# Using Cluster Analysis for Protein Secondary Structure Prediction

Reet Kamal Kaur
Assistant Professor CSE
RIMT-MAEC
Mandi Gobindgarh

Manjot Kaur
Lecturer CSE
GNDEC
Ludhiana

Amanjot Kaur
Assistant Professor CSE
Global Institute
Amritsar

## ABSTRACT
As biomedical research and healthcare continue to progress in the genomic/post genomic era, a number of important challenges and opportunities exist in the broad area of bioinformatics. In the broader context, the key challenges to bioinformatics essentially all relate to the current flood of raw data, aggregate information, and evolving knowledge arising from the study of the genome and its manifestation.

Protein structure determination and prediction has been a focal research subject in life sciences due to the importance of protein structure in understanding the biological and chemical activities of organisms. The experimental methods used to determine the structures of proteins demand sophisticated equipment and time. A host of computational methods are developed to predict the location of secondary structure elements in proteins for complementing or creating insights into experimental results.

The present work focuses on secondary structure prediction of proteins. The data mining model is implemented to predict the various parameters related to the secondary structure. These parameters include the alpha helix, beta sheets and hairpin turn. Cluster analysis is used to implement the secondary structure prediction.

*Key Words:* Data mining, Cluster analysis, Protein structure prediction.

## 1. INTRODUCTION
Bioinformatics is the application of computer technology to the management of biological information. It is the analysis of biological information using computers and statistical techniques; the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research. Bioinformatics is more of a tool than a discipline, the tool for analysis of Biological Data.

Proteins are complex organic compounds that consist of amino acids joined by peptide bonds. Proteins are essential to the structure and function of all living cells and viruses. Many proteins function as enzymes or form subunits of enzymes. Some proteins play structural or mechanical roles. Some proteins function in immune response and the storage and transport of various ligands. Proteins serve as nutrients as well; they provide the organism with the amino acids that are not synthesized by that organism[1]. Proteins are amongst the most actively studied molecules in biochemistry and they were discovered by the Swedish scientist, Jons Jakob Berzelius in 1838.

An amino acid is any molecule that contains both an amino group and a carboxylic acid group. An amino acid residue is the residuals of an amino acid after it forms a peptide bond and loses a water molecule. Since we are interested in amino acids that form proteins, it is safe to use the terms residue and amino acid interchangeably. There are 20 different amino acids in nature that form proteins.

Amino acids are the basic building blocks of proteins. Fundamentally, amino acids are joined together by peptide bonds to form the basic structure of proteins. However, owing to the many 'side groups' that are part of the amino acids other sorts of bonds may form between the amino acid units. These additional bonds twist and turn the protein into convoluted shapes that are unique to the protein and essential to its ability to perform certain functions within the human body.

Given a protein sequence with amino acids a1 a2 . . . an, the secondary structure prediction problem is to predict whether each amino acid ai is in an α−helix, a β−sheet, or neither. If you know (say through structural studies), the actual secondary structure for each amino acid, then the 3-state accuracy is the percent of residues for which your prediction matches reality. It is called "3-state" because each residue can be in one of 3 "states": α, β, or other (O). Because there are only 3 states, random guessing would yield a 3-state accuracy of about 33% assuming that all structures are equally likely. There are different methods of prediction with various accuracies[1].

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure[2]. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics [3].

## 2. TECHNIQUES USED IN LITERATURE
### 2.1 Chou Fasman Method
In this method, a helix is predicted if, in a run of six residues, four are helix favoring and the average valued of the helix

propensity is greater than 1.0 and greater than the average strand propensity. Such a helix is extended along the sequence until a proline is encountered (helix breaker) or a run of 4 residues with helical propensity less than 1.0 is found. A strand is predicted if, in a run of 5 residues, three are strand favouring, and the average value of the strand propensity is greater than 1.04 and greater than the average helix propensity. Such a strand is extended along the sequence until a run of 4 residues with strand propensity less than 1.0 is found[1]. This is a simple rule-based method dependent on finding runs of residues with preference for one type of secondary structure.

## 2.2 GOR Method

Considering the information carried by a residue about its own secondary structure, in combination with the information carried by other residues in a local window of eight residues on either side of the sequence of the residue concerned.

The accuracy of these early methods based on the local amino acid composition of single sequences was fairly low, with often less than 60% of residues being produced in the correct secondary structure state.

## 2.3 PHD

The neural net model employed by Rost and Sander was fairly complex and computationally expensive. Because of the computational demands, a 7-fold cross-validation was used in place of jack-knife testing. Accuracy was over 70% using multiple sequence alignment, but the fifth of residues with the highest reliability was predicted with over 90% accuracy. Rost and Sander also tested PHD on 26 new proteins, none with significant sequence similarity to any protein in the training set, and found comparable results. PHD, however, suffers from some of the ANN problems [2]. Rost and Sander were concerned with overtraining and therefore terminated training once the accuracy was higher than 70% for all training samples.

# 3. METHODOLOGY

A number of factors exist that make protein structure prediction a very difficult task. The two main problems are that the number of possible protein structures is extremely large, and that the physical basis of protein structural stability is not fully understood. As a result, any protein structure prediction method needs a way to explore the space of possible structures efficiently (a search strategy / retrieval strategy), and a way to identify the most plausible structure (an energy function). Progress in protein structure prediction is slow because both aspects of the problem, the energy function that must discriminate between the native structure and many decoys and the search algorithm to identify the conformation with the lowest energy, are fraught with difficulties [4]. Furthermore, difficulties in each aspect reduce progress in the other.

The proposed model uses data mining as the retrieval strategy and structure prediction algorithm to identify the structure of the given protein.

As more protein sequences become available, protein structure and function can be better studied with more accuracy and efficiency. The goal of clustering protein sequences is to get a biologically meaningful partitioning [5]. Clustering a large set of protein sequences offers several advantages: Proteins are usually grouped into families based on the sequence similarity clustering, which provides some clues about the general features of that family and evolutionary evidence of proteins; Clustering also helps to infer the biological function of a new sequence by its similarity to some function-known sequences [6]. Moreover, protein clustering can be used to facilitate protein 3-dimensional structure discovery, which is very important for understanding protein's function.

The Chou-Fasman algorithm for the prediction of protein secondary structure is one of the most widely used predictive schemes. The Chou-Fasman method of secondary structure prediction depends on assigning a set of prediction values to a residue and then applying a simple algorithm to the conformational parameters and positional frequencies. The Chou-Fasman algorithm is simple in principle [1]. The conformational parameters for each amino acid were calculated by considering the relative frequency of a given amino acid within a protein, its occurrence in a given type of secondary structure, and the fraction of residues occurring in that type of structure. These parameters are measures of a given amino acid's preference to be found in helix, sheet or coil. Using these conformational parameters, one finds nucleation sites within the sequence and extends them until a stretch of amino acids is encountered that is not disposed to occur in that type of structure or until a stretch is encountered that has a greater disposition for another type of structure. At that point, the structure is terminated. This process is repeated throughout the sequence until the entire sequence is predicted. The Chou-Fasman method of secondary structure prediction depends on assigning a set of prediction values to a residue and then applying a simple algorithm to those numbers[3]. The table of numbers is as follows:

**TABLE I: Conformational parameters and positional frequencies for α-helix, ß-sheet and turn residues.**

| NAME | P(A) | P(B) | P(TURN) | F(I) | F(I+1) | F(I+2) | F(I+3) |
|---|---|---|---|---|---|---|---|
| ALANINE | 142 | 83 | 66 | 0.060 | 0.076 | 0.035 | 0.058 |
| ARGININE | 98 | 93 | 95 | 0.070 | 0.106 | 0.099 | 0.085 |
| ASPARTIC ACID | 101 | 54 | 146 | 0.147 | 0.110 | 0.179 | 0.081 |
| ASPARAGINE | 67 | 89 | 156 | 0.161 | 0.083 | 0.191 | 0.091 |
| CYSTEINE | 70 | 119 | 119 | 0.149 | 0.050 | 0.117 | 0.128 |
| GLUMATIC ACID | 151 | 37 | 74 | 0.056 | 0.060 | 0.077 | 0.064 |
| GLUTAMINE | 111 | 110 | 98 | 0.074 | 0.098 | 0.037 | 0.098 |
| GLYCINE | 57 | 75 | 156 | 0.102 | 0.085 | 0.190 | 0.152 |
| HISTIDINE | 100 | 87 | 95 | 0.140 | 0.047 | 0.093 | 0.054 |
| ISOLEUCINE | 108 | 160 | 47 | 0.043 | 0.034 | 0.013 | 0.056 |
| LEUCINE | 121 | 130 | 59 | 0.061 | 0.025 | 0.036 | 0.070 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| LYSINE | 114 | 74 | 101 | 0.055 | 0.115 | 0.072 | 0.095 |
| METHIONINE | 145 | 105 | 60 | 0.068 | 0.082 | 0.014 | 0.055 |
| PHENYLALANINE | 113 | 138 | 60 | 0.059 | 0.041 | 0.065 | 0.065 |
| PROLINE | 57 | 55 | 152 | 0.102 | 0.301 | 0.034 | 0.068 |
| SERINE | 77 | 75 | 143 | 0.120 | 0.139 | 0.125 | 0.106 |
| THREONINE | 83 | 119 | 96 | 0.086 | 0.108 | 0.065 | 0.079 |
| TRYPTOPHAN | 108 | 137 | 96 | 0.077 | 0.013 | 0.064 | 0.167 |
| TYROSINE | 69 | 147 | 114 | 0.082 | 0.065 | 0.114 | 0.125 |
| VALINE | 106 | 170 | 50 | 0.062 | 0.048 | 0.028 | 0.053 |

The following formulae were used in the system:

$$p(t) = f(i)+f(i+1)+f(i+2)+f(i+3)$$

where the f(i+1) value for the i+1 residue is used, the f(i+2) value for the i+2 residue is used and the f(i+3) value for the i+3 residue is used.

## 4. RESULTS AND DISCUSSION

The given sequence is divided into clusters and from the table 1 the conformational parameters and positional frequencies for α-helix, ß-sheet and turn residues are established. In each cluster every region where four of six contagious amino acid residues have P(a)>100 are identified and extended until a proline is encountered (helix breaker) or a run of 4 residues with P(a)<100 is found. Similarly for regions where four of six residues have P(b).100, are extended and a beta strand is predicted if the average P(b) over all residues in the cluster are greater than 100 and $\sum P(b) > \sum P(a)$. For alpha helix prediction the $\sum P(a)$ is computed and for each cluster is >5 and the $\sum P(a) > \sum P(b)$, then the cluster is predicted to be alpha helix. The cluster is found to be either α-helix or β-sheet favoring. The turns for each residue are predicted by calculating the summation of F(i), F(i+1), F(i+2), F(i+3) and when P(t)>0.000075.

There are lots of factors that affect the protein structure formation. And because these factors proteins sometimes fold into different patterns. Hydrophobic and hydrophilic forces force the protein structure to turn and coil and hence attain a shape not in conformation with the original shape.

Similarly the Wander wall forces also change the shape of protein then the normal one. The mechanism of these forces and their significance in structure prediction is still in its infancy and hence the total accuracy of the structure prediction of protein without their consideration is still not attainable [7].

We assume that the protein folding and structure formation are independent of the factors like hydrophobic forces, Wander wall forces etc.

## 5. REFERENCES

[1] Kumar, Anil (1990) "Predicted secondary structure of maltodextrin Phosphorylase from Escherichia coli as deduced using Chou-Fasman model" Junior Biosci., Vol. 15, pp. 53-58.

[2] Chen Yonghui, Reilly Kevin D., Sprague Alan P., Guan Zhijie,(2006) "SEQOPTICS: a protein sequence clustering system" Symposium of Computations in Bioinformatics and Bioscience (SCBB06) in conjunction with the International Multi-Symposiums on Computer and Computational Sciences 2006 (IMSCCS|06), pp 1-5.

[3] Haitao Cheng, Taner Z. Sen , Robert L. Jernigan and Andrzej Kloczkowski (2005) "Consensus Data Mining (CDM) Protein Secondary Structure Prediction Server: Combining GOR V and Fragment Database Mining (FDM)" Bioinformatics journal 2007, pp 12834-12888.

[4] Ingrid Fischer and Thorsten Meinl (2004) "Graph Based Molecular Data Mining - An Overview" IEEE 0-7803-8566-7/04, pp 1-2.

[5] Eisen Michael B., Spellman Paul T., Brown Patrick O., Botstein David (1998) "Cluster analysis and display of genome-wide expression patterns" Proc. Natl. Acad. Sci. USA Vol. 95, pp. 14863–14868.

[6] Fraley Chris, Raftery Adrian E. (1998) "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis" The computer journal, Vol. 41, pp 578-587.

[7] George Tzanis, Christos Berberidis, and Ioannis Vlahavas (2002) "Biological Data Mining" Department of Informatics, Aristotle University of Thessaloniki, Greece, pp 1-8.