Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.8 No.1, 2019, pp.36-41

DOI: https://doi.org/10.51983/ajcst-2019.8.1.2121

Prediction of Secondary Structures of Human Prion Proteins using Miscellaneous Methods

Ahmad Aftab Khan[1] and Kalpana Sharma[2]
[1]Research Scholar, [2]Assistant Professor & HOD
[1&2]Department of Computer Science Engineering, Bhagwant University, Ajmer, Rajasthan, India
E-Mail: ahmedkhan311@gmail.com, kalpanasharma56@gmail.com
(Received 4 January 2019; Revised 23 January 2019; Accepted 8 February 2019; Available online 15 February 2019)

***Abstract*- The contemporary study is an attempt to predict the secondary structures of proteins, from the dataset of human prion proteins which has been acquired from NCBI repository. In this context, we have exploited PSIPRED server which is considered to be proficient and impulsive method to protein structure prediction where users can submit the query sequence of their desire and receive the results of prediction both textually and eloquently. Furthermore, Phyre2 was applied across the amino acid sequence which is among the most widely server deployed for generating consistent protein models characterizing the prediction of protein structures. Moreover, two feed forward neural networks with a sole hidden layer which was tested and trained on a 10 fold cross validation mechanism in MATLAB, and subsequently significant prediction accuracy of 71.73% and minimum mean absolute error of 12.3% was achieved.**
Keywords: **Protein Structure Prediction, PSIPRED, Phyre2, BLAST, PSIBLAST**

I. INTRODUCTION

Protein Structure prediction (PSP) is considered to be an imperative area for investigation, typically in fields including bioinformatics and biochemistry. The PSP endeavours to predict the secondary structure of proteins based on the information such as amino acid sequence and primary structure. Nevertheless, in spite of the contemporary breakthroughs in forecasting the protein secondary structures via multiple sequence alignment and various artificial intelligence algorithms, the Q3 prediction accuracy of different computational approaches has barely demonstrated significant performance that exceeds 80%. Moreover, protein structure involves four level hierarchies comprising of primary, secondary, tertiary and quaternary structure of proteins. The primary structure of proteins is formulated by the linear sequence of amino acids. The secondary structures are generated through peptide bounds by the local compositions among neighbouring amino acids, leading to the advancement of three major secondary structures visa viz. α-helix, β-sheets and coils. As a matter of consequence these compositions are responsible causing diverse kinds of forces such as attraction, repulsion and hydrophobic. These forces along with the exposition of hydrogen bounds and disulfide bridges result in the formation of balanced 3D structures known as tertiary structures. At the peak level, a more complex structure namely quaternary exists, which explains how various polypeptide chains come together to form a functional protein. The tertiary and quaternary structures are ascertained by hydrophobic and ionic interactions among amino acids [1].

Considering the significance of 3D structure which acts as foundation for predicting and discovering the protein function, however, it would be extremely a distant challenge if secondary structures are not deployed to simplify the task and exploited as an intermediate phase. Furthermore, the secondary structures act as input parameters for various bioinformatics tasks. Essentially, there are two techniques which can determine the secondary structures viz. computational and experimental approaches. The experimental means conducted by (Buxbaum, 2007) were in exercise prior to the advent of computational approaches which encompasses of electron microscopy, X-ray crystallography and nuclear magnetic resonance. The impediments of these methods are that they are costly, time consuming and sometimes may take several months or years to generate a single structure, and more importantly may not be applicable and significant to every protein. On the contrary, protein sequencing impetus produced an enormous gap among known sequences and undetermined structures. This gap compelled the researchers to come up with the development of expeditiously accurate methods, and consequently gave birth to computational approaches.

Moreover, there is a consensus among researchers (Ofer and Zhou, 2007) that literature study and various methods pertaining to proteins must have a revision in general, thereby discovering hidden patterns in datasets associated to proteins [2]. As the dataset with massive size, is more likely subjected to the issues of class imbalance and high dimensionality [3]. Different methods have been propounded to address various aspects of high complexity with the problem area. Among them, some of the techniques have been employed for multiple feature extraction, encoding systems and other pre-processing procedures. From the researchers conducted in past, it is evident that the significant efforts have been put forth to develop statistical models or various methods involving learning strategies. Moreover, pre-processing and post processing strategies have also shown improvement which are considered to be prolific when it comes to accuracy of problem solutions. As a result of such procedures, it becomes more viable to uncover limitations, bottlenecks and strengths, and

subsequently can be very constructive in identifying preeminent components.

II. BACKGROUND STUDY

The protein secondary structure prediction can be achieved based on the information of primary structure of proteins and amino acid sequence. The prominent research objective among researchers in fields such as bioinformatics and biochemistry are to predict the 3D structures of proteins from its amino acid series. It was recommended that the prediction of 3D structures of a protein cannot be succeeded in its entirety from the protein sequence in general [4]. The research community as a whole has continuously from time to time made attempts to improvise methods for predicting indispensable aspects of protein structure. Moreover, in the field of secondary protein prediction, the researcher (Rost, 2003) has observed and examined that the prediction accuracy has surpassed the threshold of 70% in entire residues of a protein. This significant achievement was made by synthesising multiple sequence alignment information and application of various artificial intelligence algorithms. In addition, Rost (2003) reported that the prediction accuracy that is likely to be attained would be around 88% as upper limit for operational forecast.

A research study was conducted by investigators namely Kabsch and Sander (1983) in which a straightforward criterion for the development of secondary structures was set, and subsequently programmed to discover patterns of geometrical features and hydrogen bonds from x-ray coordinates [5]. This particular algorithm known as Define Secondary Structure of Proteins (DSSP) has been the benchmark procedure for assigning secondary structure to the amino acid sequence of proteins (primary structure).The defines secondary structure of proteins (DSSP) is of the capability that it can categorize eight states or various kinds of secondary structures based on the bonding patterns among hydrogen's. Furthermore, these eight categories are typically graded into three groups viz. helix (G,H,I), sheet/strand (E, B) and coil/loop (other remaining).

According to authors (Fadime *et al.,* 2008), the empirical methods employed in secondary protein structures insist the exploitation of sophisticated apparatus and time [6]. Consequently, several computational techniques have been deployed to predict the accurate position of secondary structure rudiments in proteins for generating insights into empirical reports. Moreover, the performance assessment task of protein secondary structure for prediction mechanism is regarded as a tedious task, for instance, the application of different training and testing data fetched from diverse datasets, on various algorithms makes it cumbersome to draw a comparison among methods [7]. Several efforts have been put forth in this direction to develop standards in test datasets so as to corroborate the performance of miscellaneous prediction techniques. Rostet. al., selected 126 proteins as a test dataset known as RS126 that currently amounts to comparative benchmark [8]. Barton and Cuff explicated the advancement of non redundant test dataset which comprised of 396 protein instances and residual length of 80, known as CB396, where each protein resemblance with other proteins was observed as not more than 25% of amino acid sequence. In addition, the CB396 dataset was examined by various prominent researchers, each using different methods to analyse the data including PHD [8], DSC [9], PREDATOR [10] and NNSSP [11] to predict secondary structures. The subsequent results achieved afterwards in the shape of Q_3 over the exploitation of CB396 set were PHD (71.9%), DSC (68.4%), PREDATOR (68.4%) and NNSSP (71.4%). The researchers in the same study, also made efforts to contrast the RS126 dataset results with the CB396 set, and different Q_3 scores were observed viz. PHD (73.5%), DSC (71.1%), PREDATOR (70.3%) and NNSSP (72.7%) respectively.

Apart from that, researchers have made various attempts to explore other methods such as support vector machines to predict the secondary structures of proteins. In this direction, Tai *et al.,* conducted a study using support vector machine technique on RS126 dataset where Q_3 value had significant accuracy of 78.8% [12]. Park and Kim (2003), investigated two datasets including RS126 and KP480 by means of support vector machine algorithm, and thereupon noteworthy results of Q_3 on both datasets with scores of 76.1% and 78.5% respectively, were consummated [13]. Rajapakse and Nguyen (2007) propounded a two phase multi class support vector machine on RS126 and CB368 datasets, and thereby generating position specific scoring matrices through PSI-Blast [14]. The following Q_3 scores of 78.0% and 76.3% were attained on RS126 and CB396 datasets respectively.

The past research studies have witnessed growth in various computational approaches applied across heterogeneous data sources related to protein prediction of secondary structures. However, the prediction accuracy has been rolling around 80%. Therefore, it becomes imperative for researchers to develop and explore novel data mining algorithms through which greater accuracy could be accomplished. Moreover, the integration of different machine algorithms has also demonstrated meagre attempts in various research studies linked to prediction of secondary structures.

III. EMPIRICAL RESULTS ILLUSTRATING SECONDARY STRUCTURES OF PRION PROTEINS

In this research study, we have employed the power of artificial neural networks through the application of different servers (Phyre^2 and PSIPRED) and research software (Matlab), to discover information and subsequently predict secondary structures from protein sequence. The dataset exploited for this research study has been obtained from NCBI, wherein human prion proteins have been explored under artificial neural networks using various mechanisms. The results of Phyre^2 related to protein sequences are broadly categorized into three main sections

comprising of secondary structure & disorder prediction, domain analysis, template information and alignment view which are demonstrated in different underlying figures.

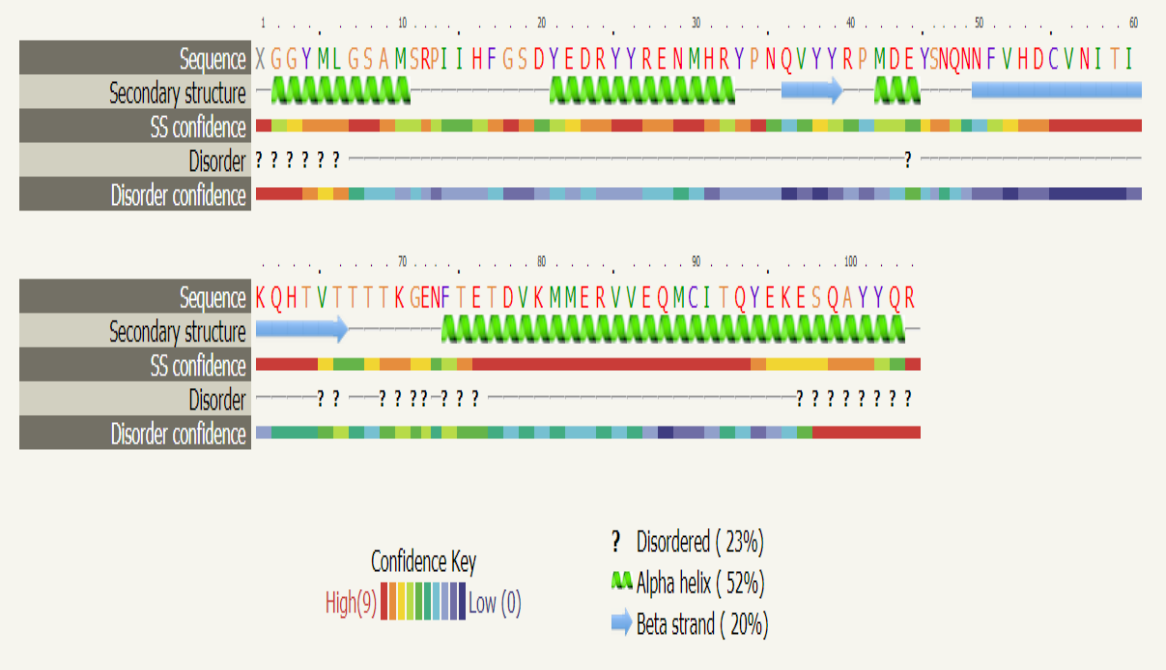

Fig.1 Secondary Structure and disorder prediction of the protein sequence

After submission of the protein sequence to an extremely massive database sequence using PSI-BLAST. The secondary structure prediction and the protein disorder prediction generated by the neural network through PSIPRED and DISOPRED respectively are highlighted in above figure 1. The presence of predicted alpha-helices, beta-sheets is graphically displayed collectively with the colour coded confidence bars, and disordered areas are represented by question marks.

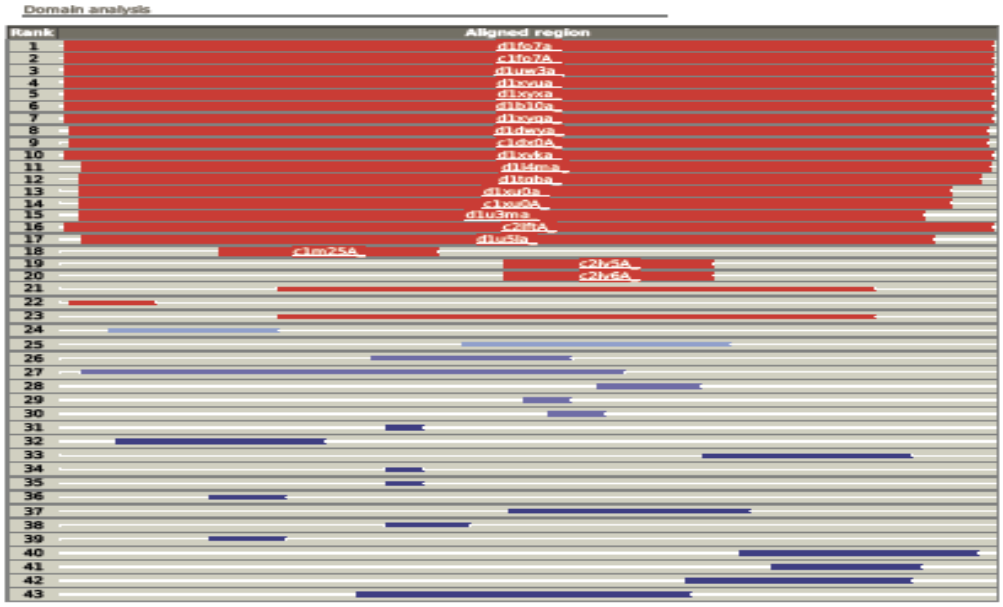

Fig.2 Output of domains of the protein sequence

From the domain analysis, it can be visualized that various proteins in which the user sequence is matching with multiple protein domains that have been colour coded by confidence bars and the same can be noticed in figure 2.The first 17 sequences among the remaining human prion proteins have shown a strong resemblance with the query sequence and rest of them have revealed limited similarity score.

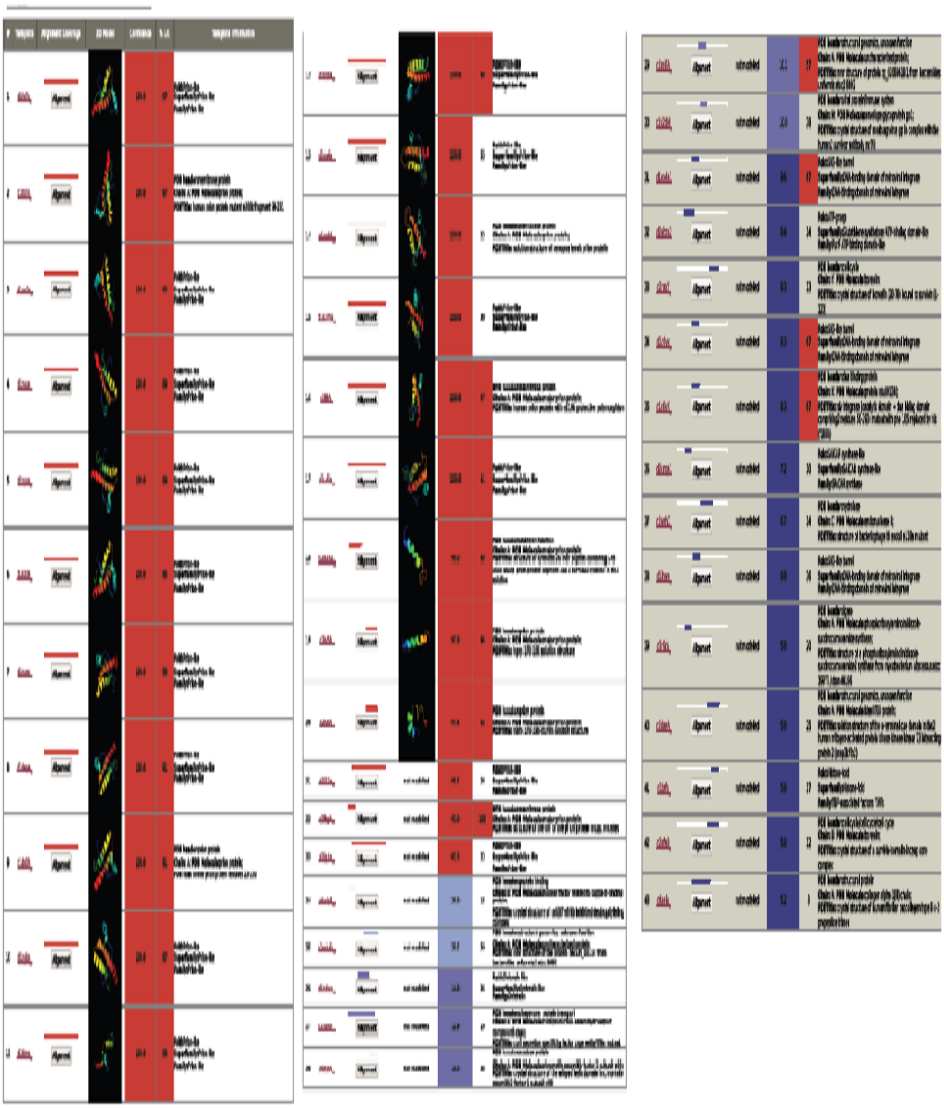

Fig. 3 Detailed information of various templates

The results produced in figure 3 cater with the following information including confidence estimates, predicted three dimensional models and template information of each model. The results have generated 20 3D models whose confidence is approximately more than 90, along with the indispensible information.

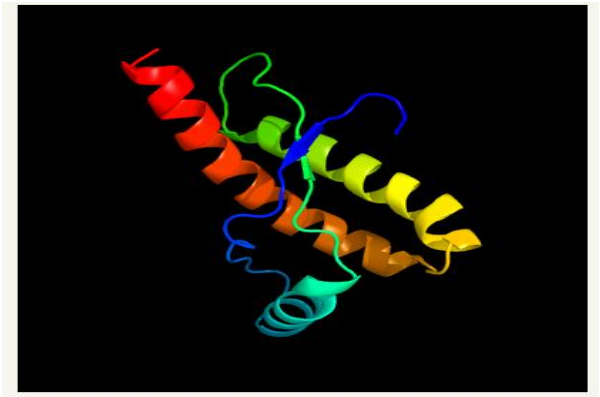

Fig. 4 Top model formulated among the entire set of model

The entire set of residues taken under investigation has exhibited an exceptional confidence level of 100% that have been modelled with the 100% confidence generated by the exclusive paramount scoring template shown in figure 4.

```
 Human_Prion_Protein                          bits  E-value  N 100.0% 2:104
1 UniRef50_P04156    Major prion protein n=533 T...  195    3e-49  1  99.0% 2:104 126:228
2 UniRef50_Q6GLY6    Major prion protein n=6 Tax...  135    3e-31  1  33.7%  4:98 114:212
3 UniRef50_P27177    Major prion protein homolog...  130    1e-29  1  37.6%  4:96 141:243
 consensus/100%
 consensus/90%
 consensus/80%
 consensus/70%
```

Fig. 5 Identities normalized by aligned length

As per the figure 5, the E-value of the first alignment is significant with an E-value of 3e-49 and with identity of 99%. Moreover, the major prion protein associated with the same alignment has got maximum score of 533 which can be noticeably visualized from the above figure. Although the e-values with the remaining alignments are significant, however the score values are extremely low.

Furthermore, we employed an exceedingly precise secondary structure prediction procedure viz. PSIPRED which consolidates two feed forward neural networks to examine the output attained from PSI-BLAST [15]. The PSIPRED has on an average demonstrated momentous Q3 value of 76.5% when its performance is evaluated based on stringent cross validation technique. In fact, this is the peak level prediction accuracy publicized for any procedure. In addition, the PSIPRED was investigated under extreme and average conditions with the submitted targets in which the method achieved a significant Q3 score of 73.4% and 77.3% respectively, over all the secondary structure predictions produced by other methods and was subsequently ranked at position first [16].

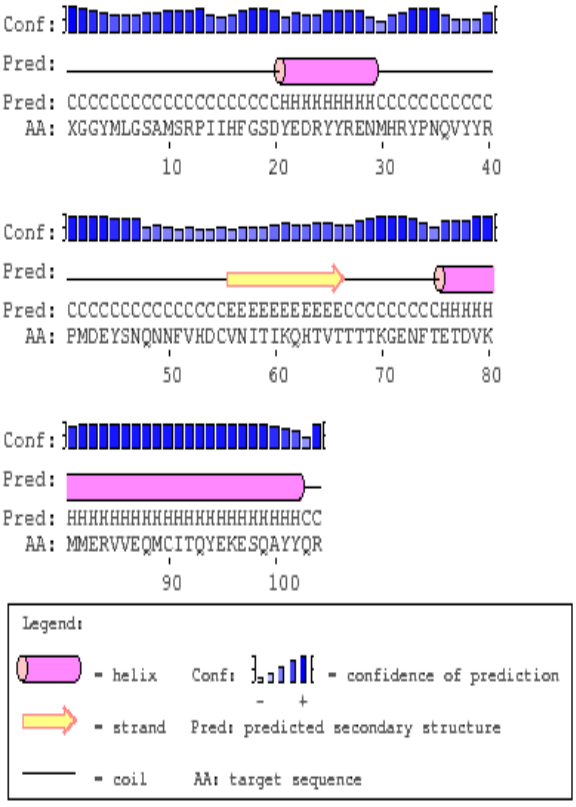

Fig. 6Results produced by PSIPRED

The figure 6 illustrates the graphical output of human prion proteins predictions generated by PSIPRED view which is a JAVA based visualization tool to construct two dimensional graphical depictions of PSIPRED predictions. As a corollary, it has become a quality graphical representation tool to cater the users with eminence publications.

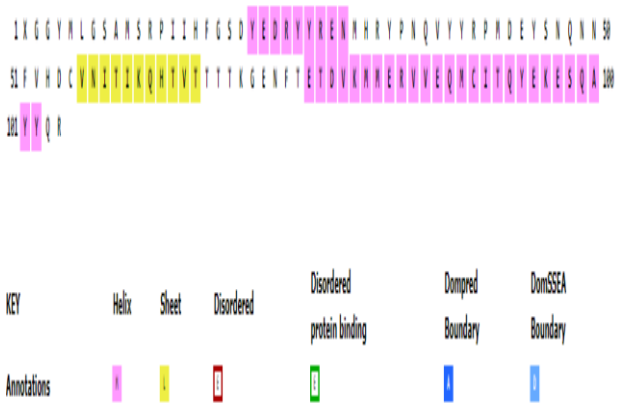

Fig. 7 Feature predictions

The above figure 7 exhibits the feature predictions which are colour coded according to the sequence feature key and have been explicated with various key annotations shown in the diagram.

IV. PERFORMANCE EVALUATION OF NEURAL NETWORKS USING MATLAB

In this case, the evaluation of the prediction method viz. neural networks was carried out on cross validation technique. The empirical results were conducted on software package which is more commonly known as MATLAB. Primarily, the human prion protein dataset was subjected to 10 fold cross validation technique for the process of training, validation and testing our sets. Moreover, we employed two feed forward neural networks with a sole hidden layer and using a window size of 10. The performance of the algorithm with such data has been measured with various widow sizes ranging from 5 to 15. Nevertheless, it was observed that the 10 residue window achieved a minimum mean absolute error of 12.3%. In addition, the prediction accuracy of neural networks evaluated on human prion protein dataset with heterogeneous sequences, demonstrated a significant accuracy of 71.73%.

V. CONCLUSION

In this research study, we have utilized two eminent tools viz. phyre2 and PSIPRED to predict the secondary structures, and as a consequence cater the researchers with straightforward and intuitive interface. With the application of these tools, secondary and tertiary structures were generated from the users protein submission query. The results demonstrated significant e-values and confidence after the query sequence was supplied. Moreover, at the later stage of our study where MATLAB was employed to further examine and predict the protein sequences. In this connection, the outcome of training and testing a two feed forward neural network algorithm with identical protein sequences for predicting protein secondary structures have also shown noteworthy results. The prediction accuracy in this case was achieved as 71.73% and minimum mean

absolute error of 12.3% was recorded. Furthermore, as a future course of our study, we would look forward to analyse and examine various amino acid sequences of human prion protein under more powerful and advanced techniques so as to explore and improve the prediction accuracy of protein secondary structures.

REFERENCES

[1] E. Buxbaum, "Fundamentals of Protein Structure and Function", *Springer*, ISBN: 978-0-387-26352-6, 2007.

[2] D. Ofer, and Y. Zhou, "Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training", *Proteins: Structure, Function, and Bioinformatics,* Vol. 66, No. 4, pp. 838-845, 2007.

[3] A. Rafał. "Dimensionality reduction of pssm matrix and its influence on secondary structure and relative solvent accessibility predictions", *World Academy of Science, Engineering And Technology,* Vol. 58, pp. 657-664, 2009.

[4] B. Rost, "Review: protein secondary structure prediction continues to rise", *J StructBiol*, Vol. 134, No. 2-3, pp. 204-18, 2001.

[5] W. Kabsh and C. Sander, "How good are predictions of protein secondary structure", *FEBS Letters*, Vol. 155, pp. 179-182, 1983.

[6] U. Y. Fadime, Y. O¨zlem, and T. Metin, "Prediction of secondary structures of proteinsnext term using a two-stage method", *Computers & Chemical Engineering,* Vol. 32, No. 1-2, pp. 78-88, 2008.

[7] J. A. Cuff, and G. J. Barton, "Application of multiple sequence alignment profiles to improve protein secondary structure prediction", *Proteins*, Vol. 40, No. 3, pp. 502-11, 2000.

[8] B. Rost and C. Sander, "Prediction of protein secondary structure at better than70% accuracy", *J MolBiol*, Vol. 232, pp. 584-599, 1993.

[9] R. D. King and M. J. E. Sternberg, "Identification and application of the concepts important for accurate and reliable protein secondary structure prediction", *Protein Sci*, Vol. 5, pp. 2298–2310, 1996.

[10] D. Frishman and P. Argos, "Seventy-five percent accuracy in protein secondary structure prediction", *Proteins*, Vol. 27, pp. 329–335, 1997.

[11] A. A. Salamov and V. V. Solovyev, "Prediction of protein secondary structure bycombining nearest-neighbor algorithms and multiple sequence alignments", *J MolBiol*, Vol. 247, pp. 11–15, 1995.

[12] H. Hu, Y. Pan, R. Harrison, and P. Tai, "Improved protein secondary structure prediction using support vector machine and a new encoding scheme and an advanced tertiary classifier", *IEEE Trans. NanoBiosci.* Vol. 3, pp. 265–271, 2004.

[13] H. Kim, and H. Park, "Protein secondary structure prediction based on an improved support vector machines approach", *Protein Eng.* Vol. 16, pp. 553-560, 2003.

[14] N. Nguyen, and J. C. Rajapakse, "Two stage support vector machines for proteinsecondary structure prediction", *Intl J Data Mining & Bioinformatics,* Vol. 1, pp. 248-269, 2007.

[15] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, (1997) , "Gapped BLAST and PSIBLAST: a new generation of Protein database search programs", *Nucl. Acids Res.* Vol. 25, pp. 3389-3402, 1997.

[16] C. A. Orengo, J.E, Bray, T. Hubbard, L. LoConte, and I. Sillitoe, "Analysis and assessment of ab initio three dimensional prediction, secondary structure, and contacts prediction", *PROTEINS Suppl.* Vol. 3, pp. 149-170, 1999.