A STUDY OF INTELLIGENT TECHNIQUES FOR PROTEIN SECONDARY STRUCTURE PREDICTION

Hanan Hendy, Wael Khalifa, Mohamed Roushdy, Abdel Badeeh Salem

Abstract: *Protein secondary structure prediction has been and will continue to be a rich research field. This is because the protein structure and shape directly affect protein behavior. Moreover, the number of known secondary and tertiary structures versus primary structures is relatively small. Although the secondary prediction started in the seventies but it has been together with the tertiary structure prediction a topic that is always under research. This paper presents a technical study on recent methods used for secondary structure prediction using amino acid sequence. The methods are studied along with their accuracy levels. The most known methods like Neural Networks and Support Vector Machines are shown and other techniques as well. The paper shows different approaches for predicting the protein structures that showed different accuracies that ranged from 50% to over than 90%. The most commonly used technique is Neural Networks. However, Case Based Reasoning and Mixed Integer Linear Optimization showed the best accuracy among the machine learning techniques and provided accuracy of approximately 83%.*

Keywords: *Bioinformatics, Machine Learning, Protein Secondary Structure Prediction.*

ACM Classification Keywords: *I.2 Artificial Intelligence, H.4 Information System Applications, H.4.2 Types of systems decision support*

Introduction

Protein structure prediction is known as predicting -getting- the secondary and/or tertiary structure from linear amino acid sequence known as Primary structure. Predicting the secondary structure of proteins helps in many domains. Some of these domains can be: knowing the functionally of the protein, drug design, the design of novel enzymes and disease detection such as “Alzheimer’s” and other diseases related to cancer [Camacho et al, 2012] and much more. Moreover, predicting the secondary structure is a basic and crucial step in the tertiary structure prediction. Tertiary structures that are known are relatively very small. In mid-2011, there were only 70,000 known tertiary structures in the PDB –Protein Data Bank- compared to 12.5 million protein sequences in the RefSeq database [Kister, 2013], so it’s very difficult to keep track of secondary and tertiary structures in the same pace of primary structures detection.

In this paper, section one presents a short biological background showing the important terminologies that are used all through the paper. Then, in section two the prediction methodologies are presented as a sequence of methodologies/ techniques. Each method is presented along with its accuracy and a brief description of the method. Methods presented can be categorized as statistical/probabilistic ones [Chou–Fasman, 2014; Garnier et al, 1996], Neural Networks which is the most common techniques used [Chandonia, 1995; Silva, 2005; Rost, 1996] and Case Based Reasoning [Glasgow et al, 2006]. Finally, the current research trends used in secondary structure prediction are presented that uses Support Vector Machines [Sui et al, 2011] and Swarm Intelligence (Bee Colony) [Li, 2014]. Finally discussing mixing more than one predictor as it is thought to be the future trend of secondary structure prediction [Wei, 2011].

Biological Background

Proteins are known to be large biological molecules built up from one or more Amino Acid residues. Proteins are responsible for many vital functions in the human body, for example: replicating DNA, responding to stimuli, metabolic functions and a lot more. There exists twenty amino acids –each has a unique shape and prefix letter – which builds up any protein. Amino acids are formed from Oxygen, Nitrogen, Hydrogen and Sulfur atoms [Protein, 2014]. The standard amino acids are shown in Figure 1.

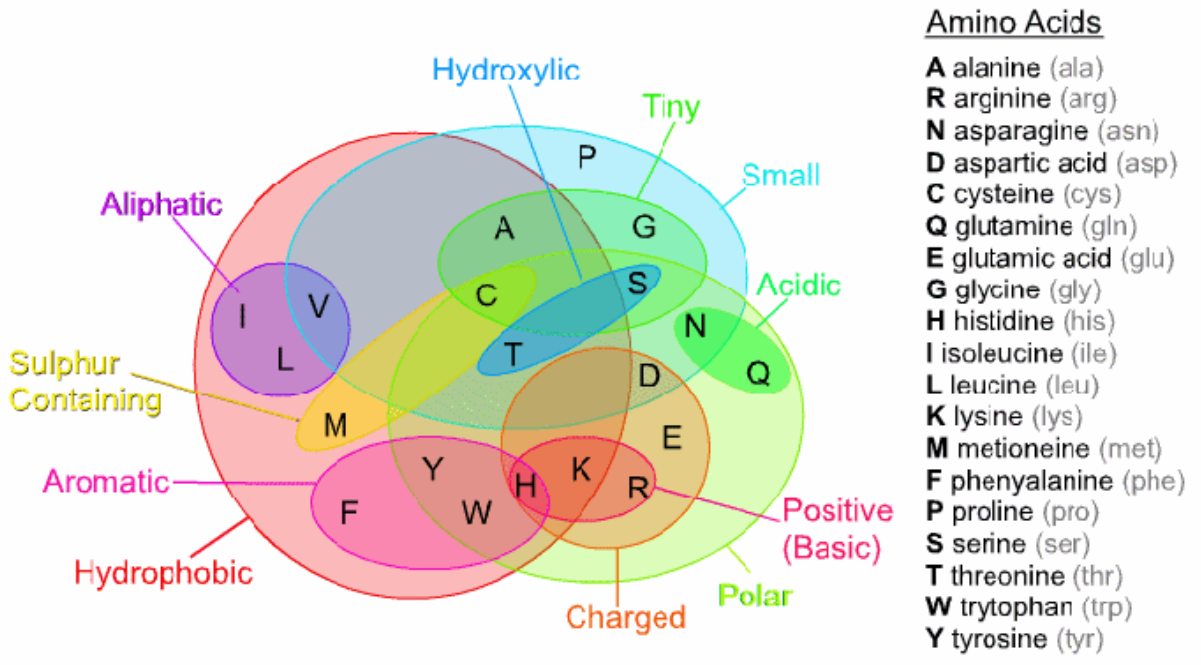

Figure 1. Venn diagram of boundaries that symbolizes the universal set of 20 common amino acids [Esquivel, 2013]

When amino acids interact together in order to be able to perform their functionalities, the result is called Structure. To understand the functions of proteins at a molecular level, it is often necessary to determine their three-dimensional structure. Protein structure can be organized into four distinct levels as shown in Figure 2.

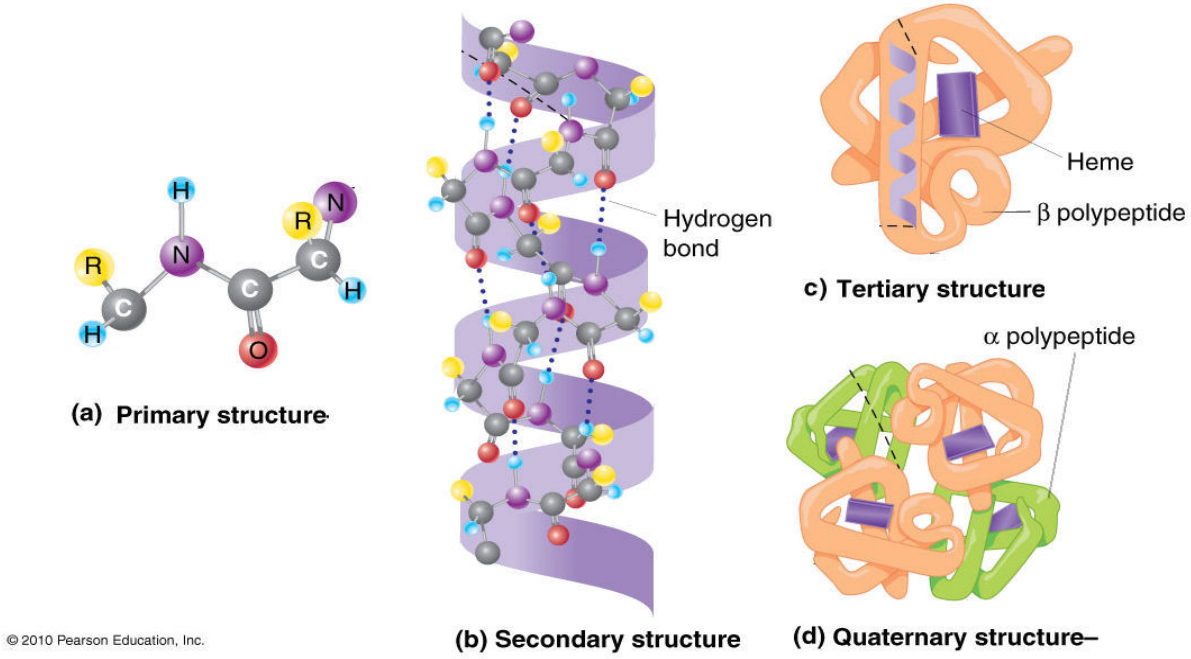

Figure 2. Four levels of Protein Structure [Russell, 2009]

The primary structure, which is the basic structure of Protein describes the linear sequence of amino acid in polypeptide chain. The primary structure is always noted to by one or three letters. Secondary structure is formed when amino acids interact together forming hydrogen bonds. According to the DSSP – Dictionary of Protein Secondary Structures – [DSSP, 2014] secondary structures can be seen as eight classes namely: H (alpha helix), G(helix-3), I (Hilex-5), E(stand), B(beta bridge), T(Turn), S(bend), -(irregular). These structures are often mapped to three levels: alpha helix (H) and beta strands/sheets (E) and coil (C) which covers S, T and - states.

The three dimensional structure –also known as Tertiary structure- is formed when alpha helix and/or beta sheet interact together forming a more complex geometrical shape forming beta-peptide. Quaternary is stabilized by the same non-covalent interactions and disulfide bonds of the tertiary structure.

Secondary Structure Prediction Techniques/Algorithms

Protein secondary structure prediction is defined as the set of techniques and algorithms in bioinformatics that aim to predict the local secondary structures based only on the knowledge of their primary structure. As stated by Sara Silva [Silva, 2005], secondary structure predication passed through generations. These generations differ from one another by the techniques used and the prior knowledge of protein, starting by pure statistical methods getting into machine learning and intelligent techniques.

A. Statistical Generation

- This generation is characterized that all its methods are based on statistical analysis of single residue. The first probabilistic method that is considered the starting point of secondary structure prediction is “Chou-Fasman Method” [Chou–Fasman, 2014];
- The method is based on analysis of the relative frequencies of each amino acid in alpha helices, beta sheets, and turns based on known protein structures solved with X-ray crystallography. From these frequencies a set of probability parameters were derived for the appearance of each amino acid in each secondary structure type, and these parameters are used to predict the probability that a given sequence of amino acids would form a helix, a beta strand, or a turn in a protein.

The second most important method in this generation is the GOR method (Garnier-Osguthorpe-Robson) [Garnier et al, 1996]. The GOR method takes into account not only the propensities of individual amino acids to form particular secondary structures, but also the conditional probability of the amino acid to form a secondary structure given that its immediate neighbors have already formed that structure. The method is therefore essentially Bayesian in its analysis.

B. Enhanced Statistical Generation

This set of techniques introduced the usage of local interactions along with segment statists in the prediction approach. The start was GOR III which was an improvement to GOR. It was the first to use local interactions between amino acids. This means that, to predict the secondary structure of a given amino acid, the information about which amino acids are following and preceding it in the sequence is used [Garnier et al, 1996; Silva, 2005].

C. Machine Learning Generation

These set of techniques are homology based, studying the local interactions and alignments. Also introducing intelligent techniques such as neural networks, case based reasoning and much more.

– Neural Networks:

Qian and Terrence [Qian et al, 1988] introduced one of the first Neural Networks used for secondary structure prediction. They worked on a network with 17 input groups having 21 units per group, 40 hidden units and three output units. The usage of Neural Networks then started evolving and different architectures were studied targeting better prediction accuracy.

Chandonia et al [Chandonia, 1995], used the standard amino acid sequence as the input to their Neural Network, then they used the output of this prediction along with other information to help predict the structural class (tertiary structure). At last they used the structural class predicted in a third network to predict again the secondary structure to reach a better accuracy.

Later on, The Profile neural network prediction from HeiDelber (PHD) [Rost, 1996] methodology was introduced. It is considered the backbone of all proceeding methods. The PHD has four processing levels the output of each level is used as input for the following. It starts with a level that has the amino acid sequence as the input and outputs the likelihood that it belongs to alpha-helix, beta-strands or others (loop). Then the second level, uses this likelihood with some global information about the protein (for example, its length) to calculate a new likelihood. The third level chooses the classification of protein. Finally, filters the result observing common errors and unreasonable results.

More advanced neural networks were then proposed by Pollastri et al [Pollastri et al, 2002]. They used bidirectional recurrent neural network. Also they introduced the Second version of the SSpro program for secondary structure prediction.

– Case Based Reasoning:

Another technique different than Neural Networks was introduced by Janice et al [Glasgow et al, 2006]. They used Case Based Reasoning technique to predict the secondary structure of protein. They present the protein by a 2D map then they use case matching to query the cases that have common features with the new case. Getting these cases they are capable of getting the structure of protein.

– Swarm Intelligence:

Swarm intelligence is also used in Protein secondary structure prediction. Bai Li et at [Li, 2014] introduced the use of Artificial Bee Colony (ABC) algorithm. They used internal feedback strategy based ABC. It was proved to be effective to improve convergence rate also it was stated that this approach is better in exploration than exploitation.

– Support Vector Machine:

Haifeng Sui et al [Sui et al, 2011] proposed that Hybrid SVM can enhance the prediction accuracy of protein secondary structure. They proposed that combining physicochemical properties of amino acid residues with position-specific scoring matrices containing evolutionary information. The accuracy of this approach was not clearly proved but it was stated that it's better that it has a better ability than other methodologies.

– Combined Methods:

A combined method was introduced by Y. Wei et al [Wei, 2011]. They combined seven secondary structure prediction methods. The prediction is accomplished using the value from each predictor these values are then combined to find out the likelihood of the amino acid sequence.

Another combined method was proposed by Camacho, R. et al [Camacho et al, 2012]. This method reached an accuracy of almost 84.9% (in the prediction of α -helices) and 99.6% (in the prediction of the inner points of β -strands). This method combined rule induction algorithms, decision trees, functional trees, Bayesian methods and other algorithms.

Other methodologies that aim to enhance the prediction accuracy are introduced as well. Some methodologies uses the prediction of the tertiary structure as input to the prediction step as shown in [Chandonia, 1995]. Others combine the result from 3 predictors. The aim is always to increase the confidence level of predicting the amino acid sequence to be alpha Helix (H), Beta strands [E] or other (Loop). The three prediction always occurs to be one of three possibilities: 3:0 which indicates that the three methods predicted the sequence the same, 2:1 which leads to majority decision, or 1:1:1 which indicates a tie in which each predictor had a different output and in this case the amino acid sequence is predicted to be L state. The first two one of the states is dominant and it is chosen to represent the sequence [Albrecht et al, 2003].

The main observations from Table 1:

1- Protein secondary structure prediction started by statistical methods at which the prediction accuracy was very low. Then the accuracy started to increase when intelligent techniques arose. Getting into a fairly better accuracy when combining more than one methodology.
2- Neural Network with its variations is the most commonly used approach for Protein secondary structure prediction.
3- Other intelligent techniques are not yet mature as Neural Networks, although they have better accuracy. The approaches that tend to have accuracy better than 90% are those which use mixed predictors. Also SVM showed 90% accuracy only for β –strands prediction.

Table 1. Comparison of Secondary Structure Prediction Methodologies

Authors	Method / Algorithm	Dataset	Accuracy
Statistical and Enhanced Statistical Techniques			
Peter Y. Chou and Gerald D. Fa sman [Chou–Fasman, 2014]	Chou-Fasman method	-	50–60%
Jean Garnier et al [Garnier et al, 1996]	GOR & GOR III	Database of 267 protein structures	60%
Machine Learning Techniques			
Qian and Sejnowski [Qian et al, 1988]	Neural Network with window size 13	106 proteins	64.3%
Chandonia and Karplus [Chandonia, 1995]	Neural Networks	Set of 62 globular proteins (69 chains)	Secondary structure prediction 62.64% Class prediction 73.9%
Rost [Silva, 2005; Rost, 1996]	PHD	-	better than 72% about 74% of the segments are correctly predicted
Gianluca et al [Pollastri et al, 2002]	Recurrent Neural Networks and Profiles	Four data sets TRAIN for training and R126, EVA, and CASP4 for testing.*	78%
Janice et al [Glasgow et al, 2006]	Case Based Reasoning	-	83%
Y. Wei et al	Mixed integer linear	3000 proteins are selected from PDB as	83.04%

Authors	Method / Algorithm	Dataset	Accuracy
Statistical and Enhanced Statistical Techniques			
[Wei, 2011]	optimization	the training set.	
Camacho,R. et al [Camacho et al, 2012]	Machine Learning (rule induction, decision trees, functional trees, Bayesian methods)	1499 protein structures from the PDB	84.9% (in the prediction of α -helices) and 99.6% for β -strands
Haifeng Sui et al [Sui et al, 2011]	HSVM	462 proteins from the CB513 for training 3 for testing RS126, CB513 and CASP9*	independent predictions for more than 55% of all amino acid residues with accuracies of up to 90%
Bai Li et at [Li, 2014]	Bee Colony	-	-

* EVA, CASP4, CASP9, CB513, R126 and RS126 are all databases of protein structures.

Conclusion and Future Work

Sequence based prediction enjoys strong interest and finds its applications in various fields. Although it started long ago with probabilistic methods, recent research tries to find suitable intelligent techniques to enhance the prediction accuracy. Finding better methodologies to predict the secondary structure helps not only in the secondary structure domain but also in the tertiary structure domains.

We showed in this paper the three different generations of protein secondary structure prediction, namely the statistical generation, Enhanced statistical generation and Machine learning. We have demonstrated some of the most used techniques in each generation. Having an objective comparison among prediction methods is very difficult and not relevant in all cases. As shown each method used a different dataset for testing, also different definition for the input sequence and topology (some used variant length while others not).

However, the highest accuracies reached are from Case Based Reasoning approach which generated an accuracy of 83% and Mixed Integer Linear Optimization generated an accuracy of 83.4%. Our future work will go towards using more than one predictor and combine their results to reach a better accuracy and confidence level.

Bibliography

[Albrecht et al, 2003] M. Albrecht, Silvio C.E. Tosatto, Thomas Lengauer and Giorgio Valle, "Simple consensus procedures are effective and sufficient in secondary structure prediction," peds -Protein Engineering design & Selection-, vol. 16, no. 7, pp. 459-462, 2003. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[Camacho et al, 2012] R. Camacho, R. Ferreira, N. Rosa, V. Guimarães, N. A. Fonseca, V. S. Costa, M. D. Sousa and A. Magalhães, "Predicting the secondary structure of proteins using Machine Learning algorithms," Int. J. Data Mining and Bioinformatics, vol. 6, no. 6, pp. 571-584, 2012.

[Chandonia, 1995] John-Marc Chandonia and Martin Karplus, "Neural networks for secondary structure and structural class predictions," Protein Science, vol. 4, no. 2, p. 275–285, February 1995.

[Chou–Fasman, 2014] "Chou–Fasman method," Available: http://en.wikipedia.org/wiki/Chou%E2%80%93Fasman_method, [Accessed November 2014].

[DSSP, 2014] "DSSP", Available: http://www.cmbi.ru.nl/dssp.html, [Accessed November 2014].

[Esquivel, 2013] Rodolfo O. Esquivel, Moyocoyani Molina-Espíritu, Frank Salas, Catalina Soriano, Carolina Barrientos, Jesús S. Dehesa and José A. Dobado, "Decoding the Building Blocks of Life from the Perspective of Quantum Information," Advances in Quantum Mechanics, pp. 641-669, 2013.

[Garnier et al, 1996] Jean Garnier, Jean-François Gibrat and Barry Robson, "[32] GOR method for predicting protein secondary structure from amino acid sequence," Methods in Enzymology, vol. 266, p. 540–553, 1996.

[Glasgow et al, 2006] J. Glasgow, Tony Kuo and Jim Davies, "Protein Structure from Contact Maps: A Case-Based Reasoning Approach," Information Systems Frontiers, vol. 8, no. 1, pp. 29-36, February 2006.

[Kister, 2013] A. E. Kister, Protein Supersecondary Structures, Humana Press, 2013.

[Li, 2014] Bai Li, Ya Li and Ligang Gong, "Protein secondary structure optimization using an improved artificial bee colony algorithm based on AB off-lattice model," Engineering Applications of Artificial Intelligence, vol. 27, p. 70–79, 2014.

[Pollastri et al, 2002] Gianluca Pollastri, Darisz Przybylski, Burkhard Rost and Pierre Baldi, "Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles," PubMed, pp. 228-235, May 2002.

[Protein, 2014] "Protein", Available: http://en.wikipedia.org/wiki/Protein, [Accessed November 2014].

[Qian et al, 1988] Ning Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," Journal of Molecular Biology, vol. 202, no. 4, p. 865–884, August 1988.

[Rost, 1996] B. Rost, "[31] PHD: Predicting one-dimensional protein structure by profile-based neural networks," Methods in Enzymology, vol. 266, p. 525–539, 1996.

[Russell, 2009] P. J. Russell, iGenetics: A Molecular Approach, Third ed., PEARSON, 2009.

[Silva, 2005] S. Silva, "Predicting Protein Secondary Structure using Artificial Neural Networks - A Short Tutorial," 2005

[Sui et al, 2011] Haifeng Sui, Wu Qu, Bingru Yan and LiJun Wang, "Improved Protein Secondary Structure Prediction Using a Intelligent HSVM Method with a New Encoding Scheme," International Journal of Advancements in Computing Technology, vol. 3, no. 3, pp. 239-250, April 2011.

[Wei, 2011] Y. Wei, J. Thompson and C. A. Floudas, "CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science, pp. 831-850, November 2011.

Authors' Information

Hanan Hendy *– Teaching Assistant at Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University; e-mail: hanan.hendy@cis.asu.edu.eg*

Major Fields of Scientific Research: Bioinformatics, Artificial Intelligence

Wael Khalifa *– Lecturer at Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University; e-mail: wael.khalifa@cis.asu.edu.eg*

Major Fields of Scientific Research: Biometrics, Bio and Medical Informatics

Mohamed Roushdy *– Professor at Computer Science Department and Dean of Faculty of Computer and Information Sciences, Ain Shams University; e-mail: mroushdy@cis.asu.edu.eg*

Major Fields of Scientific Research: Artificial Intelligence, Medical Expert Systems

Abdel Badeeh Salem *– Professor at Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University; e-mail: absalem@cis.asu.edu.eg*

Major Fields of Scientific Research: Knowledge Engineering, Artificial Intelligence, Biomedical Informatics