trend of conserved amino acids is consistent with that of the naturally occurring proteins of the family. This result supports our hypothesis that a correlated trend across protein properties are conserved and can be exploited for the classification of proteins.

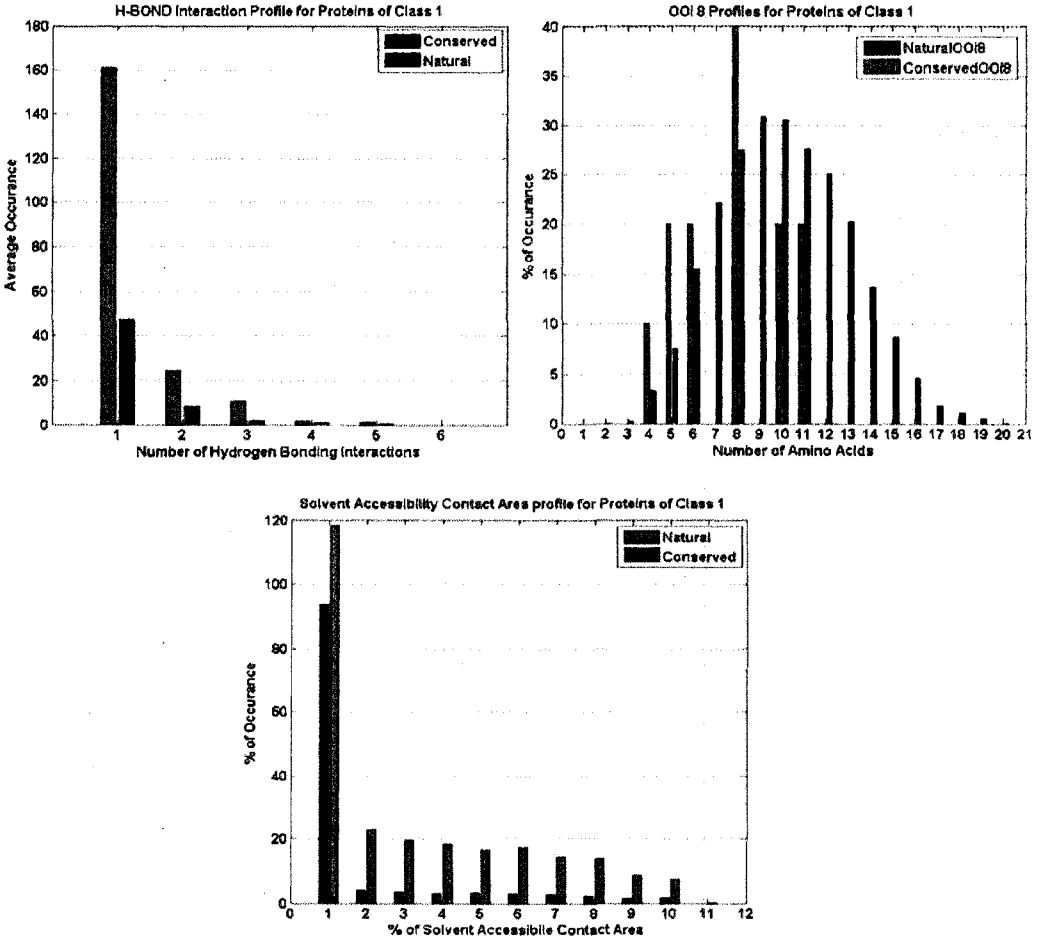


Figure 4.13 Results of analysis.

We propose a classification scheme, where the feature vector is the result of the above process. This process is shown in Figure 4.14. The dataset consists of proteins

from the UniProtKB[25], and of proteins from the Trypsin and Eukaryotic families under the UniProtKB sequence filtering constraints. UniProtKB allows one to filter out sequences based on a range of 50% to 100% sequence identity. When the search is subjected to one of these degrees of sequence similarity, the resultant is the grouping of the proteins based on proteins of UniProt50, UniProt90, or UniProt100 seed proteins. Thus, a reduced number of proteins that match these seed proteins are identified.

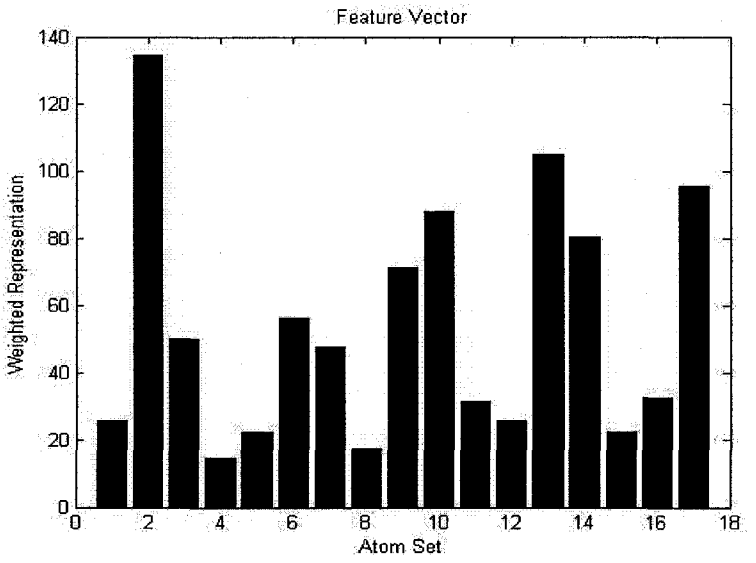


Figure 4.14 Representation of a feature vector.

This reduction results in 10,646 protein sequences which contain Trypsin in the description of the protein. The protein sequences are ordered in descending order of an identity score. Similarly, we obtain 10,995 protein sequences containing the key word Eukaryotic. From these key words, we filter out proteins that are known to be multi-domain in nature. For the purpose of training and testing we randomly choose 500 proteins from the Trypsin and 500 from the Eukaryotic descriptors.

[25] http://beta.uniprot.org/

These proteins are then subjected to classification using a Random Forest Classifier. An independent test set consisting of 100 Trypsin and 100 Eukaryotic proteins are randomly chosen from the dataset and are subjected to the trained classifier, and the results are shown in Table 4.3. The classifier consistently classifies the proteins into their corresponding families, with an average accuracy of 89%. Similarly, Table 4.4 shows the results of a 10-fold cross validation carried out on the training set, and a 90.5% accuracy is observed. These results indicate that the method can identify discriminatory domains for effectively classifying proteins that belong to the corresponding Trypsin and Eukaryotic classes of proteins.

Table 4.3 Results of classification on independent test set.

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.889	0.109	0.889	0.889	0.889	0.966	Eukaryotic
0.891	0.111	0.891	0.891	0.891	0.966	Trypsin

Eukaryotic	Trypsin	Classified as
88	11	Eukaryotic
11	90	Trypsin

Table 4.4 Results of ten fold cross validation.

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.9	0.09	0.909	0.9	0.905	0.948	Eukaryotic
0.91	0.1	0.901	0.91	0.905	0.948	Trypsin

Eukaryotic	Trypsin	Classified as
90	10	Eukaryotic
9	91	Trypsin

4.4 Conclusion

The challenges faced in the annotation of proteins have moved into a realm in which traditional sequential analysis and structural alignment techniques are not sufficient. The potential and importance of using physico-chemical properties to extract the implicit behavioral characteristics of a protein are now being realized. Through the course of this work, we envisage conservation in terms of properties rather than the residues themselves. Our contribution can be viewed as twofold. First, we aimed at creating an algorithm to provide us a means to integrate multiple physico-chemical properties in the form of a layered proteins map with each layer corresponding to a physico-chemical property. Second we proposed a wavelet-based segmentation approach that efficiently detects regions of property conservation across all the layers of the protein map. We stringently validated the reported regions using our validation schemes, and we report significant regions of accuracy to show that homologous proteins exhibit conservation of physico-chemical properties over the protein backbone.

CHAPTER 5

PROTEIN STRUCTURE CLASSIFICATION BASED ON CONSERVED HYDROPHOBIC RESIDUES

Proteins contain a large but limited number of features. Ab initio computational protein folding models assist scientists involved in molecular biology and in bioinformatics to better elucidate the intricate process of protein folding and the causal forces involved. However, no current ab initio protein folding algorithm generates a high precision rate <3.5-*Å* backbone Root Mean Squared Deviation (RMSD) from the experimental structure for the identification of regions and features in large protein structures. This low precision rate stimulates a need for more efficient computational techniques, especially those geared toward the automatic annotation and classification of newly introduced proteins.

Traditional supervised machine learning techniques compare unclassified protein sequences to classified proteins using kernel functions [52]. This method produces a low effective cut-off point for the effectual homology modeling of proteins with ~30% sequence identity, a lower bound at which the computed structure can still accurately depict the arrangement of secondary structure elements in 3-D. Largely due to impediments posed by sequence similarity, researchers focus on finding conserved regions (sub-sequences) that exhibit sequence or property conservation across structurally

related proteins by restricting the feature space [53]. To this end, as hypothesized by Kauzmann [54], hydrophobic interactions play a major role in organizing and stabilizing the architecture of proteins.

Researchers have investigated the correlation of hydrophobic interactions to similarities in 3-D structural elements, and have exhibited and exploited property conservation at these sites. A number of computational methods to this end have been proposed in the literature. Paiardini et al. [55] and Reddy et al. [56], using multiple sequence alignment (MSA) techniques, show that a significant correlation exists between the sequence, structure, and conserved hydrophobic contacts (CHC) that remain invariant during long evolutionary periods. Reddy et al. [56] present a methodology, known as conserved key amino acid positions (CKAAPs), to identify conserved residues and potential folding nuclei based on sequence and weighted homologues scoring. Tsai et al. [57] propose a method using a scoring function based on the physico-chemical properties of hydrophobicity, compactness, solvent accessibility of surface area (ASA), and segmentation to test the validity of fold unit definition based on Eigenvector analysis.

Typically, these methods lack recognition and exploitation of the structural contributions of each residue. Later models that provide insight into structure discrimination using conserved hydrophobic residues have been proposed. Particularly, the model proposed by Muppirala et al. [58] quantitatively measures the individual contributions of amino acid residues in a protein structure. Each protein is treated as a network of edges representing inter-residue interactions between hydrophobic residues. Emphasizing the relation between hydrophobic interactions and stability, Huang et al. [59], introduce a pair-wise energy function that enumerates contacts between

hydrophobic residues while weighing their sum by the total number of residues surrounding them. Although using different approaches, each model suggests a common and unexpected feature of protein packing that proteins significantly rely on based on few members of the set of conserved residues.

We propose a data mining model, which we believe will also be useful for classification purposes, for the integrated analysis of five popular hydrophobicity scales to enhance the detection of structurally conserved regions among homologous proteins. Employing the principles of graph theory and incorporating the metric of mutual information to identify compact structural units, we extract frequently occurring patterns using a discriminative weighing function. Our goal is to identify conserved hydrophobic residues among structurally related proteins, using hydrophobicity scales for classification. By doing so, we reduce our feature space and show that the reported conserved hydrophobic residues are sufficient to differentiate between native and non-native proteins at both the class and fold levels of the structural classification of proteins (SCOP) hierarchy. We test the efficacy of our model by comparing the length of the feature vector with traditional techniques. Our feature vector is significantly smaller, yet yields comparable results. The scalability analysis reaffirms that the proposed model is scalable to multiple classifiers.

5.1 Approach

Expressions of the hydrophobic effect are palpable in many facades of protein sequence-structure-function dependencies, including

1. The stabilization of the folded conformation of globular proteins in solutions;
2. The subsistence of amphipathic structures in peptides or of membrane proteins at lipid boundaries; and
3. Protein-protein interactions associated with protein subunit assembly, protein-receptor binding, and other intermolecular bio-recognition processes [11].

We hypothesize that an integrated analysis of multiple prominent hydrophobic scales can lead to better encapsulation of hydrophobic bearings on protein functional analysis. We focus on five well-known scales of hydrophobicity, the Kyte and Doolittle, the Hopp Woods, the Janin, the Rose et al. and the Eisenberg et al. scales [7]. The discussion on the scales follows in the next section 5.1.1.

5.1.1 Hydrophobicity Scales

The pioneering work of Kauzmann elucidates important attributes of the thermodynamic stabilities of proteins and suggests that hydrophobic interactions are dominant in the protein folding process. More than thirty-eight scales of hydrophobicity have been developed since the Kauzmann work [7]. These scales contain distinctive stereo-chemical hydrophobicity rankings for better understanding of protein-interaction mechanisms, which actually create confusion rather than resolution [9]. Nevertheless, the hydrophobic property of proteins is widely considered the most important underlying factor in the hierarchical structure and in the 3-D stability of proteins.

Specifically, amphipathic residues responsible for the formation of secondary structures along the backbone of the protein are also usually inconsistently ranked due to their varied nature. To correlate the hydrophobic interaction of residues and the formation of the secondary structure, we propose the creation of summary graphs (see [60] for background). These summary graphs capture behavioral similarity across hydrophobic scales, while pursuing distinct objectives: to capture the local interactions between protein residues, to reduce the feature space, and to provide an estimate of the hydrophobic behavior of the protein.

Table 5.1 shows residue ranks, in ascending order of magnitude, based on the hydrophobic propensity assigned to the residues by each scale. A wide range of hydrophobicity values exist for each amino acid. Some amino acids show a high hydrophobic ranking with one scale and a high hydrophilic ranking for another scale [9]. Though most residues are ranked consistently across scales, certain residues rank across the spectrum more than the others. Inconsistencies in the ranking of aromatic residues are attributed to the size of side chains, to the environment (solute chosen), and to the tender difficulty to use them to model protein folding [9].

Table 5.1 Amino acid ranks in hydrophobicity scales.

Rank	1	2	3	4	5	6	7	8	9	10
Kyte and Doolittle	ARG	LYS	ASP	GLU	ASN	GLN	HIS	PRO	TYR	TRP
Hopp Woods	TRP	PHE	TYR	ILE	LEU	VAL	MET	CYS	ALA	HIS
Janin et al	LYS	ARG	GLU	GLN	ASP	ASN	TYR	PRO	THR	HIS
Rose et al	LYS	ASP	GLU	GLN	ASN	PRO	ARG	SER	THR	GLY
Eisenberg et al	ARG	LYS	ASP	GLN	ASN	GLU	HIS	SER	THR	PRO
Rank	11	12	13	14	15	16	17	18	19	20
Kyte and Doolittle	SER	THR	GLY	ALA	MET	CYS	PHE	LEU	VAL	ILE
Hopp Woods	THR	GLY	PRO	ASN	GLN	SER	ASP	GLU	LYS	ARG
Janin et al	SER	ALA	GLY	TRP	MET	PHE	LEU	VAL	ILE	CYS
Rose et al	ALA	TYR	HIS	LEU	MET	TRP	VAL	PHE	ILE	CYS
Eisenberg et al	TYR	CYS	GLY	ALA	MET	TRP	LEU	VAL	PHE	ILE

5.1.2 Capturing Local Interactions between Protein Residues

With the backbone (Cα atoms) defining the overall protein structure, we use protein structure graphs *(G)*, to create a four-body nearest neighbor propensity representation of a protein using Delaunay Tessellations *(DT)* [60] (see Figure 5.1). The edges of this graph are defined for a finite set of points, satisfying the empty sphere property [19]. The corresponding adjacency matrix for the *G* is shown in Figure 5.2.

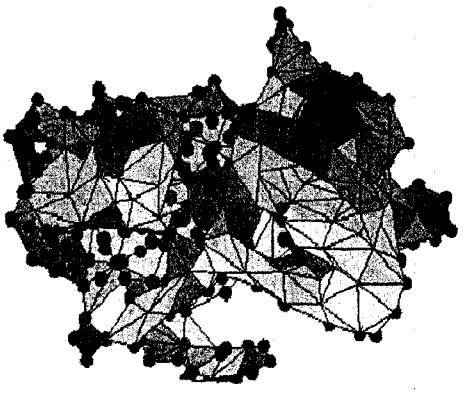


Figure 5.1 Result of applying Delaunay Tessellation.

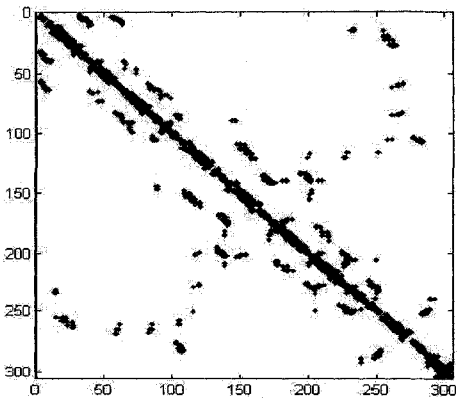


Figure 5.2 Adjacency matrix representing residues.

Each resultant tetrahedron of the tessellation identifies neighbors by capturing four natural nearest neighbor residues that fall on the circumference of a sphere of radius τ. Tetrahedra with vertex-vertex distance > τ are omitted on the grounds that significant interactions will not occur at greater distances. Thus, biases that arise from the adoption of a fixed coordination volume around a given residue can be avoided [61]. The value of τ determines the proximity for residue-residue interaction [62], and is set to 8.5 $\AA$.

5.1.3 Feature Space Reduction

The 2-D representation of hydrophobic propensities makes it difficult to observe regularity in the conformation of protein backbone that is caused by the competition between local hydrophobic interactions. The G of a protein can be viewed as an aggregate of a four-body nearest neighbor tetrahedra [60]. A weighted representation of a G, given a hydrophobicity scale *hyd*, is referred to from this point forward as a hydrophobicity scaled graph $G_{hydn}(P)$. We view each tetrahedra of $G_{hydn}(P)$, as a composition of a central residue, connected to its corresponding nearest-neighbors and located within the first coordination shell [61]. Thus, given a hydrophobicity scale, we define a hydrophobic center as that central residue that possesses the highest hydrophobic

potential. Grouping all adjacent tetrahedra coincident with the hydrophobic center, we define a neighborhood as a cluster that shares a common center of the highest hydrophobic propensity (Figure 5.3). By constraining the number of residues in the proceeding methodology, we eventually reduce our feature space.

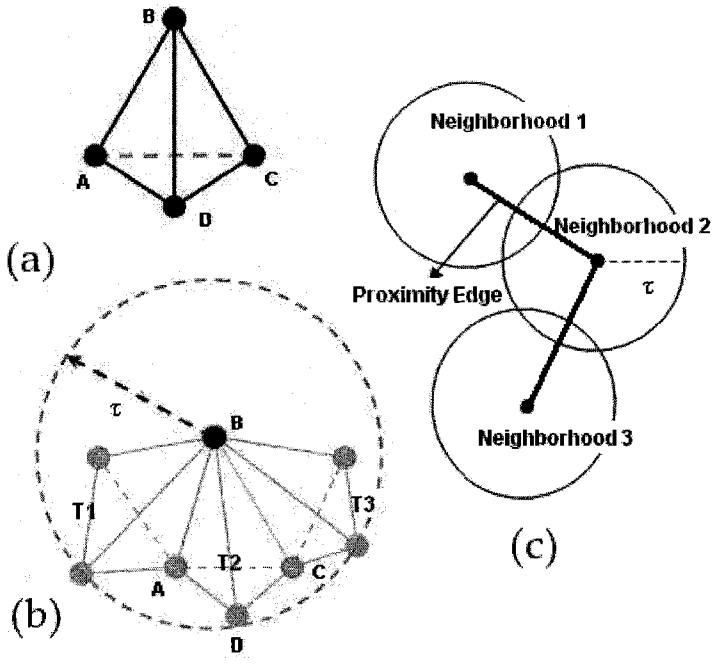


Figure 5.3 Capturing of protein structure using Delaunay Tessellation.

5.1.4 Estimation of Hydrophobic Behavior

Hydrophobic residues buried in the protein core generally display a compact structure and contain a hydrophobic interior [57]. However, in larger proteins, the collapse caused by the interaction of hydrophobic clusters with subsequent rearrangements forms secondary structure elements and tertiary structures. We interpret the interactions between the neighborhoods as long-range interactions that are captured by the proximity of hydrophobic centers in the native state (Figure.5.3.c). It is logical to

presume that the two centers of neighboring residues, associated with a central hydrophobic residue, are in close proximity if their neighboring residues are common. In this case, we say they share a proximity edge. We define a proximity edge as an edge between two hydrophobic centers that share neighboring residues.

5.2 Methodology

Figure 5.4 provides the proposed framework of the extraction and coherent subgraph mining algorithm. The following sections in this chapter are arranged as follows. Section 5.2.1, we provide a detailed description of our proposed approach to estimating hydrophobic behavior. In Section 5.2.1, we outline a detailed description of the steps involved in merging information from a set of hydrophobicity scales of a protein. In Section 5.2.2, we provide a protein partitioning scheme followed by a coherent subgraph mining schema in Section 5.2.3.

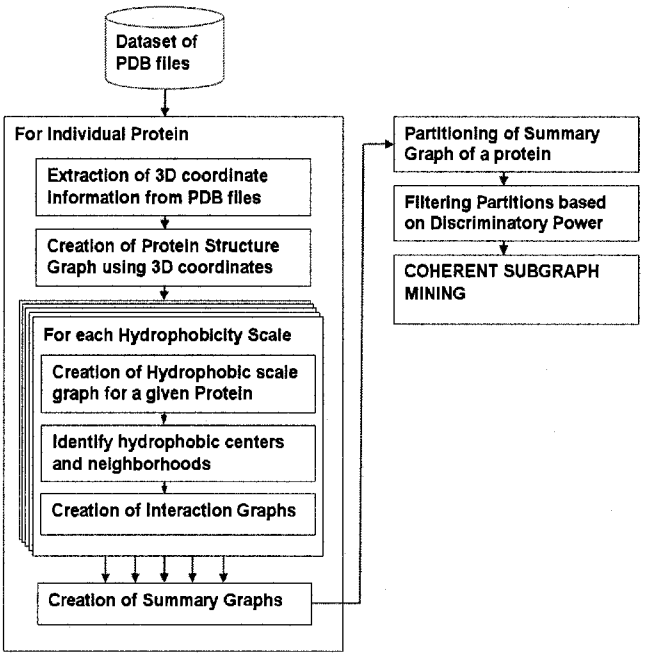


Figure 5.4 Proposed framework for the extraction of subgraphs.

5.2.1 Merging of Hydrophobicity Scales

The methodology of merging hydrophobicity scales for a given protein can be sketched as follows. We represent a protein *'P'* by an underlining graph *G(P)* called the protein structure graph, which we construct using Delaunay Tessellations. A weighted representation of the underlying graph *G(P)* is obtained from five different hydrophobicity scales (hyd_n) called the hydrophobicity scaled graph ($G_{hydn}(P)$) which we will discuss further in see Section 5.2.3. For a given scale, we attempt to reduce the feature space by considering only those residues with the highest weight (centers) among residue clusters. These centers become the vertices and edges (defined in Section 5.1.3) of the interaction graphs abstracting the behavior of the residues. An important contribution of this work is the integration of these scale representative interaction graphs in the form of summary graphs *SG*.

We first define a protein *'P'*, consisting of its set of residues, as the coordinates of Cα atoms in $\Re^3$ Euclidean space. Using this information, we define each residue as a vertex 'v.' Thus as in Eq. 5.1, let v_1 and v_2 be represented in $\Re^3$ Euclidean space

$$v_1 = (x_1, y_1, z_1) \in \Re^3, \qquad v_2 = (x_2, y_2, z_2) \in \Re^3. \tag{5.1}$$

The Euclidean distance *'d'* between the two vertices is defined as in Eq. 5.2

$$d(v_1, v_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}\,. \tag{5.2}$$

5.2.1.1 Protein structure graph and hydrophobic scales

A protein structure graph *G(P)* is defined to satisfy the constraints of Delaunay Tessellation. Let graph *G (P)* be a three-element tuple, so that $G = \{V, E, \tau\}$, where '*V (G)*' is a set of vertices that represents the Cα atoms of P. An edge $e \in E(G)$ exists between two vertices if the two vertices are spatial neighbors according to the Delaunay Tessellations "empty sphere property" [19] and [62]. Let 'τ' represent the predefined distance threshold, ranging between 8.5 and 10 Å. Thus we obtain Eq. 5.3 as

$$E(G) = \{(v_1, v_2) : v_1, v_2 \in V(G), d(v_1, v_2) \leq \tau\}. \tag{5.3}$$

The constraint results in a graph *G(P)* consisting of vertices joined by edges in a unique way to form a collection *T(G)* of non-overlapping tetrahedra [60] that can be viewed as clusters of four-body nearest neighbor residues connected by edges under the criteria specified by Delaunay Tessellations.

A hydrophobic scale on a protein structure graph *G(P)* is a function *hyd* that labels every vertex $v \in V(G)$ with a corresponding weight of hydrophobic propensity depending on the type of amino acid found at *v*. For the five scales of hydrophobicity, we use hyd_n, where $n = 1 \ldots 5$. The resultant is a protein structure graph with vertices assigned weights corresponding to a specific hydrophobic scale called a hydrophobic scale graph denoted as $G_{hydn}(P)$.

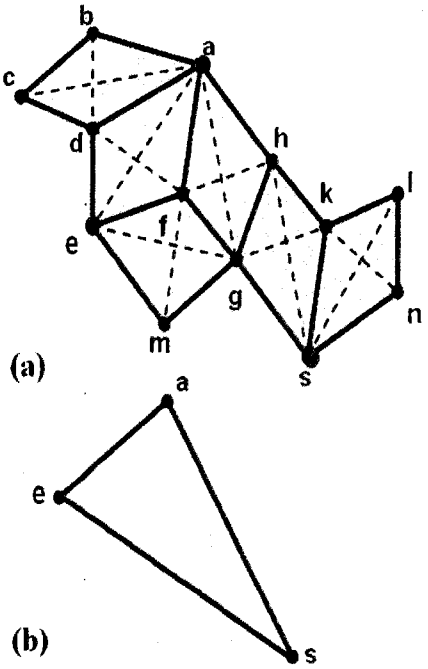
5.2.1.2 Identification of hydrophobic centers

For a given tetrahedron $t \in T(G_{hydn})$, we choose the vertex of the highest weight in hyd_n and call it the hydrophobic center *C(t)*. We cluster all tetrahedra having a common

maximum vertex, say *a*, and call the collection a neighborhood of center a denoted as *H(a)*. Thus we define

$$H(a) = \{t \in T; C(t) = a\}. \tag{5.4}$$

In this definition, as seen if Figure 5.5, a is the hydrophobic center of the neighborhood *H(a)*, $G_{hydn}(P)$. Not all tetrahedra surrounding a center belong to the same neighborhood.



t_1	a=10	b=7	c=8	d=5
t_2	a=10	e=12	f=9	d=5
t_3	a=10	f=9	g=3	h=6
t_4	e=12	f=9	g=3	m=7
t_5	s=9	g=3	h=6	k=4
t_6	s=9	k=4	l=3	n=5

$C(t_1)$=a $C(t_2)$=e $C(t_3)$=a $C(t_4)$=e
$C(t_5)$=s $C(t_6)$=s

$H(a)=\{t_1,t_3\}$ $H(e)=\{t_4,t_2\}$
$H(s)=\{t_5,t_6\}$

Figure 5.5 Example of the process of identifying centers and neighborhoods.

5.2.1.3 Interaction Graphs

We aim to identify interactions among hydrophobic centers. Two centers, *a* and *b*, are connected by an edge in an interaction graph *(IG)* if the neighborhoods of a and b share a vertex in common, such as the one shown in Figure 5.5 above. As proposed earlier, we believe the interaction graph, especially in edges among residues, is

significant in retaining the structure of the protein. Correspondingly, we term the edges in the interaction graph, *IG*, proximity edges. For a given neighborhood *H(a)* of center *a*, let $V(H(a))$ be all vertices of all tetrahedra in *H(a)*.

Definition: An interaction graph, $IG = \{V'', E''\}$ is a graph whose vertex set $V''(IG)$ represents the hydrophobic centers connected by the edges of set $E''(IG)$, and is defined in Eq. 5.5 as

$$E''(IG) = \{(a,b) : V(H(a)) \cap V(H(b) \neq \phi\}. \tag{5.5}$$

With five different hydrophobicity scales, we obtain a set of five interaction graphs (IG_n) representing individual proteins. For a given protein, the vertices for each *IG* have a common vertex set $V(G_{hydn}(P))$, but possess different edge sets. It is our objective to extract similarities among all interaction graphs of *P*.

5.2.1.4 Summary Graph

Based on our approach in Section 5.1.4, we postulate that the similarity among scales will reveal useful insight into the identification of folding units. The summary graph is an overlapping mechanism that is capable of capturing similarities across different graphs and can be used to merge information derived from the five interaction graphs (IG_n) of a protein.

Definition: A summary graph *SG (P)* with IG_n, $n=1,..,5$ is defined as an unweighted graph *SG* where *V (SG)* = *V (G(P))* and $E(SG) \subset \bigcup_{n=1}^{5}(IG_n)$ are determined by the frequency '*k*' of occurrences in the Interaction Graphs (IG_n), where $1 \leq k \leq n$ is a user defined threshold.

We aggregate (overlay) the interaction graphs obtained for a protein under the five scales of hydrophobicity. We aim to identify subgraphs in the aggregated summary graph. However, such an aggregation could result in the creation of false subgraphs that may not occur in the original interaction graphs. The frequency of occurrences '*k*' provides a threshold through which means any biases caused by the scales could be annulled. Since our approach evaluates the combined effect of the five scales, we undermine this problem.

5.2.2 Partitioning a Protein

In the following discussion, we describe the process of protein partitioning as a means of identifying significant subgraphs in the summary graph. These subgraphs contribute toward the identification of key structural characteristics embedded within the protein. Using the Trajans algorithm, we extract all possible connected components (subgraphs) of *SG*. Through the concept of mutual information, we filter insignificant components of *SG* that do not satisfy a specified threshold *(μ)*, as shown in Figure 5.6 below. The steps of our algorithm, presented in Figure 5.6, are explained in sections 5.2.2.1 and 5.2.2.2 and 5.2.2.3.

Algorithm 1 *Partitioning of Summary Graph*
Input: A Connected Summary Graph SG(P)
Output: Set of subgraphs of SG(P)

1. Identification of connected components (subgraphs) of SG(P) using Trajan's DFS algorithm.
2. Filtering components based on *Mutual Information* (μ) > *0.1*.
3. Determining partitions by sorting and finding gaps in residue locations.

Algorithm 1.3 *Identification of Gaps in* SG.
Input: *SG* and it subgraphs (U_n).
Output: *Partitions* of SG.

1. $List_{new}$ is assigned the residues of each subgraph U_n.
2. Sort residues in $List_{new}$ according to the location in protein sequence
3. If difference between residue location >2
    a. Identify as *Partition*
    b. Record beginning and end locations of *Partition*

Algorithm 2 *Frequency of occurrence of subgraphs*
Input: Protein Database *(PD)* and Subgraph List *(SL)*.
Output: Matrix *NF* containing frequency of occurrence of each subgraph U_j in *PD*.

1. Repeat for each *SG(P)* in *PD*.
2. Repeat for each subgraph U_j in *SL*.
    a. Compute *D-RRAM* between *SG(P)* and U_j.
    b. If *D-RRAM*=U_j then
        i. If location of vertices of U_j fall within the location range of *SG(P)*
            NF(Uj)=1
        else
            NF(Uj)=0

Figure 5.6 Algorithms of coherent subgraph mining.

5.2.2.1 Identification of connected components

For the first step of Algorithm 1, we use Tarjan's Algorithm [63] to find the bi-connected components, defined below, of a summary graph.

Definition: A bi-connected component of an *SG* is defined as a maximal subset of the edges of *SG* such that the corresponding induced subgraph *U* cannot be disconnected by deleting any vertex of *U*.

The connected components of an undirected graph are essentially maximal connected subgraphs. The algorithm is based on the tree structured, depth-first search, where the search begins from a root node, and strongly connected components form the subtrees of the search tree. The time complexity of this algorithm is $O(V+E)$, where V and E are the number of vertices and edges, respectively.

5.2.2.2 Filtering using mutual information

Typically, a large number of subgraphs U are produced for a single summary graph using the above process. However, since not all of the subgraphs are useful for classification, we first create a filtering process based on the information theoretic metric of mutual information that uses entropy to select the most informative collection of subgraphs.

We define function $MI(U)$ for a subgraph U of SG, which measures the marginal entropy. Similarly, $MI(SG/U)$ and $MI(U/SG)$ measure the conditional entropies. The joint entropy of SG and U are measured using the function $MI(SG,U)$. Using the functions above, the mutual information between graph SG and subgraph U is defined as in [19], using the Eq. 5.6

$$MI(SG,U) = MI(SG) + MI(U) - MI(SG,U). \quad (5.6)$$

A subgraph U is a coherent subgraph of SG if the mutual information between U and SG is above a fixed threshold μ. Selecting only coherent subgraphs offers the following advantages:

1. It filters out generic subgraphs across the protein (for those subgraphs, the mutual information tends to be low), and
2. It finds statistically significant patterns, since each coherent subgraph is strongly correlated to its own parent graph.

5.2.2.3 Partitions in protein sequence

From the previous step, we obtain a set of subgraphs U that have mutual information greater than μ with respect to the corresponding summary graph SG. We devise a simple algorithm for sorting residues and finding gaps among them, which determine partitions in the protein sequence. A detailed description of the steps involved in finding gaps in protein sequences is described in Algorithm 1 (Fig. 5.6). We call the resultant gaps partitions because they delineate portions of residues along the sequence. We validate the results of the partition algorithm by comparing them to the results achieved by Gelly et al. [64]. A detailed discussion of our results is presented in Section 5.4.1.

We test Algorithm 1.3 on a random PDB ID – 1AN2 (a) protein. The algorithm detects strongly connected components in black rectangles, as shown in Figure 5.7, by choosing those components with MI-value > μ (threshold μ is set to 0.1 in this example) and determined partitions. Figures 5.7 and 5.8 describe the location of the cuts and the corresponding MI for each component. Further validations of summary graph partitioning, such as that shown in Figure 5.7 are described in Section 5.3.1.

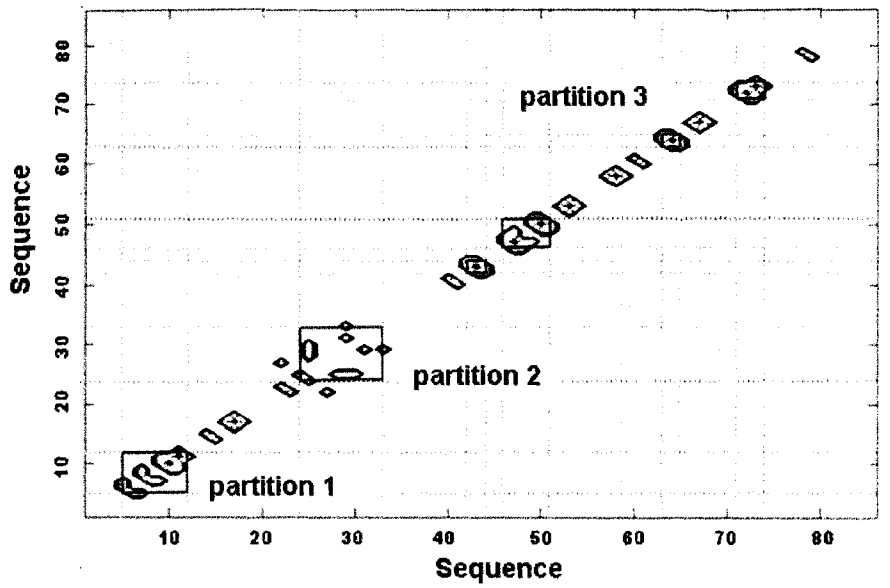


Figure 5.7 Summary Graph Representation of Protein 1AN2.

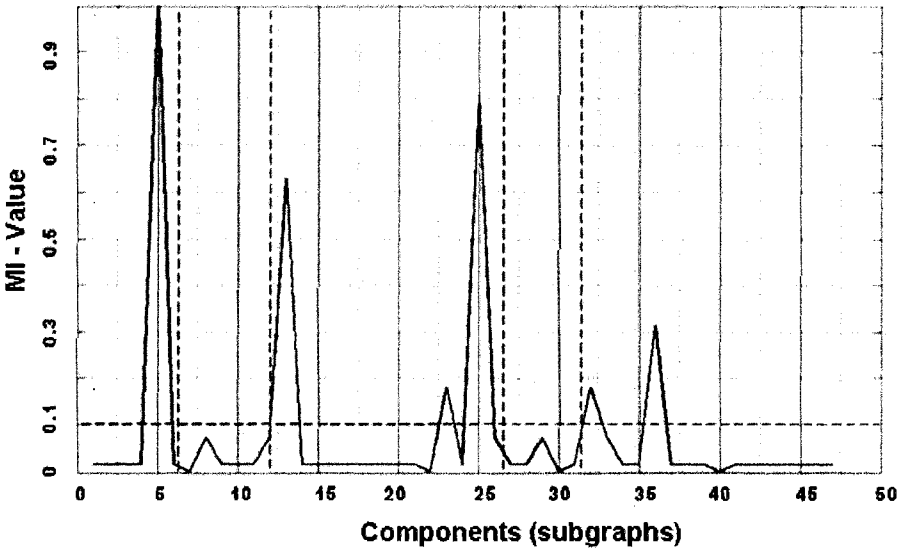


Figure 5.8 Corresponding mutual information values.

5.2.3 Coherent Subgraph Mining

Using the method of partitioning described in section 5.2.2, we are able to extract subgraphs that correspond to the structural units of a protein. We hypothesize that structurally homologous proteins exhibit conserved units dictated by the hydrophobic

behavior of the residues that belong to these units. In this section 5.2.3, we propose a means of identifying conserved residue interaction patterns within a family of proteins. Since we propose to use the frequency of interaction patterns in our classification scheme, we detail a simple approach to identify the presence of a subgraph in a summary graph in Section 5.2.1.4. In Section 5.2.1.5, we define a simple metric that estimates the discriminatory power (*DP*) of a subgraph based on its frequency of occurrence within the proteins of a family. Using the frequency patterns in Section 5.2.3.3, we provide a detailed description of the design of our feature vector for the classification of proteins.

We view a protein database (*PD*) as a collection of summary graphs *SG* corresponding to the proteins of the database. We define a subgraph list (*SL*) as a collection of all partitions (subgraphs) *U* of the summary graphs *SG* belonging to *PD*. For a comparison, we define a residue-residue adjacency matrix (*RRAM*) as a 20x20 matrix, where each row and column corresponds to the 20 known amino acid types. Thus, the *RRAM* of a *SG* or a subgraph *U* is such that *RRAM (l, m)* represents the frequency of the occurrence of the edges that have vertices of *amino acid*$_l$ and *amino acid*$_m$.

A Difference-RRAM (*D-RRAM*) is the difference operation performed between the *RRAM (SG)* and the *RRAM(U)*, defined by Eq. 5.7

$$D\_RRAM(l,m) = \min\{RRAM_{SG}(l,m), RRAM_{U}(l,m)\}. \qquad (5.7)$$

5.2.3.1 Frequency of subgraphs

Now that we have defined protein database (*PD*), a subgraph list (*SL*), and a means to compare an *SG* with a subgraph *U*, we use Algorithm 2, (Figure 5.6) to find the frequency of occurrence of each *U* of *SL* with respect to each *SG* of *PD*.

5.2.3.2 Filtering of subgraphs based on discrimination power

Our objective is to identify frequently occurring subgraphs that are capable of discriminating among proteins at the fold or class levels of the SCOP. The discrimination power (*DP*) of a subgraph is a measure of goodness used to estimate the significance of a subgraph in a family of proteins. It is used to distinguish among families of proteins, as defined by [19].

Definition: A discrimination power *(DP)* of subgraph U is defined using Eq. 5.8 as

$$DP(U) = \left| \frac{f(U)_A}{S_A} - \frac{f(U)_B}{S_B} \right|, \tag{5.8}$$

where $f(U)_A$ and $f(U)_B$ correspond to the number of proteins in family A and B having U as a subgraph (frequency of occurrence), and S_A and S_B correspond to the number of proteins in family A and B.

The greater the *DP* value, the more selective the feature. We define a threshold ξ that determines a cutoff for the selection of subgraphs. Thus, given n frequent subgraphs- $U_1, U_2, U_3 \ldots U_n$, that satisfy the threshold, we create a profile for each protein P in the dataset as an n-element vector $NF = f_1, f_2 \ldots f_n$ in feature space where f_j indicates the presence of the subgraph U in *SG*. We use the generated frequency matrix *(NF)* in the design of our feature vector (Section 5.2.3.3). The filtering process results in the reduction of the number of subgraphs, which inadvertently results in the reduction of the feature vector length used by the classifier. Though the resultant feature vector is confined to discriminating proteins that belong to two classes in *PD*, we extend this definition to suit proteins that belong to the multi-class *PD* in Section 5.3.

5.2.3.3 Feature vector design

In the feature vector designing stage, a set of descriptors capable of discriminating proteins of different classes is defined for each protein. These descriptors ensure the capture of significant, yet unique characteristics, common across a class of proteins. The presence of key interactions among residues (subgraphs) captured by *NF*, is a good discriminator. However, we believe that the nature of interacting conserved residues is unique to homologous proteins. Thus in addition to *NF*, descriptors such as

1. The connectivity of hydrophobic residues in the summary graph of a protein exhibits unique packing patterns. Connectivity, the ratio between the number of edges and the number of vertices, generally measures two aspects of interaction patterns. First, it measures which residues are interacting with one another; and second, it measures the frequency or regularity of these interactions.
2. The number of connected components in a summary graph reflects those interactions among hydrophobic residues that are prominently expressed across scales, as shown in Figure 5.9, parts a and b.
3. The relative amino acid composition of the summary graph is taken to assess the environmental-dependent parameters of conserved residues in the summary graph. Through these descriptors, we create a profile of any protein with respect to the reported conserved hydrophobic residues.

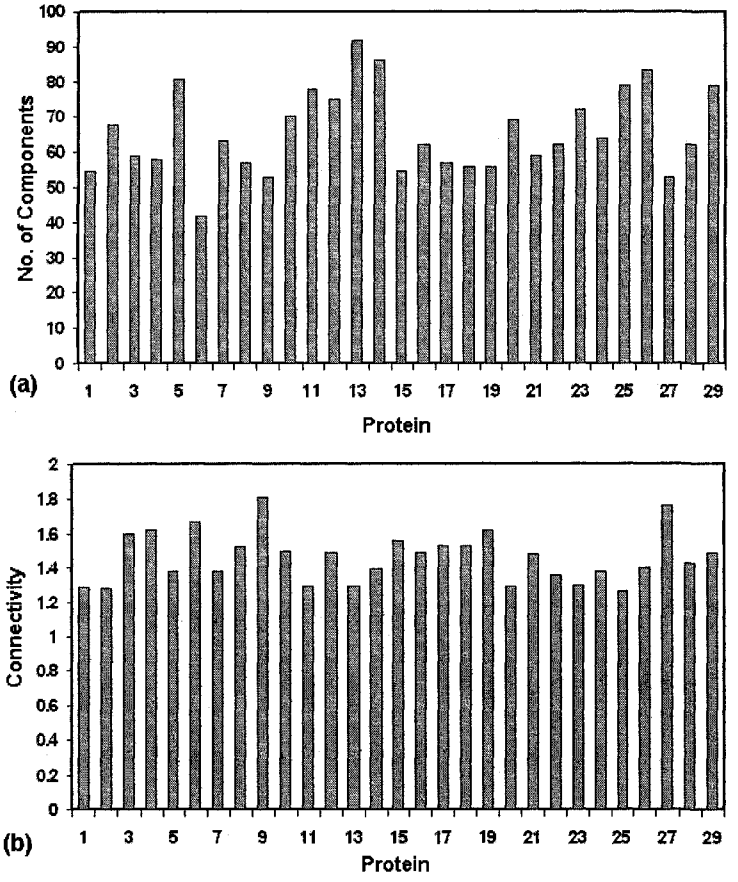


Figure 5.9 Protein of dataset C2.

5.2.3.4 Analysis of conserved residue structural environment

As in Chapter 4, we analyze the data set for all the conserved residues to compare the structural environment to amino acids in the naturally occurring proteins using packing density, hydrogen bonding, and solvent accessibility as briefly discussed below.

1. **Packing Density (Ooi Number):** A contact number with other residues within an 8 $\AA$ radius is computed using the method of Nishikawa and Ooi [47]. Because the longest distance from C^{i}_{α} *to* $C^{i+1}{}_{\alpha}$ is approximately 4 $\AA$, the nearest neighbor residues on either side of the dipeptide are omitted.

2. **Hydrogen Bond Information:** Hydrogen bond information is defined by using a donor-acceptor distance of ≤ 3.5 *Å*. Angular criteria are not considered because side-chain atoms are not equally positioned by crystallography and because not all hydrogen atom positions are fixed. Hydrogen bonding is examined from a side chain at positions *i* to the residues other than those at positions *i-1, i,* and *i+1,* and the average number of hydrogen bonds (dipole interactions) that can be formed by the residue in a given position.
3. **Solvent Accessibility:** The solvent accessible contact area of amino acids is calculated using the method of Lee and Richards [48] with a probe radius of 1.4 *Å*. The percentage of accessible contact area of the total atoms is used. The computed values are presented later in section 5.4.

5.3 Results

This section 5.3 contains the results obtained by our methodology, which is divided into two modules. In Section 5.3.1, we emphasize an efficient partitioning scheme for proteins, and in Section 5.3.2, we provide a coherent subgraph mining technique to identify discriminatory subgraphs for classification purposes. In the following sections we present the results obtained from each of the modules. Section 5.4.1 details the results obtained from our partitioning techniques, and Section 5.4.2 provides the results obtained when the feature vector is used for binary class classification and multi-class classification.

5.3.1 Protein Partitioning

The protein partitioning approach is aimed at dividing the 3-D protein structure into a limited set of compact units that identify structural units within the protein based

on the hydrophobic behavior captured by the summary graphs. Like Tsai et al. [57], we believe that these units' hydrophobic regions have the highest probability of nucleation binding domain protein (PDB ID: 1AN2 (A)), using Algorithm 1.3.

As described by Gelly et al. [64], protein peeling divides a protein into units based on the structure processed by the protein. For validation purposes, we compare the results obtained from Algorithm 1.3 to the 'Protein Peeling' approach. We consider the training proteins used by Tsai et al. [57] as the benchmark, as shown in Table 5.2. We observe similar partitions with respect to the partition's location on the proteins. For a closer look at the location of partitions, we perform an in-depth analysis of the partitioning of a protein that belongs to the family Cytochrome C (PDB ID 1AKK (A)). The partitions obtained on protein 1AAK (A) are compared to the partitions of protein peeling, as shown in Table 5.3. We observe that the partitions are consistent with those reported by [64], with the exception of the gaps reported by our method.

Table 5.2 Partitioning of proteins-dataset using protein partitioning.

Proteins	Tsai et al	Gelly et al	Ours
3cd4	2	7	5
1pph1	3	18	9
2mcm	2	5	6
1bia	3	3	1
1sgt	2	18	11
1atna1	4	7	5
1ccr	1	3	7
1fha	1	6	7
2hhba	1	5	7
2aak	1	9	9
1cus	1	14	12
104la	2	8	5
3pmga1	3	13	8
5ruba2	3	10	4
2aaib1	2	6	5

Table 5.3 Partitions of protein 1AKK (A) of Cytochrome C family.

Protein Peeling	Our Results
1-14	1-12
15-34	14-23
35-59	32-55
60-91	58-78
92-104	93-97

5.3.2 Classification of Proteins

The following experiments are carried out to test our feature vector with regard to the discrimination of proteins belonging to well-known families at the fold and class levels of the SCOP database. Datasets include both balanced and unbalanced populated classes of proteins. We evaluate and enumerate our results that capture characteristics of our feature vector to distinguish proteins when tested with both multi- and binary classes. We use the five-fold cross validation for all our classification schemes. The results are presented below.

5.3.2.1 Binary Class Classification

To test the efficacy of the feature vector on a dataset containing two classes, we choose two well-known datasets, which are shown in Table 5.4. The first dataset C1 obtained from [19] is unbalanced, consisting of distinctly related proteins from all-α class nuclear receptor ligand-binding domain proteins (NB, 16 proteins of typical length ranging between 210 to 260 residues each), against the prokaryotic serine proteases family (PSP, 10 proteins each of length averaging between 190 to 250 residues long) from the all-β classes of proteins. The second dataset C2 consists of proteins from the eukaryotic serine proteases family (ESP, 19 proteins of length between 200 to 260 residues on average) and the PSP family, belonging to the same class of all-β proteins.

Both datasets (C1 and C2) contain proteins filtered under 60% pair-wise sequence similarity to remove highly homologous proteins, with a resolution of <=3 and an R factor of <= 1.0. The datasets can be obtained from the "culled PDB list[26]". We use the Random Forest classification schema [21] on both datasets.

Table 5.4 Comparison of results of binary classification.

Dataset	Method	Features	Accuracy (%)
C1	DT	20646	100
	AD	23130-37394	96-100
	LFM-Pro	5282	100
	Proposed method	38	100
C2	DT	15895	95
	AD	18491- 32569	93-95
	LFM-Pro	2180	100
	Proposed method	29	96.55

Our methodology captures fewer discriminatory features and is more accurate than methods in [65] and [66]. As reported in Table 5.4, the length of our C1, feature vector is 38, and the length of our C2 feature vector is 29. These features represent the number of frequent coherent subgraphs augmented with additional features such as the relative amino acid composition of the coherent subgraphs, the connectivity of the summary graph (see Figures 5.9), and the number of subgraphs extracted from the summary graph. Note that the results reported correspond to the five-fold cross validation accuracy.

We use the Random Forest classification scheme in our experiments, as it offers several distinct advantages for our application. Random Forest is efficient for datasets

[26] http://dunbrack.fccc.edu/Guoli/pisces_download.php

with a large number of input variables. It can internally generate an unbiased estimate of the generalization error of the classifier scales and can balance errors in unbalanced datasets (see [23] for an excellent introduction to Random Forest). However, we are also interested in exploring the relationship of the efficacy of our feature set with the nature of the classifier employed.

To calibrate the performance of the feature vector with other classification schemes, we use the six well-known classifiers shown in Table 5.5. The dataset consists of two classes of proteins belonging to the cytochrome C fold (all-α class) and the ubiquitin fold (α+β class).

Table 5.5 Efficacy of the feature vector.

Classifiers	Accuracy (%)
Naïve Bayes	74.28
Logistic	80
Random Forest	80
K-NN (HEOM)	88.57
SVM (Polynomial)	100
C-SVC (Linear)	100

The consistency of the results obtained across the different classifiers[27] is indicative of the accuracy of the classification, which is not deterred by the nature of the classifier used.

5.3.2.2 Multi-Class Classification

The proposed multi-class classification scheme is an extension of our proposed binary class classification scheme. The choice of frequent subgraphs across the classes in the dataset is carried out as a combination of classes considered pair-wise. We choose

[27] Weka data mining suite (http://www.cs.waikato.ac.nz/ml/weka/).

those subgraphs that are common to all possible pair-wise combinations of classes and filter them based on their discriminative power across the dataset. A detailed description of the way subgraphs are chosen is found in Section 5.4. Our dataset consists of 106 proteins belonging to three structural classes, namely all-β, α/β, and α+β of the ASTRAL SCOP 1.71 database with less than 40% pair-wise identity. Table 5.6 shows the database breakdown.

Table 5.6 Multi-class classification dataset.

Structural Class	Folds	No. of proteins	Precision (%)
All Betas	Immunoglobulin-like beta-sandwich (IgFF)	38	86.8
Alpha/Beta and Alpha + Beta	CI-2 family of serine protease inhibitors, beta-Grasp (Ubiquitin-like) and Nucleotide-diphospho sugar tranferase (N)	33	87.1
All Beta	Trypsin-like serine proteases (TSP)	35	89.2
Overall		106	87.73

We consider two important fold classes of all-β proteins. The first fold class consists of 38 proteins of the immunoglobulin-like beta sandwich class of proteins (IgFF). Each protein is composed of 260 to 300 residues. The proteins from this fold exhibit a wide heterogeneity in terms of tissues and species distribution or functional implications. The domains of these proteins are far more conserved than their sequences. The second fold class of the all-β family consists of 35 Trypsin-like serine proteases proteins. The Trypsin-like serine proteases fold (TSP) has smaller than average surface areas, smaller radii of gyration, and higher Cα atom densities (approximately 238 residues in length on an average). These findings imply that proteases are, as a group,

more tightly packed than other proteins, as also evidenced in [14]. There are also notable differences in secondary structure content between the folds of these proteins.

We introduce the third random class of proteins for classification, taking into account the local bias caused by binary class dataset. This third class consists of proteins chosen at random from an unrelated structural class of proteins. In order to reduce the effect of this class on classification results, we ensure that there is no structural uniformity among these proteins. This lack of uniformity results in a class of 33 proteins, of average length 160 residues each, belonging to both the α/β and $\alpha+\beta$ structural classes. All the proteins of the dataset satisfy the criteria of < 40 % of sequence identity.

An overall accuracy of 87.73% is reported using the Random Forest classification scheme. Individual class precisions are reported in Table 5.6. In our scheme, 33 of the 38 proteins of immunoglobulin-like beta sandwich class (IgFF) are correctly classified. Table 5.7 shows the confusion matrix. Similarly, 33 of the 35 Trypsin-like serine proteases fold (TSP) proteins are correctly classified. The area under curve (AUC) of the corresponding ROC plots of Figure 5.10, are shown in Table 5.7. From the ROC and AUC for each class, we conclude that the classifier distinguishes the proteins of the three classes in the dataset.

Table 5.7 Confusion matrix.

IgFF	N	TSP		Area Under Curve (AUC)
33	4	1	IgFF	0.93
3	27	3	N	0.928
2	0	33	TSP	0.986

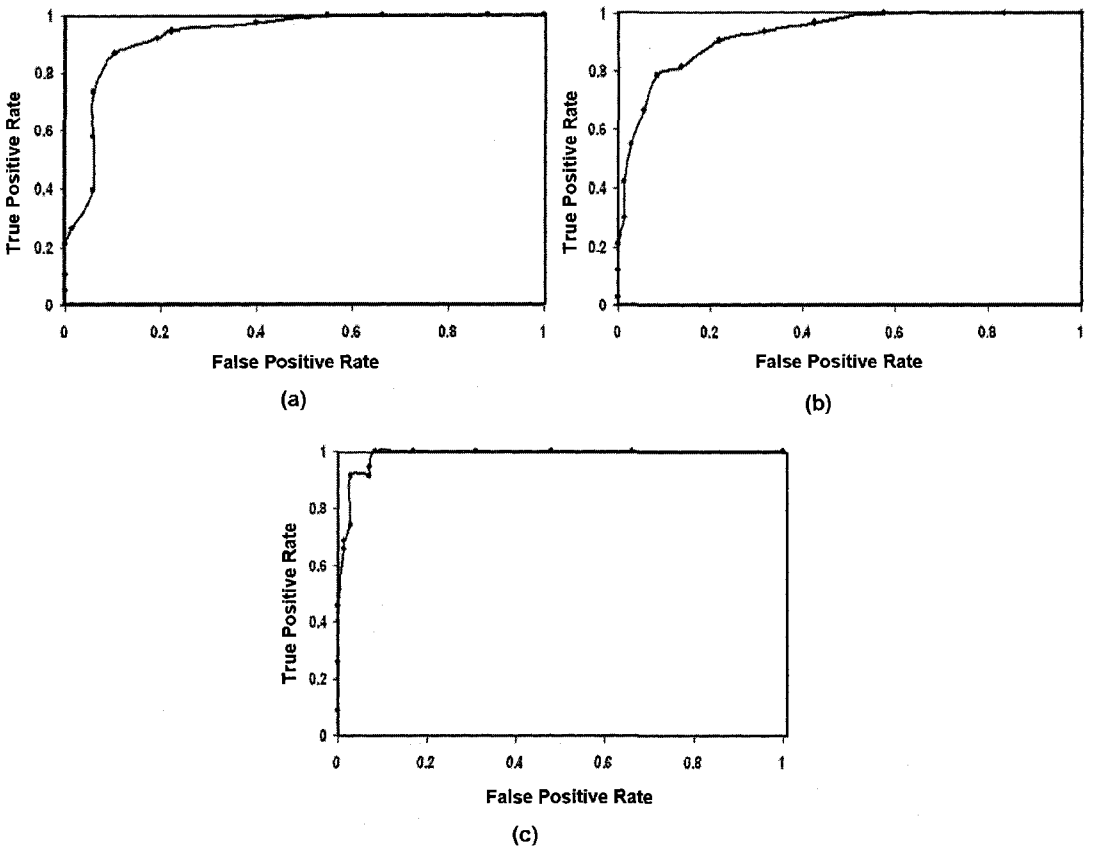


Figure 5.10 ROC Analysis using Random Forest classifier.

5.4 Discussion

5.4.1 Frequently Occurring Subgraphs

One of our objectives is to validate the accuracy of the reported frequently occurring subgraphs to discriminate between homologous proteins. To this end, we compare the residues reported by our method to those reported by Reddy et al. [56]. The proteins used in our study are located in the fssp-ckaaps-1.2 database[28] provided in [56] and belong to three structural protein classes: coiled coil, all- β, and α+β.

We have selected ten proteins from each class, resulting in a dataset consisting of 30 proteins which satisfy a RMSD of <=3.0 and a Z-score of >=4.5. We report a total of 141 coherent subgraphs from the protein database. These are further narrowed by choosing subgraphs that are common to all possible pairs of classes in the dataset and that satisfy a minimum threshold of DP. The subgraphs common to all combinations of two classes have the discriminative power to differentiate proteins in all the classes of the dataset.

Subgraphs and their residue locations, as shown in Table 5.8, are selected only if they satisfy a DP >= 0.1 on a scale of 0 to 1. We perform the multi-class classification with five-fold validation using the Random Forest Classifier, which yields commendable individual class accuracy and an overall accuracy of 90%, as shown in Table 5.9. To investigate individual residue contribution, we study the proteins that belong to the coiled coil class. We choose those subgraphs that posses the highest discriminative power (see Table 5.8). From Table 5.8, we observe that subgraphs 23 and 58 posses the highest discriminative power (0.4).

---

[28] ftp://ftp.sdsc.edu/pub/sdsc/biology/ckaap

Table 5.8 Coherent subgraphs.

Subgraph Index	Discriminative Power	Residue Locations
19	0.2	8, 9, 10
23	0.4	22, 23
24	0.2	27, 28
25	0.1	29, 30
31	0.2	19, 20, 21
33	0.2	34, 35, 36
54	0.2	19, 20
56	0.2	26, 27
58	0.4	37, 38
62	0.2	11, 12, 13
65	0.2	29, 30, 31

Table 5.9 Results of multi-class classification.

Structural Class	Precision (%)
Coiled Coil Proteins (A)	100
All Beta Proteins (B)	90.9
Alpha/Beta Proteins (C)	81.8
Overall Accuracy	90

Our results obtain a higher rank (80% of the proteins report highly ranked residues at locations 22, 23, 37, and 38) than CKAAPs [56]. We present the residues, their respective locations in protein sequences, and their associated CKAAPs ranks above, as shown in Table 5.10. We observe that though not all the proteins report conserved residue locations in CKAPPs, our results indicate that the residues at location 37 are more conserved than others. The analysis of the hydrophobic propensities of the residues across all the proteins reveals that residues at location 22 exhibit conservation of hydrophobic residues. Similarly, the residues at location 37 exhibit conservation of hydrophilic propensity.

Table 5.10 CKAAPs alphabetical rank scores.

Protein	22		23		37		38	
1fe6c	F	Ser	-	Leu	A	Ile	-	Ile
1g2ci	H	Ser	-	Leu	A	Leu	-	Lys
1czqa	-	Glu	G	Ile	B	Ile	-	Lys
2dgca	-	Met	B	Lys	-	Tyr	-	His
1qbza	-	Arg	-	Gln	-	Leu	K	Gln
1ci6b	-	Glu	-	Asn	G	Leu	J	Ser
1fe6a	B	Arg	-	Tyr	A	Leu	-	Glu
1a02j	-	Ala	-	Arg	E	Glu	A	Leu

5.4.2 Structural Environment of Conserved Hydrophobic Residues

In this study, we consider the proteins used by Paiardini et al. [55]. We compare the proteins of Trypsin-serine protease fold superfamily (TSP) to proteins from the (PLP) family, which consists of 23 proteins belonging to the 1-PLP-dependent enzymes superfamily (PLP) fold type. These proteins exhibit high structure conservation despite low sequence similarity.

To evaluate the structural environment of the conserved residues of the summary graphs of individual proteins, we compare the various environmental parameters for the conserved residues against all naturally occurring residues of the protein. As seen in Figure 5.11, the total hydrogen-bonding interactions of the conserved residues are proportional to the hydrogen-bonding interactions in the naturally occurring proteins. This plot reflects the proportional decrease of the charged group of conserved residues when compared to the overall residues in the protein. It thereby captures the integrity of the proteins in the dataset. The Ooi values (Figure 5.11) indicate that the conserved

residues are significantly more buried than naturally occurring residues. Finally, the solvent-accessible contact area, as seen in Figure 5.11, of the conserved residues does not show much difference compared with amino acids in a naturally occurring protein.

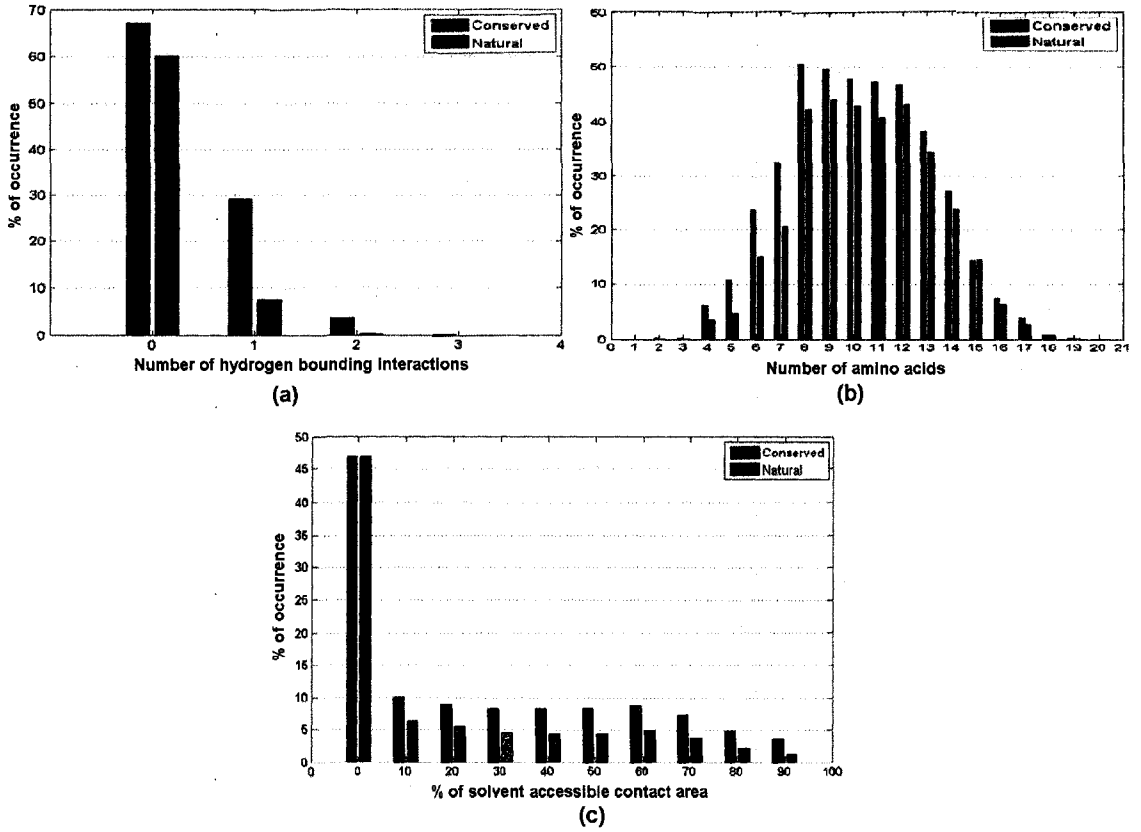


Figure 5.11 Comparison of summary graph and protein representative set.

We extract a set of 25 key positions that possess a discriminative power greater than 50%. Of those extracted, 23 conserved positions across the PLP family of proteins match the conserved residue positions reported by Paiardini et al. The matching 23 positions and their reported conservation scores are listed in Table 5.11. The biological significance of the reported conserved residue locations reported by Paiardini et al. state that positions 8, 12, 11 and 97 are involved in one of the strongest conserved

hydrophobic contacts (CHC). The position 133 of SCR- β 17, occupied mainly by the Arg residue, is the center of a cluster of interacting residues.

All residues that are in close proximity of position 133 are members of the cluster. Residues 8, 11 of SCR- α 1, and 97 of SCR- α 12 form a hinge between α 1 and α 12. These residues form a vertical strip down each side of the helices that delimit the major domain. Site 97 of SCR- α 12 is engaged in the constitution of the two most extensive hydrophobic contacts measured for the α 1– α 12 hinge. Table 5.11 shows the matching positions and conservation scores of hydrophobic contacts measured for the α 1– α 12 hinge.

Table 5.11 Matching positions and conservation scores.

Conserved residue location	Structurally conserved regions	Residues	Conservation score
5	α 1	QEIERLLKRAEQCAEKQRMEEEE	0.44
8		ALYAVKIVLLFMLVFAMLLAIIF	1.15
9		LAAAAFAKAARGVNAKCLGASKC	0.91
10		QEQEKRHKENDEADAEESKRAER	0.92
11		AWYTVRWWTILLMAHAAVMLLL	0.90
12		YLAFLEFMFTLITVHFEALYEHT	0.43
14	β 2	GRQESETWTDKTEEHDGSGKEYY	-0.46
15		VAVAEEVIIRVACMAYCQGTSIA	0.21
16		NLLAITVITALTVVLIVLVYTVV	0.31
18		YIRPSVPADSGVDNTHPAGGAVS	0.57
45	β 5	PQVTFTCNRLAGIITITVPRVQT	0.56
53	β 6	SGGGGGGGGIIGMIGGMHCGLGG	1.00
54		RDQRTAKRQEAAYVLQEASTRLA	0.40
74	β 10	VLIVATLTYLIILTALSMFVVLA	1.37
75		TVLMSVLFATYACTTTITNVLLI	0.93
76		TVRFFNTTFLAALLFMQFFQHFF	0.56
97	α 12	LLLLMLYYLLFLLLLLLIFAMLL	1.70
103		QKQEHAYVAVIGDYVYQHKDSKR	0.72
104		VLIDMYLIIMNMILRRHAHLTLR	0.60
133	β 17	RRRRRRRRRRRRIEWRRRRLRRS	2.00
134		LLILILLIAIAIYALLLFFFFLL	1.43
136		TLVTLYALCPIIMFHIIPIMFLL	0.95
137		PTGHCAGAHPYAPTRAGTCTGIF	0.67

5.5 Conclusion

A protein folds into globular structures as a response to its surrounding environment, which poses several computational challenges for the determination of causal factors involved in the folding. This (folding) behavior of proteins has been frequently governed by localized hydrophobic residue interactions. To this end, an array of hydrophobicity scales has been developed to determine the hydrophobic propensities of residues under different environmental conditions. These scales act as a relatively untapped reserve of information to provide researchers a unique perspective to observe a protein under different conditions. The similarities and discrepancies among these scales are valuable resources of information for the structural and functional behavior of the protein, and an effective abstraction strategy such as ours can lead to better elucidation of this data for functional assessment.

We have developed a graph-theory based computing framework for the identification of conserved hydrophobic residue interaction patterns using well known scales of hydrophobicity. The framework provides a means to weigh these residue-residue interaction patterns and to identify key discriminatory patterns using mutual information and a discriminative weighing function. We report that these discriminatory patterns are specific to a family of proteins, consisting of conserved hydrophobic residues that can be used for structural classification.

Our results reaffirm our hypothesis that conserved hydrophobic residues are retained in structurally homologous proteins and play a vital role in protein folding. Clearly, the success of the framework relies on three key factors: the efficient representation on the structural characteristics of a protein using Delaunay Tessellations,

the choice of hydrophobicity scales to identify hydrophobic residues, and, finally, the provision of summary graphs that prove useful in integrating information from different scales. The summary graph is vital in capturing interaction similarities across scales, eventually affecting the identification of frequent coherent subgraphs. Typically the efficacy of such a method can be compared with an appropriate random background calculated using different permutations of the given sequences. Our current focus is to propose a novel and effective approach for integrated hydrophobicity profiling and characterization. Future efforts will entail tuning and evaluating the robustness of the approach for datasets with usual sequence reshuffling or permutation, which has disrupted the biological information. We will also explore refinements when other stereo-chemical properties are included in the analysis and evaluate their affects on the integrated framework, including when the underlying biological information has been disrupted.

In conclusion, the proposed framework provides an efficient means to integrate different scales for protein analysis. This study further reinforces, with newer evidence, that the identification of conserved hydrophobic residues is vital to the exposition protein folding and further aids in the functional annotation of proteins and possible mutational studies.

CHAPTER 6

CONCLUSION

The study presented in this dissertation addresses the problem of integrating numerous physico-chemical properties (sequence based) for the structural and functional annotation of proteins. Our aim is to provide a classification mechanism that is computationally inexpensive and dependant on sequence properties.

We have presented three approaches for effective feature extraction. First, in Chapter 3, we provide an in depth look at the vital roles that different hydrophobicity scales play in the folding of protein. Our approach has overcome two major obstacles, the inherent high dimensionality of the data, and the difficulty of generalizing an n-dimensional object to a desired lower dimensional space. We deal with the first obstacle by integrating scales by providing a methodology for coherent feature extraction from the selected scales of hydrophobicity for a protein sequence. Plagued by the problem of unequal cardinality of proteins, our proposed integration scheme effectively handles the varied sizes of proteins. Here we deal with how to choose scales from a known scale space, so that we can obtain higher classification accuracies.

Since our first approach suffers from an inability to integrate two properties at a time, we build our second approach to handle multiple properties simultaneously. In the second approach, we design a schema to handle the integration of multiple physico-

chemical properties. Not limiting our choice of physico-chemical properties to hydrophobicity, we use the scales proposed by [6]. We additionally propose an integration scheme. The objective of this work is to explore this integration approach as a method of identification for conserved domains across homologous families of proteins. Theory states that the contribution of conserved residues over a protein sequence, toward determining the bio-chemical function is obtained by the interactions formed with substrates, cofactors, and other residues [15]. Thus, in Chapter 4, we hypothesize that correlated mutations of physico-chemical interactions between residues reveal residue conservation patterns that are unique to homologous proteins. We create a unique representation scheme known as protein maps for a given protein. These maps are aimed at capturing structural makers across a myriad of physico-chemical properties.

Driven by the need to identify conserved residues among homologous proteins, we further investigate and provide necessary insight to the identification of protein cores in Chapter 5. Inhibited by using features derived from sequential properties alone, we represent the sequence based properties over the 3-D structure of a protein. By using a graph theory-based data mining framework to extract and isolate protein structural features, and by applying a mutual information-based feature extraction technique, we identify those residues that exhibit sustained invariance among homologous proteins. This identification has been performed through the integrated analysis of five well-known hydrophobicity scales over the 3-D structure of proteins.

The methods proposed in Chapters 3, 4, and 5, are complementary in several aspects. All three methods are driven by a common rudiment of using sequence-based properties. Each method is aimed at improving over previous methods. In the method

proposed in Chapter 3, we integrate two physico-chemical properties. In Chapters 4 and 5, we successfully propose an integration scheme that integrates structural and multiple physico-chemical properties in a single instance.

6.1 Future Directions

The ultimate goal of this research is structural and functional annotation which we hoped to achieve by integrating various features. Several further developments can be planned for the near future. Specifically, we plan to use the tools for classifying new sequences and the proposed algorithms that we have discussed in this dissertation to explore the effects of various properties on protein structure. We expect these explorations to add to our understanding of protein properties, thereby, allowing valuable insight into protein evolution, to sequence-structure relationships, and to studies on protein function analysis.

REFERENCES

[1] M. Kanehisa, "Grand challenges in bioinformatics," *Bioinformatics,* vol. 14, no. 4, pp. 309, 1998.

[2] A. Wada, "Bioinformatics-The necessity of the quest for 'First Principles' in life," *Bioinformatics,* vol. 16, no. 9, pp. 663-664, 2000.

[3] M. Kanehisa and P. Bork, "Bioinformatics in the post-sequence era," *Nature Genetics Supplement,* vol. 33, pp. 305-310, 2003.

[4] C. Ouzounis, "Two or three myths about bioinformatics," *Bioinformatics,* vol. 16, no. 3, pp. 187-189, 2000.

[5] P. Adriaans and D. Zantinge, "Introduction to data mining and knowledge discovery," *Data Mining*, M. T. C. C. Potomac, Addison Wesley, 1996, 1999.

[6] M. Venkatarajan and W. Braun, "New quantitative descriptors of amino acids based on multi-dimensional scaling of a large number of physical-chemical properties," *Journal of Molecular Modeling,* vol. 7, no. 12, pp. 445-453, 2001.

[7] J. Cornette, K. Cease, H. Margalit *et al.*, "Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins," *Journal of Molecular Biology,* vol. 195, no. 3, pp. 659-685, 1987.

[8] M. Lienqueo, A. Mahn, and J. Asenjo, "Mathematical correlations for predicting protein retention times in hydrophobic interaction chromatography," *Journal of Chromatography,* vol. 978, no. 1, pp. 71-79, 2002.

[9] K. Biswas, D. DeVido, and J. Dorsey, "Evaluation of methods for measuring amino acid hydrophobicities and interactions," *Journal of Chromatography A,* vol. 1000, no. 1, pp. 637-655, 2003.

[10] J. Kyte and R. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology,* vol. 157, no. 1, pp. 105-132, 1982.

[11] S. Kawashima and M. Kanehisa, "AAindex: Amino Acid Index Database," *Nucleic Acids Research*, vol. 28, no. 1, pp. 374, 2000, [Online] http://www.genome.jp/aaindex/.

[12] M. Wilce, M. Aguilar, and M. Hearn, "Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides," *Analytical Chemistry,* vol. 67, no. 7, pp. 1210-1219, 1995.

[13] H. Berman, J. Westbrook, Z. Feng *et al.* "The Protein Data Bank," Nucleic Acids Research, 2000, vol. 28, no. 1, pp. 235-242. [Online] http://www.rcsb.org/pdb/home/home.do.

[14] E. Stawiski, A. Baucom, S. Lohr *et al.*, "Predicting protein function from structure: Unique structural features of proteases," *Proceedings. of the National Academy of Sciences of the United States of America,* vol. 97, no. 8, pp. 3954-3958, 2000.

[15] W. Kauzmann, "Factors in the interpretation of protein denaturation," *Advanced Protein Chemistry,* vol. 14, pp. 1-63, 1959.

[16] C. Ding and I. Dubchak, "Multi-class protein fold recognition using Support Vector Machines and Neural Networks," *Bioinformatics,* vol. 17, no. 4, pp. 349-358, 2001.

[17] A. Tan, D. Gilbert, and Y. Deville, "Multi-class protein fold recognition using a new ensemble machine learning approach," *Genome Informatics,* vol. 14, pp. 206-217, 2003.

[18] A. Chinnaswamy, W. Sung, and A. Mittal, "Protein structure and fold prediction using tree-augmented Naïve Bayesian classifier," *Pacific Symposium on Biocomputing*, no. 9, pp. 387-398, 2004.

[19] J. Huan, W. Wang, D. Bandyopadhyay *et al.*, "Mining protein family specific residue packing patterns from protein structure graphs," RECOMB '04: *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology*, San Diego, California, USA, pp. 308-315, 2004.

[20] I. Dubchak, I. Muchnik, S. Holbrook *et al.*, "Prediction of protein folding class using global description of amino acid sequence," *Proceedings of the National Academy of Sciences,* vol. 92, pp. 8700-8704, 1995.

[21] P. Welch, "The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio Electroacoustics,* vol. 15, no. 2, pp. 70-73, 1967.

[22] F. Mörchen, *"Time series feature extraction for data mining using DWT and DFT,"* 2003. [Online]. Available: http://www.mybytes.de/.

[23] C. Chen, A. Liaw, and L. Breiman, *"Using random forest to learn imbalanced data,"* 2004. [Online]. Available: http://www.stat.berkeley.edu/tech-reports/666.pdf.

[24] Y. Qi, J. Seetharaman, and Z. Joseph, "Random Forest similarity for protein-protein interaction prediction for multiple sources," *Pacific Symposium on Biocomputing,* vol 10, pp. 531-542, 2005.

[25] Z. Isik, B. Yanikoglu, and U. Sezerman, "Protein structural class determination using Support Vector Machines," *Computer Information Sciences,* vol. 3280, pp. 82-89, 2004.

[26] R. Fan, P. Chen, and C. Lin, "Working set selection using second order information for training SVM," *Journal of Machine Learning Research,* vol. 6, pp. 1889-1918, 2005.

[27] C. Apte, E. Grossman, E. Pednault *et al.*, "Insurance risk modeling using data mining technology," *Proceedings of PADD99: The Practical Application of Knowledge Discovery and Data Mining,* pp. 39-47.

[28] D. Wetlaufer, "Nucleation, rapid folding and globular inter-chain regions in proteins," *Proceedings of National Academy of Science,* vol. 70 pp. 697-701, 1973.

[29] M. Rossman and A. Lijas, "Letter: recognition of structural domains in globular proteins," *Journal of Molecular Biology,* vol. 85 pp. 177-181, 1974.

[30] M. Saraf, G. Moore, and C. Maranas, "Using multiple sequence correlation analysis to characterize functionally important protein regions," *Protein Engineering,* vol. 16 no. 6, pp. 397-406, 2003.

[31] W. Taylor, "Protein structural domain identification," *Protein Engineering,* vol. 12, no. 3, pp. 203-216, 1999.

[32] M. Fujita, M. Itoh, and M. Kanehisa, "Conservation of physicochemical properties during protein evolution," *Genome Informatics Online, Japanese Society of Bioinformatics*, 2004.

[33] A. Heger and L. Holm, "Exhaustive enumeration of protein domain families," *Journal of Molecular Biology,* vol. 328, pp. 749-767, 2003.

[34] N. Nagarajan and G. Yona., "Automatic prediction of protein domains from sequence information using a hybrid learning system," *Bioinformatics,* vol. 20, pp. 1335-1360, 2004.

[35] N. Ohsen, I. Sommer, R. Zimmer *et al.*, "Automatic protein structure prediction using profile-profile alignment and confidence measures," *Bioinformatics,* vol. 20, no. 14, pp. 2228-2235, 2004.

[36] M. Lexa and G. Valle, "PRIMEX: Rapid identification of oligonucleotide matches in whole genomes," *Bioinformatics,* vol. 19, pp. 2486-2488, 2003.

[37] D. Chivian, "Automated prediction of CASP-5 structures using the Robetta Server," *Proteins,* vol. 53 no. 6, pp. 524-533, 2003.

[38] D. Fischer, "Assigning amino acid sequences to 3-dimensional protein folds," *FASEB Journal,* vol. 10, pp. 126-136, 1996.

[39] A. Murzin, S. Brenner, T. Hubbard *et al.*, "SCOP: A structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology,* vol. 247, no. 4, pp. 536-540, 1995.

[40] R. Marsden, "Rapid protein domain assignment from amino acid sequence using predicted secondary structure," *Protein Science,* vol. 11, pp. 2814-2824, 2002.

[41] R. Linding, "GlobPlot: Exploring protein sequences for globularity and disorder," *Nucleic Acids Research,* vol. 31, pp. 3701-3708, 2003.

[42] A. Bateman, "The Pfam protein families database," *Nucleic Acids Research,* vol. 28, pp. 263-266, 2000.

[43] J. Schultz, "SMART: A web-based tool for the study of genetically mobile domains," *Nucleic Acids Research,* vol. 28, pp. 231-234, 2000.

[44] F. Corpet, "ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons," *Nucleic Acids Research,* vol. 28, pp. 267-269, 2000.

[45] D. Rigden, "Use of covariance analysis for the prediction of structural domain boundaries for multiple protein sequence alignments," *Protein Engineering,* vol. 15, no. 2, pp. 65-77, 2002.

[46] A. Natsev, R. Rastogi, and K. Shim, "WALRUS: A similarity retrieval algorithm for image databases," in *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data: SIGMOD '99*, pp. 395-406, 1999.

[47] K. Nishikawa and T. Ooi, "Prediction of the surface interior diagram of globular proteins by an empirical method," *International Journal of Peptide and Protein Research,* vol. 16, no. 1, pp. 19-32, 1980.

[48] B. Lee and F. Richards, "The interpretation of protein structures: Estimation of static accessibility," *Journal of Molecular Biology,* vol. 55, no. 3, pp. 379-400, 1971.

[49] S. Miyazaki, Y. Kuroda, and S. Yokoyama, "Characterization and prediction of linker sequences of multi-domain proteins by Neural Networks," *Genome Research*, vol. 11, pp. 1410-1417, 2002.

[50] S. Black and D. Mould, "Development of hydrophobicity parameters to analyze proteins which bear post or co-translational modifications," *Analytical Biochemistry,* vol. 193, no. 72-82, 1991.

[51] J. Qian, B. Stenger, C. AWilson *et al.*, "PartsList: A web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information," *Nucleic Acids Research,* vol. 29, pp. 1750-1764, 2001.

[52] C. Leslie, E. Eskin, A. Cohen *et al.*, "Mismatch string kernels for discriminative protein classification," *Bioinformatics,* vol. 20, no. 4, pp. 467-476, 2004.

[53] A. Shmygelska, "Search for folding nuclei in native protein structures," *Bioinformatics,* vol. 21, no. 1, pp. 394-402, 2005.

[54] W. Kauzmann, "Factors in the interpretation of protein denaturation," *Advanced Protein Chemistry,* vol. 14, pp. 1-63, 1959.

[55] A. Paiardini, F. Bossa, and S. Pascarella, "Evolutionarily conserved regions and hydrophobic contacts at the superfamily level: the case of the fold-type I, pyridoxal-5'-phosphate-dependent enzymes," *Protein Sciences,* vol. 13, pp. 2992-3005, 2004.

[56] B. Reddy, W. Li, I. Shindyalov *et al.*, "Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins," *PROTEINS: Structure, Function and Genetics*, vol. 42, pp. 148-163, 2001.

[57] C. Tsai and R. Nussinov, "Hydrophobic folding units derived from dissimilar monomer structures and their interactions," *Protein Science,* vol. 6, no. 1, pp. 24-42, 1997.

[58] U. Muppirala and Z. Li, "A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues," *Protein Engineering, Design & Selection,* vol. 19, no. 6, pp. 265-275, 2006.

[59] E. Huang, S. Subbiah, and M. Levitt, "Recognizing native folds by the arrangement of hydrophobic and polar residues," *Journal of Molecular Biology,* vol. 252, pp. 709-720, 1995.

[60] B. Krishnamoorthy and A. Torpsha, "Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations," *Bioinformatics,* vol. 19, no. 2, pp. 1540-1548, 2003.

[61] Z. Bagci, R. Jernigan, and I. Bahar, "Residue packing in proteins: uniform distribution on a coarse-grained scale," *Journal of Chemical Physics,* vol. 116, no. 5, pp. 2269-2276, 2002.

[62] T. Taylor and I. Vaisman, "Graph theoretic properties of networks formed by the delaunay tessellation of protein structures," *Physical Review E,* vol. 73, no. 4, pp. 041925-1-041925-13, 2006.

[63] R. Tarjan, "Depth-first search and linear graph algorithms," *SIAM Journal on Computing,* vol. 1, no. 2, pp. 146-160, 1972.

[64] J. Gelly, A. Brevern, and S. Hazout, "'Protein Peeling': An approach for splitting a 3-D protein structure into compact fragments," *Bioinformatics,* vol. 22, no. 2, pp. 129-133, 2006.

[65] H. Hu, X. Yan, Y. Huang *et al.*, "Mining coherent dense subgraphs across massive biological networks for functional discovery," *Bioinformatics,* vol. 21, no. 1, pp. i213-i221, 2005.

[66] A. Sacan, O. Ozturk, H. Ferhatosmanoglu *et al.*, "LFMPro: A tool for detecting significant local structural sites in proteins," *Bioinformatics,* vol. 23, no. 6, pp. 709-716, 2007.