University of Alberta

Computational Prediction of Three State Secondary Structure for Protein Structural Fragments

by

Kanaka Durga Kedarisetti ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Master of Science

Department of Electrical and Computer Engineering

Edmonton, Alberta

Fall 2005

Dedication

For my beloved Baba, who inspired me to complete this MSc program on time, provided me with a good professor for guidance and a great loving family.

Abstract

Current alignment based secondary protein structure prediction methods are close to reach accuracy limit and recent research shows that consensus methods that utilize several complimentary prediction methods are the future. To this end, a novel classification problem related to the structure prediction in three states for protein structural fragments (SF) is considered. This thesis includes investigation of a novel attribute based sequence representation that improves ability to distinguish between the structures, analysis of relation between certain sequence properties, performance comparison for several prediction algorithms and finally attribute selection for prediction of SFs. Based over 50000 experiments and using carefully prepared protein data, the results show that on average 50% error rate reduction, when compared to prediction using standard protein representation can be achieved. This research provides useful guidelines for design of prediction methods not only for structure, but also more universally for structural class and content prediction tasks and for computer assisted molecular design.

Acknowledgements

I would like to express my grateful appreciation to all the people who helped me throughout my M.Sc. program. Especially, I would like to thank my advisor Dr. Lukasz Kurgan for his guidance and helpful discussions. I was extremely lucky to find an advisor whose insight and experience were invaluable to complete this work. Next, I would like to thank Dr. Mrinal Mandal for giving me a chance to pursue my M.Sc program. In addition, I would like to thank Dr. Marek Reformat for providing me a computer to do experiments for my research. I would like to thank the research members of my advisor for their helpful discussions, especially Leila Homaeian and Wojciech Stach. Also, I would like to thank all the members of my thesis committee and the chair of the Electrical and Computer Engineering Department for their valuable advises. A special thanks to my husband Babu for supporting me all the way through my M.Sc. program and my kids Priyanka and Pradyuma for their loving co-operation.

Table of Contents

List of Tables

List of Figures

Abbreviations

Abbreviation	Expansion
1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
AA	Amino Acid
SF	Structural Fragment
PDB	Protein Data Bank
DSSP	Dictionary of secondary structure prediction
CASP	Critical Assessment Structure Prediction- CASP4 took place in year 2004
CAFASP	Critical Assessment of fully automated Structure Prediction-taking place every two years
ML	Machine Learning
MLP NN	Multi Layer Perceptron Neural Network
NB	Naïve Bayes
SLI	Slipper
RIP	Ripper
bC5.0	Boosted C5.0
DA	Dataset of all SFs
DA-s, DA-m, DA-l and DA-vl	Datasets of short, medium, large and very large size SFs
DA-2, DA-4 and DA-6	Datasets of all SFs but removed 2,4 and 6 AA on the SF ends
D1, D2, D3 and D4	Datasets of first, first two, first three and first four SFs
d1, d2, d3, d4 and d6	Datasets of first, second, thirds, fourth and sixth position of SFs

1 Introduction

Knowledge of protein structures is a key to understand the protein functions and their interactions with other molecules. Research in protein function and interactions is based on tertiary structure, which in turn depends on secondary structure conformation. Secondary structure prediction continues to be of significant impact due to large gap between number of known protein sequences and number of proteins for which the secondary and tertiary structure is known. For instance, NCBI database contains approximately 2 millions different proteins and SWISS-PROT stores over 150000 of high quality annotated protein sequences (Boeckmann et al., 2003), while only about 30000 protein secondary and tertiary structures stored in the Protein Data Bank (PDB) are known (Berman et al., 2000). Experimental methods for discovery of tertiary structures, such as X-ray crystallography and nuclear magnetic resonance spectroscopy, are time consuming, labor expensive, and cannot be applied to some proteins (Ganapathiraju et al., 2004). Computational methods for predict tertiary structure with an intermediate step of predicting secondary structure, which is usually classified in three states: helix, strand and coil became significant. Their main advantages are low cost, error-free repeatability, and relatively fast delivery of results, but they suffer from relatively low prediction quality. Given slow pace of learning new tertiary structures by experimentally (only a few thousands per year), development of reliable computational methods is of paramount importance.

Currently, existing computational structure prediction methods use multiple alignments, which is based on observation that proteins with similar protein sequences have similar secondary structure. While a query sequence can be aligned with high confidence to a set of sequences with known structure, the alignment will generate predicted secondary structure with average accuracy of no more than 88%, which is due to natural variations observed in structural families (Rost et al., 1994; MacCallum, 1997). Current alignment based protein secondary structure prediction methods achieve accuracy around 80% and soon they may reach accuracy limit. Since majority of prediction approaches are based on multiple alignment, research in methods that do not utilize this information is a viable alternative to design complementary prediction methods. Before new prediction method can be developed, solid foundations and possible architectures need to be researched and developed.

To this end, instead of following the common track of "blindly" improving accuracy of current alignment based structure prediction methods, this research defines a new problem. The problem aims to answer how the three secondary structures can be distinguished based on the primary sequence information and without using alignment. Here, protein sequences are divided into three sets of structural fragments, i.e. for helix, strand, and coil fragments. A classification model is build for each of the sets and ability of these models to distinguish between structures and correctly classify each fragment is evaluated through strict statistical tests. This research draws on several research papers to develop new and improved representation of protein sequence, to investigate which factors and prediction algorithms result in improving ability to distinguish between the different structures, and finally to propose new architectures for prediction of protein structure.

Past results show that simple representation of the sequences based on propensities of the protein residues and probability-based classification gave results for structure prediction about 60% to 66% accuracy (Chou & Fasman, 1978; Garnier et al., 1978; Gibrat et al., 1987; Rost & Sander, 1994; Rost et al., 1994). This research shows that for the analogous problem of classification of structural fragments using the same representation and similar classification methods gives about 65% accuracy. Where as improving the sequence representation and grouping structural fragments of similar length results in significant improvement in the accuracy. For the same classification methods, accuracy of over 84% was achieved, reducing the error rates by 50%, which shows that design on high quality non-alignment structure prediction methods is possible. In addition, some prediction algorithms are shown to produce superior prediction results, and thus selection of the proper algorithm is important to achieve the best possible results. We note that the research does not describe a new or compare with existing structure prediction methods, but rather analyzes different aspects of the related problem of structural fragment prediction. This problem allows addressing fundamental difficulties for all protein prediction tasks, including structure, structural class and content prediction, and since it is solved with high accuracy, it provides invaluable source of useful information.

This chapter includes a short introduction to computational biology, an overview of protein structures and protein databases, description of the related work, motivation and problem definition, overview of the research goals and finally organization of the remaining part of the thesis.

1.1 Computational Biology

Computational Biology is defined as the development and application of data by analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems (NIH, 2000). The methods have their origins in various scientific disciplines including those in physics, chemistry, engineering, and computer science. One of the promises of Computational Biology is the ability to manipulate vast amount of data in a time span not possible to achieve by traditional experimental techniques. An important sub area in Computational Biology is Molecular Biology.

There have been many fundamental changes in Molecular Biology research in the recent years due to spectacular advances in Genomics as well as computer technologies. New initiatives are now taking shape after the completion of genome sequencing projects, such as structural genomics, functional genomics or proteomics. There is now a shift in emphasis - from sequence to structure, from genes to proteins and their complexes, protein-protein interactions and post-transitional modifications.

Furthermore there is a need for characterizing structure and function at different levels to establish the link between genotype (internally coded, inheritable information) and phenotype (outward, physical manifestation). Experimental research should be conducted in coordination with theoretical and computational methods that allow for high throughput analysis and organization of biological data.

Molecular Biology is a field that encompasses a wide range of topics, ranging from molecular modeling to large-scale analysis of genome/ proteome data. For many years, an important and heavily explored problem in Molecular Biology is prediction and analysis of protein structure. Primary protein structure (protein sequence) consists of a sequence of protein residues called amino acid (AA) monomers. During synthesis protein sequence first folds into different secondary structures, which subsequently form a three dimensional (tertiary) molecule. Computationally identification of protein tertiary structure from protein sequence (Primary structure) is inaccurate. Hence, protein secondary structure prediction is used as an important intermediate step for predicting tertiary structure, protein function and protein structural change, as well as for computer-assisted molecular design (Truhlar et al., 1999). The molecular design is a basis for rational drug design and development of novel treatments for many diseases, especially genetic diseases such as cancer, cystic fibrosis, and autoimmune disorders.

1.2 Overview of Protein Structures

Proteins are required for the structure, function and regulation of the body cells, tissues and organs. Hormones, antibodies and hemoglobin are just few examples of proteins. Protein is a large bio-molecule consists of one or more amino acid (AA) chains. The order of the AA's in a protein chain and the properties of their side chains determine the 3D-structure and function of the protein.

1.2.1 Amino acids

An AA, also known as protein residue, is composed of a constant chemical group and a variable amine group as shown in Figure 1. Hence, AA chains have the same backbone structure and only differ in their side chains. Some AAs are hydrophobic, which means that they have water fearing side chains, which tend to turn themselves inward in the interior of the protein. In contrast, hydrophilic side chains tend to turn outward, to the exterior of the protein. Hydrophobic side chains can participate in hydrogen bonding. Side chains can also be charged.

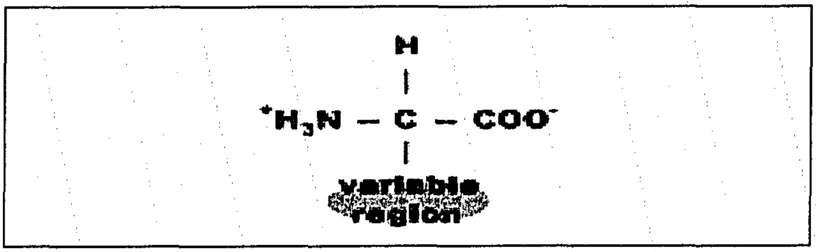


Figure 1.General Structure of an Amino acid (Bruce, 1994)

Positively charged and negatively charged side chains can attract each other forming a salt bridge. Interaction of these properties allows a chain of the AAs to fold into a unique, reproducible 3D structure. Twenty common AAs are responsible for forming proteins, which are shown in Figure 2. They are classified into four main families based on their chemical properties such as acidic, basic, uncharged polar and non polar (Bruce, 1994).

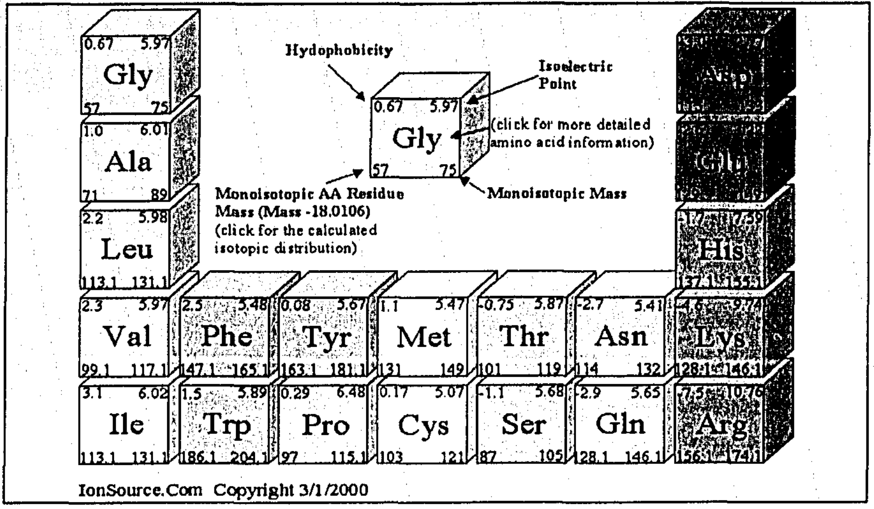


Figure 2. Amino acid chart

AAs are denoted by a single letter code in protein sequence as given in Table 1.

Table 1. Amino Acids

Amino Acid	3-Letter Code	1-Letter Code
Alanine	Ala	A
Cysteine	Cys	C
Aspartate	Asp	D
Glutamate	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	The	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

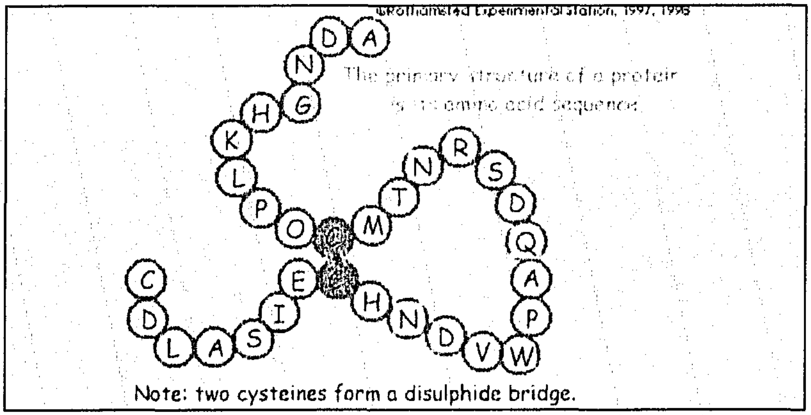
1.2.2 Protein Structures

Protein structure is usually explained in terms of a structural hierarchy, from primary to quaternary.

Primary structure: Refers to the linear sequence of AAs in a protein chain as shown in Figure 3. Protein primary structure also referred as primer, protein sequence, primary sequence or one-dimensional (1D) protein structure.

Example-1.1: For 1FV5A protein from PDB the protein sequence is:

```
GSLLKPARFMCLPCGIAFSSPSTLEAHQAYYCSHRI
```



©Rothamsted Experimental Station, 1997, 1998

The primary structure of a protein is its amino acid sequence.

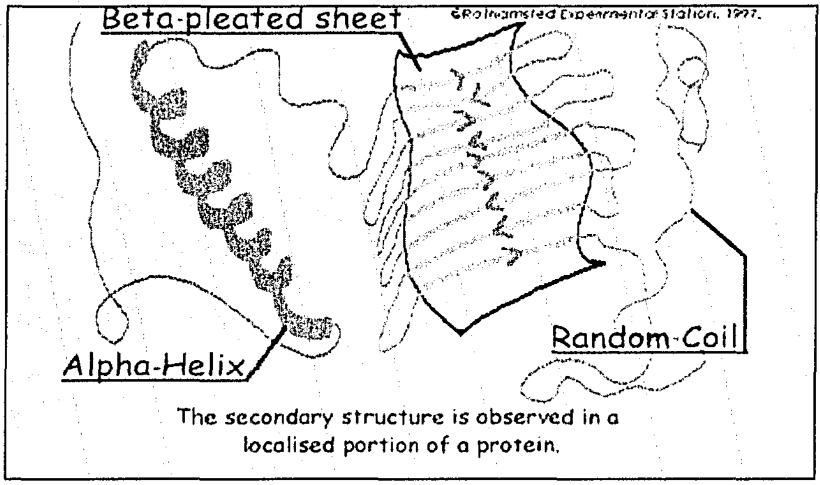
Note: two cysteines form a disulphide bridge.

Figure 3. Primary Structure

Secondary Structure: Refers to folding AA chain into one of the three states - helix or strand or coil as shown in Figure 4. Protein secondary structure is also referred as protein three-state structure, protein secondary sequence or two-dimensional (2D) protein structure. Secondary structure of a protein is determined by assigning one of the three states (Helix, Strand, Coil) to each AA in the protein sequence. Each state is represented by a single letter such as H= Helix, E= Strand and C=Coil.

Example-1.2: For 1FV5A protein from PDB the secondary sequence is:

CCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCC



©Rothamsted Experimental Station, 1997.

The secondary structure is observed in a localised portion of a protein.

Figure 4. Secondary Structure

Tertiary structure: The tertiary structure refers the way secondary structure fold with respect to each other to form a protein as a whole molecule, which is shown in Figure 5. It is also known as three-dimensional (3D) protein structure. AAs, which are very distant in the primary structure, might be close in the tertiary structure, because of the folding of the chain.

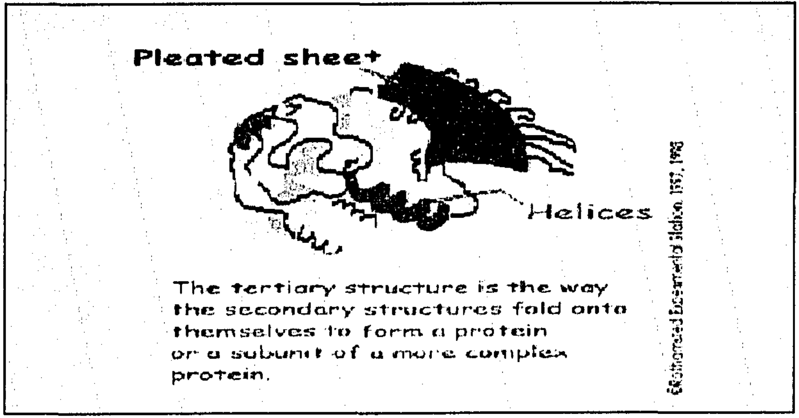

Pleated sheet
Helices
The tertiary structure is the way the secondary structures fold onto themselves to form a protein or a subunit of a more complex protein.

Figure 5.Tertiary Structure

Quaternary structure: Quaternary structure refers the association of one or more AA chains into a multi-subunit structure as shown in Figure 6.

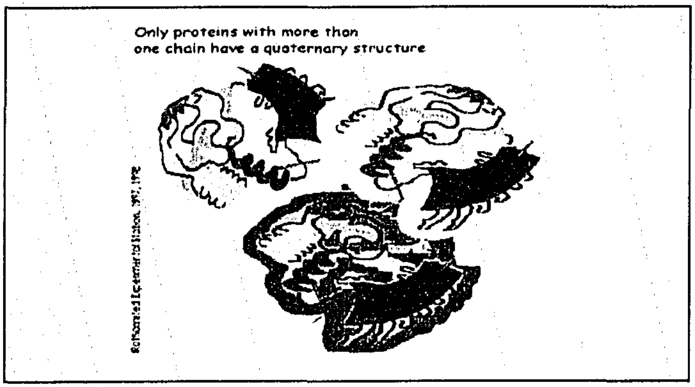

Only proteins with more than one chain have a quaternary structure

Figure 6.Quaternary Structure

1.2.3 Protein Structure Units

Domain: A Domain is a structurally distinct fragment of the protein sequence. Domain within a protein often performs different functions, and can have completely different structures and folds. Protein domains are typically 100 to 400 residues long. A protein can have one to many (10-12) domains connected by loops (coils).

Folds: A protein fold is the tertiary structure of a protein domain, i.e. the order and spatial relationship of the secondary structure elements, which form the domain. Fold is described by the number, order and relative position. The number of distinct fold appears to be limited, and is about a thousand.

Motif: A motif is a pattern of secondary structures that can be found in various proteins. For example, motifs are the coiled coil (two helices twisted around each other) and the helix-loop-helix pattern.

Structural Fragment (SF): A structural fragment is a fragment of protein sequence that has uniform secondary structure state.

1.2.4 Protein Structure Forces

The protein structure and form is a result of numerous physical forces between individual AAs and atoms (Brändén, 1991). Main forces are listed below:

Hydrogen Bonds: Hydrogen bonds occur when a pair of nucleophilic atoms such as oxygen and nitrogen shares hydrogen between them. Hydrogen bonds are directional and their strength deteriorates rapidly with changes in angle, such that they control and limit the geometry of interactions between side-chains. The pattern of hydrogen bonding is essential in stabilizing basic secondary structures such as α-helix and β-strand. Because of this, protein tends to form hydrogen bonds to maximize the number of hydrogen bonds.

Hydrophobic Effect: Proteins are composed of AAs that contain side chains of hydrophobic and hydrophilic type. In aqueous solutions, hydrophobic residues tend to concentrate towards interior of the protein due to fear of water, while hydrophilic residues stays at the surface in contact with water. This confirmation is energetically favorable. Each exposed hydrophobic residue disrupts the pattern of hydrogen bonding and destabilize the structure. The hydrophobic force is one of the strongest determinants of protein structure.

Van der Waal forces: Van der walls forces are the interactions between immediately adjacent atoms, i.e. atoms nucleus and its neighbor's electrons. They minimize the distance between atoms. Van der Waals interactions stabilize the central hydrophobic core of proteins.

Electrostatic forces: Oppositely charged side chains can form salt–bridges, which pull chains together. Charged side chains can inhibit certain folds. These electrostatic forces are relatively strong, and stabilize secondary and tertiary structure.

1.2.5 Relation between Structure and Sequence

The following points are considered to analyze the relation between structure and sequence:

Nature of the side chains: 20 AAs have different side chains that exhibit a large variation in size, polarity, charge, hydrophobicity, etc. These properties and above-mentioned forces will impact the confirmation of the protein.

Secondary structure propensities: Energy calculations and statistics shows different residues prefer helices versus strands. Such preferences are determined by analyzing the composition of α-helices and β-strands from a large sample of known structures.

Additionally, protein homology plays an important role in protein structure, i.e. two or more closely related proteins may show high sequence homology and it can be detected by sequence alignment methods. However, there are strong evolutionary relationships between proteins. Proteins can be related and have the same overall structure, even if they have 25% or less sequence homology. This observation supports that protein sequence diverged during evolution in a way that allows the conservation of structure, as well as function, while allowing sequence to change. Thus, protein structure is more highly conserved than sequence, and it is important to be able to classify proteins based on their structure. A classification scheme often used for this purpose is known as structural classification of proteins.

1.2.6 Experimental Determination of Protein Structure

As structures are solved, they are stored in a centralized database called the Protein Data Bank (PDB) (Berman et al., 2000). Two main techniques are used to determine the structure of a given protein: X-ray Crystallography and Nuclear Magnetic Resonance (NMR). Both methods are relatively slow, labor-intensive, and expensive. They are dependent on used resolution and equipment.

1.2.6.1 X-ray Crystallography

X-ray Crystallography is a technique that finds accurate 3-D positions of the atoms in the protein molecule. The protein is first isolated and purified to yield a high-concentration solution. This solution is then used to grow crystals containing the protein 'frozen' in one conformation. The resulting crystal is then exposed to an X-ray beam that diffracts regularly placed molecules in the crystal. The diffraction pattern represents the Fourier transform of the electron density in the molecule. Since electron density is higher near atoms, this information, in conjunction with the protein sequence, determines the position of the atoms within the protein and finally its structure.

A major disadvantage of X-ray Crystallography is the need to crystallize the protein, which is the most difficult task. In addition, the crystalline structure of a protein may be different from its structure in vivo. Hence, multiple maps may be needed for consensus. Many of the first structures of proteins were determined using this technique.

1.2.6.2 Nuclear Magnetic Resonance (NMR)

NMR finds atomic structure of proteins in solution rather than crystallizing the protein. A spinning tree of certain atomic nuclei generates a magnetic moment – NMR measures the energy levels of such magnetic nuclei. These levels are sensitive to the environment of the atom- what they are bonded to, which atoms they are close to spatially, what the distances are between different atoms etc. By careful measurement, the structure of the protein can be constructed.

A disadvantage of NMR is the constraint on the size of the protein. It also requires a purified protein. Finally, the process needs to be done at an unrealistic pH, while protein structure is very sensitive to pH.

Given the long and labor-intensive nature of these experiments, it would be extremely valuable to have good computational methods to predict the structure of a protein based on its sequence to reduce the sequence-structure gap.

1.3 Protein Databases

Currently, a variety of protein databases exists. Simple sequence repositories store data with little or no manual intervention in the creation of the records. Expert curate databases include the original sequence data enhanced by manual addition of further information. Due to move in research from the genome to the proteins encoded by it, the databases will play an even more important role as central comprehensive resources of protein information. The databases can be divided into *primary* and *derived* databases. Primary database records are generated from original data with or without manual intervention, whereas derived databases are derived computationally from primary databases sources as shown in Figure 7.

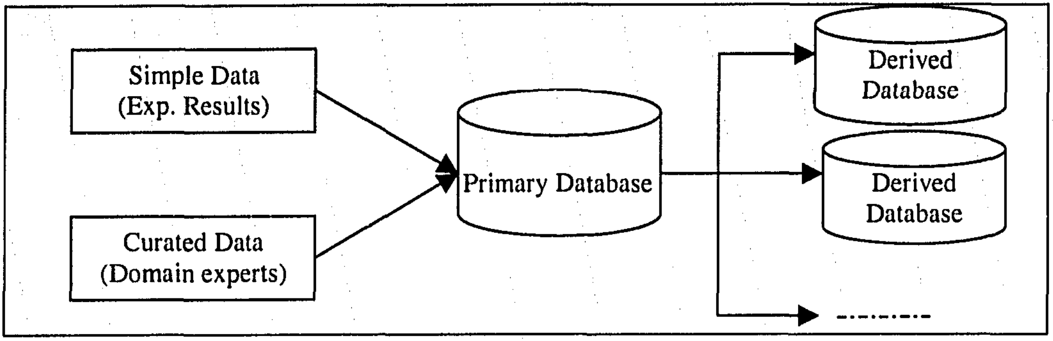


Figure 7.Overview of Protein Databases

For instance, AA as well as nucleotide sequences, protein structure and binding data are stored in the primary database. Alignments of AA sequences, protein models and phylogenetic trees are stored in derived database. One of the main sources of the primary database is PDB. A number of PDB derived, specialized secondary structure databases was created, e.g. DSSP (Kabsch & Sander 1983), PDBFINDER (Hooft et al., 1996), PDBSELECT (Hobohm & Sander, 1994), SCOP (Murzin et al., 1995; Andreeva et al., 2004), and BLOCKS (Henikoff et al., 1991). Some of the most widely used protein databases are discussed next.

1.3.1 Primary Databases

1.3.1.1 SWISS-PROT and TrEMBL

SWISS-PROT protein knowledge base was started in 1986 by Amos Bairoch in the Department of Medical Biochemistry at the University of Geneva and developed by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (Rodriguez-Tome, 1996). This database is generally considered as one of the best protein sequence databases in terms of the quality of the annotation. SWISS-PROT is a curated protein sequence database, which means that groups of designated curators (scientists) prepare the entries from literature and/or contacts with external experts. SWISS-PROT strives to provide a high level of annotation such as the description of the function of a protein, its domain structure, post-transactional modifications, variants, etc., with a minimal level of redundancy and a high level of integration with other databases. The databases can be accessed and searched through the SRS (Sequence Retrieval System) at ExPASy (Expert Protein Analysis System). One can download the entire database as one single flat file. As of May 24, 2005, SWISS-PROT (release 47.1) contains 181,821 entries. TrEMBL (Translated EMBL) is a computer-annotated supplement of SWISS-PROT that contains all the translations of nucleotide sequence entries not yet integrated in SWISS-PROT. TrEMBL provides a comprehensive and high-quality view of the current state of knowledge about proteins.

Ongoing developments of SWISS-PROT are supplemented by functional and automatic annotation in the databases, provision of additional resources such as the International Protein Index (IPI) and XML format of SWISS-PROT and TrEMBL to the user community. The SWISS-PROT database has some legal restrictions and commercial companies must pay a license fee to SIB for using SWISS-PROT.

1.3.1.2 Protein Data Bank (PDB)

Protein Data Bank (PDB) was established in the 1970s at the Brookhaven Lab on Long Island, New York State, US. Since 1999, the Research Collaboratory for Structural Bionformatics(RCSB) manages PDB, which is a joint organization between Rutgers

University, San Diego Supercomputer Center and NIST (National Institute of Standards and Technology). The PDB is a repository for 3-D structural data of proteins obtained by X-ray crystallography or NMR spectroscopy. They are submitted by biologists and biochemists from around the world (Bernstein, 1977). The PDB entries contain the atomic coordinates, some structural parameters connected with the atoms (B-factors, occupancies), computed from the structures such as secondary structure and contain some annotation, but it is not as comprehensive as in SWISS-PROT. Fortunately, there are cross-links between the databases in both file formats. There are no legal restrictions on the use of the data in the PDB. Hence, PDB is released into the public domain and can be accessed free. The PDB is a key resource in structural biology and is critical in structural genomics. Countless derived databases and projects have been developed to integrate and classify the PDB in terms of protein structure, protein function and protein evolution.

1.3.2 Derived Databases

1.3.2.1 SCOP

One of the most accurate classifications of protein structures is the SCOP (Structural Classification of Proteins) database, which is constructed in large part manually. The SCOP database started by Alexey Murzin in 1994 at the Lab of Molecular Biology, MRC (Medical research Council), Cambridge, UK (Murzin et al., 1995; Andreeva et al., 2004). Its purpose is to classify 3D protein structures in a hierarchical scheme of structural classes with four levels. Experts manually maintain SCOP database. Protein structures in the PDB are classified as shown in Figure 8 and stored in SCOP. SCOP is frequently updated as new structures are deposited in the PDB.

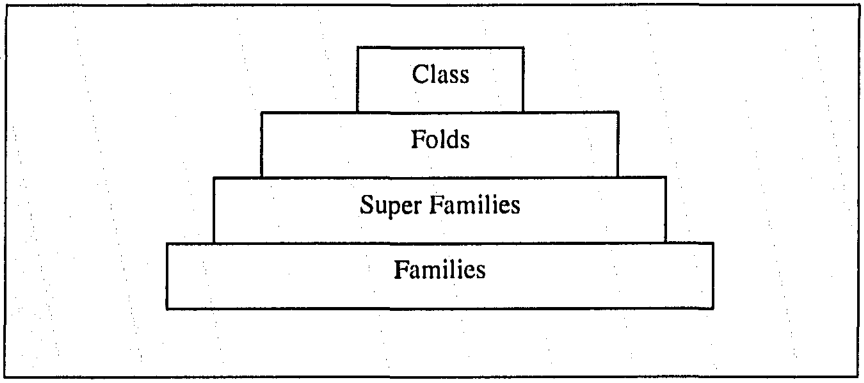


Figure 8. SCOP classification

The SCOP classes include:

- **Family:** At the bottom of the hierarchy are the individual domains of proteins, extracted from the PDB. Sets of such domains are then grouped into families, which consist of domains that have sufficient similarities in sequence, structure and functiun to imply a common evolutionary origin.

- **Super-family:** Often there are strong similarities in structure and function between families, while the sequence itself differs from family to family. Families sharing common structure and function, but lacking strong sequence similarity are grouped into super families.

- **Fold:** Super-families that have similar tertiary structures are grouped together in sets called folds.

- **Class:** Finally folds with similarities in secondary structure are grouped together into 4 classes, which is the highest level of the hierarchy. These four classes are:
    - α-helices: The secondary structure consists almost exclusively of α-helices.
    - β-sheets: The secondary structure consists exclusively of β sheets (strands).
    - α +β: The secondary structure has both α helices and mainly parallel β sheets (strands).
    - α /β: The secondary structure has both α helices and mainly antiparallel β sheets.

The first official SCOP release 9 years ago consisted of 3179 protein domains grouped into 498 families, 366 super-families and 279 folds. The seven main classes in the latest release (1.65) contain 40452 domains organized into 2327 families, 1294 super-families and 800 folds. These domains correspond to 20619 entries in the Protein Data Bank (PDB) (Westbrook J., 2002) and 1 literature reference to a structure with unpublished coordinates. Statistics of the current and previous releases, summaries and full histories of changes and other information together with parsable files encoding all SCOP data are available from the SCOP website (http://scop.mrc-lmb.cam.ac.uk/scop/) (Lo Conte L., 2002).

1.3.2.2 PDBSELECT

The PDBSELECT database is a subset of the structures in the PDB that does not contain (highly) homologue sequences. The representative lists of protein chains are intended for anyone interested in working with currently known protein structures. They are intended to save time and effort by offering a representative selection that is currently about a factor of five or six smaller than the entire database (Hobohm & Sander, 1994). Typical uses are introductory browsing, analysis of protein architecture, development of prediction methods, and model building by moduiar construction. To use the lists, a user needs access to datasets from the Protein Data Bank and software that reads protein structure files.

1.3.2.3 DSSP

The Dictionary of Secondary Structures of Proteins (DSSP) (Kabsch & Sander 1983) database mainly contains secondary structure assignments for all protein entries in the PDB. The DSSP program defines secondary structure, geometrical attributes and solvent exposure of proteins, given atomic coordinates in PDB format. The DSSP annotates each protein residue belonging to one of the eight secondary structure types: H (alpha-helix), G (3-helix or 310 helix), I (5-helix or π-helix), B (residue in isolated beta-bridge), E (extended strand), T (hydrogen bond turn), S (bend), and "_" (any other). Typically they are reduced to 3 groups: helix (H, which includes "H" and "G"), strand (E, which includes "E" and "B"), and coil (C, which includes remaining types) (Moult et al., 1997).

1.3.2.4 PDBFINDER

The PDBFINDER database is a database that is constructed using a PERL script from the PDB, DSSP and HSSP (Homology derived Secondary Structure of Protein) databases (Hooft R.W.W, 1996). View Objects, View Sequences or Import to workbench are aids to displays the original database annotation, associated sequences and sequence information for comparison and alignment. Many of the fields contained in the PDBFINDER database are difficult to access from the original databases. Some information is retrieved from the original literature.

1.3.2.5 PROSITE

PROSITE (Database of protein families and domains) is a method of determining what is the function of uncharacterized proteins translated from genomic sequences. It consists of a database of biologically significant sites, patterns and profiles that help to reliably identify to which known family of protein (if any) a new sequence belongs (Bairoch, 1993).

1.3.2.6 BLOCKS

BLOCKS are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. The blocks for the BLOCKS database are made automatically by looking for the most highly conserved regions in groups of proteins represented in the PROSITE database. These blocks are then calibrated against the SWISS-PROT database to obtain a measure of the chance distribution of matches. These calibrated blocks that make up the BLOCKS database (Henikoff, S, 1991).

Block Searcher, Get Blocks and Block Maker are aids to detection and verification of protein sequence homology. They compare a protein or DNA sequence to a database of protein blocks (current version), retrieve blocks, and create new blocks, respectively.

1.4 Related Work

Ultimate goal of computational approaches is to determine 3D protein structure based purely on protein sequence. Several approaches have been designed to predict protein structure. One of the important computational problems in predicting protein structure is the secondary structure prediction, which is the key focus in this research. The DSSP annotates each AA that constitutes the primary structure as belonging to one of the eight secondary structure types, which are typically reduced to three states: helix (H), strand (E) and coil (C). Three-state secondary structure (secondary sequence) of a protein aims to assign one of the three states of secondary structure to each AA in the protein sequence.

Example-1.3: 1FJNA protein from PDB has:

Protein sequence:	**GFGCPNNYQCHRHCKSIPGRCGGYCGGWHRLRCTCYRCG**
Three-state structure:	**CCCCCCHHHHHHHHHHCCCCCEEEECCCCCCCEEEECCC**

Secondary structure prediction currently incorporates structure prediction, content prediction and structural class prediction tasks. In structure prediction, each AA in a protein sequence is assigned to one of the three states of the secondary structure, which is usually performed using database search methods. In content prediction, prediction of the secondary structure content of a given protein sequence is done based on sequence properties. In structural class prediction, folds with similarities in secondary structure are grouped together, which is usually performed using data mining methods. Predictions performed with use of multiple alignment profiles (described later) are known as *alignment* methods and methods that are directly based on sequence is known as *attribute* methods, see Figure 9.

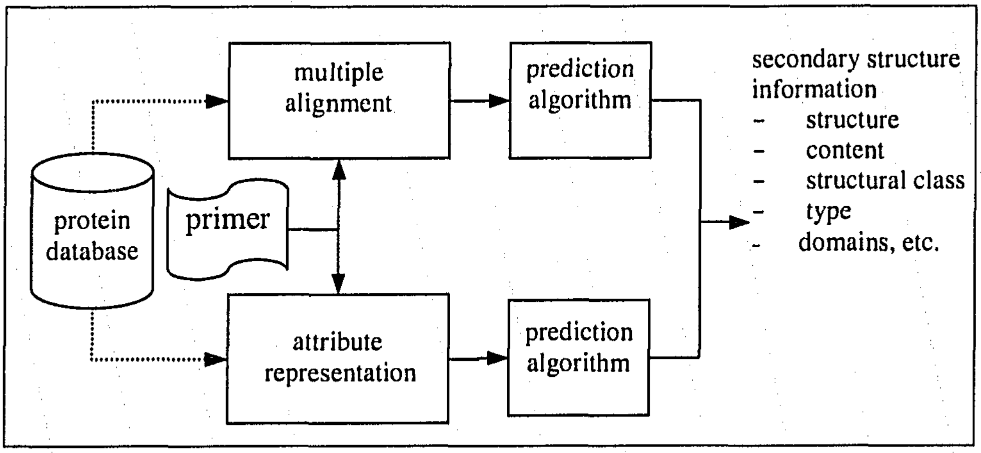


Figure 9.Overview of the protein structure prediction approaches

Structure prediction is mainly performed by alignment methods, while majority of the structural class and content prediction is performed by attribute methods. The later represent different protein sequences by an attribute vector of the same length, which allows using standard machine learning and data mining methods for prediction. Structure prediction methods use different datasets and different approaches. The methods most often use PSI-BLAST profiles (Altschul, 1997) as input and usually do not use other information derived from the protein sequence. Over the last couple of years, their underlying architecture stays the same.

1.4.1 Structure Prediction Methods

Over the last 30 years numerous prediction methods have been developed and continue to improve accuracy, from early results of about 60% to state-of-the-art methods that achieve about 80% accuracy (Rost, 2001). The first generation prediction methods were based on single AA propensities (Chou & Fasman, 1978; Garnier et al., 1978). Second-generation methods have been based on 3-51 adjacent residues propensities (Gibrat et al., 1987; Rost & Sander, 1994), and achieved accuracy of less than 66%. The third generation methods use evolutionary information and large protein databases, and consider global properties associated with protein families. They use multiple alignment and position specific profiles, and commonly apply PSI-BLAST (Altschul et al., 1997; Hargbo & Elofsson, 1999; Jones, 1999; Rost & Sander, 2000). Recent years brought advancements in prediction evaluation with the CASP (Moult et al., 2003) CAFASP (Fisher et al., 2003), and EVA (Eyrich et al., 2001) procedures, and integrated tools for prediction using multiple servers (Kurowski & Bujnicki, 2003). Strong interest is continuously observed, e.g. EVA monitors quality of 19 third generation methods. Most of the early structure prediction methods used single sequence information while recent methods use multiple sequence alignment information as input for prediction.

1.4.1.1 Attribute Methods

The principal idea behind early methods for secondary structure prediction is the fact that segments of consecutive residues have preferences for certain secondary structure states (Brändén, 1991). The goal is to predict whether the residue at the center of a segment of typically 13-21 sequence-consecutive residues is in a helix, strand or in none of the two (non-regular secondary structure, often referred to as the 'coil' or 'loop') (Rost, 2001). Thus, the predictive information is sequence-local. Many different algorithms have been applied to tackle this simplest version of the protein structure prediction problem: physico-chemical principles, rule-based devices, expert systems, graph theory, linear and multi-linear statistics, nearest-neighbor algorithms, molecular dynamics, and neural networks (Rost, 1996). The early methods by Nagano (Nagano, 1973), Chou and Fasman (Chou, 1974) and Garnier and colleagues (Garnier et al., 1978) relied on statistical treatment of compositional information for predicting three-state secondary structure.

Following these prediction achievements, attempts have been made by using sequence information at a more abstract and general level, such as the hydrophobicity rules applied in Lim's method (Lim, 1974). Prediction is also achieved by training, in most cases, a Neural Network (NN) on fixed-size sequence windows classified by the state of the central residue. However, until 1992 performance accuracy for three-state prediction seemed to have been limited to about 60%. In general, attribute methods represent the sequence by a fixed length vector that describes certain properties, such as AA composition, hydrophobicity, weight, etc. The limited accuracy was argued to result from the fact that all methods used only local information in sequence (window of less than 20 adjacent residues). Since last decade, the main improvement in accuracy was due to using family-derived profiles for both training and input.

1.4.1.2 Alignment methods

The principle idea behind sequence alignments is to optimally align the strings of amino (or nucleic) acid sequences for similarity. In protein sequences, finding the best alignment usually requires introduction of gaps in one sequence, or insertions in the other. Each alignment scheme assigns a penalty for each change and optimizes the alignment so that the total cost is minimized. Similarities between two sequences are represented by pairwise alignments as shown in Figure 10 and the similarities between three or more sequences are represented by multiple alignments as shown in Figure 11.

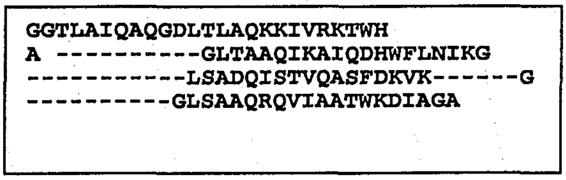
```
ACGGCTTACTAC
ACGGCATACTAC
```

Figure 10. Pair-wise alignment

Pairwise Alignment Methods: Needleman and Wunsch (Needleman & Wunsch, 1970) created the first automated global pairwise alignment algorithm that optimizes alignment over the entire length of the two sequences. Smith and Waterman introduced local sequence alignments (Smith & Waterman, 1981), which find more highly conserved subsequences. Dayhoff (Dayhoff, 1978) and Henikoff (Henikoff, 1991) assign different penalties for substituting AAs that are similar than for substituting AAs that are very different using scoring matrices, which are based on probabilistic principles (Durbin, 1998). In addition to developments in optimal sequence alignment, work on approximate alignments has been very important for practical use. The

FASTA (Lipman and Pearson, 1985) and BLAST (Altschul, 1990) algorithms both advanced the practicality of searching sequence databases for local alignments because they have greatly increased search speed over optimal alignments. Both work on the principle of heuristic elimination of sequence regions that are unlikely to produce good local alignments. BLAST (Basic Local Alignment Search Tool) has been the dominant sequence alignment program because of its superior speed and a well-developed statistical interpretation of the results. Modifications to the algorithm, such as PSI-BLAST(Position Specific Iterative BLAST) (Altschul, 1997), have also been significant for finding more distantly related sequences.

Multiple Alignment Methods: Multiple sequence alignment is useful for displaying the commonalities in a family of sequences that may have common structure or function see Figure 11.



```
GGTLAIQAQGDLTLAQKKIVRKTWH
A ----------GLTAAQIKAIQDHWFLNIKG
-----------LSADQISTVQASFDKVK------G
----------GLSAAQRQVIAATWKDIAGA
```

Figure 11. Multiple alignments

Although pairwise alignment may provide some of this information, it does not optimize the comparison across all the sequences in the family. Optimal multiple sequence alignment is extremely costly and most practical methods use a variety of heuristics to perform multiple alignments. As with pairwise alignments, many variations and improvements have been made to multiple alignment methods (Gusfield, 1997). ClustalW (Thompson, 1994) is a popular multiple sequence alignment program that illustrates the type of approximation algorithms that may be used to speed up the alignment. ClustalW first performs pairwise alignments between all the pairs of sequences to be aligned. The similarity scores are used to create a tree by clustering sequences that are more similar. The sequences are then progressively aligned starting with the most similar sequences. Sequences are merged into profiles that can then be merged with other sequences or profiles. A variety of other heuristic refinements improves the performance of the algorithm. By observing the aligned portion of the sequences, many patterns can be easily observed. The first method that reached a sustained level of three-state prediction accuracy above 70% was the profile-based neural network system PHD that uses evolutionary information derived from multiple sequence alignments as input (Rost, 1996). By stepwise incorporation of particular evolutionary information, prediction accuracy was pushed above 72% accuracy (Rost & Sander, 1993a, 1993b, 1994, 1996). At the beginning of twenty first century methods, such as PROFsec (Rost, 2000) and PSIPRED (Jones, 1999) reached a level of 76% three-state per-residue accuracy (Eyrich, 2001; Rost, 2001). Furthermore, significantly fewer residues are confused between the helix and strand states, which is the most frequent mistake.

Other Multiple Alignment Methods: Gibbs sampling techniques have been used to accelerate multiple alignment (Lawrence, 1993). This method is commonly used to find short local alignments that are motifs in relatively divergent groups of sequences. Many other patterns, such as consensus sequences or HMMs (Eddy, 2001), utilize multiple sequence alignments as a preprocessing step.

1.4.2 Content and Structural Class Prediction Methods

Since the last decade, an increased interest in prediction of other structural aspects, such as structural class and content is observed. In the former case the protein sequence is used to predict structural class (Wang & Yuan, 2000) usually defined based on the SCOP method (Murzin et al., 1995). In the latter case the amount of each of the three-state secondary structures in a given protein sequence are predicted (Lin & Pan, 2001). Each of these tasks gives insight into protein structure, and provides invaluable information that can be used to improve structure prediction. Other prediction tasks include protein type prediction (membrane and soluble), and protein domain partition.

1.5 Motivation

Since predicting the complete protein 3D structure is difficult, many researchers have focused on trying to predict the secondary structure of a protein. Unfortunately, predicting secondary structure of a protein is also a very difficult problem because it depends on the overall 3D structure of the fold. Most of these existing computational prediction methods for protein secondary structures use alignment approach, which is based on observation that protein chains of similar primary structure have similar secondary structure. The prediction of secondary structure by the alignment methods is limited, on average to no more than 88% accuracy. The state-of-the-art alignment prediction methods currently achieve around 80% accuracy for the three-state (H, E, C) secondary structure prediction (Petersen et al., 2000; Pollastri & McLysaght, 2005), and soon they will reach the accuracy barrier. Research shows that prediction accuracy is limited by accuracy of the template alignments, and that present methods do not overcome misalignments (Schonbrun et al., 2002). Further breakthrough can be expected by combining results produced by several different and complimentary methods. This was shown in the recent CASP4 study, when CAFASP-consensus method, which combines results from several, automated servers, performed better than any individual method (Sippl et al., 2001). Since majority of current prediction approaches are based on multiple alignment and recent research shows that consensus methods that utilize several complimentary prediction methods are the future. To this end, this thesis defines a novel classification problem related to the structure prediction in three states for protein structural fragments (SF) are considered based on the protein sequence information and without using alignment information.

1.6 Research Aim

A novel classification problem related to the structure prediction in three states is considered. Protein sequences are divided into three sets of structural fragments, defined as the longest fragments of protein sequence that correspond to the same secondary structure state, and an automated classification of the fragments to the corresponding secondary structure states without using sequence alignment is performed. The classification problem allows addressing several important hypotheses and questions: 1) How well helix, strand, and coil fragments can be distinguished using state of the art advancements from the existing protein structure, content and structural class prediction fields? 2) Which factors and prediction algorithms result and do not result in improving the ability to distinguish between the three secondary structures? 3) Are methods that do not utilize alignment are feasible to reliably distinguish between different secondary structures? and 4) How novel non-alignment based structure prediction methods can be developed?

1.7 Overview of the Research Goals

The main goals of research include investigation of a novel attribute based sequence representation, analysis of relation between certain sequence properties like length, position and information at the edges, comparison of several prediction algorithms performance and finally optimum attribute selection to improve ability to distinguish between the three secondary structures. This work performs comprehensive sets of experiments using wide range of Machine Learning based classification systems, to address multi-goal investigation using carefully selected large protein database. Detailed description of the goals is provided later in the thesis.

1.8 Organization

The remaining thesis is organized as follows. Chapter 2 includes background information related to Machine Learning including basic definitions, concepts related to classification, various classification systems and performance measures. Chapter 3 describes attribute representation of protein sequences used by current methods and a new proposed representation. Chapter 4 defines the problem of structural fragment prediction, dataset preparation and detailed goals. Chapter 5 discusses comprehensive experimental results and analysis of the goals. Finally, Chapter 6 concludes with the summary and future work of the considered prediction problem with respect to commonly performed protein structure, content and structural class prediction tasks.

2 Background in Machine Learning

Machine Learning (ML) is defined as a "computer program that can learn from experience with respect to some class of tasks and performance measure" (Mitchell, 1997). Prediction is the key for learning, which is the essence for intelligence. Many problems in biological systems cannot be defined well except by using examples (experimental data). Humans can specify the input/output pairs, but the relationship between the inputs and outputs are unknown (e.g. the protein folding mechanism). As pointed out in (Baldi, 1998), ML methods, such as neural networks, hidden Markov models, and belief networks, are ideally suited for areas where there is a lot of data but little theory. This is exactly the situation in Molecular Biology. In particular, molecular biologists are constantly faced with induction and inference problems, where they are building models from available data. ML methods are suitable for Molecular Biology data due to the learning algorithm's ability to construct hypotheses that can explain complex relationships in the data. Finally, the hypotheses can then be interpreted and validated by a domain expert, who either accepts or refuses them. In this section, basic definitions and concepts used in ML and most popular classification systems, which are applied was described.

2.1 Definitions

Attribute is a property, used to distinguish individuals. Its value is nominal or numerical. Attribute data can be continuous or discrete.

Example-2.1: 1FV5A protein has a sequence length 35.

Protein Id and sequence length are attributes of a protein.

Protein Id =1FV5A is a nominal value

Sequence length =35 is a numerical value.

Discrete Attribute has only a finite number (countable) of values. The values cannot be subdivided meaningfully. Discrete information can be categorized into a classification.

Example-2.2: Secondary structure state belongs to {helix, strand, coil}. Hence, secondary structure state data is discrete.

Continuous attribute has larger number of unique values. The values can be broken down into smaller parts and still have meaning. Different discretization methods are available to convert continuous into discrete form.

Example-2.3: Protein length varies between several to several hundreds of AAs. Hence, protein length is continuous.

Tuple is set of attributes that describe of an individual, also referred as object, example, sample or instance. Most commonly, tuple is represented as a row in a table.

Example-2.4: As shown in Table 2, each row represents a tuple described by attributes of a protein.

Table 2. Protein Dataset

Protein Id	Protein sequence	Secondary sequence	
1EDSA	TTLYTSLHGYFVFGPTGCNLEGFFATLGGEI	CCCCCCCCCCHHHHCCCCCCCCCCCCCCCCC	← tuple
1QK7A	GCLGDKCDYNNGCCSGYVCSRTWKWCVLAGPW	CCCCEEECCCCCCCCCEEEECCCCEEEECCCC	← tuple
1JDMA	MGINTRELFLNFTIVLITVILMWLLVRSYQY	CCCCCCCHHHHHHHHHHHHHHHHHHHCCCCC	----

Dataset is a named collection of data that contains tuples having the same attributes. Its most popular representation is a table, which consists of rows and columns. Row represents a tuple and column represents an attribute.

Example-2.5: Above **Error! Reference source not found.** is a dataset of proteins that contains protein objects.

Class attribute is defined by the user and usually is discrete. Its value is defined by a condition that involves combination of the predicting attribute values, which referred as class or class label.

Predicting attributes used to define a class, i.e. all attributes except the class attributes.

Training Data is a finite subset of dataset, in which class is predefined. Training dataset is used for generating a model in a learning process.

Test Data is a finite subset of dataset, in which class is not defined (unknown). Test data is used in a classification (prediction) process, and is not used in the learning process.

Example-2.6: In Table 3, where Structure_type is a class attribute. Length, Molecular_weight and Average_hydrophobicity are predicting attributes. Training data contains first three rows where class is predefined and test data contains last three rows.

Table 3. Protein subsequences dataset

Predicting attributes			Class attribute	
Length	**Molecular _weight**	**Average_hydrophobicity**	**Structure _ Type**	
14	134.36	-0.22	H	Training Data
9	118.36	0.28	E	
5	120.75	-0.64	C	
11	124.36	-0.22	--	Test Data
17	108.36	0.64	--	
15	110.75	-0.22	---	

Positive and Negative examples objects in a dataset that satisfy the class condition are called positive (+ve) examples and remaining are called negative (-ve) examples.

Example-2.7: Protein subsequences dataset shown in Table 4, where class 'H' is defined by a condition If length>5 then 'H'. Positive examples are the first three rows that satisfy the given condition and negative examples are last three rows that do not satisfy the given condition.

Table 4. Protein Dataset with +ve and –ve examples

Length	Molecular_weight	Average_hydrophobicity	Structure_ Type	
6	95	0.42	H	+ve examples
9	150	0.32	H	
7	136	-0.54	H	
4	124	-0.62	E	-ve examples
3	108	0.44	E	
2	110	0.92	C	

2.2 Machine Learning Concepts

Learning Schemes: Generally there are two types of ML schemes. *Supervised learning* where the goal is to learn patterns from the training data in order to predict class labels of unseen (unlabeled) test data; and *unsupervised learning* where the goal is to group or cluster unlabeled data based on observed patterns or associations. The overall tasks for the learning system (learner) are to classify, characterize, and group the input data. Both supervised and unsupervised learning are used extensively in computational biology research. General ML topics can be found in (Mitchell, 1997). This section is focused on ML in the protein secondary structure prediction problem, which is generally applied in a classification problem framework.

Classification Problem: Classification problems are supervised learning tasks in which the classification system learns patterns from a *training set.* A training set T=(A, C) consists a subset of objects $A = (A_1 \ldots A_n)$ belonging to class *labels* $C = (C_1 \ldots C_n)$. Each object of the training set (A_j, C_j) may correspond to an individual protein (or its structural fragment) A_j, and its known class label C_j. Each protein is represented as a set of *attributes* $A_j = (a_{j1}, a_{j2}, \ldots . a_{jm})$. The learning process generates *classification model* that can consequently be used to make a class label prediction $\overline{C}_j$ for each object A_t from a *test set.* Finally, evaluation of the performance by comparing the predicted class labels with the original class labels of the test set objects is performed using metrics and testing procedures which are described later. The process is shown in Figure12.

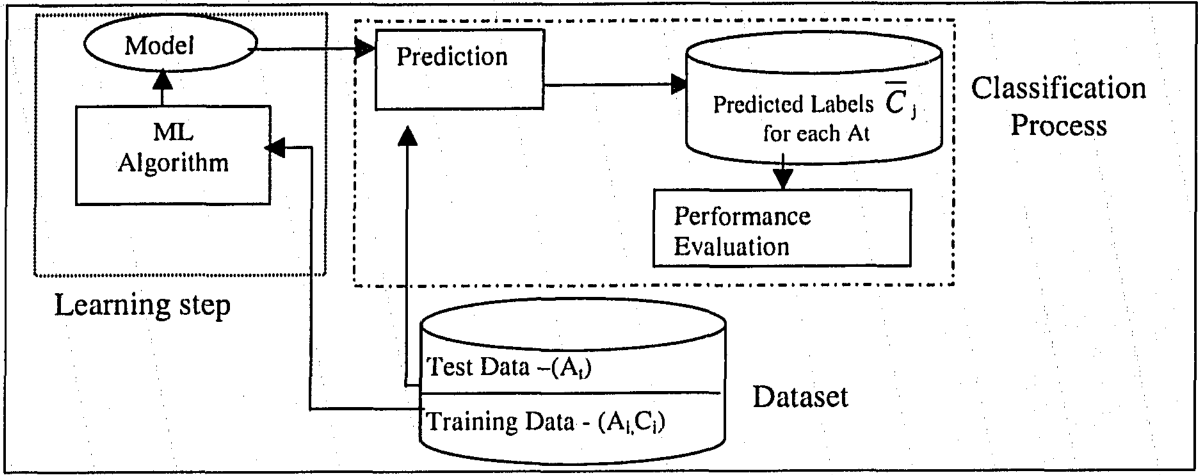


Figure 12. Learning and Classification Process

2.3 Testing and Evaluation Techniques

While performing classification task on test set (unseen during training), several performance metrics and testing procedures are used to validate and evaluate the generated model.

Testing: Generally, two testing procedures are utilized to validate the model.

- **Single-split testing:** In this procedure, the original data is divided into training and testing sets. These sets are disjoint, and the first one is used to derive data model, and the second to test it.
- **k-fold cross validation testing:** In this procedure, the data is divided into k subsets of approximately equal size. Model is generated k times by a ML algorithm, each time leaving out one of the subsets from training, but using only the omitted subset to calculate tests. Test results report mean values and

standard deviations of the prediction quality, averaging through k tests. Ten-fold cross validation is the most often used cross validation procedure.

K-fold cross validation method is more reliable test procedure than the single-split method. It shows the true performance in terms of validity and of tested classification system. Using the single-split procedure may lead to fitting the generated model to the test set, which may lead to falsifying the true performance of the tested algorithm. Therefore, a 10-fold cross validation testing procedure is used in this research.

Performance of classification system is commonly evaluated using a *confusion matrix.* A confusion matrix contains information about actual and predicted classifications done by a classification system (Kohavi & Provost, 1998). The following Table 5 shows confusion matrix that results from the prediction of a binary classification, in which 2 classes are defined. Several performance metrics have been defined using this confusion matrix. The entries in the confusion matrix have the following meaning:

Table 5. Confusion Matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	a	b
Actual Positive	c	d

- *a is the number of correct predictions that an example is negative,*
- *b is the number of incorrect predictions that an example is positive,*
- *c is the number of incorrect predictions that an example negative, and*
- *d is the number of correct predictions that an example is positive.*

Performance metrics:

The *accuracy* (*AC*) is the proportion of the total number of predictions that were correct:

$$AC = \frac{a+d}{a+b+c+d}$$

The *sensitivity* (*true positive rat*e *or TP or recall*) is the proportion of positive cases that were correctly identified:

$$TP = \frac{d}{c+d}$$

The *false positive rate* (*FP*) is the proportion of negatives cases that were incorrectly classified as positive:

$$FP = \frac{b}{a+b}$$

The *specificity* or *true negative rate* (*TN*) is defined as the proportion of negatives cases that were classified correctly:

$$TN = \frac{a}{a+b}$$

The *false negative rate* (*FN*) is the proportion of positives cases that were incorrectly classified as negative:

$$FN = \frac{c}{c+d}$$

The precision (*P*) is the proportion of the predicted positive cases that were correct:

$$P = \frac{d}{b+d}$$

The accuracy metric may not be an adequate performance measure when the number of negative cases is much greater than the number of positive cases (Kubat et al., 1998). Suppose there are 1000 cases, 995 of which are negative cases and 5 of which are positive cases. If the system classifies them all as negative, the accuracy would be 99.5%, even though the classification system missed all positive cases. Hence, metrics like sensitivity measures how many of the examples described by the rules as positive were truly positive and specificity measures how many of the examples described by the rules as negative were truly negative. Sensitivity and specificity enable evaluation of how the rules perform on the positive and negative data separately, which is very important in case when the numbers of positive and negative examples are different. Only the results with high values for accuracy, sensitivity and specificity can assure high confidence in the generated model.

2.4 Classification systems

The classification systems can be divided based on the generated model into the following families:

- **Black-box systems**, generate models, which cannot be interpreted by the user.
- **White-box systems**, generate interpretable models. These systems can be further divided into:

- *Rule-based systems* - generate models that consist of production rule sets.
- *Decision tree systems* - generate models that consist of decision trees.
- *Probabilistic systems* - generate probabilistic models.

2.4.1 Neural Network Classification Systems

Neural Networks (NNs) are one of the widely used ML algorithms and they are the earliest technique applied to the field of biological analysis (Stormo *et. al.*, 1982). NN models are based on the operation of synaptic connections in neurons of the brain, where input is processed on several levels and mapped to a final output. NNs are built from multi-layer of nodes linking each other. Generally, there are three layers in the network, the input layer, the output layer and a hidden layer(s) in between them. The most common NN model is the multilayer perceptron (MLP) that is as a supervised network because it requires a desired output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using training data, so that the model can be used to produce the output when the desired output is unknown. A graphical representation of an MLP is shown in Figure 13. The MLP and many other NNs learn using an algorithm called back-propagation. With back-propagation, the input data is repeatedly presented to the NN. With each presentation, the output of the NN is compared to the desired output and an error is computed. The error is then fed back (back-propagated) to the NN and used to adjust it such that the error decreases with iteration and the neural model gets closer to producing the desired

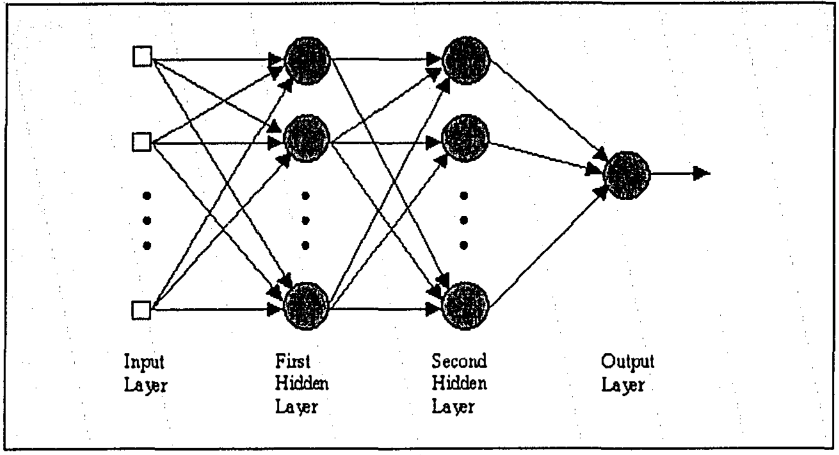


Figure 13. two hidden layer multiplayer perceptron (MLP)

output. MLP NNs are capable to learn and solve many real-world problems, but lack of explanatory power is their main limitation. It is hard to interpret the decisions and approaches of each node in the networks and thus validation of the networks becomes infeasible.

Neural Networks have been widely used in the protein structural and functional prediction (Hirst & Sternberg, 1992; Qian & Sejnowski, 1988; Sasagawa & Tajima, 1993; Nakata, K. 1995; Macklin & Shavlik 1993) and protein classification (Wu *et. al.* 1995; Ferran & Ferrara, 1992).

2.4.2 Decision Tree Classification Systems

The decision tree (Quinlan, 1986) is a supervised learning technique, which is one of the widely used ML algorithms due to simplicity. The decision tree allows classification of a test data by following series of decisions from the root to a leaf of the tree. Each node in the tree represents a decision on one of the predicting attributes and a leaf node represents a class label. Decision about which branch to follow is made on basis of the value of a predicting attribute. The classification returned for an object is found as the decision process reaches the leaf node containing the appropriate class label. A graphical representation of a simple decision tree is shown in Figure 14.

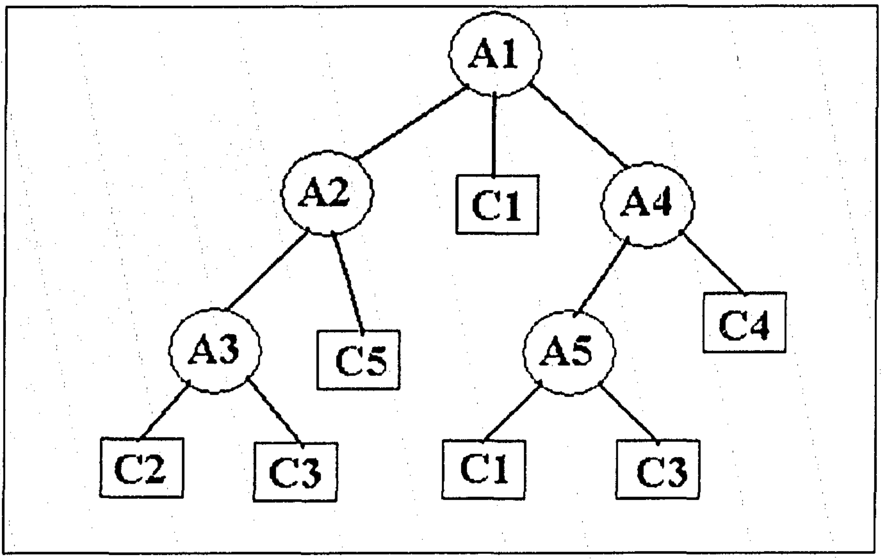


Figure 14. A decision tree with attributes A1 ... A5 and classes C1 ... C5

Some decision tree classification systems are CART (Breiman, L.,1984), ID3 (Quinlan, J.R., 1986), C4.5 (Quinlan, J.R, 1993), T1 (Holte, R.C, 1993), and C5.0 (RuleQuest Research, 2003). Among all decision tree learners, the most scalable, and at the same time most accurate in terms of the generated rules and trees is the C5.0 learner (Cohen, 1999)

The advantages of decision trees are that they are easy to use, practical, robust to noise and capable of learning disjunctive expressions. Over fitting of the data and overlapping in the classes are some possible drawbacks of decision trees. Furthermore, the decision trees are hard to optimize.

Shimozono et. al. (Shimozono et. al., 1992) used the decision trees to classify membrane protein sequences according to functional classes. Cherkauer, Shavlik (Cherkauer and Shavlik, 1993) and Selbig et al. (Selbig et al., 1999) applied them for protein structure prediction, and Salzberg (Salzberg, 1995) used decision trees to locate protein-coding genes.

2.4.3 Probabilistic Classification systems

Probabilistic classification generates models based on probability. Naïve Bayes (NB) is one of the most popular probabilistic classification system used in the ML community. NB considers all attributes as random variables and assumes independence among attributes. Given a object with attributes (A_1, A_2,…,A_n) and goal is to predict class C, it finds the value of C that maximizes $P(C| A_1, A_2,\ldots,A_n)$. Naïve Bayes classification systems can be applied only with discrete attributes.

Conditional probability:

$P(A_1, A_2, \ldots, A_n |C) = P(A_1| C) P(A_2| C)\ldots P(A_n| C)$, where $P(A_i | C) = |A_i|/ N_c$.

If some attributes are dependent, then other techniques such as Bayesian Belief Networks (BBN) can be used. Detailed descriptions of Bayesian networks can be found in Mitchell (Mitchell, 1997).

Probabilistic classification systems are robust to isolated noise points, handle missing values and are robust to irrelevant attributes.

Probabilistic systems have been used by Schmidler *et al.* (Schmidler *et al.*, 2000) to predict protein secondary structure and Cai *et al.* (Cai *et al.*, 2000) to model the splice sites.

2.4.4 Rule-based Classification Systems

Rule based classification systems are widely accepted due to their easy understandability and interpretability. Rule discovery has been studied for more than twenty years and a number of methods have been proposed. Rule discovery algorithms use the divide and conquer approach, which finds a first "best" rule from a dataset and then all objects covered by the rule is removed from the dataset. This procedure repeated until there are no objects left in the dataset. The "best" rule is usually found using heuristics. Some typical systems in this category are AQ family of algorithms (Kaufman, 1999), INDUCE (Dietterich, 1981), FOIL (Quinlan, 1990), REP (Cohen, 1993), IREP (Furnkranz, 1994), RISE (Domingos, 1994), RIPPER (Cohen, 1995, 1996), SLIPPER (Cohen, 1999), LAD (Boros, 2000), LERILS (Chisholm, 2002) and IREP++ (Dain, 2004). RIPPER (Cohen, 1995) learner was shown to have very competitive accuracy, and better complexity due to using a divide-and-conquer approach combined with a greedy set-covering based search procedure during the rule induction process. After a rule is generated, it is immediately pruned in a greedy manner. SLIPPER (Schapire, 1998) is one of the most advanced rule learners. It is shown to improve upon the accuracy of RIPPER learner by applying a boosting strategy in the induction process. SLIPPER learner is characterized by low complexity, which is asymptotically identical to RIPPER's complexity. The RIPPER and SLIPPER learners are selected as a representative of rule learners in this research for the experimental comparison.

The disadvantage of rule-based systems is their heuristic nature. Many traditional rule-based classification systems prefer small rule sets to large rule sets, and small classification systems may be sensitive to the missing values in unseen test data.

Rule-based classification systems have been applied in several research areas in computational biology such as protein structure prediction (Sternberg *et. al.*, 1992) and learning drug properties (King *et. al.*, 1995).

2.4.5 Other Classification Systems

Other classification systems such as Hidden Markov Models (HMMs), Support Vector Machines (SVMs) and Genetic Algorithms (GAs) are also become popular in computational biology.

HMMs are statistical models that predict the class based on the probabilities of the model states. Every probability in the states is summed up to give a final score and the prediction is based on the score. Background about HMM can be found in (Durbin *et. al.*, 1998; Baldi and Brunak, 1998). The states in the model are associated with meaningful biological attributes (Durbin *et. al.*, 1998). Profile HMMs (Gribskov &

Veretnik, 1996) are most popular to treat the gaps in systematic way for the sequence alignments.

The main idea of an SVM (Vapnik, 1995) is to separate classes with a surface that maximizes the margins between them. Also, it is a powerful classification learning approach, which applies a concept that non-linear input vectors are mapped through a very high dimension attribute space where the linear decision of the input vectors can be computed. Although SVMs have good generalization performance, the disadvantages of this method are time intensive test phase and lack of expressive power. SVMs have been used in several applications in molecular biology such as protein fold recognition (Ding and Dubchak, 2001), classification of microarray data (Furey *et. al.*, 2000) and recognition of translation initiation sites (Zien *et. al.*, 2000).

The main idea of GAs (Davis, 1991) is to maintain a population of data that represent the candidate solutions to the problem. The population undergoes recombination (crossover between two strings) and mutation (changes in a string) processes to adapt the new environment. The ultimate goal of the candidate solution is to become the fittest (best solution) in the environment. The main disadvantage of GAs is non-transparent learning process and high computational complexity. GA has been widely used in DNA fragment assembly (Parsons *et al.*, 1995; Cedeno, & Vemuri, 1993; Fickett & Cinkosky, 1993) and were applied in multiple alignments (Zhang & Wong, 1997).

3 Protein Representation

The prediction of secondary structure is performed with an intermediate step that transforms the protein sequences into their attribute space representation. This is due to differences in the length of the protein sequences for different proteins, i.e. the protein sequences length can vary between several to several hundred residues (amino acids), while ML algorithms usually assume input data of fixed length. The usual attributes that describe a protein sequence are the amount and the position of AAs that compose a given protein, which are referred as composition vector and composition moment vector. Researchers already recognized that the most commonly used composition vector is not sufficient for prediction purposes and therefore alternative solutions were sought (Dubchak et al., 1997; Zhang et al., 2001; Lin and Pan, 2001; Cai et al., 2003; Ruan et al., 2005; Luo, Feng and Liu, 2002; Chou and Cai, 2004). The main drawback of the current representations is insufficient number of used attributes. This research is the first to use a comprehensive attribute representation of a protein (sub) sequence that draws from numerous recent papers, and to perform selection of a subset of best attributes in terms of improving prediction accuracy. This chapter explains current protein representations and proposes a new comprehensive representation.

3.1 Current Representation

Majority of structural class and content prediction is performed by attribute methods. A typical attribute space representation consists of just a few attributes, e.g. composition vector and molecular weight. Recent methods use advanced representation that includes hydrophobicity and higher order composition moment vectors. Some content prediction methods use structural class (Zhang et al., 2001) as an input. The structural class prediction enjoys the widest selection of prediction algorithms and attribute representations. Recent representative prediction methods are summarized in Table 6. An integrated PROSPECT-PSPP server was developed recently. It applies several prediction tasks including homology, protein type, domain partition and to perform integrated secondary structure prediction (Guo et al., 2004).

Table 6. Summary information for representative secondary structure, content and structural class prediction methods

1 (Rost, 1996), 2 (Jones, 1999), 3 (Ouali & King, 2000), 4 (Karplus, 2001), 5 (Przybylski & Rost, 2002), 6 (Pollastri et al., 2002), 7 (Guo et al., 2004), 8 (Pollastri & McLysaght, 2005), 9 (Eisenhaber et al., 1996), 10 (Zhang et al. 1998), 11 (Zhang et al., 2001), 12 (Lin & Pan, 2001), 13 (Cai et al., 2003), 14 (Ruan et al., 2005), 15 (Wang & Yuan, 2000), 16 (Li & Lu, 2001), 17 (Cai et al., 2003), 18 (Chou & Cai, 2004)

Pred. task	Method	Ref	Prediction algorithm	Multiple alignment	Protein representation
Structure	PHD	1	neural network	BLAST	N/A
	PSIred	2	neural network	PSI-BLAST	N/A
	PROF	3	neural net & discriminants	BLAST	N/A
	SAM-T	4	hidden markov models	BLAST	N/A
	PHDpsi	5	neural network	PSI-BLAST	N/A
	SSpro	6	neural network	PSI-BLAST	N/A
	PMSVM	7	support vector machines	PSI-BLAST	N/A
	Porter	8	neural network	PSI-BLAST	N/A
Content	SSCP	9	vector decomposition	N/A	composition vector
	IMLR-1	10	multiple linear regression	N/A	composition vector, autocorrelation.
	IMLR-2	11	multiple linear regression	N/A	same as above
	IMLR-3	12	multiple linear regression	N/A	composition vector, autocorrelation., hydrophobicity
	---	13	neural network	N/A	pair coupled composition vector
	---	14	neural network	N/A	composition moment vector
Structural class	---	15	Bayesian	N/A	composition vector
	---	16	diversity measure	N/A	protein sequence
	---	17	support vector machines	N/A	composition vector
	---	18	intimate sorting predictor	N/A	functional domain composition

3.2 Proposed Representation

The main difference between proposed and the existing representations lies in comprehensiveness of attribute sets used for describing a protein sequence. In contrast, existing representation consider very limited attribute space representation. The proposed representation considers a large and diverse attribute sets and performs attribute sets selection to find an optimal representation in terms of prediction quality. This investigation assumes attribute representation that is based on protein and AA properties. The considered attribute sets together with detailed description, motivation for introduction and references are summarized in Table 7. Each attribute set contains one or more attributes and explained in detail below. A subset of the attribute sets in the proposed representation was used in a recent method for content prediction (Kurgan & Homaeian, 2005).

Table 7. Attribute representation for a protein sequence

1 (Ruan et al., 2005), 2 (Lin & Pan, 2001), 3 (Muskal & Kim, 1992), 4 (Syed & Yona, 2003), 5 (Eisenhaber et al., 1996), 6 (Zhang et al. 1998), 7 (Zhang et al. 2001), 8 (Wang & Yuan, 2000), 9 (Luo, Feng & Liu, 2002), 10 (Cai et al., 2003), 11 (Ganapathiraju et al., 2004), 12 (Nelson & Cox, 2000), 13 (Wang, J., et al, 2000), 14 (Yang & Wang, 2003), 15 (Hobohm & Sander, 1995)

Attribute set name	Description	Motivation	References
length	# of residues	may be related to content	
composition vector	normalized composition percentage of each AA in the protein sequence	considered as standard for most content and structural class methods	2, 5, 6, 7, (content) 8, 9, 10 (struct class)
composition moment vector	1st order composition vector that incorporates position of AAs in the sequence	supplementing composition with position was shown to improve content prediction	1 (content)

hydrophobicity	average and accumulated (summed) average hydrophobicity computed using a hydrophobic index	hydrophobic force is one of the strongest determinant factors of a protein structure	1, 2 (content)
molecular weight	sum of molecular weights of neutral, free AAs	may be related to content and function	3 (content), 4 (function)
Auto-correlation	autocorrelation value computed using hydrophobic index	reflects profile of hydrophobic indices along the protein sequence	2, 6, 7 (content)
electronic group	divides AAs into neutrals, electron donors or acceptors	electrostatic forces stabilize structure	11 (structure)
R group	combines hydropathy, molecular weight and pI	may be related to structure and content	12 (structure/content)
exchange group	some AAs can be substituted by other without impact on the structure	represents conservative replacements through evolution	13 (family), 14 (structure)
hydrophobic group	divides AAs into hydrophobic and hydrophilic	the same as for hydrophobicity	4, 15 (function)
other groups	considers the following mixed classes: charged, polar, aromatic, small, tiny, bulky, and polar uncharged	may be related to function	4, 15 (function)
chemical group	chemical groups are associated with AAs	may be related to structure	11 (structure)

Length defines the number of AAs that constitute a sequence.

Composition vector and composition moment vector provide information about AA propensities and position in a sequence (Ruan et al., 2005). Composition vector and composition moment vector values for each AA are calculated using the following equation (1):

$$x_i^{(k)} = \frac{\sum_{j=1}^{o_i} n_{ij}^{\ k}}{\prod_{d=0}^{k}(N-d)} \qquad (1)$$

Where N is the length of the AA sequence, n_{ij} is the j^{th} position of the i^{th} AA, o_i represent the occurrence (composition) of the i^{th} AA in a sequence, and k is the order of the composition moment vector. Zero and first order composition moment vector values are used in the proposed representation. Note that, zero order composition moment vectors reduce to the composition vector.

Hydrophobicity can be used to represent a protein sequence by using a hydrophobic scale, or its transformation, in which each AA of a sequence is replaced by its hydrophobic index value. Average hydrophobicity H_{avg} and accumulated average hydrophobicity $\overline{H}_{avg}$ values are calculated using the following equations (2) and (3) respectively.

$$H_{avg} = \frac{\sum_{i=1}^{N} h_i}{N} \qquad (2)$$

$$\overline{H}_{avg} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{j} h_i = \frac{1}{N}\sum_{i=1}^{N}(N+1-i)h_i \qquad (3)$$

where hi is the hydrophobicity index value for each AA and N is the length of the sequence. Two hydrophobic indexes are considered in proposed representation: Esienberg's hydrophobic index (Eisenberg, 1984) as shown in Table 8 and the Fauchere et al hydrophobic index (Fauchere & Pliska, 1983) as shown in Table 9.

Table 8. Eisenberg's hydrophobicity index values

Amino Acid	A/M	C/N	D/P	E/Q	F/R	G/S	H/T	I/V	K/W	L/Y
Hydrophobicity index value	0.62	0.29	-0.90	-0.74	-1.19	0.48	-0.40	1.38	-1.50	1.06
	0.64	-0.78	0.12	-0.85	-2.53	-0.18	-0.05	1.08	0.81	0.26

Table 9. Fauchere et al hydrophobicity index values

Amino Acid	A/M	C/N	D/P	E/Q	F/R	G/S	H/T	I/V	K/W	L/Y
Hydrophobicity index value	0.42	1.34	-1.05	-0.87	2.44	0.00	0.18	2.46	-1.35	2.32
	1.68	-0.82	0.98	-0.30	-1.37	-0.05	0.35	1.66	3.07	1.31

Molecular weight refers to the sum of the atomic weight of all of the atoms within the AA. Average molecular weight of a sequence is calculated using the equation (4), which used AA molecular weight information shown in Table 10 (Black & Mould, 1991).

$$M_{avg} = \frac{\sum_{i=1}^{N} m_i}{N} \qquad (4)$$

Where m_i represents the molecular weight of the i^{th} neutral free AA and N is the length of the sequence.

Table 10. Molecular weight for AA's (Black & Mould, 1991)

Amino Acid	A/M	C/N	D/P	E/Q	F/R	G/S	H/T	I/V	K/W	L/Y
Hydrophobicity	0.42	1.34	-1.05	-0.87	2.44	0.00	0.18	2.46	-1.35	2.32
index value	1.68	-0.82	0.98	-0.30	-1.37	-0.05	0.35	1.66	3.07	1.31

Auto-correlation refers to the profile of the hydrophobicity indices of AAs along the sequence. Autocorrelation function r_n of a sequence is calculated using the following equation (5) (Zhang et al., 2001):

$$r_n = \frac{\sum_{i=1}^{N-n} h_i h_{i+n}}{N-n} \qquad (5)$$

where $h_1,h_2,\ldots..h_n$ is the numerical sequence that is created by replacing each AA by its Fauchere et al hydrophobicity index value, h_i is the hydrophobicity index for the i^{th} AA and N is the length of the sequence. Note that value of n=1,2...10 is used in this research, i.e. 10 auto-correlation functions are computed.

Attribute groups divide the AAs into groups related to specific characteristics of an individual AA. Each attribute group's vector value is calculated using the equation (6) in the proposed representation.

$$A_i = \frac{a_i}{N} \qquad (6)$$

where a_i represents the occurrence of the i^{th} group in a sequence and N is the length of a sequence. Attribute groups, such as electronic group, r-group, exchange group, hydrophobic group, other groups and chemical groups, are explained below.

Electronic group divides AAs based on their electronic properties, i.e. if they are neutral, electron donor or electron acceptor. Five electron groups are used in the proposed representation, see Table 11.

Table 11. Electronic group for AAs

Electronic Groups	AAs
Electron donor	D, E, P, A
Weak electron donor	V, L, I
Electron acceptor	K, N, R
Weak electron acceptor	F, Y, M, T, Q
Neutral	G, H, W, S

R-group divides AAs by combining the hydropathy index, molecular weight and pI value together (Nelson, 2000). Five R-groups are used in proposed representation, see Table 12.

Table 12. R-group for AAs

R-groups	AAs
Nonpolar aliphatic	A, V, L, I, M, G
Polar uncharged	S, P, T, C, N, Q
Positively charged	K, H, R
Negative	D, E
Aromatic	F, Y, W

Exchange group represents conservative replacements of AAs through evolution. Three exchange groups are used in proposed representation, see Table 13.

Table 13. Exchange group for AAs

Exchange Groups	AAs
B	A, G, P, S, T
C	D, E, N, Q
F	I, L, M

Hydrophobicity group divides AAs into hydrophobic, which are insoluble or slightly soluble in water, in contrast with hydrophilic, which are water-soluble. Two types of hydrophobicity groups are used in representation see Table 14.

Table 14. Hydrophobicity group for AAs

Hydrophobicity Groups	AAs
Hydrophobic	V, L, I, M, A, F, P, W, Y, C, G
Hydrophilic polar with uncharged side chain	S, T, N, Q

Other groups divides the AAs by considering the seven types, such as, charged, polar, aromatic, small, tiny, bulky, and polar uncharged, see Table 15.

Table 15. Other group for AAs

Other Groups	AAs
Charged	DEKHRVLI
Polar	DEKHRNTQSYW
Aromatic	FHWY
Small	AGST
Tiny	AG
Bulky	FHWYR
Polar uncharged	NQ

Chemical group is associated with individual AAs. Ten chemical groups are used in the proposed presentation; those are CH, CO, NH, CH_3, CH_2, CAROM, CHAROM, CH_2RING, C, and OH.

The following example shows all the attribute values of a protein subsequence to illustrate the proposed representation.

Example-3: 1NBLA protein from PDB

Protein sequence:

KSCCRNTLARNCYNACRFTGGSQPTCGILCDCIHVTTTTCPSSHPS

Protein structural fragments based on uniform secondary structure:

KSCCR, NTLARNCYNACRFT, SQPTCGILC, GG, DCIHVTTTTCPSSHPS

For fragment KSCCR the considered attribute values are shown in Table 16.

Table 16. Attribute representation Example

Attribute set name	Attribute values for Subsequence KSCCR
Length	5
composition vector	0.0,0.4,0.0,0.0,0.0,0.0,0.0,0.0,0.2,0.0,0.0,0.0,0.0,0.0,0.2,0.2,0.0,0.0,0.0,0.0
composition moment vector	0.0,0.35,0.0,0.0,0.0,0.0,0.0,0.0,0.05,0.0,0.0,0.0,0.0,0.0,0.25,0.1,0.0,0.0,0.0,0.0
Hydrophobicity	-0.726, -1.860, -0.018, -0.324
molecular weight	133.56
Auto-correlation	-0.009925, -1.237267, -0.870250, 1.849500, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
electronic group	0.0, 0.0, 0.4, 0.0, 0.2
R group	0.0, 0.6, 0.4, 0.0, 0.0
exchange group	0.2, 0.0, 0.0
hydrophobic group	0.4, 0.2
other groups	0.4, 0.6, 0.0, 0.2, 0.0, 0.2, 0.0
chemical group	1.0, 1.0, 1.2, 0.0, 2.0, 0.0, 0.0, 0.0, 0.2, 0.2

4 Prediction Scenario and Detailed Goals

This research investigates impact of protein representation, certain sequence properties and prediction algorithms on the quality of prediction of secondary structure for structural fragments (SF). SF prediction problem allows to evaluate accuracy limits and characteristics of the attribute based secondary structure prediction, and simultaneously to lay foundations for the development of a new family of protein structure prediction methods. Before describing the considered problem, protein structural fragment (SF) is defined as the longest fragments of protein sequence that correspond to the same secondary structure state as shown in the example below.

Example-4: 1FJNA protein from PDB

Protein sequence:	`GFGCPNNYQCHRHCKSIPGRCGGYCGGWHRLRCTCYRCG`
Secondary sequence:	`CCCCCCHHHHHHHHHHCCCCCEEEECCCCCCCEEEECCC`
SFs:	`GFGCPN,NYQCHRHCKS,IPGRC,GGYC,GGWHRLR,CTCY,RCG`

4.1 Structural Fragment Prediction

4.1.1 Problem Definition

Structural fragment prediction is defined as prediction of secondary structure state for a given protein structural fragment based on models inferred from SFs for which the corresponding structure state is known; see Figure 15. Solid lines denote how the models are generated, while dotted lines show how they are used to perform prediction.

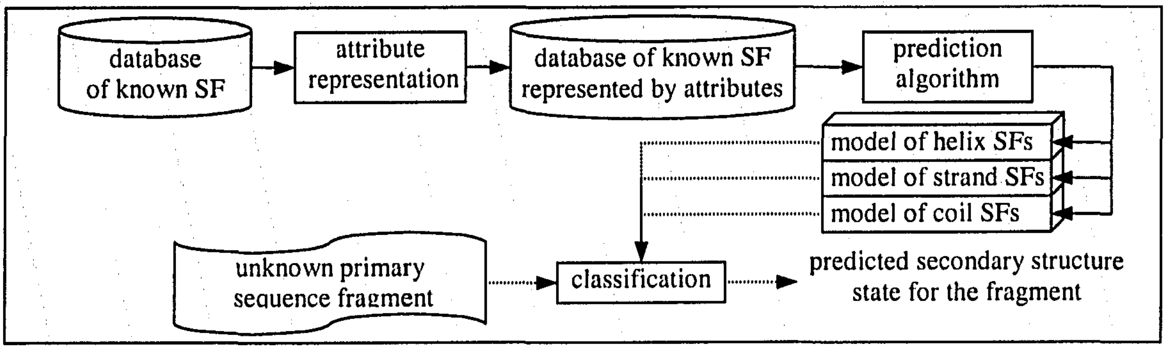


Figure 15. Structural fragment prediction

The SF prediction allows investigating how difficult it is to distinguish between different secondary structure states, and analyzing how to improve quality of the prediction by using certain protein sequence properties, prediction algorithms, and protein representations. Information about specific ways in which the three state secondary structures can be better distinguished. This prediction will provide invaluable help in performing prediction of not only protein secondary structure, but also content, structural class, and other structure related prediction tasks.

4.1.2 Selection of SFs

The assessment of the quality of the prediction for SFs is based on two assumptions. First, the PDB is used as a source of data that includes primary and secondary protein structure. A custom set of filters, which are described later, is used to guarantee high quality of the input data. Second, this research divides the protein sequences into SFs. This is an idealized situation where the fragments are uniform and span the entire corresponding primary subsequence. The actual known method for dividing the sequences into fragments guarantees uniformity, but at the same time the resulting fragment usually are only subsequences of the entire corresponding primary subsequences (Ruan et al., 2005). For instance, based on primary and secondary protein sequences shown in the example-4, the idealized scenario realized in this research divides the protein sequence into GFGCPN, NYQCHRHCKS, IPGRC, GGYC, GGWHRLR, CTCY, and RCG, while the actual division might be GFGCPN, NYQC, HRHCKS, IPGRC, GGYC, GGW, HRLR, CTCY, and RCG. Application of the idealized scenario may results in overestimating the accuracy of the resulting classification, but it provides certain benefits such as direct evaluation of the accuracy of classification of the SFs into three state secondary structures. It also allows performing additional studies related to investigation of the relation between SFs and their secondary structure. Figure 16 describes the method used to assess quality of the prediction of SFs.

To accommodate the considered scenario, the proposed representation given in Table 7 was supplemented with following attribute sets:

Number of duplicates, which equals to number of occurrences of a given SF in the entire dataset. Bigger value gives higher confidence that the SF is associated with a given secondary structure.

Relative position, which defines SF's position in terms of which quarter of the sequence a majority of the SF resides. This allows studying relationship between SF position and structure. In example-4, KSCCR fragment belongs to the first quarter of 1NBLA protein sequence.

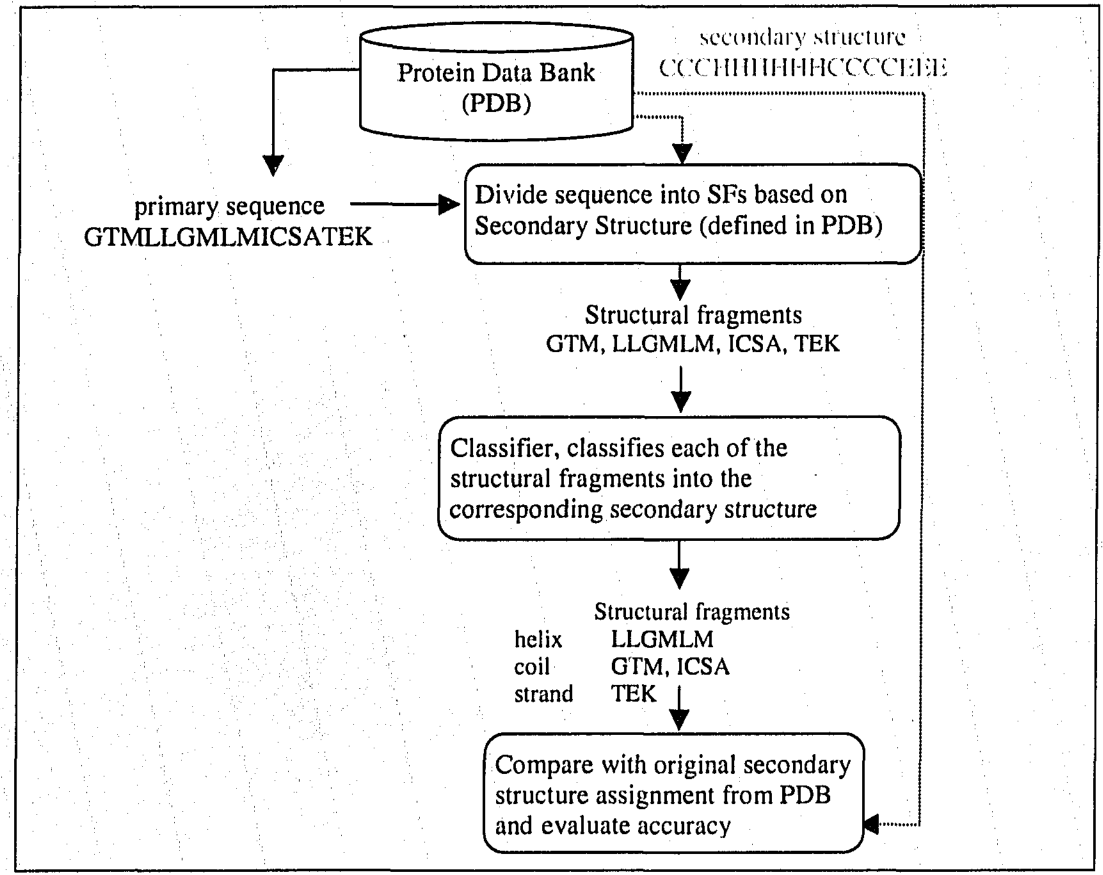

Protein Data Bank
(PDB)
secondary structure
CCCHHHHHHCCCCEEE
primary sequence
GTMLLGMLMICSATEK
Divide sequence into SFs based on Secondary Structure (defined in PDB)
Structural fragments
GTM, LLGMLM, ICSA, TEK
Classifier, classifies each of the structural fragments into the corresponding secondary structure
Structural fragments
helix LLGMLM
coil GTM, ICSA
strand TEK
Compare with original secondary structure assignment from PDB and evaluate accuracy

Figure 16. The method used to assess quality of prediction of SFs

4.1.3 Dataset Preparation

Since the SF prediction task was not considered in the past research, a suitable database of SFs was created. First, the primary and corresponding protein sequences were extracted. Next, the sequences were spliced into SFs, which were further converted into attribute representation is grouped by their corresponding secondary structure label.

Preprocessing: The main goal was to assure high quality of the used protein sequences. The data was extracted from PDB, release as of August 12, 2004. For proteins with multiple chains, the last one was selected. Next, the proteins were filtered to eliminate errors and inconsistencies. The proteins with missing primary or secondary sequence, with sequence length < 5, with sequence containing unknown or incorrect residues, with helices of length < 3, and with strands of length < 2 were filtered out. After filtration, 5834 proteins were left. Among them, a subset of 539 high quality non-homologous proteins was selected using 25% PDB SELECT list (Hobohm and Sander, 1994). The 25% PDB SELECT list is about a factor of fifteen smaller than the PDB database and includes only high quality non-homologous proteins, i.e. proteins scanned with high resolution and with low about 22% - 45% sequence identity. Detailed description of the filtering procedure can be found in (Kurgan and Kedarisetti, 2005). Next, the 539 proteins were spliced into helix, strand, and coil SFs, and put into separate sets for each structure. In each set duplicates and inconsistent SFs were counted and removed, and helix SFs of length ≤ 3, and strand and coil SFs of length ≤ 2 were eliminated to improve quality of the data. Total of 7056 SFs were generated. Schematic diagram for the entire dataset preparation procedure is shown in Figure17.

Datasets generation, In order to investigate impact of protein sequence properties on the ability to distinguish between different corresponding secondary structure states, the following datasets with the corresponding number of SFs were created:

- DA (7056), which includes all SFs,

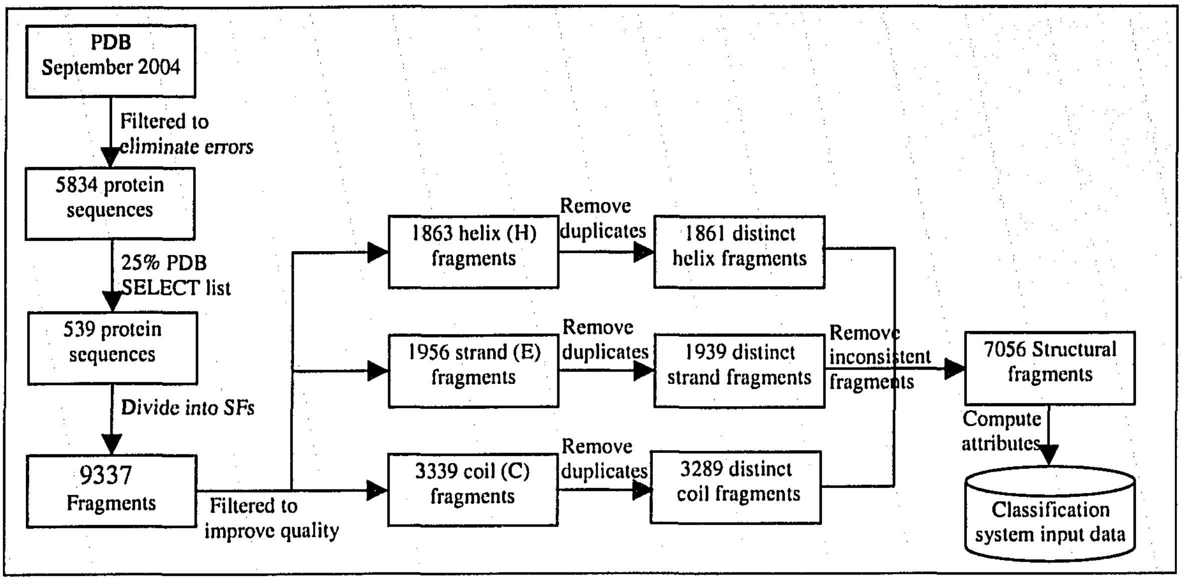

PDB
September 2004
Filtered to eliminate errors
5834 protein sequences
25% PDB SELECT list
539 protein sequences
Divide into SFs
9337 Fragments
Filtered to improve quality
1863 helix (H) fragments
1956 strand (E) fragments
3339 coil (C) fragments
Remove duplicates
Remove duplicates
Remove duplicates
1861 distinct helix fragments
1939 distinct strand fragments
3289 distinct coil fragments
Remove inconsistent fragments
7056 Structural fragments
Compute attributes
Classification system input data

Figure 17. The dataset preparation procedure

- D1 (423), D2 (924), D3 (1338), and D4 (1800), which include only first, first two, first three, and first four SFs, with respect to the protein head, for each protein,

- d2 (501), d3 (414), d4 (462), and d6 (429), which include only first, second, third, and fourth SF respectively,

- DA-2 (5284), DA-4 (3590) and DA-6 (2557), which include all SFs, but with removed 2, 4, and 6 residues on the SF ends (1, 2, and 3 residues are removed on each side),

- DA-s(hort) (2213), DA-m(edium) (2309), DA-l(ong) (1841), and DA-vl(verylong) (693) that include subsets of DA with different lengths defined to ensure similar distributions, see Table 17.

Table 17. Division of SFs by their length

dataset	helix		strand		coil	
	length	# SFs	length	# SFs	length	# SFs
DA-short	4÷9	629	3÷4	591	3÷4	993
DA-medium	10÷14	587	5÷6	694	5÷7	1028
DA-long	15÷22	467	7÷9	469	8÷15	905
DA-very-long	23÷68	176	10÷26	170	16÷74	347

Distribution of each SF type, i.e. helices, strands, and coils, for each dataset is shown in Figure 18. Datasets D1, d3, and D3 were discarded due to highly skewed distribution, which would result in inaccurate prediction results and remaining 13 datasets were kept further experiments. SFs in these sets were converted into the proposed attribute representation.

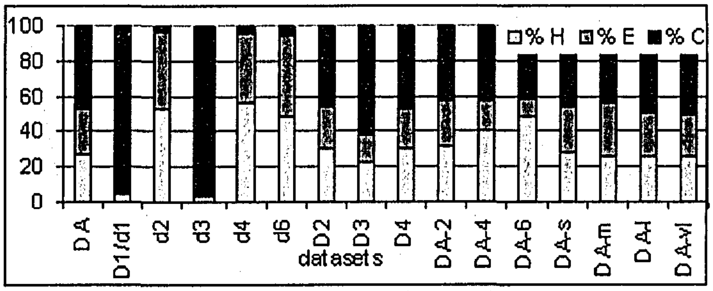


Figure 18. Distribution of the 3 secondary structures for considered datasets

The summary information for all the considered datasets of SFs is shown in Table 18.

Example-4 that splices 1FJNA protein, shown earlier, and puts resulting SFs into corresponding datasets is given in Table 19. Since the protein contains relatively short SFs, no fragments for datasets DA-l and DA-vl were extracted.

Table 18. Summary information for the considered datasets of SFs

Dataset	State	# fragments	length		Dataset	State	# fragments	length	
			Min.	Max.				Min.	Max.
DA	H	1863	4	68	d3	H	13	5	24
	E	1956	3	26		E	3	5	10
	C	3339	3	74		C	398	3	15
D2	H	282	4	68	d4	H	258	4	59
	E	225	3	19		E	184	3	17
	C	417	3	74		C	20	3	18
D4	H	553	4	68	d6	H	204	4	59
	E	412	3	19		E	203	3	17
	C	835	3	74		C	22	3	18
DA-2	H	1667	4	66	DA-s	H	629	4	9
	E	1334	3	24		E	591	3	4
	C	2238	3	72		C	993	3	4
DA-4	H	1432	4	64	DA-m	H	587	10	14
	E	640	3	22		E	694	5	6
	C	1518	3	70		C	1028	5	7
DA-6	H	1231	4	62	DA-l	H	467	15	22
	E	278	3	20		E	469	7	9
	C	1048	3	68		C	905	8	15
d2	H	267	4	68	DA-vl	H	176	23	68
	E	223	3	19		E	170	10	26
	C	11	3	15		C	347	16	74

Table 19. Results of splicing 1FJNA protein into SFs (all SFs are in italics)

dataset	structural fragments
DA	GFGCPN, NYQCHRHCKS, IPGRC, GGYC, GGWHRLR, CTCY, RCG
D2	GFGCPN, NYQCHRHCKS
D4	GFGCPN, NYQCHRHCKS, IPGRC, GGYC
d2	NYQCHRHCKS
d4	GGYC
d6	CTCY
DA-2	FGCP,YQCHRHCK, PGR, GWHRL
DA-4	QCHRHC, WHR
DA-6	CHRH
DA-s	GGYC, CTCY, RCG
DA-m	GFGCPN, NYQCHRHCKS, IPGRC, GGWHRLR

4.1.4 Prediction Algorithms

The SF prediction can be performed using a wide range of prediction algorithms. In this research, eight representative prediction algorithms were considered based on a generated model. These can be divided into two categories. Black-box algorithms generate models that cannot be interpreted by a user, and white-box algorithms generate interpretable models. The latter are further divided based on specific models into rule-based, decision trees and probabilistic algorithms. Representative prediction algorithms shown in Table 20 for each of the categories are used.

Table 20. Representative algorithms used to perform structural fragment prediction

algorithm type		algorithm name	reference
black-box		Multiple layer perceptron neural network (MLP)	(Hornik et al., 1989)
white-box	rule-based	RIPPER (RIP)	(Cohen, 1996)
		SLIPPER (SLI)	(Cohen & Singer, 1999)
	decision trees	ID3	(Quinlan, 1986)
		CART	(Breiman et al., 1984)
		C5.0	(RuleQuest, 2003)
		bC5.0 (C5.0 with boosting)	(RuleQuest, 2003)
	probabilistic	Naïve Bayes (NB)	(Duda and Hart, 1973)

To accommodate for ensemble methods, which were recently applied in context of structural class prediction (Tan et al., 2003), C5.0 algorithm was used with boosting, which generates and combines several models to increase accuracy (Schapire and Singer, 1998). Data generated for predicting attributes in this research are continuous. For Naïve Bayes algorithm, predicting attributes were discretized using equal-frequency discretization.

Implementation of these algorithms was obtained from the authors, and in case of the ID3, CART, MLP, and NB systems, the TANAGRA version 1.1.3 software, available at http://eric.univ-lyon2.fr/~ricco/tanagra/ was used; The C5.0 system combined with boosting option was obtained at http://rulequest.com/; RIPPER and SLIPPER systems can be obtained at http://www-2.cs.cmu.edu/~wcohen/.

4.2 Detailed Specific Goals

The general problem of SF prediction was used to address a number of specific goals related to how the three secondary structures can be distinguished (predicted) based on the primary sequence:

GOAL 1: Investigation of quality of different prediction algorithms. Eight different algorithms from four distinct families were tested and their accuracy was compared.

GOAL 2: Investigation of impact of particular sequence properties on the prediction accuracy. Several properties were investigated, including length of SFs, SF position in the sequence, and impact of the quality of secondary structure information for the outer most SF residues, i.e. residues on both SF ends. These goals directly translate into hypotheses that are more general:

SUB-GOAL 2.1 investigates if prediction should be performed using all known SFs as a reference, or if they should be separated into sub-groups, in this case based on length, to improve the prediction accuracy

SUB-GOAL 2.2 answers if the quality of prediction depends on the position of a given SF in the sequence.

Finally, **SUB-GOAL 2.3** questions if structure prediction for residues on the edge between different SFs suffers from decreased reliability of the secondary information, and if discarding these residues would results in changes with respect to prediction accuracy. It also allows addressing a situation when SFs would be extracted from a protein with some mistakes.

GOAL 3: Optimal attribute sets selection for representation. A comprehensive representation consisting of attribute sets described in section 3.2 and 4.1 were used for selection. An optimal subset of attribute sets was chosen with respect to prediction accuracy. Finally, three different representations, i.e. composition vector representation, selected attribute sets representation and the comprehensive representation, were compared.

Original dataset DA was used to investigate for goal 1. Size-wise datasets, DA-s, DA-m, DA-l, and DA-vl were used to investigate the sub-goal 2.1 and Goal 3. Position corresponding datasets, d2, d4, d6, D2, and D4 were used to investigate the sub-goal 2.2. Information missing at edges datasets, DA-2, DA-4, and DA-6 were used to investigate the sub-goal 2.3.

These above goals address issues related to investigation of specific factors and algorithms related to improving the ability to distinguish between secondary structure states, and thus lay foundations for development of a new family of structure prediction methods. Next chapter presents experimental results in support of each the defined goals.

5 Experiments, Results and Goal Analysis

In this chapter, specific goals defined in Section 4.2 were verified experimentally. The experiments are divided into two parts. The first set of experiments relates to the Goal 1 and Goal 2 and investigates the performance of the prediction. The second set of experiments relates to Goal 3 that concentrates on selection of attribute sets for prediction and compares the results with other representations.

5.1 Prediction Experiments

The first major set of experiments covers the SF prediction for the considered 13 datasets with eight prediction algorithms. To ensure statistical validity experiments were performed using ten-fold cross-validation. Over 1000 experiments were performed. Results report average accuracy and standard deviations. For all prediction algorithms, except RIPPER and SLIPPER that do not provide sufficient reports, average sensitivity and specificity were computed to give further insights. The results are summarized in Table 21. Discussion of the results is divided by the corresponding goals defined in section 4.2.

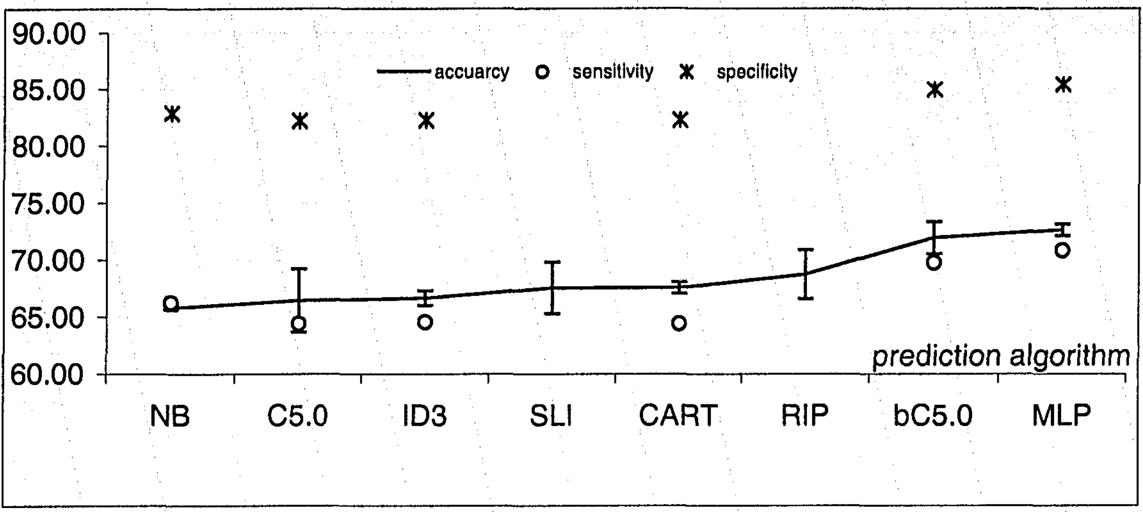
5.1.1 Goal 1: Prediction Algorithm Selection

Average accuracy for the eight prediction algorithms for DA dataset is 68.4%, which shows that simple prediction for general sets of SFs is similar with respect to accuracy of the first and second generation methods for protein structure prediction, which was between 60% and 66%. The slightly higher accuracy is a result of considering simpler problem of SF prediction and using more advanced protein sequence representation.

Table 21. Experimental results for prediction of protein SFs structure

	Dataset	MLP	RIP	SLI	ID3	CART	C5.0	bC5.0	NB	Avg.
accuracy±standard deviation	DA	72.6±0.5	68.8±2.1	67.6±2.3	66.7±0.6	67.6±0.5	66.5±2.8	72.0±1.4	65.8±0.2	**68.4**
	DA-s	79.6±0.8	78.3±1.4	76.7±1.9	77.5±0.6	77.4±0.5	77.3±2.3	78.8±3.5	78.4±0.2	**78.0**
	DA-m	85.0±0.5	84.5±2.6	84.0±3.3	87.7±1.1	87.5±1.5	84.0±2.1	86.9±1.0	85.1±0.6	**85.6**
	DA-l	87.5±0.6	89.7±1.5	87.5±3.2	87.8±0.8	88.1±0.3	89.2±1.7	89.9±1.9	87.9±0.4	**88.5**
	DA-vl	86.7±0.7	88.3±3.1	87.0±4.1	87.0±0.4	87.2±1.0	88.0±4.5	90.8±3.0	88.6±0.4	**87.9**
	d2	83.8±1.1	80.0±5.3	78.4±4.2	80.2±0.4	80.3±0.6	79.2±4.3	79.2±6.0	79.7±0.7	**80.1**
	d4	79.0±0.7	77.3±6.3	76.2±6.1	78.3±0.6	76.8±1.4	73.0±2.3	77.9±4.0	74.7±0.6	**76.6**
	d6	80.9±1.1	77.1±6.8	77.0±6.1	78.9±0.5	78.0±1.2	73.4±5.5	79.9±5.0	78.0±1.2	**77.9**
	D2	75.2±1.1	68.0±4.4	68.3±2.8	68.8±1.0	69.4±0.8	69.3±5.1	75.4±5.1	67.7±0.3	**70.3**
	D4	72.9±0.4	66.1±2.0	66.2±3.4	65.8±0.6	66.3±1.0	65.0±3.4	72.0±3.1	67.6±0.2	**67.7**
	DA-2	73.5±0.4	69.0±3.0	67.0±2.3	67.3±0.7	68.0±0.5	68.0±2.5	72.3±2.8	67.8±0.1	**69.1**
	DA-4	73.3±0.2	68.8±2.8	67.2±2.8	67.5±0.7	67.5±0.7	67.1±3.6	73.9±1.8	67.1±0.2	**69.0**
	DA-6	75.0±0.6	71.2±3.8	70.5±3.6	69.5±0.7	70.1±0.7	68.0±1.8	73.7±2.1	67.7±0.3	**70.7**
Average		**74.1**	**71.5**	**70.5**	**71.7**	**71.5**	**69.8**	**73.6**	**69.8**	
Sensitivity/specificity	DA	71 / 85	---	---	65 / 82	64 / 82	64 / 82	70 / 85	66 / 83	**67/83**
	DA-s	78 / 89	---	---	77 / 88	76 / 88	76 / 88	78 / 89	78 / 89	**77/88**
	DA-m	86 / 92	---	---	85 / 91	85 / 91	85 / 91	88 / 93	86 / 92	**86/92**
	DA-l	87 / 93	---	---	89 / 94	89 / 94	89 / 94	90 / 94	89 / 94	**89/94**
	DA-vl	87 / 93	---	---	87 / 93	87 / 93	88 / 93	91 / 95	89 / 94	**88/93**
	d2	57 / 89	---	---	55 / 88	55 / 88	54 / 87	54 / 87	54 / 87	**55/88**
	d4	55 / 86	---	---	55 / 86	54 / 86	55 / 83	57 / 86	55 / 83	**55/85**
	d6	61 / 88	---	---	56 / 87	55 / 86	60 / 84	66 / 88	65 / 87	**60/86**
	D2	73 / 87	---	---	67 / 83	68 / 84	68 / 84	73 / 87	69 / 84	**70/85**
	D4	71 / 86	---	---	64 / 82	63 / 81	63 / 81	70 / 85	69 / 84	**67/83**
	DA-2	72 / 86	---	---	66 / 83	66 / 83	67 / 83	71 / 86	68 / 84	**68/84**
	DA-4	69 / 82	---	---	62 / 82	63 / 83	63 / 83	69 / 86	67 / 84	**66/84**
	DA-6	64 / 82	---	---	59 / 82	57 / 82	57 / 82	62 / 85	67 / 84	**61/84**
Average		**72 / 88**	**---**	**---**	**68 / 86**	**68 / 86**	**68 / 86**	**72 / 88**	**71 / 87**	

The eight algorithms were ranked using average accuracy. MLP is the most accurate and bC5.0 is the second best, see Figure 19 (standard deviations are shown using vertical bars). Average accuracy, sensitivity and specificity shown in Figure 19, also confirm superiority of the two methods. High average specificity value of 83% means that false positives are very low and thus low accuracy is a result of relatively low sensitivity, i.e. low number of true positives.



vertical bars show confidence intervals equal to doubled standard deviation computed based on average over 10-folds cross-validation

Figure 19. Ranking of prediction algorithms on the DA dataset

This means that algorithms generate very selective models that can be further improved by relaxing some constrains, e.g. by pruning, to shrink gap between sensitivity and specificity and thus increase accuracy. For further insights, the average sensitivity and specificity of the models generated for each secondary structure state against prediction algorithms is plotted, see Figure 20. This graph shows that again MLP and bC5.0 prediction algorithms are generating better models for each secondary structure state when compared to other prediction algorithms. MLP and bC5.0 prediction algorithms are best suited for prediction of each of secondary structure states. In addition, independent of the prediction algorithm, specificities of helix and strand models are higher than the corresponding sensitivities. High specificity is most desirable than high sensitivity in medical data, i.e. low false positives are preferred. Hence, models generated for predicting helix and strand secondary structure states are more suitable for prediction when compared to the model generated for coil structure.

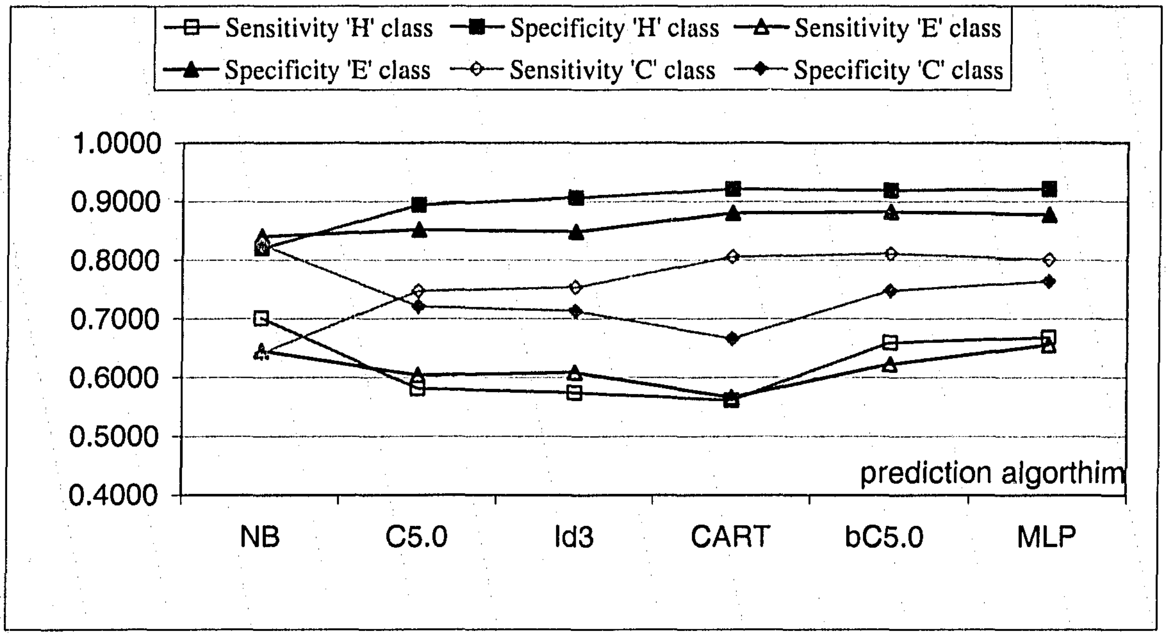
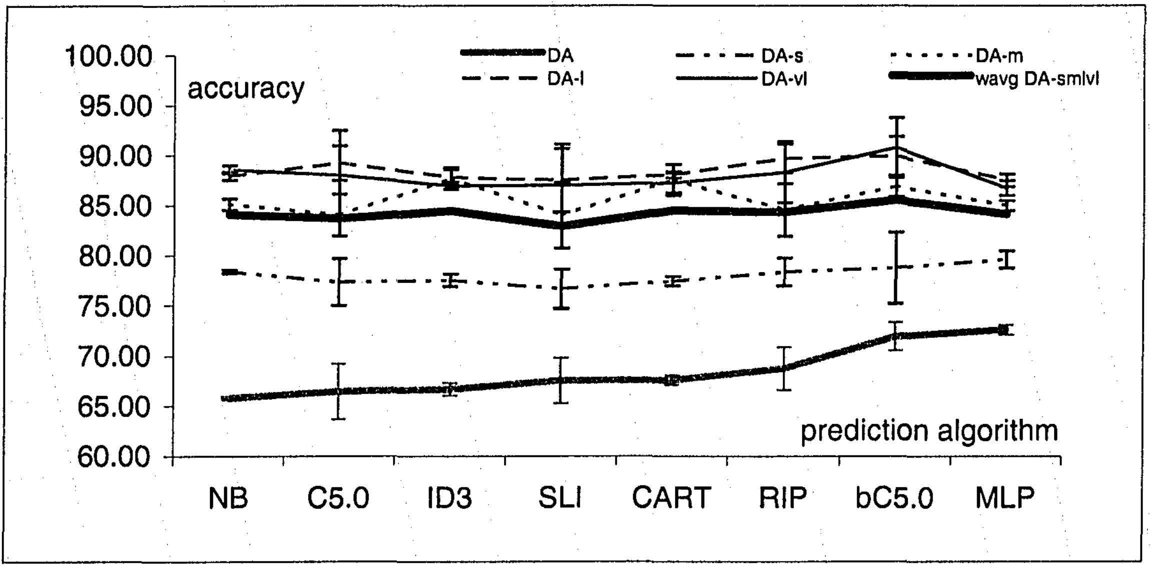


Figure 20. Sensitivity and specificity of each state model vs. prediction algorithms

5.1.2 Goal 2: Sequence Properties

Sub-Goal 2.1:

Results of prediction using all SFs. i.e. dataset DA, were compared to prediction when the data is separated by SF length into DA-s, DA-m, DA-l, and DA-vl (DA-size in short). Figure 21 shows comparison of results where wavg is a weighed, by the datasets size, average of accuracies when using DA-size datasets. The results show that prediction from SFs of similar length on average reduces errors by 50%, i.e. from 68% accuracy for DA to 84% when using subsets of similar length SFs. The specific amount of improvement is independent of the prediction algorithm, but the algorithms perform with more similar accuracy when compared to prediction from all SFs. Also, as expected the best accuracy is achieved for long and very long SFs since they contain more information than short and medium SFs. At the same time, prediction accuracy for short and medium SFs is still substantially better when compared to using dataset DA.



vertical bars show confidence intervals equal to doubled standard deviation computed based on average over 10-folds cross-validation

Figure 21. Results for prediction when using SFs of similar length

Closer analysis based on average, over all algorithms, sensitivity and specificity, shows that both true and false positive scores are always better than in case of DA-size datasets, see Figure 22. The best results are achieved for helical SFs, with specificity of 99% and sensitivity of 95%, which shows that 95% of helical SFs were classifies as helices and only 1% of other SFs was misclassified as helices. For coils, the sensitivity and specificity are similar and equal to about 85%, while on average sensitivity and specificity are the worst for strand fragments. These results show a significant pattern. In general, when the original data is divided into subsets the overall prediction accuracy evaluated using cross-validation is similar or worse. In contrast, the above results show that dividing SFs by their length allows distinguishing better between the structures.

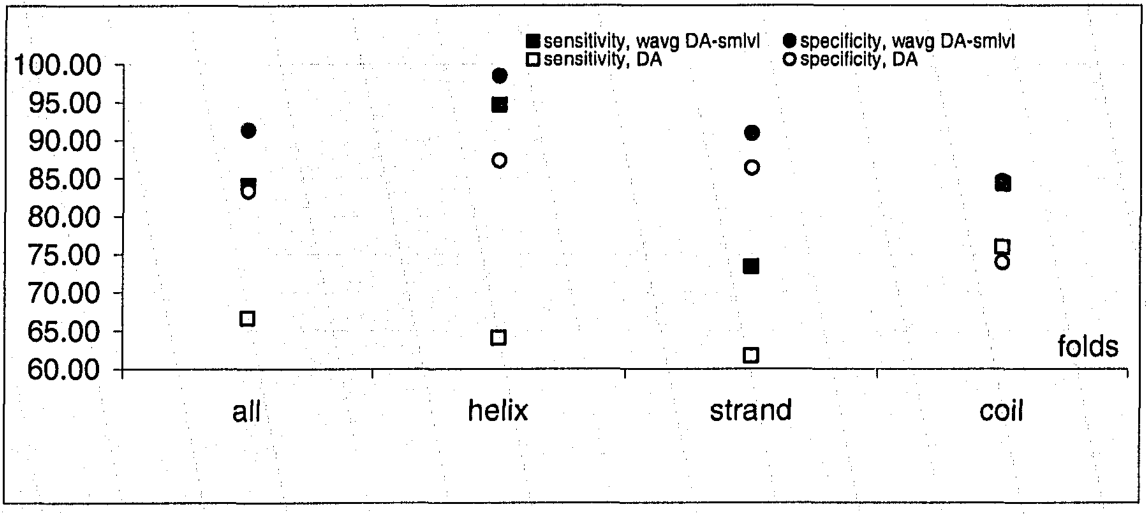
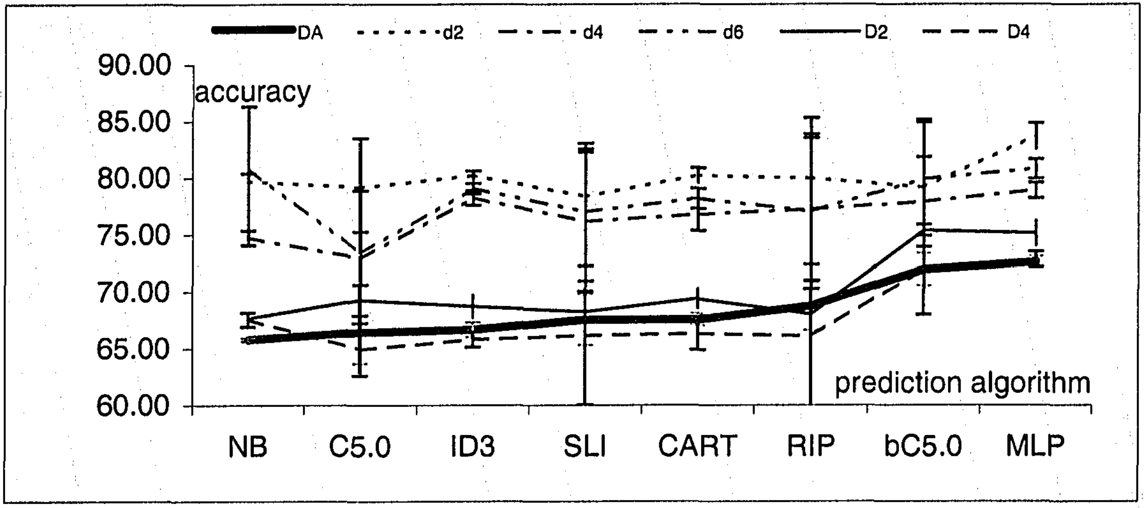


Figure 22. Sensitivity and specificity when using SFs of similar length

Most importantly, the average accuracy of SF prediction in this case is over 84%, which is significantly better than current accuracy of about 80% of the third generation alignment based structure prediction methods. This shows that the non-alignment based prediction, if properly performed, can potentially generate relatively high quality results.

Sub-Goal 2.2

Results when dataset DA was used were compared with results using d2, d4, d6, D2, and D4 datasets to evaluate if quality of prediction depends on the residue position, see Figure 23. The results show that dividing SFs into subsets of the same relative position in the protein sequence results in some improvements. For the d2, d4, and d6 the mean accuracies are 80%, 77%, and 78% respectively, which reduces the error by 30% compared to results when using all SFs.



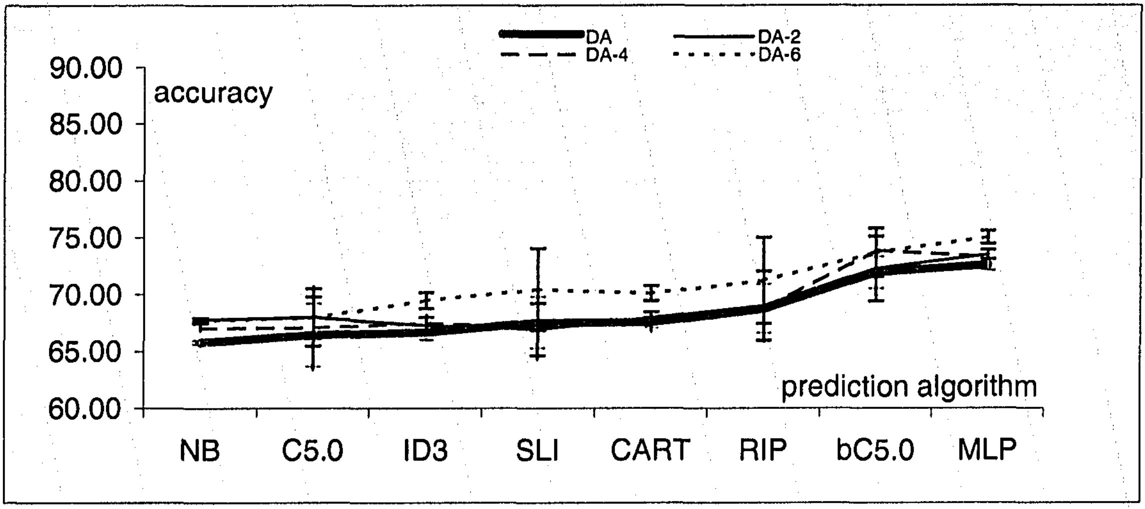
vertical bars show confidence intervals equal to doubled standard deviation computed based on average over 10-folds cross-validation

Figure 23. Results of prediction when using sequence position specific SFs

On the other hand, merging adjacent sequences starting at the protein head, i.e. D2 and D4 datasets, does not result in significant improvements. Thus, in contrast to results for sub-goal 2.1, the results are inconclusive and show no strong evidence of difference in prediction accuracy relative to where in the sequence, with respect to the protein head, a SF is positioned.

Sub-Goal 2.3

Results when dataset DA was used were compared with results on DA-2, DA-4, and DA-6 datasets to analyze if information for residues on the edge between different SFs suffers from decreased reliability with respect to SF prediction, see Figure 24. The four datasets use all SFs, but without respective residues on the outer edges. The results show that independently of the used prediction algorithm there is no negative impact of information located on the SF edges when compared to the results when the entire SFs were used, i.e. for DA dataset.



vertical bars show confidence intervals equal to doubled standard deviation computed based on average over 10-folds cross-validation

Figure 24. Results of prediction when removing residues on SF edges

This indicates that performing prediction on SFs that are incomplete or containing errors, i.e. some residues on the fragment edges are missing, would lead to similar ability to distinguish between different structures as in case when complete fragments would be used.

Average sensitivity and specificity values are computed on different groups of datasets to give further insights, see Table 22. Results show that a group containing DA-size datasets obtains on average the best results in case of both true and false positive scores, see the bolded values in columns five and nine. Helix models perform best when compared with coil and strand models independently of the dataset group, see bolded values in the last row. Out of all helix models, those generated from DA-size datasets are the best.

Table 22. Average sensitivity and specificity for each state model on dataset groups

Group	Sensitivity				Specificity			
	Helix	Strand	Coil	Average	Helix	Strand	Coil	Average
d2,d4,d6	80	81	08	56.33	83	78	99	86.67
D2,D4	67	63	75	68.33	88	88	76	84.00
DA-2,4,6	75	49	72	65.33	81	91	80	84.00
DA-s,m,l,vl	94	77	84	**85.00**	98	92	86	**92.00**
DA	62	62	76	66.67	90	86	74	83.33
Average	**76**	66	63	68.33	**88**	87	83	86.00

In addition, on average, models generated by considering SF-position (d2, d4, d6, D2 and D4) and information at edges (DA-2, DA-4, and Da-6) has the similar average sensitivity and specificity values when compared to dataset DA. This means that both of these factors have both no positive or negative impact on prediction accuracy.

5.2 Attribute Sets Selection Experiments and Results

The second major set of experiments was performed using ten-fold cross-validation with a total number of over 50000. These comprehensive experiments use four DA-size datasets, as they give the most accurate prediction, and three representative prediction algorithms, i.e. ID3, MLP, and NB. There are a total of 15 attribute sets considered for the selection, which is shown in Table 23.

Table 23. Attribute sets considered for representation

Attribute set Id#	Attribute set name	# of set attributes	Attribute set Id#	Attribute set name	# of set attributes
1	SF length	1	9	autocorrelation	10
2	# duplicates	1	10	electronic group	5
3	relative position	1	11	R group	5
4	Eisenberg's hydrophobicity	2	12	exchange group	3
5	Fauchere's hydrophobicity	2	13	hydrophobic group	2
6	molecular weight	1	14	other group	7
7	composition vector	20	15	chemical group	10
8	composition moment vector	20			

The goal is to find a subset of the attribute sets representation described in section 3.2 and 4.1 that uses only selected attribute sets while giving comparably good results. Attribute set selection was performed iteratively, where in each step each attribute sets was individually tested, and the best one was selected. The first iteration uses one attribute set at a time for prediction and results are shown in Table 24. Bolded values in the table show the best results. Each value in this table represents the average prediction accuracy of the 10 fold-cross validations. Results show that prediction using composition moment vector attribute set has the similar average accuracy as the prediction with SF length attribute. The composition moment vector was selected because of it was best more often. On average up to 25% worse accuracy is achieved by using the single attribute set for prediction when compared with prediction using all attribute sets. Hence, the second selection iteration is executed. The second iteration of experiments is done by considering a pair of attribute sets in each step for the prediction, which includes the first selected attribute set and one attribute set from the remaining attribute sets.

Table 24. Experimental results for prediction accuracy in iteration 1

Attribute set Id /Dataset	DA-short			DA-medium			DA-long			DA-very long			Avg. Accu.
	NB	ID3	MLP	NB	ID3	MLP	NB	ID3	MLP	NB	ID3	MLP	
1	70	69	**70**	69	68	68	**84**	82	**84**	**78**	73	72	**73.92**
2	45	42	45	44	42	44	49	n/a	49	50	n/a	50	46.00
3	45	43	45	44	39	45	49	48	50	50	49	50	46.42
4	56	49	49	52	56	50	53	56	53	54	56	52	53.00
5	45	57	54	61	64	61	60	64	59	62	66	63	59.67
6	69	44	45	44	46	47	49	49	50	50	50	49	49.33
7	62	65	56	73	72	72	72	73	74	66	72	**83**	70.00
8	**75**	**78**	55	64	**83**	68	66	**84**	77	64	**87**	79	**73.33**
9	67	73	69	**79**	79	**76**	68	65	70	58	65	65	69.50
10	68	66	59	68	62	63	61	62	63	62	60	65	63.25
11	61	61	58	65	64	56	62	59	60	57	58	61	60.17
12	65	65	61	66	60	55	60	58	62	56	57	60	60.42
13	56	45	57	53	47	49	54	49	51	50	50	51	51.00
14	62	67	56	61	69	60	62	62	61	54	57	58	60.75
15	67	63	64	73	72	64	72	71	70	63	70	73	68.50
Average	61	59	56	61	62	59	61	63	62	58	62	62	60.50
All Attributes	78	80	78	88	85	85	88	87	88	87	87	89	85

Experimental results are shown in Table 25, which resulted in the selection of sequence length attribute because of its superior accuracy when compared with any other pair of attribute sets. The third selection iteration was executed since still 2% to 19% less accuracy is achieved by using the pair of attribute sets when compare with using all attribute sets.

Table 25. Experimental results for prediction error in iteration 2

Attribute set Id /Dataset	**DA-short**			**DA-medium**			**DA-long**			**DA-very long**			**Avg. Accu.**
	NB	ID3	MLP	NB	ID3	MLP	NB	ID3	MLP	NB	ID3	MLP	
1	**76**	**79**	**76**	**76**	**83**	**82**	**86**	**88**	**89**	**88**	**87**	**86**	**83.00**
2	62	78	56	56	64	68	65	n/a	77	64	n/a	79	66.90
3	61	78	55	55	64	68	64	84	77	64	**87**	79	69.67
4	61	**79**	56	56	64	67	65	86	76	62	86	79	69.75
5	62	**79**	58	58	66	69	67	86	78	66	86	81	71.33
6	62	78	54	54	64	68	64	84	77	64	**87**	78	69.50
7	69	78	55	55	73	73	72	84	80	66	84	85	72.83
8	---	----	----	----	----	----	----	----	----	---	----	-----	-----
9	**76**	78	71	71	81	81	73	83	82	67	84	82	77.42
10	67	78	60	60	72	72	67	84	78	68	86	80	72.67
11	66	**79**	64	64	66	71	65	84	78	65	86	78	72.17
12	63	**79**	63	63	67	68	64	84	78	64	86	78	71.42
13	64	**79**	63	63	65	69	65	84	76	64	**87**	79	71.50
14	64	**79**	60	60	64	72	64	84	77	66	85	79	71.17
15	66	79	65	65	73	71	70	85	79	67	86	81	73.92
Average	66	79	61	61	69	71	68	85	79	67	86	80	72.67
All Attributes	78	80	78	88	85	85	88	87	88	87	87	89	85.00

The third iteration was performed by considering three attribute sets at each step for the prediction, which includes the selected 2 attribute sets and the third attribute set from the remaining attribute sets. Experimental results are shown in Table 26, chemical group attribute set was selected due to its high average accuracy and large number of times it resulted in achieving highest accuracy.

Table 26. Experimental results for prediction error in iteration 3

Attribute set Id /Dataset	DA-short			DA-medium			DA-long			DA-very long			Avg. Accu.
	NB	ID3	MLP	NB	ID3	MLP	NB	ID3	MLP	NB	ID3	MLP	
1	---	---	---	---	---	---	---	---	---	---	---	---	---
2	76	79	76	83	83	82	86	n/a	**89**	88	n/a	86	82.80
3	76	79	76	83	84	82	86	87	**89**	**89**	87	87	83.75
4	76	**80**	76	83	84	80	88	87	88	88	87	86	83.58
5	77	**80**	77	84	**85**	82	**89**	**88**	88	88	87	86	**84.25**
6	75	79	76	83	84	82	86	87	88	**89**	87	86	83.50
7	76	**80**	75	83	**85**	**84**	86	**88**	87	88	87	**88**	83.92
8	---	---	---	---	---	---	---	---	---	---	---	---	---
9	76	79	76	84	84	83	88	87	87	88	**88**	**88**	**84.00**
10	**78**	79	77	84	**85**	83	86	**88**	88	**89**	87	87	**84.25**
11	75	79	77	83	84	83	86	87	88	88	86	86	83.50
12	76	79	77	83	84	82	87	**88**	87	88	87	86	83.67
13	77	79	76	84	84	82	87	87	88	88	87	85	83.67
14	75	79	76	83	84	82	87	87	87	88	87	85	83.33
15	77	**80**	**78**	**85**	**85**	83	88	**88**	88	87	87	87	**84.42**
Average	76	79	76	83	84	82	87	87	88	88	87	86	83.58
All Attributes	78	80	78	88	85	85	88	87	88	87	87	89	85.00

5.2.1 Goal 3: Selection of Attribute Representation

Summary of attribute selection results for the above three iterations is shown in Table 27. Table 23 explains numbering of attribute set ids. Selected attribute sets are shown in gray color and best five in each iteration are shown in bold. The attribute sets are ranked based on prediction accuracy and the rank values are shown in Table 27.

Table 27. Attribute selection results

Iteration/ Attribute set id#		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	avg
1	avg accuracy	**74**	46	46	53	60	49	**70**	**73**	**70**	63	60	60	51	61	**69**	60.5
	# times best	**4**	0	0	0	0	0	**1**	**5**	**2**	0	0	0	0	0	**0**	
	rank	**1**	15	14	11	10	13	**3**	**2**	**4**	6	9	8	12	7	**5**	
2	avg accuracy	**84**	70	72	72	73	72	**75**	**n/a**	**78**	**75**	74	73	73	73	**75**	74.5
	# times best	**9**	0	0	0	0	0	**0**	**n/a**	**1**	**0**	1	0	0	0	**1**	
	rank	**1**	11	14	12	7	13	**4**	**n/a**	**2**	**5**	6	8	9	10	**3**	
3	avg accuracy	**n/a**	83	84	84	**84**	83	**84**	**n/a**	**84**	**84**	84	84	84	83	**84**	83.8
	# times best	**n/a**	1	0	0	**3**	1	**3**	**n/a**	**1**	**1**	0	0	0	0	**3**	
	rank	**n/a**	13	10	9	**2**	12	**5**	**n/a**	**3**	**4**	8	7	6	11	**1**	

Selection process stopped at the third iteration since the selected attribute sets already gave accuracy comparable with accuracy when using all attribute sets. The attribute set rank in all iterations given in Table 27 shows that autocorrelation, electronic group, and composition vector also contribute to improved accuracy. For virtually all experiments, hydrophobicity attribute set computed using Fauchere's index was superior to the Eisenberg's index. The composition moment vector gave on average significantly better results than commonly used composition vector, which confirms results in (Ruan et al, 2005a). In short, the results show that only a handful of attributes is needed to distinguish between the three types of SFs, but at the same time the selected attributes are different than the commonly used attribute representations. As such, the results provide useful guidelines for attribute selection.

Additionally, prediction accuracies when using 1) all attribute sets, 2) the selected attribute sets consists of composition moment vector, SF length and chemical group, and 3) most commonly used representation that includes composition vector, were compared. The experiments were performed using eight representative prediction algorithms applied on DA and DA-size datasets, see Table 28. Bolded values in bold help to perform side-by-side comparison of the best accuracies for different representations and for using all SFs vs. dividing the SFs by their length. The main finding shows that division of SFs by length does not bring significant improvements when insufficient attribute representation is used, e.g. see results for the representation 1 for the most accurate MLP algorithm. This result confirms problems of the first and second-generation protein structure prediction methods. Although these methods usually used polypeptide sequences of similar size to perform prediction, apparently lack of sufficient information and use of neural networks were their limiting factors.

Table 28. Comparison of prediction with different attribute representations

Note: 1 (composition vector only), 2 (selected best attributes), 3 (all attributes), wavg (weighted average for DA-size datasets)

Prediction	**DA-s**			**DA-m**			**DA-l**			**DA-vl**			**wavg**			**DA**		
algorithm	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
MLP	66	80	80	72	85	85	72	89	87	72	86	87	**70**	**85**	**84**	**68**	**72**	**73**
RIP	76	76	78	84	84	85	82	89	90	73	74	88	80	82	84	65	68	69
SLI	71	76	77	80	84	84	76	86	88	70	70	87	75	81	83	63	67	68
ID3	68	77	78	73	84	88	72	88	88	66	87	87	70	83	85	64	67	67
CART	72	77	77	78	84	88	76	88	88	70	87	87	75	83	84	64	67	68
C5.0	73	76	77	81	84	84	76	89	89	71	87	88	76	83	84	64	67	66
bC5.0	78	79	79	85	87	87	85	90	90	82	91	91	**83**	**86**	**86**	**68**	**71**	**72**
NB	56	78	78	72	85	85	74	88	88	83	87	89	69	84	84	65	67	66
Average	**70**	**78**	**78**	**78**	**84**	**86**	**77**	**88**	**88**	**73**	**84**	**88**	**75**	**83**	**84**	**65**	**68**	**68**

The results show that on average, 10% accuracy was gained by dividing SFs by their length, and additional 9% was gained by using the new attribute representation. Using just improved attribute representation without dividing by SF length gives relatively small, about 3%, average improvement in accuracy. In short, methods, which combine both using fragments of relatively similar length and apply the proposed protein sequence representation, are expected to be able to distinguish between different secondary structures with high accuracy.

6 Summary and Future Work

6.1 Summary

Protein structure is the key to understand protein functions and interactions with other molecules. Research in computational methods for prediction of protein structures from sequences is in high demand due to huge known-sequence structure gap and availability of different protein databases. Although many advanced prediction methods exists, we are still far from being able to achieve a highly accurate solution for this problem. Most of the current methods for predicting 3-state secondary structure are alignment-based, while this study explores attribute representation based on protein sequence information to predict the 3-state secondary structure for protein structural fragments (SF).

This study proposes and analyzes a novel SF prediction problem to investigate how well 3-state secondary structures can be distinguished based on sequence information using a comprehensive attribute representation. Three goals are defined. Goal 1 investigates performance, in terms of the SF prediction accuracy, of the eight representative prediction algorithms Goal 2 investigates the impact of three different factors on SF prediction accuracy. Those factors are structural fragment length; structural fragment position in the protein sequence and removal of information at structural fragment edges. Goal 3 performs feature selection to find optimal, in terms of the trade-off between accuracy and number of features, protein sequence representation.

Based on comprehensive experimental study this research investigates the defined goals. Results for goal 1 show that two prediction algorithms, namely multiple layer feed-forward neural network and boosted C5.0 decision tree, perform best among the considered eight algorithms. Therefore, selection of a suitable method for a prediction application that needs to distinguish between the three secondary structures is one of critical considerations. Results for goal 2 show that significant improvements can be made by grouping SFs of similar length. On the other hand, the results show that the ability to recognize SFs does not depend on their position in the sequence, as well as it does not depend on the availability of structural information on the fragment's edges. Finally, investigation of the goal 3 shows that currently used attribute based representation of a protein sequence should be modified to include composition moment vector, SF length, chemical group, hydrophobic autocorrelation, and electronic group information to improve ability to distinguish between different secondary structure states. In short, results for different goals show that on average considering SFs of similar length increases prediction accuracy by 10%, and using proposed in this thesis attribute representation provides further increase of 9%. Combining best attribute sets selection and SFs of similar length reduces the prediction error by 50%.

The considered SF prediction problem task provides useful guidelines to improve the existing attribute methods and design of new prediction methods for structure, content, and structural class. Although different prediction algorithms and representations are used by current prediction methods, they fail to provide major improvements due to insufficient attribute representation and using all SFs at the same time.

6.2 Future Work

The SF prediction problem can be seen as generalization of other protein structure prediction tasks, i.e. structure, content and structural class prediction. All these problems aim to predict one of the three state protein secondary structures, and thus using the SF prediction task to discover a set of guidelines concerning ability to better distinguish between the structures would provide researchers in the structure prediction domain with very important information.

The SF prediction can be also directly applied to perform protein structure prediction. Figure 25 shows how it can be applied to perform protein structure prediction.

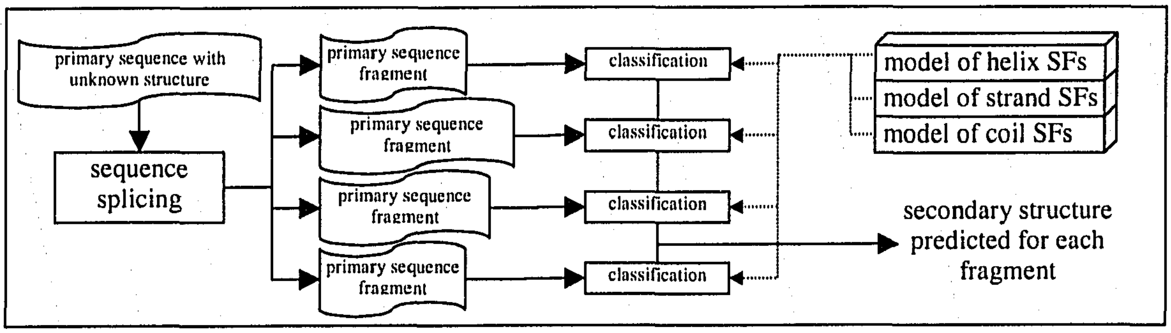

primary sequence with unknown structure
sequence splicing
primary sequence fragment
primary sequence fragment
primary sequence fragment
primary sequence fragment
classification
classification
classification
classification
model of helix SFs
model of strand SFs
model of coil SFs
secondary structure predicted for each fragment

Figure 25. Structure prediction task based on the considered SF prediction

In this case SFs models are used to classify protein fragments. To increase accuracy, appropriate models generated for four DA-size datasets should be used by matching the size of the protein fragment with the sizes of the SFs from the datasets. The main challenge in this application is to splice the protein into structurally uniform fragments, which can be done in at least two ways:

- by prediction of structure sequence turn points (from helix or strand to coil and from coil to helix or strand). This is due to regularity in which secondary structure aligns, i.e. helix and strand SFs are very rarely directly connected, but rather are connected through a coil SF.
- by using idea of protein fingerprint (Ruan et al., 2005b). The protein fingerprint consists of a set of four hydrophobicity based curves that are used to splice the sequence into structurally uniform fragments.

The SF prediction can be also applied to perform content and structural class prediction. In this case, a sliding window of fixed size will be applied to classify the corresponding sequence fragments to one of the structures. The sequence fragment under the window will be classified to one of the structures, and the number of times each of the structures was recognized will be used to compute the content values and the corresponding structural class.

The size of the window should correspond to average sizes of SFs from one of the DA-size datasets, and models generated for this dataset should be used for classification. To improve accuracy several passes with windows of different sizes, which correspond to different DA-size datasets, should be performed. The overall diagram showing this prediction application is shown in Figure 26. Results for goal 2.3 have important implications in case of both applications. The spliced sequence fragments and window sizes do not have to precisely correspond to sizes of SFs from the datasets that were used to generate the helix, strand, and coil models.

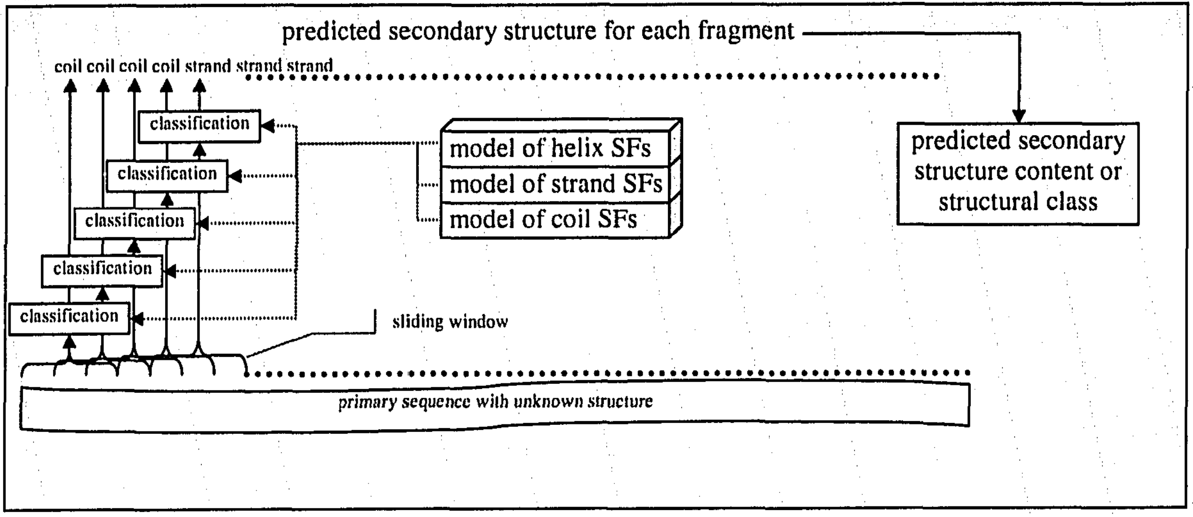

predicted secondary structure for each fragment
coil coil coil coil strand strand strand
classification
classification
classification
classification
classification
model of helix SFs
model of strand SFs
model of coil SFs
predicted secondary structure content or structural class
sliding window
primary sequence with unknown structure

Figure 26. Content and structural class prediction task based on the SF prediction

Results for goal 2.3 show that even if a sequence would be spliced into fragments that are slightly smaller or the window size will be slightly too small with respect to the SFs size, the prediction of the corresponding structural class should be still reliable. These applications are beyond the scope of this research, and will be addresses as the future work.

References

1. Altschul S, Gish W, Miller W, Myers EW, and Lipman D, (1990), A basic local alignment search tool, *Journal of Molecular Biology*, 215(3): 403–410.

2. Altschul, S., Madden, T., et al., (1997), Gapped blast and psi-blast: a new generation of protein database search programs, *Nucleic Acids Research*, 25, 3389-3402.

3. Andreeva A. et al., (2004), SCOP Database in 2004: Refinements integrate structure and sequence family data, *Nucleic Acid Research*, 32, D226-9.

4. Bairoch A. and Boeckmann B., (1993), The SWISS-PROT protein sequence data bank, recent developments, *Nucleic Acid Research*, 21:3097-3103.

5. Bairoch, A. and Apweiler, R., (2000), The SWISS-PROT protein sequence database and its supplement trembl , *Nucleic Acid Research*, 28, 45-48.

6. Baldi P and Søren Brunak, (1998), *Bioinformatics: The Machine Learning Approach*, MIT Press.

7. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D. et al. (1977), The protein data bank: A computer based archival file for macromolecular structures, *Journal of Molecular Biology*, 112, 535-542.

8. Berman, H. M., Westbrook, J., Feng,Z., Gillliland, G., Bhat, T. N. et al. (2000), The protein data bank, *Nucleic Acids Research*, 28, 235-242.

9. Black, S.D., and Mould, D.R., (1991), Development of hydrophobicity parameters to analyze proteins, Which Bear Post- Or Co-Transitional Modifications, *Analytical Biochemistry*, 193, 72-82.

10. Boeckmann B. et al. (2003), The SWISS-PROT protein knowledgebase and its supplement trembl, *Nucleic Acids Research*, 31, 365-370.

11. Boros, E., Hammer, P.L., Ibaraki, T., Kogan, A., Mayoraz, E., and Muchnik, I., (2000), An implementation of logical analysis of data, *IEEE Transactions on Knowledge and Data Engineering*, 12:2, 292-306.

12. Brändén, C. and Tooze, J. (1991), *Introduction to Protein Structure*, Garland Publ., New York, London

13. Breiman, L., Friedman, J., et al., (1984),*Classification and Regression Trees*, Chapman and Hall.

14. Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, James D. Watson, (1994), *Molecular Biology of the Cell*, Garland Pub, 3rd edition.

15. Burkhard Rost, Jinfeng Liu, Dariusz Przybylski, Rajesh Nair, Kazimierz O. Wrzeszczynski, Henry Bigelow & Yanay Ofran, (2002), Predicting protein structure and function through evolutionary information, *Chemoinformatics - From Data to Knowledge*, J Gasteiger & T Engel (eds.), Wiley.

16. Bystroff, C., Thorsson, V. and Baker, D., (2000), HMMSTR: A Hidden Markov model for local sequence-structure correlations in proteins, *Journal of Molecular Biology*, 301, 173-190.

17. Cai, D., Delcher, A., Kao, B. and Kasif, S., (2000), Modeling splice sites with Bayes networks, *Bioinformatics*,16: 152-158.

18. Cai Y. et al. (2003) Support Vector Machines for Prediction of Protein Domain Structural Class, *J Theoretical Biology*, 221, 115-120

19. Cedeno, W., and Vemuri, V., (1993), An investigation of DNA mapping with genetic algorithms: preliminary results, *In: Proc. Of the Fifth Workshop on Neural Networks,* Vol. 2204 of SPIE

20. Chisholm, M., Tadepalli., P., (2002), Learning decision rules by randomized iterative local search, *Proc. of the Intern. Conference on Machine Learning*, 75-82.

21. Chou K-C and Cai Y-D, (2004), Predicting Protein Structural Class by Functional Domain Composition, *Biochem. and Bioph. Research Comm.*, 321, 1007-1009

22. Chou P.Y., and Fasman G.D., (1978), Prediction of the Secondary Structure of Proteins from Their Amino Acid Sequences, *Advances in Enzymology*, 47, 45-148

23. Cohen, W., (1993), Efficient pruning methods for separate-and-conquer rule learning systems, *Proc. of the 13th Intern. Joint Conference on Artificial Intelligence,* 988-994.

24. Cohen, W., (1994), Grammatically biased learning: learning logic programs using an explicit antecedent description language, *Artificial Intelligence*, 68, 303-366.

25. Cohen, W., (1995), Fast effective rule induction, *Proc. of the 12th Intern. Conf. on Machine Learning*, 115-123.

26. Cohen, W., (1996), Learning trees and rules with set-valued attributes, *Proc. of the 13th National Conf. on Artificial Intelligence*, 709-716.

27. Cohen, W., and Singer, Y., (1999), A simple, fast and effective rule learner, *Proceedings of the 16th National Conference on Artificial Intelligence*, 335-342.

28. Cherkauer, K.J. and Shavlik, J.W., (1993), Protein structure prediction: selecting salient features from large candidate pools, *In: Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*,Bethesda, MD: AAAI Press.

29. Chou, P. Y., Fasman, G. D. (1974), Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins, *Biochemistry*, 13, 211.

30. Chou, P.Y., and Fasman, G.D., (1978), Prediction of the secondary structure of proteins from their amino acid sequences, *Advances in Enzymology*, 47, 45-148.

31. Cornette, J.L., Cease, K., Margalit, H., et al., (1987), Hydrophobicity scales and computational techniques for detecting amphipathic structures in protein, *Journal of Molecular Biology*, 195, 659-685.

32. Cox, D.R. and Miller, H.D., (1965), *The Theory of Stochastic Processes*, Chapman and Hall.

33. Dain, O., Cunningham, R., and Boyer, S., (2004), IREP++ a faster rule learning algorithm, *Proc. of the 4th SIAM Intern. Conference on Data Mining*, 138-146, Lake Buena Vista, FL.

34. Davis, L. ed., (1991), *Handbook of genetic algorithms*. New York: Van Nostrand Reinhold.

35. Dayhoff MO, Schwartz RM, and Orcutt BC., (1978), A model of evolutionary change in proteins, *Matrices for Detecting Distant Relationships*, 5, 345–358.

36. De Jong, K.A. (1990), Genetic-algorithm-based learning. In: *Machine Learning: An Artificial Intelligence*

37. Dietterich T.G., and Michalski R.S., (1981), Inductive learning of structural descriptions: evaluation criteria and comparative review of selected methods, *Artificial Intelligence*, 16:3, 257-294.

38. Ding, C.H.Q. and Dubchak, I., (2001), Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics*, 17: 349-358

39. Domingos, P., (1994), The RISE system: conquering without separating, *Proc. of the 6th IEEE Intern. Conference on Tools with Artificial Intelligence*, New Orleans, LA, 704-707.

40. Donald G. Truhlar et al , (1999), Pharmaceutical drug design reference from CHIPS, *Rational Drug Design.*

41. Dubchak I et al., (1997) Protein Folding Class Predictor for SCOP: Apprach Based on Global Descriptors, *Proc of 5th ISMB*, 104-107

42. Duda, R.O., and Hart, P.E., *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 1973

43. Durbin R, Eddy S, Krogh A, and Mitchison G., (1998), *Biological Sequence Analysis: Probabilistic Models Of Proteins And Nucleic Acids,* Cambridge University Press.

44. Eddy, SR., (2001), HMMER: Profile hidden markov models for biological sequence analysis, *WWW Document*, http://hmmer.wustl.edu.

45. Eisenberg, D., Weiss, R.M., Terwillinger, T.C., (1984), The hydrophobic moment detects periodicity in protein hydrophobicity, *Proceedings of the National Academy of Science*, 81:1,140-144.

46. Eisenberg, D., Wilcox, W., McLachlan, A.D., (1986), Hydrophobicity and amphiphilicity in protein structure, *Journal of Cell Biochemistry*, 31:1, 11-17.

47. Eisenhaber, F., et al., (1996), Prediction of secondary structural contents of proteins from their amino acid composition alone: I. New analytic vector decomposition methods, *Proteins*, 25:2, 157-168.

48. Engel R, (1982), Autoregressive conditional heteroskedastcity with estimates of the variance of united kingdom inflation, *Econometrica*, 50, 987-1008.

49. Eeyrich, V., Martí-Renom, M.A., Przybylski, D., Fiser, A., Pazos, F. et al. (2001), EVA: Continuous automatic evaluation of protein structure prediction servers, *Bioinformatics,* 17, 1242-1243.

50. Fauchere, J.L and PlISKA, V., (1983), Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides, *European journal of medicinal chemistry*, 18, 369-375.

51. Ferran, E.A. and Ferrara, P., (1992), Clustering proteins into families using artificial neural networks, *Comput. Appl. Biosci*, 8: 39-44

52. Fickett, J., & Cinkosky, M., (1993), A genetic algorithm for assembling chromosome physical maps, *In: Proc. Of the Second International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis*, 272-285

53. Fisher D. et al., (2003), CAFASP3: 3rd Critical assessment of fully automated structure prediction methods, *Proteins*, 53, 503-516.

54. Forster MJ, (2002), Molecular modeling in structural biology, *Micron*, 33(4): 365-84.

55. Furey, T.S., Cristianini, N., Duffy, N.,Bednarski, D.W., Schummer, M. and Haussler, D., (2000), Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics,*16: 906-914.

56. Furnkranz, J., and Widmer, G., (1994), Incremental reduced error pruning, *Machine Learning: Proc. of the 11th Annual Conference*, New Brunswick and New Jersey.

57. Ganapathiraju, M.K. et al. (2004), Characterization of protein secondary structure, *IEEE Signal Processing Magazine*, 78-87.

58. Garnier, J., Osguthorpe, D.J., and Robson, B., et al., (1978), Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *Journal of Molecular Biology*, 120:1, 97-120.

59. Garnier, J., Gibrat, J.-F. & Robson, B. (1996), GOR method for predicting protein secondary structure from amino acid sequence, *Methods in Enzymology*, 266, 540-553.

60. Gibrat, J.F., Garnier, J., and Robson, B., (1987), Further developments of protein secondary structure prediction using information theory: New parameters and consideration of residue pairs, *Journal of Mòlecular Biology*, 198:3, 425-443.

61. Gribskov, M. and Veretnik, S., (1996), Identification of sequence patterns with profile analysis, *Methods in Enzymology*, 266: 198-227.

62. Gusfield D, (1997), *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press.

63. Guo J. et al., (2004), PROSPECT-PSSP: An Automatic Computational Pipeline for Protein Structure Prediction, *Nucleic Acid Research*, 32, W522-W525

64. Hargbo, J., and Elofsson, A., (1999), Hidden Markov Models that use predicted secondary structures for fold recognition, *Proteins*, 36, 68-76.

65. Henikoff, S and Henikoff, JG (1991), Automated assembly of protein blocks for database searching, *Nucleic Acids Research*, 19:6565-6572.

66. Hirst, J.D. and Sternberg, M.J.E., (1992), Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks, *Biochemistry*, 31: 7211-7218.

67. Hobohm, U., Scharf, M., et al., (1992), Selection of a representative set of structures from the Brookhaven protein data bank, *Protein Science*, 1, 409-417.

68. Hobohm, U., and Sander, C., (1994), Enlarged representative set of protein structures, *Protein Science*, 3, 522.

69. Hornik, K., Stinchcombe, M., and White, H., (1989), MLP's are universal approximations, *Neural Networks*, 2, 359-366.

70. Holte, R.C., (1993), Very simple classification rules perform well on most commonly used datasets, *Machine Learning*, 11, 63-90.

71. Hooft R.W.W., C. Sander and G. Vriend., et al., (1996), The PDBFINDER database: A summary of PDB, DSSP and HSSP information with added value, *CABIOS*, 12, 525-529.

72. Jones, D. T., (1999), Protein secondary structure prediction based on position-specific scoring matrices, *Journal of Molecular Biology*, 292, 195-202.

73. Kabsch W, and Sander C., (1983), Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical attributes, *Biopolymers*, 22:12, 2577-637

74. Karplus, K. et al., (2001), What is the Value added by Human Intervention in Protein Structure Prediction? *Proteins*, 45 (Supp 5), 86–91

75. Kaufman, K.A., and Michalski, R.S., (1999), Learning from inconsistent and noisy data: The AQ18 approach, *Proc. of the 11th Intern. Symposium on Methodologies for Intelligent Systems*, Warsaw, Poland, 411-419.

76. Kihara, D., and Skolnick, J., (2003), The PDB is a covering set of small protein structures, *Journal of Molecular Biology*, 223, 793-802.

77. King, R. D., Srinivasan, A. and Sternberg, M. J. E., (1995), Relating chemical activity to structure: an examination of ILP successes, *New Gen. Computing*, 13: 411-433

78. Kodratoff Y. and Michalski eds. San Mateo R., Morgan Kaufmann., (1990), *Machine Learning: An Artificial Intelligence Approach,* Vol. III, San Mateo, CA, Morgan Kaufmann Publishers, 63-111.

79. Kohavi, R. & Provost, F., (1998), Editorial for the special issue on applications of Machine Learning and the knowledge Discovery process, *Glossary of Terms*, 30:2/3.

80. Kubat, M., Holte, R. C., Matwin, S., (1998), Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30 (2/3).

81. Kurgan, L., and Homaeian, L., (2005), Prediction of secondary protein structure content from protein sequence alone - a feature selection based approach, *Proceedings of the International Conference on Machine Learning and Data Mining (MLDM 2005),* 334-345, Leipzig, Germany, Springer Verlag, LNAI 4587, 2005.

82. Kurgan, L.A., Kedarisetti, K., (2005), Secondary protein structure Fragments – Feasibility study in prediction and analysis, Proceedings of the symposium on Human-Centric Computing (HC_205), 26-36

83. Kurowski M. and Bujnicki J. (2003), GeneSilico protein structure prediction meta-server, *Nucleic Acids Research*, 31:13, 3305-07

84. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, and Wootton JC, (1993), Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment, *Science*, 262(5131):208–214.

85. Li Q-Z and Lu Z-Q, (2001), The Prediction of Structural Class of Protein: Application of the Measure of Diversity, *J Theoretical Biology*, 213, 493-502

86. Lim, V. I., (1974), Structural principles of the globular organization of protein chains secondary structure, *Journal of Molecular Biology* 88, 857-894.

87. Lin Z and Pan X-M, (2001) Accurate Prediction of Protein Secondary Structural Content, *J Protein Chemistry*, 20:3, 217-220

88. Lipman DJ and Pearson WR, (1985), Rapid and sensitive protein similarity searches, *Science*, 227:1435–1441.

89. Lo Conte L., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2002), SCOP database in 2002: Refinements accommodate structural genomics, *Nucleic Acids Research*, 30, 264–267.

90. Luo R., Feng Z. and Liu J., (2002) Prediction of Protein Structural Class by Amino Acid and Polypeptide Composition, *European J Biochemistry*, 269, 4219-4225

91. MacCallum, R.M., (1997), *Computational Analysis of Protein sequence and Structure*, Ph.D thesis, Department of Biochemistry and Molecular biology, University College London.

92. Maclin, R. & Shavlik, J. W. (1993), Using knowledge-based neural networks to improve algorithms: refining the Chou-Fasman algorithm for protein folding, *Machine Learning*, 11, 195-215.

93. Mitchell, T.M., (1997), *Machine Learning*, McGraw-Hill International, Singapore.

94. Moult J. et al. (1997), Critical assessment of methods of protein structure prediction (CASP): Round II, *Proteins*, 29,2-6.

95. Moult J. et al. (2003), Critical assessment of methods in protein structure prediction (CASP) – Round V, *Proteins*, 53, 334-339.

96. Mount, D.W. (2001), *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

97. Muskal S.M. and Kim S-H., (1992) Predicting Protein Secondary Structure Content: a Tandem Neural Network Approach, *J Mol Biology*, 225, 713-727

98. Murzin A.G. et al., (1995), SCOP: A structural classification of proteins database for the investigation of sequences and structures, *Journal of Molecular Biology*, 247, 536-540

99. Nagano, K., (1973), Normalized frequency of alpha-helix, beta-structure and coil, *Journal of Molecular Biology*, 75, 401-420.

100. Nakata, K., (1995), Predictiton of zinc finger DNA binding protein, *Comput. Appl. Biosci.*, 11: 125-131.

101. Nelson D. and Cox M., Lehninger., (2000), *Principles of Biochemistry Amino*, Worth Publish.

102. Needleman SB and Wunsch CD, (1970), A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of Molecular Biology*, 48, 443–453.

103. NIH (National Institute of Health), (2000), Working Definition of Bioinformatics and Computational Biology, *WWW document*, http://www.bisti.nih.gov/CompuBioDef.pdf.

104. Ouali, M. and King R.D. (2000) Cascaded Multiple Classifiers for Secondary Structure Prediction, *Protein Science*, 9, 1162–1176

105. Parsons, R.J., Forrest, S. and Burks, C., (1995), Genetic algorithms, operators, and DNA fragment assembly, *Machine Learning*, 21: 11-33.

106. Pearson W. R. (1990), Rapid and sensitive sequence comparison with FASTP and FASTA, methods, *Enzymology*, 183, 63 – 98

107. Petersen T. et al. (2000) Prediction of Protein Secondary Structure at 80% Accuracy, *Proteins*, 41, 17-20

108. Pollastri, G. et al., (2002), Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles, *Proteins*, 47, 228-335

109. Pollastri G and McLysaght A. (2005) Porter: A New, Accurate Server for Protein Secondary Structure Prediction, *Bioinformatics*, in print

110. Przybylski D and Rost B., (2002), Alignments Grow, Secondary Structure Prediction Improves, *Proteins*, 46, 197-205

111. Qian, N. and Sejnowski, T.J., (1988), Predicting the secondary structure of globular proteins using neural network models, *Journal of Molecular Biology*, 202: 865-884

112. Quinlan, J.R., (1986), Induction of decision trees, Morgan Kaufmann, *Machine Learning*, 1, 81-106.

113. Quinlan, J.R., (1990), Learning logical definitions from relations, *Machine Learning*, 5, 239-266.

114. Quinlan, J.R., (1993), C4.5 *Programs for Machine Learning*, Morgan-Kaufmann.

115. Richard O. Duda, Peter E. Hart, David G. Stork (2001), *Pattern classification* (2nd edition), Wiley, New York.

116. Rodriguez-Tome P., Stoehr P.J., Cameron G.N. and Flores T.P., (1996), *Nucleic Acids Res*. 24, 6-12.

117. Rost, B. & Sander, C. (1993a) Prediction of protein secondary structure at better than 70% accuracy, *Journal of Molecular Biology*, 232, 584-599.

118. Rost, B. & Sander, C. (1993b), Improved prediction of protein secondary structure by use of sequence profilesand neural networks, *Proc. Natl. Acad. Sci.* U.S.A., 90, 7558-7562.

119. Rost, B., and Sander, C., (1994), Combining evolutionary information and neural networks to predict protein secondary structure, *Proteins*, 19:1, 55-72.

120. Rost, B., Sander, C., and Schneider, R., (1994), Redefining the goals of protein secondary structure prediction, *Journal of Molecular Biology*, 235, 13-26.

121. Rost, B., (1996), PHD: Predicting one-dimensional protein structure by profile based neural networks, *Methods in Enzymology*, 266, 525-539

122. Rost WWW, B., (2000), Better secondary structure prediction through more data, Columbia University, *WWW document,* http://cubic.bioc.columbia.edu/predictprotein.

123. Rost, B., and Sander, C., (2000), Third generation prediction of secondary structure, In: Webstar, D., (Ed.), *Protein Structure Prediction: Methods and Protocols*, Human Press Clifton, NJ, 71-95.

124. Rost, B., (2001), Review: Protein secondary structure prediction continue to rise, *Journal of Molecular Biology,* 134:2-3, 204-18.

125. Ruan, J., Wang, K., Yang, J., Kurgan, L., and Cios, K.J., (2005), Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences, *Artificial Intelligence in Medicine*, special issue on Computational Intelligence Techniques in Bioinformatics, in print.

126. Ruan J., Shen S., Wang K., Kurgan L., and Tuszyski J., (2005), Exploring a fourth generation of prediction methods for protein secondary structure in three states, submitted.

127. Rule Quest Research, (2003), C5.0, *www document*, http://www.rulequest.com/see5-info.html.

128. Salamov A.A. & Solovyev V.V. (1995), Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiply sequence alignments, *Journal of Molecular Biology*, 247,1.

129. Salzberg, S., (1995), Locating protein coding regions in human DNA using a decision tree algorithm, J.Comp. Biol.,2: 473-485.

130. Sasagawa, F. and Tajima, K., (1993), Prediction of protein secondary structures by a neural network, *Comput. Appl. Biosci.*, 9: 147-152.

131. Schapire, R.E, and Singer, Y., (1998), Improved boosting algorithms using confidence-rated predictions, *Proc. of the 11th Annual Conf. on Computational Learning Theory*, 80-91.

132. Schmidler; S.C., Liu, J.S. and Brutlag, D.L., (2000), Bayesian segmentation of protein secondary structure, *J.Comp. Biol,* 7: 233-248.

133. Schonbrun, J., Wedemeyer W.J. and Baker D., (2002), Protein Structure Prediction in 2002, *Current Opinion in Structural Biology*, 12:3, 348-354

134. Selbig, J. Mevissen, T. and Lengauer, T., (1999), Decision tree-based formation of consensus protein secondary structure prediction, *Bioinformatics,* 15: 1039-1046.

135. Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., and Arikawa, S., (1992), Finding alphabet indexing for decision trees over regular patterns: an approach to bioinformatical knowledge acquisition, *Technical Report* RIFIS-TR-CS-60, Research Institute of Fundamental Information Science, Kyusha University.

136. Sippl M.J. et al., (2001), Assessment of the CASP4 Fold Recognition Category, *Proteins*, 45, 55-67.

137. Smith TF and Waterman MS, (1981), Identification of common molecular subsequences, *Journal of Molecular Biology*, 147:195–197.

138. Sternberg, M., Lewis, R., King, R. and Muggleton, S., (1992), Modelling the structure and function of enzymes by machine learning, *Proceedings of the Royal Society of Chemistry: Faraday Discussions,* 93: 269-280.

139. Stormo, G., Schneider, T., Gold, L. & Ehrenfeucht, A., (1982), Use of the perceptron algorithm to distinguish translational initiation in E.coli., *Nuclei Acids Research,* 10: 2997-3011.

140. Syed U. and Yona G., (2003), Using a Mixture of Probabilistic Decision Trees for Direct Prediction of Protein Function, *Proc of RECOMB 2003*, 224-234

141. Tan A., Gilbert D. and Deville Y., (2003), Multi-class protein fold classification using a new ensemble machine learning approach, *Genome Informatics*, 14, 206 - 217

142. Thompson JD, Higgins DG, and Gibson TJ., (1994), CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research,* 22(22): 4673–4680.

143. Truhlar, D.G. et al. (1999), The IMA volumes in mathematics and its applications, *Rational Drug Design*, 108.

144. Vapnik, V., (1995), *The Nature of Statistical Learning Theory,* Springer-Verlag, New York.

145. Westbrook J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W., Weissig,H., Greer,D.S. et al., (2002), The protein data bank: unifying the archive, *Nucleic Acids Res.*, 30, 245–248.

146. Wang, J. et al. (2000) Application of Neural Networks to Biological Data Mining: a Case Study in Protein Sequence Classification, *Proc of the 6th ACM SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining*, 305-309

147. Wang Z-X.and Yuan Z., (2000), How Good is Prediction of Protein Structural Class by the Component-Coupled Methods, *Proteins*, 38, 165-175

148. Wu, C., Berry, M., Shivakumar, S. and Mcarty, J., (1995), Neural networks for full-scale protein sequence classification: sequence encoding with singular value decomposition, *Machine Learning,* 21: 177-193.

149. Yang X. and Wang B. (2003) Weave Amino Acid Sequences for Protein Secondary Structure Prediction, *Proc of the 8th ACM SIGMOD workshop on Research issues in Data Mining and Knowledge Discovery*, 80-87

150. Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T., and Müller, K.-R., (2000), Engineering support vector machine kernels that recognize translation initiation sites, *Bioinformatics*, 16: 799-807.

151. Zhang, C. and Wong, A.K., (1997), A genetic algorithm for multiple molecular sequence alignment, *Comput. Appl. Biosci*, 13: 565-581

152. Zhang CT. et al., (1998), Prediction of Helix/Strand Content of Globular Proteins Based on Their Primary Sequences, *Protein Engineering*, 11:11, 971-979

153. Zhang, Z.D., Sun, Z.R., and Zhang, C.T., (2001), A new approach to predict the helix/strand content of globular proteins, *Journal of Theoretical Biology*, 208, 65-78.

154. Zvelebil, M.J.J.M., Barton, G.J., Taylor, W.R. & Sternberg, M.J.E., (1987), Prediction of protein secondary structure and active sites using the alignment of homologous sequences, *Journal of Molecular Biology*, 195, 957-961.