Reviewing the Methods of Predicting Protein Secondary Structure

Shivani Agarwal
Dept. of Computer Science
IMS Engineering college
Ghaziabad, U.P., INDIA

Rishabh Kaushik
B. Tech 2nd year/CSE
IMS Engineering College,
Ghaziabad, U.P., INDIA

Atul Kumar
Associate professor
IMS Engineering college
Ghaziabad, U.P., INDIA

ABSTRACT
Not just will a good understanding of the protein structure assist in deciphering the biochemical mechanisms of proteins which would in turn result in diagnosing deficiencies, diseases, but also in treating them by giving humans the ability to create very explicit drugs targeted to perform a specific function. It is one of the key problems of molecular biology in this century. Predicting the protein secondary structure is an important step in this direction, as the structure of a protein is amalgamated with its function and characteristics. Although, this problem of predicting the protein structure with very high accuracy still lies unsolved even after decades of tedious research. But with the advancements in machine learning in the recent years have provided us with new tools that offer a ray of hope to tackle this problem. This review paper brings to the light the advancements in machine learning to predict the protein secondary structure.

General Terms
α-helix, β-sheet, loops, machine learning

Keywords
Neural networks, Protein secondary structure prediction, Hidden Markov Model, Bioinformatics, ANN

1. INTRODUCTION
Proteins are basically polymers; macromolecules that are made of amino acids. There are basically 20 amino acids that more or less make up all the types of proteins ever found. The amino acids are generally specified by a single alphabet. Thus, we can say that out of the 26 alphabets of the English language, 6 are silent when it comes to symbolization of amino acids. Proteins are specified by 3D arrangement of these amino acids. Determining the structure would help in understanding the traits and functions of these proteins. Theoretical biology still cannot predict the protein structure of a DNA sequence with 100% certainty unto this time but some classic physics techniques are used for this purpose; for a 100% surety of the determination of the protein structure, NMR (Nuclear Magnetic Resonance) or X-Ray Crystallography are used. But these methods are very tedious and time consuming and may take months or even years to provide the complete result. Protein structure prediction using machine learning intends to tackle this problem and give quick, yet reliable results. But this has not reached a level advanced enough to guarantee the predicted structure due to the fact that numerous similar structures can be generated in many similar ways with essentially the same sequencing. Table 1 lists the basic 20 amino acids and Table 2 lists the three main classes of amino acids. [1]

Table 1- The 20 amino acids and their corresponding nucleotide sequences

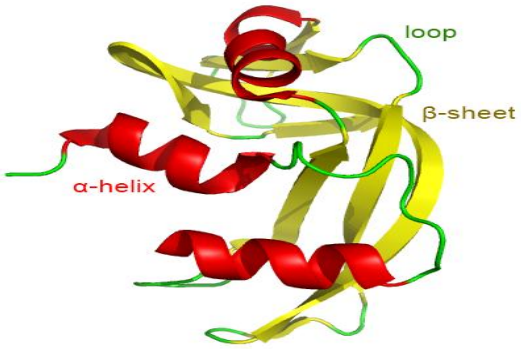
Amino Acid	**Tryptophan**	**Mathionine**	**Tyrosine**	**Cysteine**	**phenylalanine**
Nucleotide Sequences	TGG	ATG	TAT TGC	TGT TGC	TTT TTC
Amino Acid	**Histidine**	**Glutamine**	**Asparaqine**	**Lysine**	**Aspartic acid**
Nucleotide Sequences	CAT CAC	CAA CAG	AAT AAC	AAA AAG	GAT GAC
Amino Acid	**Glutamic acid**	**Tsoleucine**	**Glysine**	**Alamine**	**Valine**
Nucleotide Sequences	GAA GAG	ATT ATC ATA	GGT GAC GGA GGG	GCT GCC GCA GCG	GTT GTC GTA GTG
Amino Acid	**Threonine**	**Proline**	**Serine**	**Leucine**	**Arginine**
Nucleotide Sequences	ACT ACC ACA ACG	CCT CCC CCA CCG	TCT TCC TCA TCG AAT AGC	TTA TTG CTT CTC CTA CTG	CGC CGC CGA CGC AGA AGG

Table2:- The three main classes of amino acid

Class	Amino Acid
Hydrophobic(repels water)	**Alamine, Valine,, phenylalanine, Proline, Mathionine, Tsoleucine, Leucine, Glysine**
Charged Residues	**Aspartic acid, Glutamic acid, Lysine, Arginine**
Polar (Hydrophilic-attracted to water)	**Serine, Threonine, Tyrosine, Histidine, Cysteine, Asparaqine, Glutamine, Tryptophan**

Being able to accurately determine the protein structure will serve us with the ability to understand the structure-function relationship of proteins which in turn will help in creating custom drugs which would affect very specific organ, cells, or other proteins without interacting with other cells. Thus helping in curing diseases, repairing deficiencies in genes or dna; create humans with no weaknesses, no flaws. It will give us the ability to snipe at a cellular level which is a great requiremment as diseases are evolving too and getting resistant to existing drugs. Although, not very later as a very high accuracy in predicting the protein strucure is achieved, the need to establish correlated ethical laws will come up as this bio-technology will also make available the power to create *"custom viruses"*. You just need to get access to dna of the target (which is far too easy), run it for finding the deficiencies or flaws, create a virus to use this flaw to take that person down. This may seem fictitious and fabricated at this point, but soon enough, this will be what assassins will be using. The ultimate goal is to understand function of the proteins, link it to it's structure and the design proteins with required qualities.

2. α-HELICES, β-SHEETS, AND LOOPS

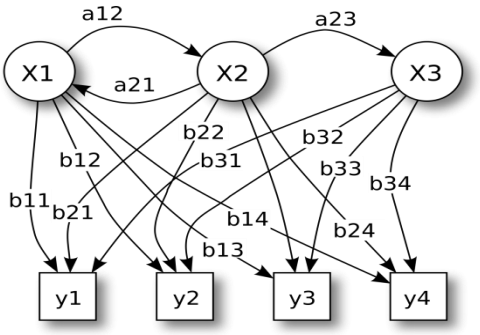


Secondary structure of proteins has 3 major conformations: alpha-helix, beta-sheet and coils (a.k.a. loops or reverse turns). The inputs fed to the system are the amino acid sequences while the output is the structure predicted which may or may not be correct. A typical protein contains about 32% alpha helices, 21% beta sheets, and 47% loops or non-regular structure. [1]

3. ALGORITHM [2]

The input provided is the amino acid sequence, where n is the length of the sequence. Initially we assume the probability for the amino acid sequence as the input to Hidden Markov Model.

Firstly, a basic diagram is shown to pictorially explain in brief the Hidden Markov Model (HMM):

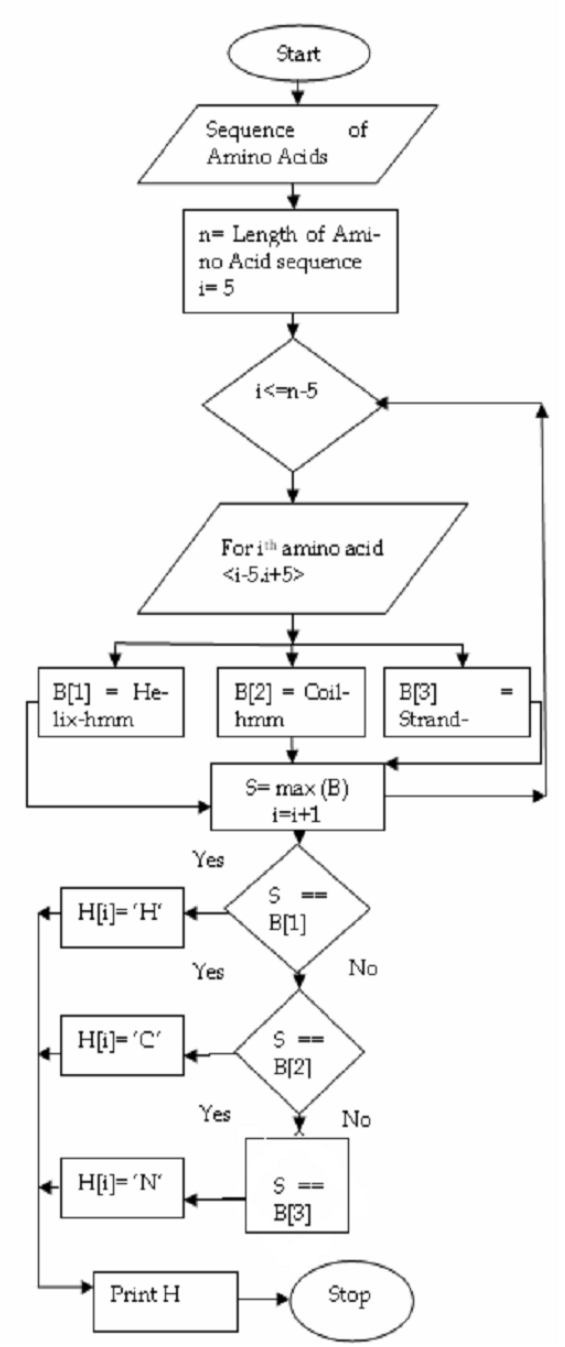


In HMM, there are intermediate "hidden" states between the inputs and the output(s). It uses the probabilistic model and is considered the simplest **dynamic Bayesian network [3].**

Bayesian Networks or simply Bayes nets is a probabilistic graphical model (hybrid of probability and graph theory) that represents a set of random variables (as nodes on the graph) with their conditional dependencies (as edges between the nodes) [3] The algorithm is described as SSP_H (Secondary Structure Prediction using Hidden Markov Model)

```
SSP_H(S, n)
{

S=sequences of Amino Acid/Input

n= The length of the amino acid/Input

H=Secondary Structure Sequence/initially it is null

For i=5 to n-5

{

B[1]=Helix_Hmm(S[i-5,i+5])

B[2]=Coil_Hmm(S[i-5,i+5])

B[3]= strand_Hmm(S[i-5,i+5])

C=max(B)//B is an array

If C=B[1]

H*i+='H'//Helix

Else if C=B[2]

H*i+='C'//Coil

Else

C=B[3]

H*i+='N'//strand

}

Print array H//output

//This  is the secondary structure of the given amino acid sequences
```

Flowchart for prediction of protein secondary structure:



4. WORKING OF NEURAL NETWORK

Neural network, as the name suggests, is a study of neurons, a study of nervous system of biological beings and implement it to create electronic models with a presumption that "similar causes have similar effects". Also, biological machines or "living things" take too much of time to evolve. For instance, average intelligence of humans is pretty much the same for the last 500+ years. But, electronic systems, being discovered just about 100 years ago have evolved so fast that they are mimicking their biological counterparts. This is the conclusive proof that they will overtake all kinds of biological beings in the not-so-far future.

Neural network, the study of building a computer model which is made of large number of simple,

highly interconnected computational units (neurons) operates parallel. Each unit combines its input and according to

some threshold value it generates output. Initially randomconnection strengths (weights) and thresholds (biases) are modified in repeated cycles by maximizing the accuracy of secondary structure assignment using the dataset of known protein structure. This is called the "training" phase. After this phase, the learned "knowledge" (which is actually derived weight and threshold value) is used in "test" phase to predict the unknown protein secondary structure. The network is composed of one input layer, one or more hidden layer and one output layer. Input layer encodes a moving window into amino acid sequence and central residue of the window is predicted. The computation is done in each input layer and output layer. The total input "Ei" to unit "i"

is,

$$E_i = \sum_j W_{ij} S_j + b_i$$

Where "bi" is the bias of the unit and the output of each unit "i" is generated by:

$$S_i = F(E_i) = \frac{1}{1+e^{-E_i}}$$

After calculating each time the error is predicted using the function,

$$E = \sum_c \sum_j (O_{J,C} - D_{J,C})^2$$

until the error is reduced to some satisfactory value. The basic neural network model is given on the top-right of the page:

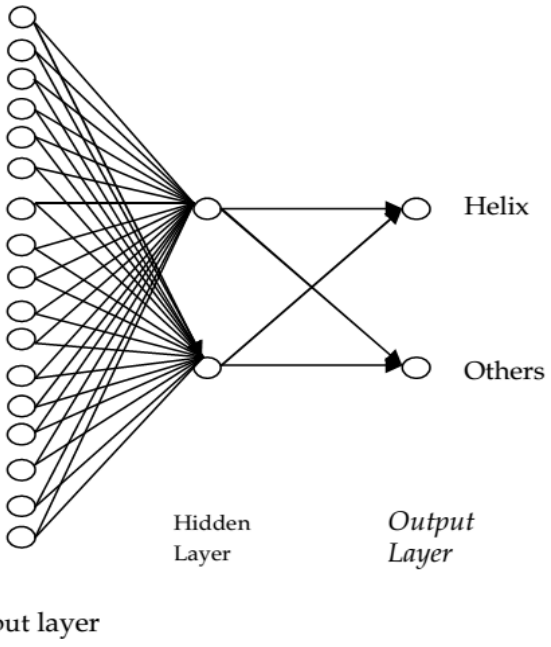


Fig. 1 Basic neural network

5. CONCLUSION

We talked about Amino acids, how they make up proteins, the prediction of protein secondary structure with a briefing of Hidden Markov Model (HMM) and Artificial Neural Networks (ANN). There are new state-of-the-art technologies being developed and current technologies like Feed-Forward Network being optimized continuously that are resulting in small and big breakthroughs. This is the problem of NP hard problem and time and space complexities are high.In this we use the encoding scheme for prediction the secondary structure of proteins. Although the accuracy of the current models like HMM still remains an issue but consistent and persistent research would increase the odds of a highly trustworthy protein prediction system.

6. ACKNOWLEDGMENTS

Our praises and obligations to the experts who have helped by providing us with their research material and authorities of IMSEC for providing the opportunity and for the valuable time they took out to guide us.

7. REFERENCES

[1] Prediction of Protein Secondary Structure by S.N. Vel Arjunan, Safaai Deris, Rosli Md Illias.

[2] Algorithm for predicting Protein Secondary Structure, K.K. Senapati, G. Sahoo, D. Bhaumik

[3] Final Project/ BIOMEDIN 231: Computational Molecular Biology Artificial Neural Network in Protein Secondary Structure Prediction: A Critical Review of Present and Future Applications

[4] Protein Secondary Structure Prediction using Feed-Forward Neural Network by: M. A. Mottalib, Md. Safiur Rahman Mahdi, A.B.M. Zunaid Haque, S.M. Al Mamun, and Hawlader Abdullah Al-Mamun

[5] Design and Implementation of an Algorithm to Predict Secondary Structure of Proteins using Artificial Neural Network, Shivani Agarwal, Ms. Arushi Baboota, Ms. Deepali Mendiratta.

[6] A Comparative Study of the Protein Secondary Structure Prediction methods: Shivani Agarwal, Ms. Arushi Baboota,Mr.AtulKumar.