Int. J. Adv. Res. Sci. Technol. Volume 2, Issue2, 2013, pp 98-105.

IJARST

International Journal of Advanced Research in Science and Technology

journal homepage: www.ijarst.com

ISSN 2319 – 1783 (Print)

ISSN 2320 – 1126 (Online)

Effective Prediction of Proteins Secondary Structure using Efficient Integrated Signal Processing and Neural Network Methods Induced by Physico-Chemical Parameters

Jayakishan Meher*

Department of Computer Science and Engg, Vikash College of Engineering for Women, Bargarh, Odisha, India.

****Corresponding Author's Email: jk_meher@yahoo.co.in***

A R T I C L E I N F O

Article history:

Received 05 Sept. 2013

Accepted 30 Sept. 2013

Available online 03 Oct. 2013

Keywords:

Digital signal processing,
protein secondary structure prediction,
protein folding,
α-Helix,
β-Strand,
Wavelet transforms.

A B S T R A C T

Protein structure analysis and prediction is a core area of research in bioinformatics. Prediction of protein secondary structure from amino acid sequences is one of the most important problems in molecular biology, because the structure of a protein is related to its function. Thus high prediction accuracy of protein structure from its sequence is highly desirable. Considerable research effort has been devoted to predicting the secondary structure of proteins from their amino acid sequences that typically have 76% approximate level of accuracy on an average. Thus, there is a considerable room for improvement. Recently digital signal processing (DSP) tools have been successfully applied in solving problems in the field of bioinformatics. In this paper we have proposed an effective feature extraction method based on discrete wavelet transform (DWT) to detect informative proteins and radial basis function neural network (RBFNN) classifier is used to efficiently predict the sample class which has a low complexity than other classifier in which effective numerical representation based on physico-chemical parameters induces the prediction more accurately. The potential of the proposed approach is evaluated through an exhaustive study by benchmark non-redundant dataset and a prediction accuracy of 93% is achieved.

Introduction:

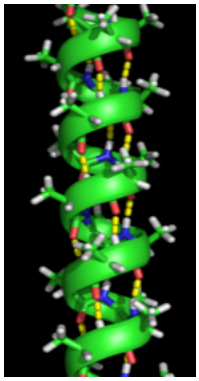
Determining the protein structure from amino acid sequence leads to better understanding of the functionality of the protein resulting in faster drug discovery. Proteins are fundamental components of all living cells, performing a variety of biological tasks. Each protein has a particular structure that determines its function. Protein structure is more conserved than protein sequence, and more closely related to function. Proteins are macromolecules that are responsible for a wide range of vital biochemical functions, which include acting as catalysts, oxygen transport, cell signaling, antibody production, nutrient transport and building up muscle fibers [1-2]. More specifically, proteins are chains of amino acids, of which there are twenty different types, joined by peptide bonds.

Proteins have a three-level structural hierarchy, typically referred to as primary, secondary and tertiary structure [3]. The higher-level structures determine the function of the protein and consequently, the knowledge of the structure provides insight into its function. Proteins are large polypeptides which consist of 20 amino acid residues. Chemical properties that distinguish these amino acids cause the protein chains to fold up into specific structures that define their particular functions in the cell.

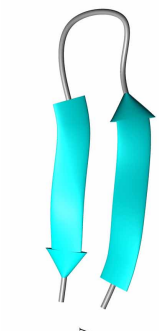
The shape of a protein is specified by its amino acid sequence. There are four levels of protein structure [4]. The primary structure refers simply to the linear sequence of amino acids. The primary structure of a protein consists of amino acids linked by peptide bonds to form polypeptide chains. The code for the primary structure is in DNA. The secondary structure is the locally ordered

structure created by hydrogen bonding within the protein backbone. The amino acids in a polypeptide chain form hydrogen bonds between the N-H and C=O groups. The chain twists around on itself and forms a three-dimensional structure. Most common folding patterns are the α-helix and the β-sheet [5]. Tertiary structure refers to the global folding of a single polypeptide chain, and quaternary structure involves the association of two or more polypeptide chains into a multisubunit structure. The final structure of a protein is the one in which the free energy is minimized. Hydrophobic amino acid chains are buried on the inside of a protein and hydrophilic amino acid chains gather on the outside. Sulphur bridges stabilize the structure. Prediction of the secondary structure is important as it provides insights into the function of the protein. By jointly comparing amino acid and secondary structure sequences, it is possible to improve the prediction of protein function [6]. In addition, secondary structure prediction is a step towards the prediction of the 3-D structure of a protein.

The fundamental elements of the secondary structure of proteins are α-helices, β-sheets, coils and turns. Thus, the secondary structure prediction can be analyzed as typical pattern recognition or classification problem, where the secondary structure class of a given amino acid residue in a protein is predicted based on its sequence features. *α*-helices are strengthened by hydrogen bonds between every fourth amino acid so that the protein backbone adopts a helical configuration as shown in Figure 1. Likewise in loops (e.g., turns or bends), the hydrogen bonding is mostly local. For example, the turn segment has a hydrogen bond between the first and the fourth amino acids. The hydrogen bonding structure in *β*-strands is slightly different, where both local and nonlocal interactions are observed. In *β*-strands, the most common local hydrogen bonding is between every two amino acids, and nonlocal interactions are due to hydrogen bonds between amino acid pairs positioned in interacting *β*- strand segments [7].



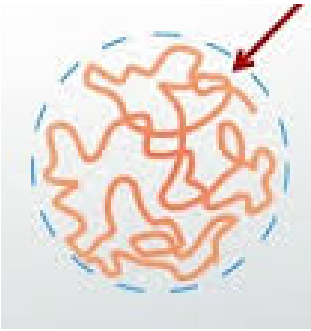
(a) α-helix



(b) β-strand

(c) Coil

(d) Beta-alpha-beta unit



(e) Random (Unstructured peptide chain)

Figure: 1.Secondary structure organization in proteins

Various statistical, machine learning and signal processing algorithms have been used to predicting the secondary structure of proteins from their amino acid sequences. A simple goal in the secondary structure prediction is to predict whether an amino acid residue of a protein is in a helix, strand or coil [8]. The first generation of secondary structure prediction techniques emerged in the 1960s and were based on single amino acid propensities and, for each amino acid, calculated the probability of it belonging each of the secondary structural elements. The secondary generation of prediction methods extended this concept by taking into account the local environment, of an amino acid, into consideration. Prediction accuracies with the second generation methods seemed to stall at around 60% accuracy, seemingly because these methods were local in that only information in a window of adjacent residues were used in predicting the secondary structure of an amino acid. [9] Local information accounts for approximately 65% of secondary structure information [10]. Since the early 1990s, third generation prediction methods achieved prediction accuracies around 70% and such methods incorporate machine learning techniques, evolutionary knowledge about proteins and with relatively more complex algorithms. [10-11].

Homology modeling bases the prediction for an unknown target protein, on the known secondary structures of proteins of similar amino acid sequence [12]. The basis of threading is that a limited number of unique protein folds exist in nature and structure prediction of a target sequence can be performed by consulting a database of known folds and determining which fold-model best fits the sequence. Both homology modeling and threading rely on the existence of known structures and the disadvantage of such approaches is that accurate prediction relies on proteins of similar structure already being solved. Another approach, namely the ab initio techniques [13] or prediction from first principles, bases structure prediction on known biochemical and biophysical facts related to the proteins. In general they are computationally very expensive methods. Machine learning methods such as neural network and nearest neighbor techniques, utilize a localized prediction methodology in the sense that a window, typically of less than 20 amino acids, is presented to the prediction system with the aim of predicting secondary structure. However, local information accounts for approximately 65% of secondary structure formation [8]. Therefore, prediction can potentially be improved by incorporating a more global prediction scheme [9]. Secondary structure prediction methods often employ neural networks (NNs) [14], SVMs [15], and hidden Markov models (HMMs) [16, 17]. In neural networks and SVMs utilize an encoding scheme to represent the amino acid residues by numerical vectors. On the other hand, in HMM methods, hidden states generate segments of amino acids that correspond to the non-overlapping secondary structure segments. There are two types of protein secondary structure prediction algorithms. A single sequence algorithm does not use information about other similar proteins. The algorithm should be suitable for a nonhomologous sequence with no sequence similarity to any other protein sequence. Algorithms of another type explicitly use sequences of homologous proteins, which often have similar structures. The accuracy (sensitivity) of the best current single sequence prediction methods is below 70%. The prediction accuracy of the best prediction methods that employ information from multiple alignments is close to 82.0% [18].

The genomic and proteomic information are digital in nature and thus makes it suitable for the application of signal processing techniques to better analyze and understand the characteristics of DNA, proteins, and their interaction. Therefore, signal processing offers a variety of methods from pattern recognition and network analysis for the diagnosis and therapy of genetic diseases [19-20]. It is possible to map protein into a digital signal by assigning numeric values to each amino acid. DSP techniques relating to protein structure analysis assign numeric values - often their hydrophobocity values [21-24], to the amino acids, and analyze the resulting sequence via Fourier analysis, wavelet processing or some other DSP techniques. Helix kink prediction has been made using DSP tools using polarizability property as feature vector [25]. A methodology that is primarily targeted for any given query protein rather being trained over a pre-determined training set is used by D. Mitra and M. Smith based on homology-modeling to improve the accuracy [26]. For some query proteins our prediction accuracies are predictably higher than most other methods, while for other proteins they may not be so, but we would at least know that even before running the algorithms. When a significantly homologous protein with known structure is available in the database the prediction accuracy could be even 90% or above. This uses digital signal processing technique that is of global nature in assigning structural elements to the respective residues. An automated approach for the secondary structure prediction based on the Digital Signal Processing (DSP) techniques which involve two DSP operators, Convolution and Deconvolution are used by D. Mitra and M. Smith for the purpose of predicting secondary structures [26]. Mappings between an amino acid sequences and the corresponding numerical time-series or "signals" are processed. Convolution is a method of applying a filter on an incoming signal, producing an outgoing signal. Deconvolution is the inverse operation of convolution and permits the filter to be recovered if the outgoing signal and the incoming signal are known. This method predicts three states (helix, strand, and coil) for the secondary structure.

Secondary structure of a protein can be predicted efficiently by combining continuous wavelet transform (CWT) and Chou-Fasman method. The authors in [27] have selected a protein with ID 1gca from PDB database, and substituted every amino acid of the protein with corresponding hydrophobic value. Then CWT was used to get the nucleated residues of certain type of secondary structure. The regions were extended along the protein sequence in each direction based on Chou-Fasman rules. The prediction accuracy for alpha-helix, beta-sheet, loop is 80% on an average. Support vector machines (SVM) have shown strong generalization ability in a number of application areas, including protein structure prediction. In [28] a new tertiary classifier is introduced that makes use of support vector machines as neurons in a neural network architecture. This network is optimized using genetic algorithms. The novel tertiary classifier is better than most available techniques. In [29] Golem takes, as input, examples and background knowledge described as Prolog facts. It produces, as output, Prolog rules which are a generalization of the examples. Golem was applied to learning secondary structure prediction rules for alpha domain type proteins Golem learned a small set of rules predicting which residues is part of α-helices based on

their positional relationships and chemical and physical properties. This representation is more easily understood by molecular biologists. Performance of the learned rules was 81%.

The rest of the paper is organized as follows. Section 2 describes materials and methods for protein secondary structure of proteins. This section includes dataset preparation and then how to represent the amino acid sequence in numerical sequence. The section 3 describes the proposed method of prediction of secondary structure of proteins. This section includes the wavelet transform based feature extraction and followed by a radial basis function neural network classifier to predict the major classes of secondary structure of proteins.

Materials and Methods:

A. Dataset Preparation:

Co-ordinate files of proteins are obtained from the protein data bank. For the study in this work we have chosen a non-redundant set of PDB files. From the non-redundant PDB chain set and their coordinate files obtained from the public domain are used for preparing a database of elements of protein secondary structure. Using MAPMAK [30] program the major classes of secondary structure of protein such as α-Helix, β, Turn and Random are found out.

B. Numerical Representation of Amino acid Sequence:

The amino acid sequence is a string consisting of 20 residues represented by 20 distinct alphabets. Proteomic signal processing deals with numerical sequence. Hence the protein sequence can be represented in numerical sequence. There are various numerical representations, but the methods having high sensitivity and specificity is obtained by using physico-chemical properties of amino acids. It is seen that physico-chemical properties of amino acid residues such as polarizability, dipole moment and alpha are correlated with the structure and function of the proteins and hence help in the classification of major classes of protein secondary structure of protein effectively. If we substitute the dipole moments for amino acids, we get a numerical sequence which represents the distribution of polarity of a chemical bond within a molecule along the protein sequence. Now the resulting numerical representation is subjected for analysis. The amino acid sequence of each character is converted into numerical sequence by substituting its physico-chemical parameter of residues (Table 1) from a known dataset.

Table: 1. Physico-chemical properties of amino acids

Amino acid	Polarizability	Dipole moment	Alpha
A	4.44	5.937	1.489
R	14.16	37.5	1.0224
N	7.72	18.89	0.772
D	6.55	29.49	0.924
C	7.44	10.74	0.996
Q	11.39	39.89	1.164
E	8.38	42.52	1.504
G	2.61	0.0	0.510
H	11.84	20.44	1.003
L	9.95	3.782	1.236
I	9.95	3.371	1.003
K	10.72	50.02	1.172
M	11.11	8.589	1.363
F	14.10	5.98	1.195
P	8.69	7.916	0.492
S	5.08	9.836	0.739
T	6.92	9.304	0.785
W	19.37	10.73	1.090
Y	14.74	10.41	0.787
V	8.11	2.692	0.990

Proposed method of Prediction:

A. Wavelet based feature extraction method:

Protein data is very rich and complex. Such data can be analyzed with wavelet transform to extract the important features. This transformation method is used for feature extraction for a machine learning approach to the protein secondary structure prediction problem. Recently, the use of wavelet transform in the Bioinformatics field is promising. A wavelet is a waveform that is localised in both time and frequency domains. This wavelet is dilated and translated along the signal to perform the analyses. The commonly used wavelets in practice are Haar, Daubechies, Gaussian wave, Mexican hat and Morlet wavelets. The selection of particular wavelet for any analysis depends on the kind of signal being studied and kind of signal variation to be captured.

An important attribute of wavelet methods is that, due to the limited duration of every wavelet, local variations of the signal are better extracted and information on the location of these local features is retained in the constituent waveforms. In discrete wavelet transform a subset of scales and positions are chosen, in which the correlation between the signal and the shifted

and dilated waveforms are calculated. Consequently, the signal is decomposed into several groups of coefficients, each containing signal features corresponding to a group of frequencies. Small scales refer to compressed wavelets, depicted by rapid variations appropriate for extracting high frequency features of the signal.

Wavelet transform has been applied for transmembrane structure prediction [31]. In this work, the wavelet transform is used to determine kink in segments of amino acid sequences of α-helical membrane proteins. Protein sequence similarity has also been studied using DWT of a signal associated with the average energy states of all valence electrons of each amino acid [32]. DWT has been applied on hydrophobicity signals in order to predict hydrophobic cores in proteins [33].

Wavelet transform proposed by Grossman and Morlet [34] is an efficient time-frequency representation method which transforms a signal in time domain to a time-frequency domain. The basic idea is that any signal can be decomposed into a series of dilations and compressions of a mother wavelet ($\Psi(t)$). Hence the continuous wavelet transform of a signal is defined as:

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi \left(\frac{t-b}{a} \right) dt \qquad (1)$$

where $$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left(\frac{t-b}{a} \right), a \in R^{+}, b \in R$$

The resolution of the signal depends on the scaling parameter 'a' and the translation parameter 'b' determines the localization of the wavelet in time. The CWT can be realized in discrete form through the discrete wavelet transform .The DWT is capable of extracting the local features by separating the components of the signal in both time and scale. In the protein data the amino acid sequence is considered as a signal which can be represented as a sum of wavelets at different time shifts and scales using the DWT.

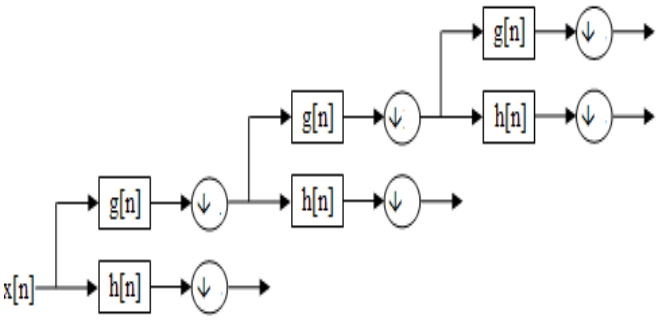


Fig: 2. Wavelet decomposition

The wavelets can be realized by iteration of filters with rescaling which was developed by Mallat [35] through wavelet filter banks. The resolution of the signal, which is a measure of the amount of detail information in the signal, is determined by the filtering operations, and the scale is determined by up sampling and down sampling operations. The approximation coefficients obtained by the decomposition at a particular level is used as the features for further study. The discriminate feature set has been obtained from discrete wavelet transform into level 2 using db7 wavelet to get the approximation coefficients as the extracted feature set. The Wavelet decomposition is shown in figure 2.

B. Radial basis function neural network classifier:

The main goal of a secondary structure prediction algorithm needs a classifier having a feature set that is comprehensive enough to capture the essential correlations from the available data. Wavelet transform method is used for feature extraction for a machine learning approach to the protein secondary structure prediction problem. For function approximation and pattern classification problems the radial basis function network (RBFNN) has been used in this paper which is a neural structure because of their simple topological structure and their ability to learn in an explicit manner.

The radial basis function neural network is simple in structure. In the RBF network, there is an input layer, a hidden layer consisting of nonlinear node function, an output layer and a set of weights to connect the hidden layer and output layer. Due to its simple structure it reduces the computational task as compared to conventional multi layer perception (MLP) network. In RBFNN, the basis functions are usually chosen as Gaussian and the number of hidden units are fixed using some properties of input data. The structure of a RBF network is shown in Fig. 3.

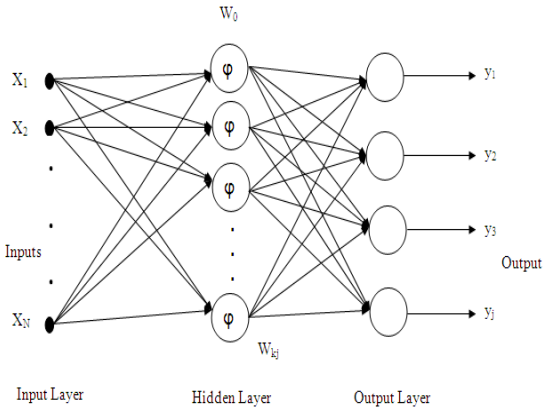


Fig: 3. The RBFNN based classifier

For an input feature vector x, the output of the jth output node is given as.

$$y_j = \sum_{k=1}^{N} w_{kj}\varphi_k = \sum_{k=1}^{N} w_{kj} e^{-\frac{\|x(n)-C_k\|}{2\sigma_k^2}} \quad (2)$$

The error occurs in the learning process is reduced by updating the three parameters, the positions of centers (C_k), the width of the Gaussian function (σ_k) and the connecting weights (w) of RBFNN by a stochastic gradient approach as defined below:

$$w(n+1) = w(n) - \mu_w \frac{\partial}{\partial w} J(n) \quad (3)$$

$$C_k(n+1) = C_k(n) - \mu_c \frac{\partial}{\partial C_k} J(n) \quad (4)$$

$$\sigma_k(n+1) = \sigma_k(n) - \mu_\sigma \frac{\partial}{\partial \sigma_k} J(n) \quad (5)$$

Where $J(n) = \frac{1}{2}|e(n)|^2$, e (n) =d (n) - y (n) is the error, d (k) is the target output and y (k) is the predicted output. μ_w μ_C And μ_σ are the learning parameters of the RBF network. The complete process of the proposed feature extraction based protein secondary structure prediction process is presented in Fig. 4.

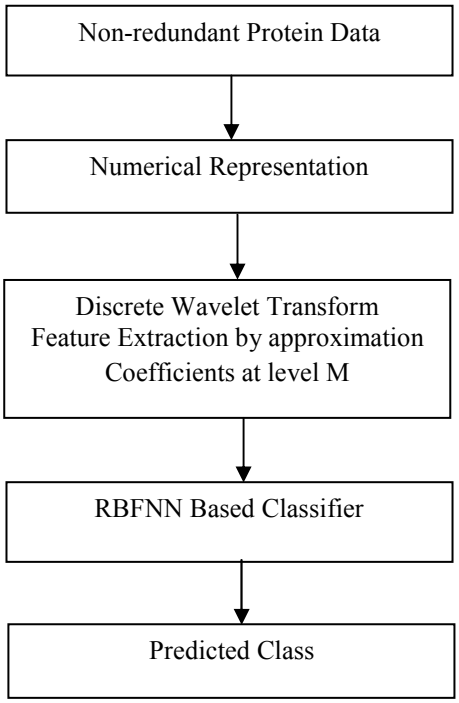


Fig: 4. Flow graph of the proposed feature extraction based protein secondary structure prediction method

Results and Discussion:

All the datasets categorized into multi class to assess the performance of the proposed method. The feature selection process proposed in this paper has two steps. First the protein data is decomposed and optimally choose the discriminate feature set then using discrete wavelet transform into level 2 using db7 wavelet to get the approximation coefficients as the extracted feature set. The performance of the proposed feature extraction method is analyzed with the well studied neural network classifiers such as RBFNN. The leave one out cross validation (LOOCV) test is conducted by combining all the training and test samples for the classifiers with the dataset.

In order to compare the efficiency of the proposed method in predicting the class of the protein structure we have used standard non-redundant datasets. All the datasets are categorized into four classes such as α-Helix, β, Turn and Random to assess the performance of the proposed method. 100 sequences from each class of protein dataset are taken as training set. The feature selection process proposed in this paper includes polarizability, dipole moment and alpha as shown in the Table 1. To implement the RBFNN classifier, we first read the file of protein sequence which is represented with numerical values. The performance of the proposed feature extraction method is analyzed with the neural network classifiers called as radial basis function neural network (RBFNN). The leave one out cross validation (LOOCV) test was conducted by combining all the training and test samples for the classifiers with datasets [16]. LOOCV is a technique where the classifier is successively learned on n-1 samples and tested on the remaining one. i.e., it removes one sample at a time for testing and takes other as training set. It involves leaving out all possible subsets so the entire process is run as many times as there are samples. This is repeated n times so that every sample was left out once. Repeating these procedure n times gives us n classifiers in the end. Our error score is the number of mispredictions. Out of 400 sequences from of protein dataset, 378 samples are detected as true positive whereas 22 samples are detected as false negative.

The prediction accuracy has been analyzed in terms of three measuring parameters such as accuracy (A), precision (P) and recall (R). These are defined in terms of four parameters true positive (t_p), false positive (f_p), true negative (t_n) and false negative (f_n). t_p denotes the number of protein secondary structure and are also predicted as protein secondary structure, f_p denotes the number of actually Non protein secondary structure but are predicted to be protein secondary structure, t_n is the number of actually Non protein secondary structure and

also predicted to be Non protein secondary structure, and f_n is the number of actually protein secondary structure and predicted to be Non-protein secondary structure.

A. *Accuracy*

The accuracy of prediction of protein secondary structure in amino acid sequence is defined as the percentage of protein secondary structure correctly predicted of the total sample sequences present. It is computed as follows:

$$A = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \quad (6)$$

B. *Precision*

Precision is defined as the percentage of protein secondary structure correctly predicted to be one class of the total protein secondary structure predicted to be of that class. It is computed as:

$$P = \frac{t_p}{t_p + f_p} \quad (7)$$

C. *Recall*

Recall is defined as the percentage of the protein secondary structure that belongs to a class that is predicted to be that class. Recall is computed as:

$$R = \frac{t_p}{t_p + f_n} \quad (8)$$

Table: 2. Measuring parameters for prediction accuracy

Actual → Predicted ↓	Secondary Structure (ST)	Non Secondary Structure (NST)
ST	378 (t_p)	29 (f_p)
NST	22 (f_n)	271(t_n)

The accuracy, precision and recall are 0.93, 0.92, and 0.94 respectively. The accuracy of sequence based classifiers reported so far is about 76%. Hence the present classifier appears to have high accuracy compared to existing sequence based classifiers. It needs to be extended for all the protein dataset found in biosystems before it can be used at proteomic level.

Conclusions

In this paper a feature extraction method using the wavelet transform has been used to effectively select the discriminative proteins on dataset. A simple RBFNN based classifier has been used to classify the major protein classes such as α-Helix, β, Turn and Random efficiently. Digital signal processing plays an important role in prediction of secondary structure of proteins. Again the physico-chemical properties such as polarizability, dipole moment and alpha have been used for numerical representation that is correlated with the protein sequences and thus induces better classification. The simulation results elucidated that the proposed approach is a better predictor with less computational complexity. Above all it fulfills the need of a classifier for protein secondary structure keeping in view the growing database of protein secondary structure along with the escalating interest of scientific community in the field during last decade.

Acknowledgement:

The author would like to thank the Management members and Principal for establishment of R & D Centre in the College and providing the required infrastructure and other supports to carry out the research work.

References

[1] Sitbon, E.; Pietrokovski, S. Occurrence of protein structure elements in conserved sequence regions BMC Struct. Biol.,2007, 7, 1-15.
[2] Chothia, C. Proteins. One thousand families for the molecular
biologist. Nature, 1992, 357, 543-544.
[3] Alexandrov, N., Solovyev, V. Effect of secondary structure prediction on protein fold recognition and database search. Genome Informatics 7, 119-127, 1996
[4] Brandon C., Tooze J., Introduction to Protein Structure. Garland Publishing. New York, 1991
[5] Rost B., Protein Structure Prediction in 1D, 2D, and 3D. The Encyclopedia of Computational Chemistry (eds. PvRSchleyer, NL Allinger, T Clark, J Gasteiger, PA Kollman, HF Schaefer III and PR Schreiner), 3, 1998, 2242-2255
[6] Anfinsen, C. B. Principles that govern the folding of protein chains. Science. 181, 223-230, 1973.
[7] Chou, P., Fasman G., Prediction of the secondary structure of proteins from their amino acid sequence. Advanced Enzymology, 47, 45-148, 1978.
[8] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, "Exploiting the past and the future in protein secondary structure prediction," Bioinformatics, vol. 15, no. 11, pp. 937–946, 1999.
[9] Rost, B., Review: Protein Secondary Structure Prediction Continues to Rise. Journal of Structural Biology, 134, 204-218,2001.

[10] Murzin, A.G.; Brenner, S.E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J.Mol. Biol.,1995, 247, 536-540.

[11] Pollastri, G., Przybylski, D., Rost, B., Baldi, P., Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks. Protein: Structure, Function and Genetics. 47:228-235, 2002.

[12] Abagyan, R., Batalov S., Cardozo,T., Totrov, M., Webber, J., Zhou, Y. Homology Modeling With Internal Coordinate Mechanics: Deformation Zone Mapping and Improvements of Models via Conformational Search. PROTEINS: Structure, Function and Genetics. 1:29-37,1997.

[13] Xia, Y., Huang, E., Levitt, M., Samudrala, R. 2000. Ab Initio Construction of Protein Tertiary Structures Using a Hierarchical Approach. Journal of Molecular Biology 300: 171-185,2000.

[14] Riis, S. K. and Krogh, A. 1996. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. J. Comp. Biol., 3: 163-183.

[15] J. Guo, H. Chen, Z. Sun, and Y. Lin, "A novel method for protein secondary structure prediction using dual-layer SVM and profiles," *Proteins*, vol. 54, no. 4, pp. 738–743, 2004.

[16] S.C. Schmidler, J.S. Liu, and D.L. Brutlag, "Bayesian segmentation of protein secondary structure," *J. Comp. Biol.*, vol. 7, no. 1/2, pp. 233–248, 2000.

[17] Z. Aydin, Y. Altunbasak, and M. Borodovsky, "Protein secondary structure prediction with semi Markov HMMs," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing 2004 (ICASSP'04), 2004, vol. 5, pp. 577–580.

[18] B. Rost, "Rising accuracy of protein secondary structure prediction," in *Protein Structure Determination, Analysis, and Modeling for Drug Discovery*, D Chasman, Ed. New York: Marcel Dekker, 2003, pp. 207–249.

[19] J. Chen, H. Li, K. Sun, and B. Kim, "How will bioinformatics impact signal processing research," *IEEE Signal Processing Mag.*, vol. 20, no. 6, pp. 16–26, 2003.

[20] E.R. Dougherty and A. Datta, "Genomics signal processing: Diagnosis and therapy," *IEEE Signal Processing Mag.*, vol. 22, no. 1, pp. 107–112, 2005.

[21] Hirakawa, H., Kuhara, S., 1997. Prediction of Hydrophobic Cores of Proteins Using Wavelet Analysis. Genome Informatics, 8, 61-70

[22] Irback, A., Sandelin, E., 2000 On Hydrophobicity Correlations in Protein Chains. Biophysical Journal, 79, 2252-2258

[23] Irback, A., Peterson, C., Potthast, F., 1996. Evidence for nonrandom hydrophobicity structures in protein chains. Proc. Natl. Acad. Sci., 93, September, 9533-9538

[24] Kyte, J., Doolittle, R., 1982. A Simple Method for Displaying the Hydropathic Character of a Protein. Journal of Molecular Biology, 157, 105-132

[25] J.K.Meher, N.Mishra, P.K.Mohapatra, M.K.Raval, P.K.Meher and G.N.Dash. "Signal Processing Approach for Prediction Kink in Transmembrane α-Helices", *Springer CCIS, ISBN 978-3-642-20572-9* (*AIM-2011*), pp. 170-177, April-2011,

[26] Debasis Mitra and Michael Smith, Digital Signal Processing in Predicting Secondary Structures of Proteins, Innovation in applied artificial intelligence, Vol 3029/2004, 40-49, DOI: 10.1007/978-3-540-24677-0_5

[27] Chen, Hang, Predicting protein secondary structure using continuous wavelet transform and Chou-Fasman method, Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference , 2603 – 2606

[28] Zhang, Yan-Qing, Protein Secondary Structure Prediction Using Genetic Neural Support Vector Machines, Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference, 1355 – 1359

[29] Muggleton, S., Using logic for protein structure prediction, System Sciences, 1992. Proceedings of the Twenty-Fifth Hawaii International Conference 1992,Volume1, Page(s): 685 -696

[30] Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configuration, *J. Mol. Biol.* 7, 95-99.)

[31] Murray, K.B., Gorse, D., Thornton J.: Wavelet Transforms for the Characterization and Detection of Repeating Motifs. J. Mol. Biol. 316, 341--363 (2002)

[32] de Trad, C., Fang, Q., Cosic, I.: Protein Sequence Comparison Based on the Wavelet Transform Approach. Protein Eng. 15, 193--203 (2002)

[33] Hirakawa, H., Muta, S., Kuhara, S.: The Hydrophobic Cores of Proteins Predicted by Wavelet Analysis. Bioinformatics 15,141--148(1999)

[34] Grossmann A. and Morlet J. Decomposition of Hardy functions into square integrable wavelets of constant shape. SIAM Journal on Mathematical Analysis, 1984, vol. 15, no. 4, pp.723–736.

[35] Mallat S. G. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, vol. 11, no. 7, pp. 674–693.