Chou-Fasman Method for Protein Structure Prediction using Cluster Analysis

Rajbir Singh, Sumandeep Kaur Deol, Parvinder S. Sandhu

***Abstract*—**The research in bioinformatics has accumulated large amount of data. As the hardware technology advancing, the cost of storing is decreasing. The biological data is available in different formats and is comparatively more complex. Knowledge discovery from these large and complex databases is the key problem of this era. Data mining and machine learning techniques are needed which can scale to the size of the problems and can be customized to the application of biology. In the present research work, the Chou-Fasman Method is implemented with the help of data mining. Protein structure determination and prediction has been a focal research subject in the field of bioinformatics due to the importance of protein structure in understanding the biological and chemical activities of organisms. The experimental methods used by biotechnologists 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. Cluster analysis is used as data mining model to retrieve the results.

***Keywords*—**Amino-Acid, Protein , Polypeptide, clusters, DNA, RNA, PHD, GOR

I. Introduction

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. Proteins are amongst the most actively studied molecules in biochemistry.

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.

Er. Rajbir Singh, Asstt. Prof. & Head, Department of Information Tech., Lala Lajpat Rai Institute of Engineering & Technology, Moga , Punjab, INDIA (phone: +91-9417977061; e-mail: cheema_patti@yahoo.com).

Er. Dheerajpal Kaur, Lecturer. He is now with the Department of Electronic & Comm. Engg., Lala Lajpat Rai Institute of Engineering. and Technology, Moga, Punjab, INDIA (e-mail: dheeraj_pl@yahoo.co.in).

Dr. Parvinder S. Sandhu is working as Professor with the Rayat & Bahra Institute Of Engineering & Bio-Technology, Mohali-Sahauran14004. E-Mail: parvinder.sandhu@gmail.com,

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.

Structure of Protein

$-NH_2$	+	-COOH	=	-CONH-
Amino group (Amino acid 1)		Carboxylic group (Amino acid 2)		Peptide Bond

A chain of such peptide bonds is called *polypeptide* and is a protein.

Examples of proteins:

a) Protective Proteins, for example, keratin (nails).
b) Defense Proteins, for example, antibodies.
c) Toxins, for example, snake venom.
d) Structural Proteins, for example, collagen of bones.
e) Enzymes (biocatalysts), for example, pepsin, trypsin.
g) Hormones, for example, insulin is a protein.

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.

Amino acids play central roles both as building blocks of proteins and as intermediates in metabolism. The 20 amino acids that are found within proteins convey a vast array of chemical versatility. The precise amino acid content, and the sequence of those amino acids, of a specific protein, is determined by the sequence of the bases in the gene that encodes that protein. The chemical properties of the amino acids of proteins determine the biological activity of the protein. Proteins not only catalyze all (or most) of the reactions in living cells, they control virtually all cellular process. In addition, proteins contain within their amino acid sequences the necessary information to determine how that protein will fold into a three dimensional structure, and the stability of the resulting structure.

As amino acids bind together in chains to form the stuff from which our life is born. It's a two-step process: Amino acids get together and form peptides or polypeptides. It is from these groupings that proteins are made. Commonly recognized amino acids include glutamine, glycine, phenylalanine, tryptophan, and valine. Three of those

phenylalanine, tryptophan, and valine are essential amino acids for humans; the others are isoleucine, leucine, lysine, methionine, and threonine. The essential amino acids cannot be synthesized by the body; instead, they must be ingested through food.

Amino acids make up 75% of the human body. They are essential to nearly every bodily function. Every chemical reaction that takes place in your body depends on amino acids and the proteins that they build.

Humans can produce 10 of the 20 amino acids. The others must be supplied in the food. Failure to obtain enough of even 1 of the 10 essential amino acids, those that we cannot make, results in degradation of the body's proteins muscle and to obtain the one amino acid that is needed.

Amino acids are carbon compounds that contain two functional groups: an amino group (NH2) and a carboxylic acid group (COOH). A side chain attached to the compound gives each amino acid a unique set of characteristics.

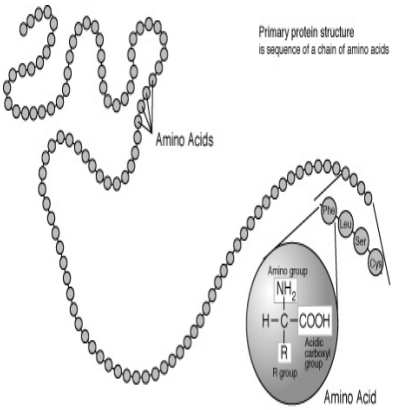

Fig. 1 A generic Amino acid Structure

Structures of proteins are investigated under four primary groups:

- Primary Structure is the sequence of amino acids in the protein. Counting of residues always starts at the N-terminal end (NH_2-group), which is the end where the amino group is involved in a peptide bond. The primary structure of a protein is determined by the gene corresponding to the protein.
- Secondary Structure is the composition of common patterns in the protein. Some patterns are frequently observed in the native states of proteins. This structure class includes regions in the protein of these patterns but it does not include the coordinates of residues.
- Tertiary Structure is the native state, or folded form, of a single protein chain. This form is also called the functional form. Tertiary structure of a protein includes the coordinates of its residues in three dimensional spaces. The elements of secondary structure are usually folded into a compact shape using a variety of loops and turns.
- Quaternary Structure is the structure of a protein complex. Some proteins form a large assembly to function. This form includes the position of the protein subunits of the assembly with respect to each other.

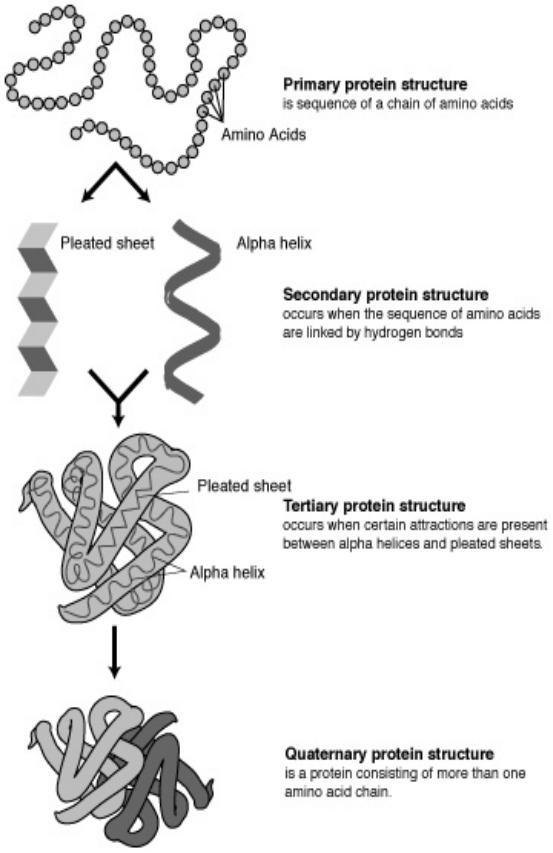

Fig. 2 Different representations of protein structure

Protein Synthesis

While the genetic code itself resides in DNA, DNA is never used directly in the synthesis of proteins. DNA is "transcribed" into messenger RNA (mRNA) that carries information from DNA and it is mRNA that is then used to "translate" this information into a specific sequence of amino acids that constitute proteins. Similarly, amino acids cannot recognize codons directly, an adapter molecule is necessary, a transfer RNA (tRNA). Amino acids are incorporated into a protein in an order predetermined by the mRNA sequence. A tRNA can recognize more than one codon very often, the first 2 nucleotides of the codon are sufficient to specify an amino acid and the third nucleotide varies.

(i)Transcription (DNA > RNA)

RNA is synthesized on a DNA template in the process known as DNA transcription. Transcription generates the mRNA containing the information to synthesize a specific protein and also the other RNA molecules, ribosomal RNA, tRNA's involved in the process. The key enzyme(s) involved in the process is RNA polymerase, an incredibly complex enzyme of molecular mass of 500,000kDa. DNA is transcribed by RNA polymerase binding to a specific start site or "promoter" on the DNA and proceeding until it reaches a termination signal.

The DNA double helix is partially unwound by the polymerase and transcription always proceeds in a 3' to 5' direction on the DNA template so that the RNA produced is extended in a 5' to 3' direction.

(ii) Translation (RNA > Protein)
The ribosome binds to the mRNA at the start codon (AUG) that is recognized only by the initiator tRNA. The ribosome proceeds to the elongation phase of protein synthesis. During this stage, complexes, composed of an amino acid linked to tRNA, sequentially bind to the appropriate codon in mRNA by forming complementary base pairs with the tRNA anticodon. The ribosome moves from codon to codon along the mRNA. Amino acids are added one by one, translated into polypeptidic sequences dictated by DNA and represented by mRNA. At the end, a release factor binds to the stop codon, terminating translation and releasing the complete polypeptide from the ribosome.

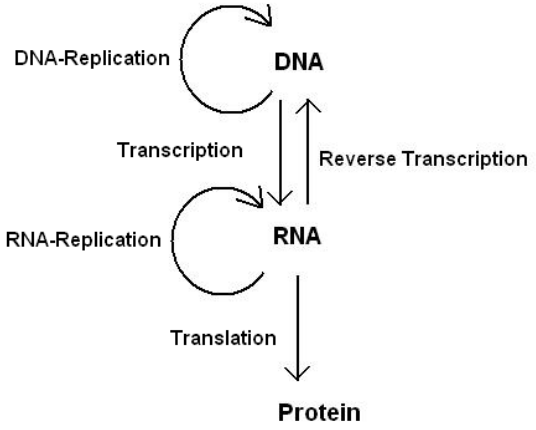

Fig. 3 Protein Synthesis from DNA to RNA to Protein.

Secondary Structure Prediction
Given a protein sequence with amino acids $a_1\ a_2\ .\ .\ .\ a_n$, the secondary structure prediction problem is to predict whether each amino acid a_i is in a 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. Some of these methods are:
(i) GOR Method
The GOR method, named for the three scientists who developed it – *G*arnier, *O*sguthorpe, and *R*obson. 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.

(ii)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. Rost and Sander were concerned with overtraining and therefore terminated training once the accuracy was higher than 70% for all training samples.

(iii) Chou- Fasman Method
The Chou-Fasman method was among the first secondary structure prediction algorithms developed and relies predominantly on probability parameters determined from relative frequencies of each amino acid's appearance in each type of secondary structure. 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 100 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 100 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 100 is found.

II. METHODOLOGY

Data Mining Model used for implementation of the CHOU-FASMAN method
As part of the larger process known as knowledge discovery, data mining is the process of extracting information from large volumes of data. This is achieved through the identification and analysis of relationships and trends within commercial databases. Data mining is used in areas as diverse as space exploration and medical research.

This model makes use of Clustering as the data mining method and uses conceptual clustering as the type of clustering. Clustering can be considered the most important unsupervised learning problem.

Clustering: Cluster analysis is an exploratory data analysis tool for solving classification problems. Its object is to sort cases into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster thus describes, in terms of the data collected, the class to which its members belong; and this description may be abstracted through use from the particular to the general class or type.

Cluster analysis is thus a tool of discovery. It may reveal associations and structure in data which, though not previously evident, nevertheless are sensible and useful.

Applying Cluster Analysis Technique for identifying the Secondary structure of a given amino acid sequence to be in α-helix β-Sheet or Turn.

- Input :- Amino acid sequence (Plain Text Format)
- Output: - Clusters of α-helix, β-sheet and Turn.

Chou-Fasman Method For Protein Structure Prediction

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.

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.

TABLE I CONFORMATIONAL PARAMETERS AND POSITIONAL FREQUENCIES OR A-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 algorithm contains the following steps:

(a) Assign parameter values to all residues of the Peptide.

(b) Scan the peptide and identify regions where 4 out of 6 contiguous residues have P(α)>100.Theseregions nucleate α- helices. Extend these in both directions until a set of four contiguous residues have an average P(α)<100.This ends the helix.

(c) Scan the peptide and identify regions where 3 out of 5 contiguous residues have P(β)>100.These residues nucleate β- strands. Extend these in both directions until a set of four contiguous residues have an average P(β)<100.This ends β- strand.

(d) Any region containing overlapping α and β assignments are taken to be helical or β depending on if the average P(α) and P(β) for that region is largest. If this residues an α or β- region so that it becomes less than 5 residues, the α or β assignment for that region is removed.

(e) To identify a β-turn at residue number i,the product p(t) = f(i)f(i+1)f(i+2)f(i+3) is calculated. To predict a β-turn, the following three conditions have to be simultaneously fulfilled:

p (t)>0.000075

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

- The average value for P (turn)>100 for four amino acids.
- The average P (turn) is larger than the average P(α) as well as P(β).

(f) The remaining part of the sequence without Assignment = are considered as coils.

III. CHOICE OF SEQUENCE FORMAT

There are various formats of Amino acid sequences, and each has its own set of characters and utility. To get a deeper understanding and better results it is essential to choose a valid input format. The various formats are:

- Plain text format
- FASTA format
- Genetic Computer Group Format (GCG)
- NEXUS
- NBRF & PIR

After studying the different Amino Acid formats, plain text was chosen for the present problem. The plain text format is comparatively simple than the other formats available. The plain text sequence format is typically generated by word processors (saved as text file with line breaks). Moreover the other details regarding DNA are not necessary for the present research work.

Plain text format: Plain Text format looks like the following:
MAYPMQLGFQDATSPIMEELLHFHDHTLMIVFLISSLVL
YIISLMLTTKLTHTSTMDAQEVETIWTILPAIILILIALPSL
RILYMMDEINNPSTVKTMGHQWYWSYEYTDYEDLSF
DSYMIPTSELKPGELRLLEVDNRVVLPMEAAQQEEEE
MAYPMQLGFQDATSPIMEELLHFHDHTLMIVFLISSLVL
YIISLMLTTKLTHTSTMDAQEVETIWTILPAIILILIALPSL
RILYMMDEINNPSTVKTMGHQWYWSYEYTDYEDLSF
DSYMIPTSELKPGELRLLEVDNRVVLPMEAAQQE.

IV. RESULTS AND DISCUSSION

For a given sequence of amino acids, this technique first clusters the amino acids and then these amino acid clusters are analyzed to predict the structure of protein. The user inputs the primary structure of the protein i.e. the amino acid sequence. The clusters of amino acids are extended till a alpha-helix, beta helix or a turn are predicted using the conformational parameters and positional frequencies for α-helix, ß-sheet and turn residues.

The whole detailed method is explained below:
Example:

INPQAIFDIQIKRLHEYKRQHHDKQVHMANLCVVGGFA
VNGVAALHSDLVVKDLFPEYHQLWPNKFHNVTNGITP
RRWIKQCNPALAALLDKSLQKEWANDLDQLINLVKLA
DDAKFRQLYRVIKQANKVRLAEFVKVRTIDLNLLHILA
LYKERIRENP

The above sequence is divided into clusters and from the table the conformational parameter and positional frequencies for α-helix, ß-sheet and turn residues are established

Hence the final secondary structure of the given sequence is:

TTTBBBBBBBBBBBBTTTTAAAAAAABBBBBBTTTT
TTTTTTTTTTTTTTTTTBBBBBTTAAAAAAAAAAAAA
AAATTBBBBBBTTTTTTTTTTTTBBBBBBTTTTAAAAA
AAATTTTTTTTBBBBTTTT

V. CONCLUSION

It attempts to classify amino acid in protein sequence according to their predicted local structure, which can be subdivided into three states: α-helix, β-sheet or turn.

- This research work will assist researchers to predict tertiary quaternary structure and will also act as a base for protein fold and function prediction.
- Protein fold can be predicted with better accuracy with this technique.
- Protein Structure and Function Prediction can be included in the same system.
- Various other data mining techniques can be used to determine an optimum result.
- Choice of various formats of amino acid sequences can be utilized.
- Protein structure and protein function prediction can be done based on improved Chou-Fasman method which includes 4 amino acids enabling a reverse β-turn.

VI. FUTURE SCOPE OF WORK

Following improvements regarding the developed model of bioinformatics can be made:

- The system can be extended to predict the tertiary structure of the protein.
- Various different mining techniques can be utilized to determine the optimum result.
- Different formats of amino acids can be utilized.
- Protein fold can be predicted with better accuracy with this technique.
- This technique can be further extended for multiple sequence alignment.

REFERENCES

[1] András Fiser, Andrej Sali (2000) *"Comparative protein structure modeling"* Pels Family Center for Biochemistry and Structural Biology ,The Rockefeller University, pp 82-88.

[2] Andreas Rechtsteiner, Jeremy Luinstra, Luis M Rocha, Charlie E M Strauss (2006) *"Use of Text Mining for Protein Structure Prediction and Functional Annotation in Lack of Sequence Homology"* Center of Genomics and Bioinformatics, Indiana University, Bloomington, IN 47401, pp 1-4.

[3] Ben Blum, Michael I. Jordan (2007) *"Feature Selection Methods for Improving Protein Structure Prediction with Rosetta"* Department of Electrical Engineering and Computer Science University of California at Berkeley, CA 94305, pp1-7.

[4] Chen Yonghui, Reilly Kevin D., Sprague Alan P., Guan Zhijie, *"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) Hangzhou, China. June 20–24, 2006, pp 1-5.

[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, No. 8, 1998 pp 578-587.

[7] Fraley Chris, Raftery Adrian E. (2000) *"Model based clustering, Discriminant Analysis, and density estimation"* Working Paper no II, Center for statics and social science, University of Washington, USA, pp 1-28.
[8] George Tzanis, Christos Berberidis, and Ioannis Vlahavas (2002) *"Biological Data Mining"* Department of Informatics, Aristotle University of Thessaloniki, Greece, pp 1-8.

Singh R is an Assistant Professor, Department of Information Technology of Lala Lajpat Rai Institute of Engineering & Technology Moga, India. He received his B.E (Honor) degree in Computer Science and Engineering from MD University, Rothak, Haryana and M-Tech degree in Computer Science and Engineering from Punjab Technical University, Jalandhar Pb. (INDIA). He has authored 04 books on Computer Science. His main field of research interest is Bio-Informatics and Data mining. He works on the Gene Expression, Phylogenetic Trees and Prediction of Protein Sequence & Structure.

Sumandeep Kaur Deol is a Faculty with the Department of Computer Science & Engineering of Lala Lajpat Rai Institute of Engineering & Technology Moga, India. She received her B.Tech in Computer Science & Engineering and M-Tech degree in Computer Science & Engineering from Punjab Technical University, Jalandhar Pb. (INDIA). Her research interests include Neural Networks, Genetics Algorithm and Data Mining.