Int. J. Adv. Res. Sci. Technol. Volume 1, Issue1, Oct-2012, pp 86-89.

IJARST

International Journal of Advanced Research in Science and Technology

journal homepage: www.ijarst.com

ISSN 2319 – 1783

Prediction of secondary Structure of Proteins using Signal Processing Methods

J.K.Meher[1*], M.K.Raval[2], G.N.Dash[3], P.Mishra[4]

[1]*Department of Computer Science and Engg, Vikash College of Engg for Women, Odisha, India*
[2]*Department of Chemistry, Gangadhar Meher College, Sambalpur, Odisha, India*
[3]*School of Physics, Sambalpur University, Burla, Odisha, India*
[4]*Department of Electronics and Communication Engg, Vikash College of Engg for Women, Odisha, India*

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

ARTICLE INFO

Article history:
Received 26 Sept. 2012
Accepted 30 Sept. 2012
Available online 01 October 2012

Keywords:
Digital Signal Processing,
Secondary structure
Protein folding
α-Helix,
β-Strand,
Coils.

ABSTRACT

The prediction of protein folding is important because the structure of a protein is related to its function. The study of protein structure therefore produces valuable practical benefits for medicine, agriculture and industry. The understanding of enzyme function allows the design of drugs which inhibit specific enzyme targets for therapeutic purposes. Structural information can provide insight into protein function, and therefore, high- accuracy prediction 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. Present methods of prediction based on the statistical methods and machine learning methods typically have 76% approximate level of accuracy on an average. Thus, there is a considerable room for improvement. Digital Signal Processing (DSP) is an Engineering discipline concerning the creation, manipulation and analysis of digital signals. New approach for the secondary structure prediction based on the DSP techniques can take major role for fast and accurate result. Unknown secondary structure of a targetprotein can be predicted by using the appropriate digital signal processing tools for a baseprotein of a significant amino acid sequence-similarity and whose secondary structure is known. In this study we present an extensive review of existing methods of secondary structure prediction of proteins.

Introduction:

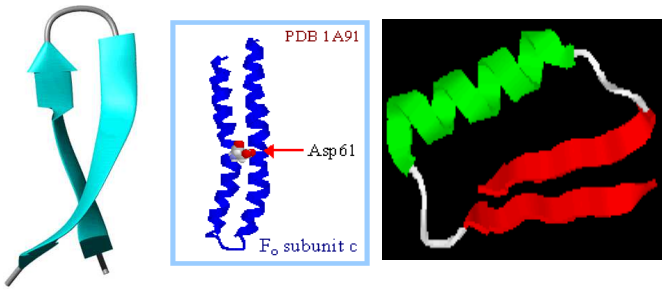
Proteins are fundamental components of all living cells, performing a variety of biological tasks. Each protein has a particular structure that determines its function. Protein structure is more conserved than protein sequence, and more closely related to function. Proteins are macromolecules that are 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-tiered structural hierarchy, typically referred to as primary, secondary and tertiary structure [3]. Being able to determine the structures of proteins is of tremendous value to the biological community. This is because 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 amino acid residues. Chemical properties that distinguish the 20 standard 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. The prediction of the 3-D structure of proteins is far in the future, because proteins are generally self-folding.

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 toward the prediction of the 3-D structure of a protein. For instance, secondary structure predictions can be included in fold recognition methods, in which a target amino acid sequence with unknown structure is compared against a library of structural templates (folds) and the best scoring fold is assumed to be the one adopted by the sequence.

The three major secondary structure states are the α-helix {H}, the β- strand {E}, and the turn, coil or loop {L} [7]. 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 in Figure 1(c) 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. A β-sheet is a set of such segments, in which the interacting segment pairs adopt either a parallel or an antiparallel conformation.



(a) **(b)** **(c)**

Fig: 1. Secondary structure organization in proteins: (a) Beta hairpin, (b) Helix hairpin and (c) Beta-alpha-beta unit.

Protein Secondary Structure Prediction:

Considerable research effort has been devoted 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].

Despite the existence of varying techniques, there are broadly three main approaches to structure prediction. 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. The methodology of such an approach is not to predict the structure from a primary sequence, but rather to fit a known structural-model to a sequence. Typically, steps are taken to align the target sequence to a known set of folds and a scoring function is employed to determine the best fitting structure. 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. The third 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. However, progress has been relatively slow as the physical processes by which a protein folds are not completely understood. [10]. In general they are also computationally very expensive methods.

Various secondary structure prediction methods, particularly some 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 of the central element, using only the information gleaned from the amino acids within the window. 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. It must be mentioned that this ideology has been documented and various prediction methods are adapting a more global view of structure prediction [9]. Secondary structure prediction methods often employ neural networks (NNs) [14], SVMs [15], and hidden Markov models (HMMs) [16], [17]. In HMM methods, hidden states generate segments of amino acids that correspond to the nonoverlapping secondary structure segments, and the goal is to find the most likely hidden-state sequence representation under the probabilistic model defined by the HMM. On the other hand, neural networks and SVMs utilize an encoding scheme to represent the amino acid residues by numerical vectors.

To convert the amino acids into vectors, the amino acid sequence is partitioned into overlapping segments by a sliding window of size n (typically between 13 and 17). Then each segment is represented (as a vector) in the $21 \times n$-dimensional input space to predict the secondary structure class of the central residue. Here the first 20 dimensions are allocated for the amino acid types, and the 21st dimension is added to be able to extend the window over the sequence ends. The secondary structure prediction problem then becomes the classification of points in a multidimensional vector space. This is achieved by partitioning the

space into disjoint regions of secondary structure classes. NN methods perform the classification in the space of the input vector by defining decision boundaries. On the other hand, SVM methods first map the input vectors into a higher dimensional Hilbert space by a transformation kernel and then perform the classification in that space by finding separating hyperplanes.

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. Prediction accuracy of such an algorithm should be higher than one of a single sequence algorithm due to incorporation of additional evolutionary information from multiple alignments. 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].

Signal Processing Approach:

The digital nature of genomic information makes it suitable for the application of signal processing techniques to better analyze and understand the characteristics of DNA, proteins, and their interaction. Prediction of genes, protein structure, and protein function greatly utilize pattern recognition techniques, in which hidden Markov models, neural networks, and support vector machines (SVMs) play a central role. Moreover, the subsequent analysis of microarray data seeks to extract meaningful results from the noisy measurements and reliably infer gene regulatory networks. In that respect, genomic research greatly benefits from the signal processing theory for the detection of genes with strong or weak expression patterns; classification of genes according to their similarity in expression levels; and prediction, control, and statistical-dynamical modeling of gene networks. Therefore, signal processing offers a variety of methods from pattern recognition and network analysis for the diagnosis and therapy of genetic diseases [19-20].

Digital Signal Processing (DSP) is an area of science and engineering undergoing rapid development, largely due to the advances in computing and integrated circuits. We map a protein into a digital signal by assigning numeric values to each amino acid. DSP techniques relating to protein structure analysis, such as [21-23], assign numeric values - often their hydrophobocity values [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 polarisability 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. Smithfor 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.

Conclusions:

Considerable research effort has been devoted to predicting the secondary structure of proteins from their amino acid sequences. Despite the plethora of prediction techniques, present methods typically have 76% approximate level of accuracy on an average. Thus, there is a considerable room for improvement. The main goal of a secondary structure prediction algorithm should be to design a classifier having a feature set (dependency structure) that is comprehensive enough to capture the essential correlations yet simple enough to allow reliable parameter estimation from available training data. In single-sequence prediction, one issue limiting the prediction accuracy is the small sample size. Digital signal processing plays an important role in prediction of secondary structure of proteins.

References:

[1] Brandon C., Tooze J., Introduction to Protein Structure. Garland Publishing. New York, 1991

[2] R.M. Karp, "Mathematical challenges from genomics and molecular biology," *Notices AMS*, vol.49, no. 5, pp. 544–553, 2002.

[3] Alexandrov, N., Solovyev, V. Effect of secondary structure prediction on protein fold recognition and database search. Genome Informatics 7, 119-127, 1996

[4] Alexandrov, N., Solovyev, V. Effect of secondary structure prediction on protein fold recognition and database search. Genome Informatics 7, 119-127,1996

[5] Anfinsen, C. B. Principles that govern the folding of protein chains. Science. 181, 223-230, 1973.

[6] Chou, P., Fasman G., Prediction of the secondary structure of proteins from their amino acid sequence. Advanced Enzymology, 47, 45-148, 1978.

[7] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, "Exploiting the past and the future inprotein secondary structure prediction," *Bioinformatics*, vol. 15, no. 11, pp. 937–946, 1999.

[8] 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

[9] Rost, B., Review: Protein Secondary Structure Prediction Continues to Rise. Journal of Structural Biology, 134, 204-218,2001.

[10] Bourne, Philip E., Weissig, Helge, 2003. Structural Bioinformatics. John Wiley & Sons.

[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] 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.

[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

[21] signal processing: Diagnosis and therapy,"*IEEE Signal Processing Mag.*, vol. 22, no. 1, pp. 107–112, 2005.

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

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

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

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

[26] 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,

[27] DebasisMitra 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