Int J Pharm Bio Sci 2015 July; 6(3): (B) 549 - 561

Review Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

A STUDY ON PREDICTING PROTEIN SECONDARY STRUCTURE USING VARIOUS DATA MINING APPROACHES

ANBARASI M* AND SALEEM DURAI M. A

School of Computing Science and Engineering,
VIT University, Vellore, Tamilnadu, India.

ABSTRACT

The protein structure prediction is one of the major areas of interest and development of bioinformatics, theoretical chemistry, biotechnology and it is highly important in medicine. Protein secondary structure (PSS) plays a major role in 3D structure prediction, in the past two decades, an enormous number of works have been developed for protein secondary structure prediction (PSSP), where the accuracy for that particular research problem are not more than 90%. In this case there is a need to improve high accuracy prediction methods. The emanate methods and strategies are analyzed and generalized probabilistic approach. The main aim of this study is how data mining techniques have applied for predicting the protein secondary structure.

KEYWORDS: Protein secondary structure prediction, Data Mining, Classification, Neural Network and Clustering.

*Corresponding author

ANBARASI M
School of Computing Science and Engineering,
VIT University, Vellore, Tamilnadu, India.

INTRODUCTION

A protein is a polymeric macromolecule made of amino acid building blocks given in a linear sequence and attached together by peptide bonds. The primary structure of the proteins is represented by a linear polypeptide chain. The primary structure is typically represented by a sequence of letters over a 20 letter alphabet associated with the 20 native environment occurring amino acids. Proteins are built up by amino acids that are linked by peptide bonds to form a polypeptide chain. The amino acid sequence plays major role in predicting the secondary structure prediction such as α-helices, β-strands and c-coils. A protein consists of 20 amino acids, in each amino acid consists of structural component that includes central carbon atom (C) which is attached to an amino group (NH2), a hydrogen atom (H), a carboxyl group (COOH) and a side chain (R). A typical protein contains 200 to 300 amino acids, but some proteins sequences may have up to 30,000 amino acids. Protein structures are decomposed into five different levels: 1-D structure is the sequence of residues in the polypeptide chain. 2-D structures are classified into α-helices, β- strands and c-coils. 3-D structure created by assembly of secondary structures and 4-D structure formed by more than one polypeptide chain. The PSS plays an important role in prediction of protein structure and function [1,2,3]. Moreover, PSS provides useful input for prediction of fold type[4]. Over the past two decades, the secondary structure prediction accuracy has improved significantly through the various approaches, including data mining, pattern recognition, soft computing and machine learning methods have been developed to solve this problem. Today, there are a broad array of approaches to secondary structure prediction, including physic chemical properties, statistical information, sequence patters, graph theory, Chou Fasman[5], Garnier- Osguthorpe-Robson (GOR) techniques[6,7], a Profile network from HeiDelberg method[8,9] and several more methods. But previous methods for predicting PSS were mainly based on statistical analysis of single residue and its neighboring residues. However, the accuracy has not crossed more than 70%. Data mining methods are also applied for PSS using the following algorithm: Decision trees[10], Hidden Markov Models (HMMs)[11,12] clustering [13], multiple linear regression [14,15], nearest neighbor methods (NNSSP) [16], SSpro (Pollastri,s bidirectional recurrent neural network (PBRNN)) and association rules[17]. With the advance machine learning methods, such as an Artificial Neural Network (ANN) architectures have been used in that feed-forward ANNs [18,19,] Bidirectional Recurrent ANNs (BRNNs) [20,21,22] cascade-correlation ANNs [23,24] and Support Vector Machines (SVMs) have been proven successful over the past decade [25, 26, 27,28] the prediction rates increased to 80%. In the recent publications [43-45] the success rates of prediction range from 82 to 90%. The purpose of this study is to predict the protein secondary structure using data mining approaches.

(a) SSDICPGFLQVLEALLLGSESNYEAALKPFNPASDLQNAGTQLKRLVDTLPQETRINIVKLTEKI
(b) ----------HHHHHHHHHH----HHHHHHHHCCC---HHHHHHHHHHHHHHHH---HHHHHH

(c)

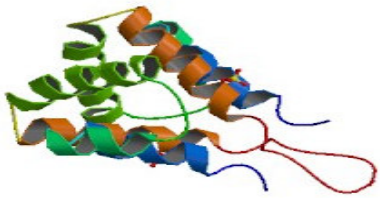

Figure 1
(a) Primary Structure (b) Secondary Structure (c) Tertiary Structure

SECONDARY STRUCTURE ASSIGNMENT (SSA)

First of all, it is necessary to consider the general concepts related to the SSA. The secondary structure is usually obtained from the experimentally determined tertiary structure by DEFINE, STRIDE or DSSP. The DSSP is the most commonly used annotation for protein secondary structure prediction. In DSSP classify the residues into eight different secondary structure classes are G, I, H, B, E S, T and C and these eight classes are reduced into three classes (H-helix, E-strand, C-coil) as follows: {G, H, I}→ H, {B, E}→ E, and {T, C, S}→ C.

METHODS FOR DATA MINING OF PROTEIN SECONDARY STRUCTURE

In this review we use the SSA approach which is considered as classes often also external groups. The methods used for bioinformatics, computational and systems biology and the features of this processing the following key approaches to formation of PSSP, which is one of the most fundamental issues in structural biology[29] and bioinformatics.

(i) Decision Trees

Joachim Selbig et al [30] proposed method CoDe (Consenus formation by decision tree learning) and compared with 10 methods such as SOPMA, PREDATOR, DPM, HNN, LEV, GORI, SIMPA96, GOR2, GOR4 AND SOPM which yields reliable results than individual methods then the PREDATOR method yielded accuracy of 80.1% for PSS. Nitesh Chawala et al [31] proposed bagging method that showed improved results but only for small data set. For the large set of data the sub sampling method was proposed and the accuracy yielded 74.1% but a single classifier was 78.6%. Leong Lee et al [32] proposed a new method RT-RICO (Relaxed Threshold Rule Induction from Coverings) that achieved better than earlier reported methods and yielded accuracy of 81.7%. Minh, N et al [33] searched for the relevant rules using c4.5 and developed prediction model using two stages SVM. The generated rules are higher confidence levels when compared with other rule extraction techniques. In each method, special installations are employed differing in engineering solution of particular decision trees. Decision trees are quite familiar in expert systems and they are generally considered to be an ideal approach. Decision Tree can be set of rules and produce interpretable rule and to improve human readable form. A decision tree contains decision leaves and nodes. Leaf nodes assign a class to objects. The allocation can be probabilistic and can specify a degree of confidence. A decision node indicates a test on an attribute of an object. The test performed by n-way divide on a discrete data set or can be a disparity on a continuous value. The tree is to predict specified object in the class, a path is traced from the root to leaf node of the tree subsequent arcs as specified by the outcome of test. A path in a tree represents a conjunction of tests: a&b&c&.... For example, the object <T,F,T,F> is in c2(class 2) because starting from the root the subsequent directions are taken: left, right, left, right. The leaf predicts c2(class2) by the results of tests.

The function to be learned

class 1 = ($1 & $2) or ($3 & $4)

class 2 = ~ 1

The binary attributes are $1, $2, $3 and $4.

The disadvantage of decision tree is over–sensitivity to the training set for irrelevant attributes and noise data, Sometimes decision trees can be difficult to understand. The rule sets consist of simple that are derived from decision tree in which we need to write a rule from root to a leaf node.

HICSS 26V1 p671 (1993)

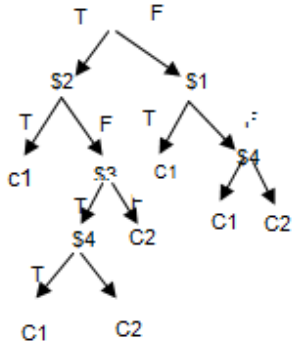

Figure 2
Decision Tree

In Jieyue He et al[10] proposed a rules if A481>0 then ,~E, [96.8%] and IF seq[6]=,A, THEN Seq[0] =, ~E, [96.8%].

(ii) Neural Network

Protein secondary structure forms from a sequence of amino acid. Some properties of a protein can be determined from knowing the secondary structure of the protein. A subsequence string of a long amino acid string can be formed into three distinct structures namely α – helix, β – sheet and coil. There are so many possible combinations of amino acid that could fold into distinct structure that is why automatic predicting secondary structure of protein can be of great use. The neural network (NN) as its name suggests, simulator to PSSP. The recent advances made using this method is when an element of the neural network fails, it can continue without any problem, allows for qualitative data, can handle noisy data and its user friendly. NN were first in use for PSSP in Ning Qian [34] proposed method globular proteins based on non-linear NN model able to achieve the accuracy 64% and tested with non homologous protein, the success rate of homologous protein were higher than that of non-homologous proteins. Soren Kamaric Riis [35] proposed separate networks and designed using prior knowledge of amino acid properties and the characteristic periodicity, the accuracy yielded 66.3% using seven fold cross validation for non homologous globular proteins. After applying multiple sequence alignments the accuracy improved to 71.3%. Ashraf Yaseen et al [37] proposed template-based and incorporated features with sequence and evolutionary information approach to improve 8-state secondary structure prediction accuracy 78.85%. Anureet Kaur Johal et al [36] have achieved the maximum prediction accuracy of about 81%. Haifeng Sui et al proposed multi-modal BP method which had a greater accuracy rate of approximately 85%. The neural network[62] is defined with input layer (IL), hidden layer (HL) and output layer (OL). The protein sequence is represented as a sliding window of size w (from 15 to 29) and the prediction is prepared on the structural assert of the middle residue of the window. For a protein sequence of window size w is represented as 20xw. Hence the IL consists of 20xw input units, i.e., w is a collection of 20 inputs each one is for each window. All the proteins are used for training the neural network that are encoded and stored in vector. Each target is represented as a Boolean array of size 3, which represents one of the secondary structural asserts of the amino acid at that location in the protein sequence. The secondary structure states are defined based on DSSP assignment. Thus H is represented as 100, E is represented as 010 and finally L is represented as 001. Thus the OL of the neural network consists of three units and one for each class. The target matrix is also prepared. The size of the HL is considered as 2xw+1. By each training

sequence, one residue at a time through the sliding window the training sequences is presented to the network. Each hidden unit changes the signal that is received from the IL by using transfer function (i.e log sigmoid) to predict the output signal either 0 or 1. Weights are corrected so that the inaccuracy between the observed output and the desired output is minimized. In neural network the number of parameters adjusted based on the training set size on the variable and by changing number of hidden units. The disadvantage over this method is risk of overfitting, requires a definition of architecture, this process is time consuming and expensive.

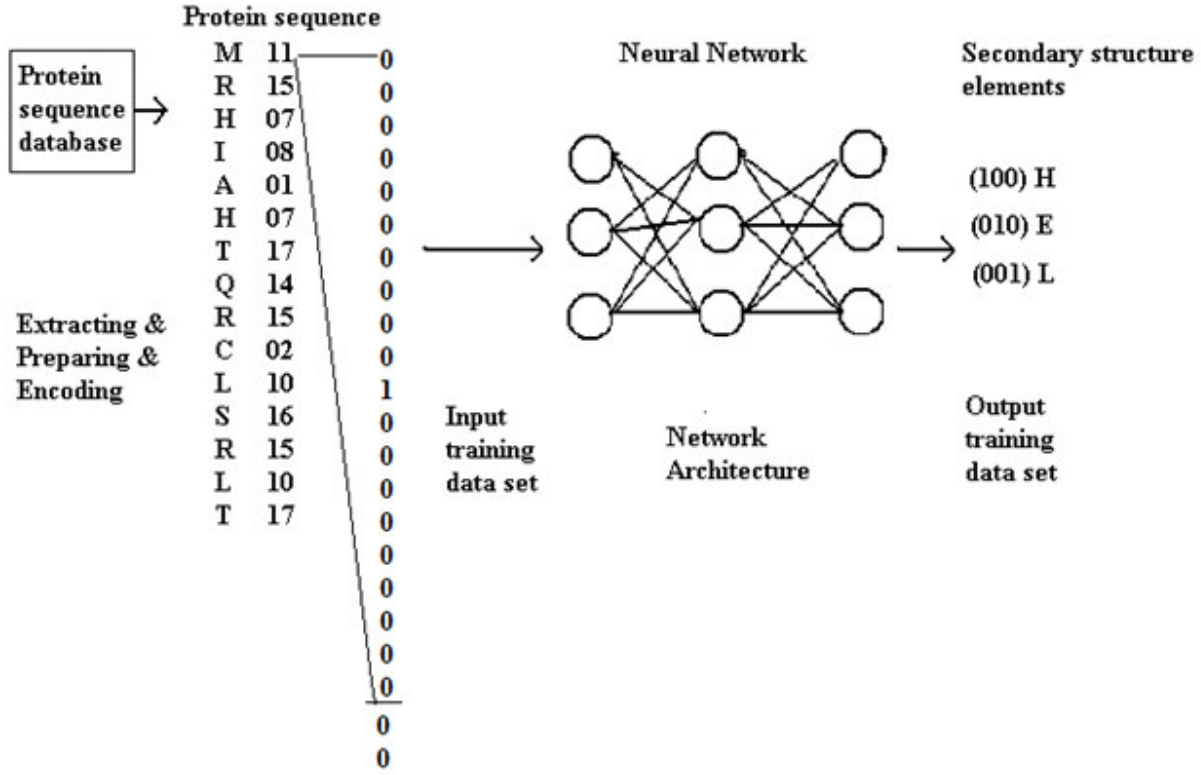

Figure 3
Neural network for Protein secondary structure prediction[61]

(iii) Support Vector Machine (SVM)

Prediction of protein structure has been a successful technique in the data mining approach. Bioinformatics techniques include protein fold recognition, protein to protein interaction, microarray gene expression analysis and protein remote homology which uses the SVM. The non linear classification problem can be solved by mapping of input feature vectors into higher dimensional feature space this is an advantage of using SVM not just that SVM also produces very accurate classifiers, less over fitting and robust to noise data. To increase the potential of SVM they used the optimal window size of protein structure for helix, stand and coil and best feature vector. The drawback of using this method is long training time, difficult to understand the learned function, not easy to incorporate domain knowledge and if this method is binary classifier. To do a multi class classification for that pair wise classification can be used (one against all others, for all classes) not only because it is computationally expensive, the performance as well is slow. Geoffrey et al [38] prediction for PSSP based on automatic and semi automatic methods and yielded an accuracy rate of 70% approximately. Then in Minh N. et al [39] proposed two stages SVM performs better than single stage SVM techniques and tested with two data sets

obtained accuracy of 79.5%. Later, in Long- Hui Wang et al [40] proposed method based on SVM and considered structure, physical chemical properties of amino acids showed up an accuracy of 78%. Then in Nguyen M.N [41] proposed two stage multiclass SVM approach and used position specific scoring matrices obtained accuracy of 78.0%. Mohammad Shoyaib et al [42] developed method based on machine learning. In that first identified frequent pattern of consecutive amino acids and set of frequent word, then binary or tertiary classification is used for classification of proteins and the accuracy is better than the previous method. Finally in Bingru Yang et al [43] proposed mixed model SVM method for each amino acid considered physicochemical properties and position specific scoring matrix accuracy achieved 85.58%.A sliding window scheme is used to train the SVM with structural information and protein sequence. In sliding scheme, a window is one training sample to predict the structure of residue in center of the window. The benefits of this training pattern, is that the sequence about the local interactions between residues are embedded. In (Figure 4) shows sliding window scheme having window size of 5. In this example we need to predict the amino acid, S, for the sequence, Q N S P I, and that will be considered as one input pattern. Similarly if you consider next input pattern, P, the next group of sequence,NSPIS, and so on.

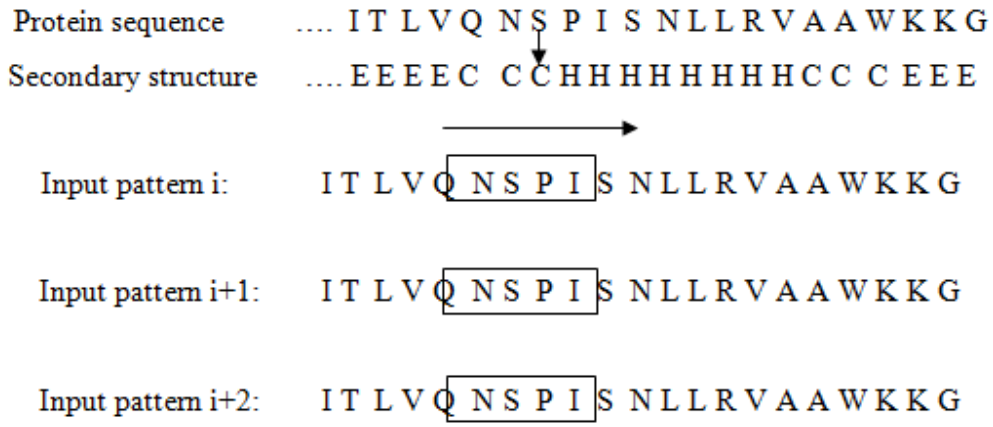

Figure 4
Sliding window with window size of 5

The above single sliding window scheme is used to form the data for binary classifier that is challenged with multiple windows scheme. Both the binary classifiers show approximately the same accuracy. The optimal window size is resolute by Hu,s method [57] that is applied for binary classifiers. Then, two tertiary classifiers have been introduced. In both tertiary classifiers shows better accuracy when compared to Hu method. Based on the performance of binary classifier the orthogonal and BLOSUM62 matrix was not suitable but for two tertiary classifiers the performance was improving and the prediction accuracy has been optimized.

(iv) Hidden Markov Model (HMM)

Kiyoshi Asai et al [46], for analyzing the protein sequence using stochastic model. The result was reasonable even though they implemented ,without grammar,. In Jeanette Hargbo et al [47] using fold recognition that yielded a good accuracy when compared with the ordinary HMM and by using substitution matrix they have improved the performance of HMM. Later in Juliette Martin et al [48] which achieved an accuracy of 78.2 %.Then Zafer Aydın [49] proposed the method hidden semi-Markov model for a single-sequence and accuracy rate of 88%. In Qi Dai et al [50] constructed transition matrices to predict the yesterdays, today,s and tomorrows of 20 amino acids that method has the advantages as well as the disadvantages.

Then in K. K Senapati et al [51] proposed algorithm and used sliding window for PSSP and achieved 64% accuracy for helical structure. In Liang K, Wang X [52] proposed deterministic sequential sampling based algorithm for single sequence, to locate the most likely secondary structure conformation of protein. In Mechelke M et al [53] used chemical shift parameters for PSS and improved accuracy. This algorithm is based on a windowed observation HMM they have improved conformation scored by considering an average over the expected conformations within a window.In this method, a markov sequence, the character appearing location n in the sequence depends only on the previous character at location (n-1). Hence, a markov chain is fully defined by transition matrix. In HMM, the transition matrix states that hidden state generated the observation at time n. HMM can be used to uncover the hidden process which are most possible generated the observed sequence with the viterbi and forwards-backwards algorithms. For the protein sequences, the hidden process to be recovered is the protein secondary structure and the observed process is the amino acid sequence.

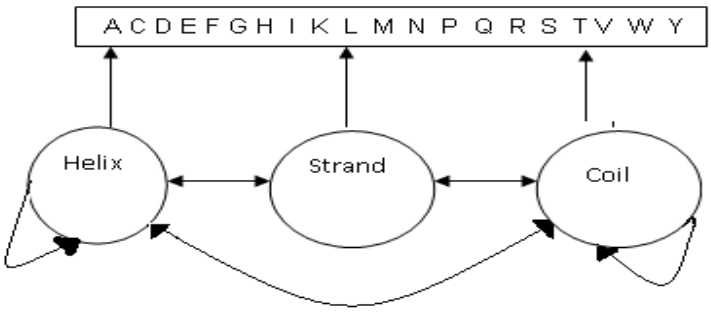

Figure 5
Protein secondary structure 3-State of HMM.

Arrows are connecting the hidden and observable states are the probability of considering a certain amino acid (observed state), given that the Markov Chain is in a particular hidden state. It is assumed that the residues within a sequence can be categorized into three secondary structures: helix, sheet and coil. 20 distinct amino acids are associated with each secondary structure of protein. In order to build an HMM, the parameters λ need to be expected from a set of observation sequences. This involves likelihood of HMM to be defined. Once, The model can be used to pick up the most possible sequence of hidden states from an observation sequence, then the parameters of HMM have been estimated This is the main objective of protein secondary structure prediction in HMM. To use the model to pick up from a protein,s amino acid sequence, what state of the amino acid sequence folds into secondary structures. Using this method maximum they have achieved 84% accuracy.

(v) Nearest Neighbor (NN)

Asaf et al [54] proposed combined approach nearest-neighbor algorithms and multiple sequence alignments and reached the accuracy of 72.2%, which was better than multilayered neural-network approach, which tested on non-homologous protein. Later, in Saejoon kim et al [55] showed up 75% accuracy which is better than the neural network. In Ashish Ghosh et al used three distanced based classifiers (minimum distance, KNN and Fuzzy KNN) among these they found minimum distance classifier performs better results for window size 11. In [52] proposed distance metric learning technique and the energy based calculation rule for predicting PSSP which is better than previous nearest neighbor methods such as PREDICT and NNSSP. Still, our

prediction get overall accuracy obtained 75.44%. In the same year same authors [56] proposed DPred to use homologous and non homologous information for PSSP which is compared with the following methods Porter_H, PROTEUS and CDM and the accuracy achieves 87.51%.This method is used for choosing the closest subsequences to a window in the order of the amino acid which is being predicted. The secondary structure prediction is made using the known secondary structure of aligned sequences. However, this method is simple but, there are number of undefined parameters that allow the method to be useful in a broad range of ways. In all nearest neighbor approach, different set of parameters have been used for instance, Sequences are selected based on their similarity, but which method is used for defining the similarity? How many close sequences should be selected? What could be the window size sequence? This method can be prepared in a number of ways, using distance measure suggestive of the probabilities used in the information theory, SSPAL method is much more successful. This algorithm gives better results when homologous protein is not available. In this method, a large list of short sequence portions is prepared by sliding window of different length approximately 100-400 sequences of known structures. The minimal sequence similarity to each other and the secondary structure of the amino acid in each window is recorded. In the given query sequence, a sliding window of same size is selected and compared with other sequence portion and identified the best 50 matching portions. The rates of recurrence of known secondary structure of the center amino acid in each of the matching portions are used to predict the secondary structure of the center amino acid in the query window. Rules are used to make a final prediction for each amino acid position.

(vi) Association Classification (AC)

In Zhun Zhou et. al , modify the amino acid sequence in the form of window, If the current window is unknown, then in rule set helix H, search for satisfied set h subset H, else search in the rule set E else search in rule set C. Based on the output H, E and C run CMAR to build a new classification. If classification is achieved then predicts the current window as H, E and C. otherwise mark the window as unpredictable and checks it's a last window. If not go to the next window until the last window is reached. The accuracy ratio exceeded 85% when confidence threshold value was 70% and 90%.

(vii) Clustering

Shing H et al [58] proposed SVM and Clustering method achieves good performances on redundant and non-redundant sets and maintaining good accuracy rate. Rajbir Singh et al [59] attempts to classify amino acid in protein sequence according to that predicted local structure. In Lu Zhao et al [60] proposed a new cluster merging method to increase the precision of protein folding prediction by innovative idea to balance over fitting and under fitting and showed up the accuracy of 82.5%. Clustering provides several advantages, based on the sequence similarity clustering proteins are usually grouped into families, which provides some hints on the common features of that family and evolutionary confirmation of proteins. This method also helps to understand the biological function of a new sequence by its similarity to some function-known sequences. In addition, protein clustering can be used to find 3-dimensional structure of protein.The given sequence is divided into clusters and in reet kamal et al proposed propensities values for helix, stand and coil. 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 amino acid residues have $P(i)>100$ are identified and extended until a proline is meet (helix breaker) or a run of 4 residues with $P(i)<100$ is found. Similarly for regions where four of six residues have $P(j)<100$, are extended and a beta strand is predicted if the average P(j) over all residues in the cluster are greater than 100 and $\Sigma P(j) > \Sigma P(i)$. For alpha helix prediction the $\Sigma P(i)$ is computed and for each cluster is >5 and the $\Sigma P(i) > \Sigma P(j)$, 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.

PERFORMANCE EVALUATION

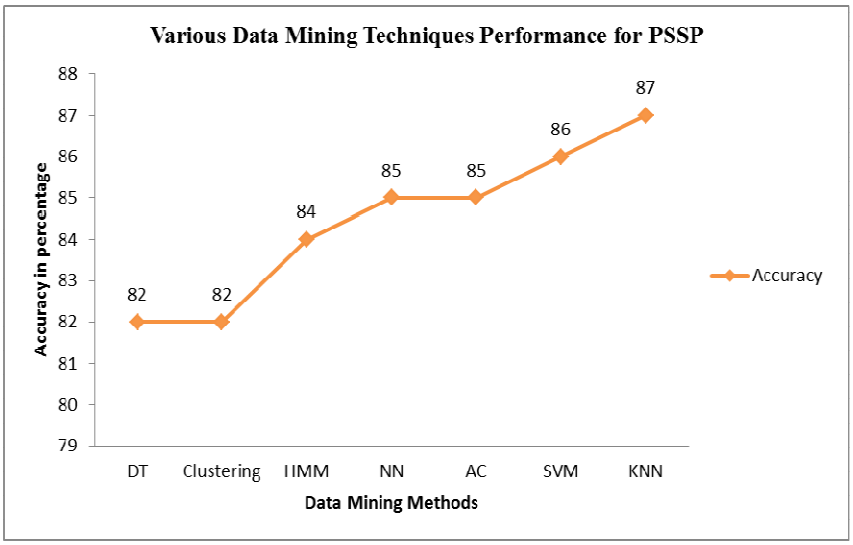

Figure 6
Prediction results achieved for PSSP using Data Mining Techniques

(Figure 6) shows performance of various data mining techniques for PSSP. The classification techniques show better performance than other data mining techniques. In classification techniques for PSSP, KNN gives better results in terms of accuracy and accurate information. When compared with other classification algorithms like SVM, NN, HMM and DT.

CONCLUSION

Protein secondary structure plays an important role in the rapidly developing branches of engineering specializing in the study of objects in 3D structure prediction. Despite the fact that the extensive use of (especially for biological applications) developed for predicting protein secondary structure. But many of the researches they have applied probabilistic approach. There are many decision tree algorithm the results of [32] show that c4.5 provide good classification accuracy are used more and more often in every day practice. Some companies have already used various architectures, but three layer architectures gives good performance with minimum number of nodes. SVM has shown much better performance than most traditional machine learning approaches such as neural networks. K-Nearest Neighbor methods give a relatively better performance than Hidden Markov models or Neural Networks. Association classification breaks the limitations of algorithms built on homological analysis and classic artificial intelligence, such as decision tree and the SVM and by pursuing a different way; it develops, out of rules of association classification. Up to now, for this problem maximum 90% achieved with less similarity. In our opinion, it is time for extensive search for the ways of practical use PSS accuracy and speed consideration are likely to remain important as genomic, proteomics and protein engineering projects continue to generate great challenges and opportunities in this area. According to the result mentioned in (Figure 6), we conclude that the predict system applies to protein secondary structure prediction get very good results for the classification techniques. The protein 2D structure prediction is one of the most hopeful works in the future.

REFERENCES

1. Lim, T. S. Loh, W. Y. Shih, Y. S , A comparison of prediction accuracy, complexity and training time of thirty tree old and new classification algorithm, Machine Learning, 40(3), 203–228, (2000).
2. Pollastri, G. Przybylski, D. Rost, B. Baldi, P.,Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles, PROTEINS: Structure, Function, and Genetics, 47, 228–235,(2002).
3. Rost, B ,Review: protein secondary structure prediction continues to rise, Journal of Structural Biology, 134, 204–218, (2001).
4. Chen,K. Kurgan,L, PFRES: protein fold classification by using evolutionary information and predicted secondary structure, Bioinformatics, 23 (21), 2843–2850, (2007).
5. Chou PY, Fasman GD, Prediction of protein conformation, Biochemistry, 13(2), 211–215, (1974).
6. Gibrat, J.F. Garnier, J. Robson, B, Further Developments of Protein Secondary Structure Prediction Using Information Theory, Journal of Molecular Biology, 198(3), 425-443, (1987).
7. Garnier J, Osguthorpe DJ, Robson B, Analysis and implications of simple methods for predicting the secondary structure of globular proteins, Journal of Molecular Biology, 120 (1), 97–120, (1978).
8. Jones, D.T, Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices, Journal of Molecular Biology, 292 (2), 195-202, (1999).
9. Rost B. Sander, C., Prediction of Protein Secondary Structure at Better than 70% Accuracy, Journal of Molecular Biology, 232 (2), 584-599, (1993).
10. Jieyue He, Hae-Jin Hu, Robert Harrison, Phang C. Tai, Yi Pan, Rule Generation for Protein Secondary Structure Prediction With Support Vector Machines and Decision Tree, IEEE Transactions on Nanobioscience, 5(1), 46-53, (2006).
11. K. Karplus, C. Barrett, M. Cline, M. Diekhans, L. Grate, R. Hughey, Predicting Protein Structure Using Only Sequence Information, Proteins, 37, 121-125, (1999).
12. Lin, K. Simossis, V.A. Taylor, W.R. Heringa J.,A Simple and Fast Secondary Structure Prediction Method using Hidden Neural Networks, Bioinformatics, 21 (2), 152-159, (2005).
13. Reet Kamal Kaur, Manjot Kaur, Amanjot Kaur, Using Cluster Analysis for Protein Secondary Structure Prediction, International Journal of Computer Applications, 4(12), 20-22, (2010).
14. Pan X. M, Multiple Linear Regression for Protein Secondary Structure Prediction, Proteins, 43 (3), 256-259, (2001).
15. Qin, S. He, Y. Pan X.M, Predicting Protein Secondary Structure and Solvent Accessibility with an Improved Multiple Linear Regression Method, Proteins, 61(3), 473-480, (2005).
16. Salamov A. A. and Solovyev V.V, Prediction of Protein Secondary Structure by Combining Nearest-Neighbor Algorithms and Multiple Sequence Alignments, Journal of Molecular Biology, 247(1), 11-15, (1995).
17. Zhun Zhou a,Bingru Yang b and Wei Hou b,Association classification algorithm based on structure sequence in protein secondary structure prediction, Expert Systems with Applications, 37 (9), 6381–6389, (2010).
18. Holley HL, Karplus M, Protein secondary structure prediction with a neural network, Proceeding of the National Academy of Science of the United states of America, 86, 152-156, (1989).
19. Baldi, P. Brunak, S. Frasconi, P. Soda, G. Pollastri. G, Exploiting the Past and the Future in Protein Secondary Structure Prediction, Bioinformatics, 15(11), 937-946, (1999).
20. Pollastri, G. and McLysaght, A, Porter: A New, Accurate Server for Protein

Secondary Structure Prediction, Bioinformatics, 21(8), 1719-1720, (2005).
21. Wood M.J. and Hirst J.D, Protein Secondary Structure Prediction with Dihedral Angle, Proteins, 59(3), 476-481, (2005).
22. Dor O, Zhou Y, Achieving 80 % ten-fold cross-validated accuracy for secondary structure prediction by large-scale training, Proteins, 66 (4), 838–845, (2007).
23. Faraggi E, Zhang T, Yang YD, Kurgan LK, Zhou YQ, SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles, Journal of Computation Chemistry, 33 (3), 259–267, (2012).
24. Karypis, G, YASSPP: Better Kernels and Coding Schemes Lead to Improvements in Protein Secondary Structure Prediction, Proteins, 64(3), 575-586, (2006).
25. Hua, S. and Sun, Z, A Novel Method of Protein Secondary Structure Prediction with, High Segment Overlap Measure: Support Vector Machine Approach, Journal of Molecular Biology, 308(2), 397-40, (2001).
26. Kountouris, P. and J.D. Hirst, Prediction of Backbone Dihedral Angles and Protein Secondary Structure Using Support Vector Machines, BMC Bioinformatics, 10(1), 1-14, (2009).
27. Zhang, Q. Yoon, S. and Welsh, W. J, Improved method for predicting beta-turn using support vector machine, Bioinformatics, 21(10), 2370–2374, (2005).
28. Zimmermann, O. and Hansmann, U.H, Support vector machines for prediction ofdihedral angle regions, Bioinformatics, 22 (24), 3009–3015, (2006).
29. Swapnil G Sanmukh, Waman N Paunikar, Tarun K Ghosh, Tapan Chakrabarti, Structural & Functional Prediction Of Hypothetical Proteins In Bacteriophages Against Halophilic Bacteria- An In Silico Approach, International Journal of Pharma and Bio Sciences, 2(2), 61-70, (2011).
30. Joachim Selbig, Theo Mevissen ,Thomas Lengauer, Decision tree based formation of consensus protein secondary structure prediction, Bioinformatics, 15 (12), 1039-1046, (1999).
31. Nitesh Chawala, Thomas, E. Moore Kevin. Jr, Bowyer, W, Bagging-Like Effects for Decision Trees and Neural Nets in Protein Secondary Structure Prediction, Workshop on Data Mining in Bioinformatics, 3(2), 50-59, (2001).
32. Leong Lee, Jennifer L. Leopold, Cyriac Kandoth and Ronald L. Frank, Protein Secondary Structure Prediction Using RT-RICO: A Rule-Based Approach, The open Bioinformatics Journal, 4, 17-30, (2010).
33. Minh, N. Nguyen. Jacek, M. Zurada. Jagath. Rajapakse, C, Toward Better Understanding of Protein Secondary Structure: Extracting Prediction Rules, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(3), 856-864, (2011).
34. Ning Qian. Terrence, J. Sejnowski, Predicting the Secondary Structure of Globular Proteins Using Neural Network Models, Journal of Molecular Biology, 202(4), 865-884, (1988).
35. Soren Kamaric Riis, Improving Prediction of Protein secondary structure using structured Neural Networks and Multiple sequence Alignments, Journal of Computational Biology, 3(1), 163-183, (1996).
36. Anureet Kaur Johal, Prof. Rajbir Singh, “Secondary Structure Prediction Using Improved Support Vector Machine And Neural Networks”, International Journal Of Engineering And Computer Science ISSN:2319- 7242,3(1), Jan,(2014).
37. Ashraf Yaseen, Yaohang Li,”Template-based C8-SCORPION: a protein 8-state secondary structure prediction method using structural information and context-based features”, From Third IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS 2013) New Orleans, LA, USA. 12-14 June. (2013).

38. Geoffrey J Barton, Protein secondary structure prediction, Current Opinion in Structural Biology, 5, 372-376, (1995).
39. Minh N. Nguyen and Jagath C. Rajapakse, Multi-Class Support Vector Machines for Protein Secondary Structure Prediction, Genome Informatics, 14, 218–227, (2003).
40. Long Hui Wang and Juan Liu, Predicting Protein Secondary Structure by a Support Vector Machine Based on a New Coding Scheme, Genome Informatics, 15 (2), 181–190, (2004).
41. Nguyen M.N. and Rajapakse J.C, Two-Stage Multi-Class Support Vector Machines, Pacific Symposium on Biocomputing, 10, 346-357, (2005).
42. Mohammad Shoyaib. Syed Murtuza Baker. Taskeed Jabid. Firoz Anwar. Haseena Khan, Protein Secondary Structure Prediction with High Accuracy using Support Vector Machine, 1-4, (2007).
43. Bingru Yang, Qu Wu, Zhou Ying and Haifeng Sui, Predicting protein secondary structure using a mixed-modal SVM method in a compound pyramid model, Knowledge- Based Systems, 24 (2), 304–313, (2011).
44. Yin-Fu Huang and Shu-Ying Chen, Extracting Physicochemical Features to Predict Protein Secondary Structure, IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 9-15, (2013).
45. Yang W, Wang K, Zuo W, Prediction of protein secondary structure using large margin nearest neighbour classification, International Journal of Bioinformatics Research and Applications, 9(2), 207-219, (2013).
46. Kiyoshi Asai, Satoru Hayamizu and Ken ichi Handa, Prediction of Protein Secondary Structure By The Hidden Markov Model, Oxford University Press, 9(2),141 -146, (1993).
47. Jeanette Hargbo and Arne Elofsson , Hidden Markov Models That Use Predicted Secondary Structures For Fold Recognition, PROTEINS: Structure, Function and Genetics 36, 68–76, (1999).
48. Juliette Martin, Jean-François Gibrat, and François Rodolphe, Choosing the Optimal Hidden Markov Model for Secondary-Structure Prediction, Data Mining in Bioinformatics, 1541-1672/05 IEEE Published by the IEEE Computer Society,19-25, (2005).
49. Zafer Aydın, Y ucel Altunbas¸ak, Mark Borodovsky, Protein Secondary Structure Predictionwith Semi Markov Hmms, ICASSP 2004 0-7803-8484-9/04, 577-580, (2004).
50. Qi Dai a, Xiao-qing Liu b, Tian-ming Wang a, Analysis of Protein Sequences and their Secondary Structures Based on Transition Matrices, Journal of Molecular Structure: THEOCHEM, 803, 115–122, (2007).
51. K. K Senapati, G. Sahoo and D. Bhaumik, Algorithm for Predicting Protein Secondary Structure, Journal of Computer Science And Engineering, 1(1), 68 – 71, (2010).
52. Liang K, Wang X, Protein Secondary Structure Prediction using Deterministic Sequential Sampling, Journal of Data Mining in Genom Proteomics, 2(107), pp 1-7, (2011).
53. Mechelke M, Habeck M, A probabilistic model for secondary structure prediction from protein chemical shifts, Proteins, 81(6), 984-93, (2013).
54. Asaf A. Salamov and Victor V. Solovyev, Prediction of Protein Secondary Structure by Combining Nearest-neighbor Algorithms and Multiple Sequence Alignments, Journal of Molecular Biology, 247(1), 11–15, (1995).
55. Saejoon Kim, Protein β-turn prediction using nearest-neighbor method, Bioinformatics, 20(1), 40–44, (2004).
56. Wei Yang, Kuanquan Wang and Wangmeng Zuo, A fast and efficient nearest neighbor method for protein secondary structure prediction, 3rd International Conference on Advanced Computer Control, 18-20 January 2011, Harbin, China, (2011).

57. Hu, H. Yi, P, Improved Secondary Structure Prediction Using Support Vector Machines with a New Encoding Scheme and an Advanced Tertiary Classifier, IEEE Transaction on Nanobioscience, 3(4), (2004).
58. Shing H. Doong Chi Y. Yeh, Secondary Structure Prediction Using SVM and Clustering, Proceedings of the Fourth International Conference on Hybrid Intelligent Systems, Kitakyushu, Japan, (2004).
59. Rajbir Singh, Sumandeep Kaur Deol and Parvinder S. Sandhu, Chou-Fasman Method for Protein Structure Prediction using Cluster Analysis, World Academy of Science, Engineering and Technology, 48, 980-985, (2010).
60. Lu Zhao. Ngyuen - Quang Phuoc. Sung - Ryul Kim, A new approach to protein secondary structure prediction based on Cluster Merging, 2, 326-329, (2012).
61. Zikrija Avdagic, Elvir Purisevic, Emir Buza1, Zlatan Coralic, Neural Network Algorithm for Prediction of Secondary Protein Structure, 17(2), 67-70, (2009).
62. P.V. Nageswara Rao, T. Uma Devi, DSVGK Kaladhar, Protein Secondary Structure Prediction using Pattern Recognition Neural Network, International Journal of Engineering Science and Technology, 2(6), 1752-1757,(2010).