International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 246-253

Predicting Secondary Structure of Amino Acids Using Information Theory-Based Method

Amandeep Singh
Research Scholar Desh Bhagat Foundation Group of Institutions Moga

Deepak Sharma
Assistant Professor Desh Bhagat Foundation Group of Institutions Moga

Gurjeet Singh Mattu
Assistant Professor Desh Bhagat Foundation Group of Institutions Moga

Sukhdeep Sharma
Research Scholar Desh Bhagat Foundation Group of Institutions Moga

Abstract
Bioinformatics is the application of computer science and information technology to the field of biology and medicine. Bioinformatics provide the information related to bio-molecules on large scale. Bioinformatics deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory, software engineering, image processing, signal processing, discrete mathematics, for generating new ideas related to biology and medicine, and improving & discovering new methods of computation. In this research we observed that the GOR method which is based on information theory and Bayesian Statistics is quite successful in accurately prediction of secondary structure of Amino acids. During this research work we predicted Probabilities of three conformational states for each residue in the sequence using GOR method. The developed method is highly stable and consistent when tested against the different DSSP secondary structure reduction methods. Information regarding the secondary structure elements such as helix, sheet and coil that form for a particular sequence of amino acid is distributed across whole window. This information is retrieved from database of 267 proteins. Different types of input formats of sequences are used to determine the accuracy of secondary structure prediction GOR method. In the developed process, cclassification trees predict responses to data. To predict a response, decisions in the tree from the root (beginning) node down to a leaf node is followed. The leaf node contains the response. Classification trees give responses that are nominal, such as 'true' or 'false'. Each step in a prediction involves checking the value of one predictor (variable).

Keywords: *GOR method, secondary structure of protein, information theory*

Date of Submission: 25-05-2022 Date of acceptance: 05-06-2022

I. Introduction to bioinformatics:

Bioinformatics is the application of computer science and information technology to the field of biology and medicine. Bioinformatics provide the information related to bio-molecules on large scale. Bioinformatics deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory, software engineering, image processing, signal processing, discrete mathematics, for generating new ideas related to biology and medicine, and improving & discovering new methods of computation as shown in Figure 1. The emphasis here is on the use of computers because most of the tasks involved in genomic data analysis are highly repetitive or mathematically complex. Computers are the part of bioinformatics field which help indispensable in mining genomes for information gathering and knowledge building.

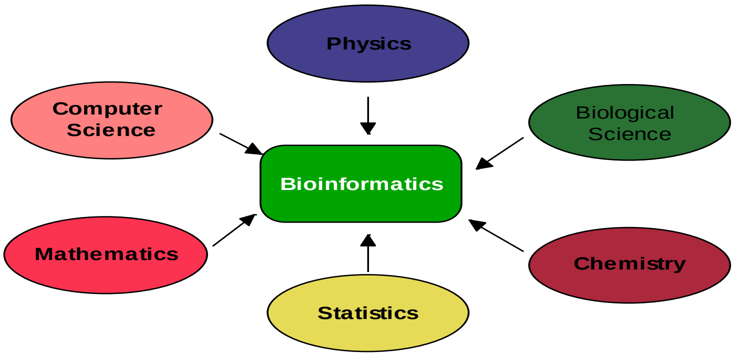


Figure 1: Bioinformatics Combination

Bioinformatics differs from a related field known as computationalbiology. The area of the bioinformatics is limited to sequence, structural, and functional analysis of genes and genomes and their corresponding products. However, computational biology defines the computation encompasses in all biological areas. For example, mathematical modelling of ecosystems, population dynamics, application of the game theory in behavioural studies, and phylogenetic construction using all employ computational tools, but do not necessarily involve biological macromolecules.Beside this distinction, there are some terms presents which define how these two terms are related. For example, one version defines bioinformaticsas the development and application of computational tools in managing all kindsof biological data, whereas computational biologyis more confined to the theoretical development of algorithms used for bioinformatics.

II. Literature Review:

Gerhart et.al (2011) discussed method must be found to handle how the drastically increasing amount biological data is stored and accessed. Currently, this data is spread across numerous sources, making analysis of said data more difficult. The BioDB tool will serve as a solution. By creating programmable objects from the current biological data tables and using them in conjunction with a local database that is populated on an as-needed basis, BioDB combines the various data sources into a single fast and effective platform.

Chang et.al (2010) studied the intellectual property (IP) for bioinformatics databases play a key role in accelerating development of biological sciences and biotechnological industry. This paper presents current and global position of IP protection in bioinformatics database. A protection method has been proposed after analyzing characteristics of bioinformatics database and considering different database protection methods. Further, the paper seeks to analyze the diffusion process of biological information and develops an argument that bioinformatics primary database should be put in public domain, though they may be given financial subsidies by the government or other public funds according to the diffusion phase of biological information. Suitable methods of IP protection in the bioinformatics secondary database have been suggested.

Chen et al. (2011) discussed that classification trees are nonparametric statistical learning methods that incorporate feature selection and interactions, possess intuitive interpretability, are efficient, and have high prediction accuracy when used in ensembles. This paper provided a brief introduction to the classification tree-based methods, a review of the recent developments, and a survey of the applications in bioinformatics and statistical genetics.

Fallahiet. al (2011) revealed that different amino acids have different preferences for taking part in alpha helices and beta sheets conformations. Different substitution matrices have been prepared to compare the proteins' secondary and tertiary structures. Unfortunately, in many cases these matrices are unable to produce reliable results, due to the complexity of these conformations. In this work, following dissection of beta sheets with different size, the amino acids compositions of each position in each class of beta sheets were extracted and compared. The amino acid contents of the same position in different beta sheets were also compared. Our results indicate great differences in amino acid contents between beta sheets with different size. Individual substitution matrices might be required in order to do alignment and comparison studies for different types of beta sheets. These results might also imply alternative evolutionary route for beta sheets with different size; longer beta sheets could be a result of merging the smaller one together in different combinations, rather than simple expansion of the smaller sheets. Based on these findings hopefully we would be able to propose an improved substitution matrix (possibly more than one) that could be used for all secondary structures, regardless of their size.

Lin et. al (2011) discussed the adsorption and desorption of three basic amino acids (L-His, L-Lys and L-Arg) with a spherical cellulose adsorbent containing the carboxyl group. Static experiments were carried out as function of pH, initial amino acid concentration, adsorption time and temperature. The adsorbent showed excellent adsorption capacities for the mentioned basic amino acids. The equilibrium adsorption data of the amino acids on adsorbent were analyzed by Langmuir and Freundlich models. The kinetics of basic amino acids adsorption nicely followed the second-order rate expression which demonstrated that chemical adsorption plays a significant role in the adsorption mechanism. Moreover, dynamic experiments were performed to determine the breakthrough curves. Regeneration experiments were tried four cycles and the results implied that the cellulose adsorbents had such advantages as simple regeneration, high recovery and reusability.

Zamaniet. al (2011) studied the efficiency of a number of commonly used amino acid encodings by using artificial neural networks and substitution scoring matrices. An important step in many machine learning techniques applied in computational biology is encoding the symbolic data of protein sequences reasonably efficient in numeric vector representations. This encoding can be achieved by either considering the amino acid physicochemical properties or a generic numerical encoding. In order to be effective in the context of a machine learning system, an encoding must preserve information relative to the problem at hand, while diminishing superfluous data. To this end, it is important to measure how much an encoding scheme can conserve the underlying similarities and differences that exist among the amino acids. One way to evaluate the effectiveness of an amino acid encoding scheme is to compare it to the roles that amino acids are actually found to play in biological systems. A numerical representation of the similarities and differences between amino acids can be found in substitution matrices commonly used for sequence alignment, since these substitution matrices are based on measures of the interchangeability of amino acids in biological specimens. In this study, a new encoding scheme is also proposed based on the genetic codon coding occurs during protein synthesis. The experimental results indicate better performances compared to the other commonly used encodings.

Greene et.al (2012) studied the isolated polypeptides comprising or consisting essentially of specific structural motifs (e.g., three beta-sheets and two alpha-helices) are provided, wherein the polypeptides exhibit at least one cell signaling or other non-canonical activity of biological relevance, Also, provided are polynucleotide’s encoding such polypeptides, binding agents that bind such polypeptides, analogs, variants and fragments of such polypeptides, etc., as well as compositions and methods of identifying and using any of the foregoing.

Exarchoset. al (2007) determines the distinction between cis peptide bonds from trans isomers in protein sequences facilitates the exploration of protein structures and functions. In this study, authors evaluated the effect of a large and informative feature vector, towards the reliable prediction of peptide bond conformation between any two amino acids. We used multiple sequence alignment, secondary structure information, real valued solvent accessibility predictions for each amino acid and physicochemical properties of the surrounding residues. A three stage schema was developed, comprising of feature extraction, feature selection and peptide bond classification between any two amino acids. Authors also explored the performance achieved when using the full feature vector without performing feature selection. The best discriminating ability was achieved using a Naive Bayes classifier, combined with wrapper feature selection. The proposed approach yielded prediction accuracy 86%, sensitivity 82% and specificity 90% in discriminating cis and trans peptide bond conformations.

An et.al (2009) discussed the development of the protein secondary structure prediction. Some concerned secondary structure prediction methods are introduced. Then we propose a novel method to predict protein secondary structure , which substantially improves the prediction accuracy both over 80% in CB513 and RS126 database. At the end, we point out some possible trends in the protein secondary structure prediction in the future.

Shi et. al (2009) discussed the protein prediction, the location of disulfide bonds can strongly reduce the search in the conformational space. Therefore the correct prediction of the disulfide connectivity starting from the protein residue sequence may also help in predicting its 3D structure. In this paper, we describe a method to predict disulfide connectivity in aprotein given only the amino acid sequence, using neural network, and given input of symmetric flanking regions of N-terminus and C-terminus cystines augmented with residue secondary structure (helix, sheet, and coil) as well as evolutionary information. 252 protein sequences were selected from the SWISS-PROT database. From the results of 4-fold cross validation, authors find that merging protein secondary structure allows us to obtain significant prediction accuracy improvements.

Hateganet. al (2007) studied the problem of jointly encoding the amino acid sequence and the secondary structure information of proteins, in the current format in which more and more proteins are stored in Swiss-Prot database. The new method, dubbed ProtCompSecS, combines the compressor ProtComp previously designed only for amino acid sequences, with a dictionary based method, where the dictionary containing the patterns for representing the secondary structure is obtained by suitably processing the Dictionary of Protein Secondary Structure (DSSP) data base. We experimented with protein sequences of 14 complete proteomes. When comparing the performance of ProtCompSecS algorithm with that of ProtComp algorithm, for those sequences that have annotated secondary structure information, it surprisingly appeared that encoding both sequence and secondary structure information is more efficient than encoding the protein sequence alone (without knowledge of the secondary structure). This is a strong argument for claiming that the secondary structure has a high descriptive value for modeling and understanding the primary structure (the amino acid sequence) of a protein.

Miyazawa (2011) stated advantages of a mechanistic codon substitution model for evolutionary analysis of protein-coding sequences. Each codon substitution rate is proportional to the product of a codon mutation rate and the average fixation probability that depending on the type of amino acid replacement and has advantages over nucleotide, amino acid and empirical codon substitution models in evolutionary analysis of protein-coding sequences. It can approximate a wide range of codon substitution processes. If no selection pressure on amino acids is taken, it will become equivalent to a nucleotide substitution model. If mutation rates are assumed not to depend on the codon type, then it will become essentially equivalent to an amino acid substitution model. Mutation at the nucleotide level and selection at the amino acid level can be separately evaluated.

Ismail et.al (2010) studied the sequence-structure relationship is the key step in protein modeling and de novo protein design. Although almost 55,000 protein structures are solved and stored in proteindatabank, elucidating sequence-structure relationship is still a challenging task. To understand sequence-structure relationship better, a statistical analysis of amino acid residues in four major structural classes of protein viz. α proteins, β proteins, α/β proteins and α+β proteins is performed. We use non-homologous proteins from (<; 30% identity) October 2008 release Brookhaven ProteinDataBank (PDB) with resolution better than 2.5 angstrom. Interestingly, in comparison to the helical protein, the helical propensities of hydrophobic residues in mix proteins (containing both α helix and β sheet) are increased significantly. On the other hand, the helical propensities of hydrophilic residues are reduced in mixed proteins. A reverse trend is observed in strand propensity. The difference in helical propensity of hydrophobic and hydrophilic residues in different fold may be due to differential folding mechanism. The size of protein may also play a crucial role. A position specific analysis of helices is also done in all α and α/β proteins. The detailed analysis of helix dissection revealed that, the presence of β sheet influences the individual preference of amino acids in different positions within helix. This result indicates that the preference of amino acid in different positions (N terminus, C terminus and middle) within α helix are influenced by long range interactions with other structural elements.

Kumar et.al (2010) discussed in this paper a new method for the prediction of the protein secondary structure from the amino acid sequence. The method is based on the most recent version of the standard GOR algorithm. A significant improvement is obtained by combining multiple sequence alignments with the GOR method. Additional improvement in the predictions is obtained by a simple correction of the results when helices or sheets are too short, or if helices and sheets are direct neighbours along the sequence. The imposition of the requirement that the prediction must be strong enough, i.e. that the difference between the probability of the predicted (most probable) state and the probability of the second most probable state must be larger than a certain minimum value also improves significantly secondary structure predictions.

Singh et.al (2010) discussed the research in bioinformatics has accumulated large amount of data. As the hardware technology advancing, the cost of storing is decreasing. The biological data is available in different formats and is comparatively more complex. Knowledge discovery from these large and complex databases is the key problem of this era. Data mining and machine learning techniques are needed which can scale to the size of the problems and can be customized to the application of biology. In the present research work, the Chou-Fasman Method is implemented with the help of data mining. Protein structure determination and prediction has been a focal research subject in the field of bioinformatics due to the importance of protein structure in understanding the biological and chemical activities of organisms. The experimental methods used by biotechnologists to determine the structures of proteins demand sophisticated equipment and time. A host of computational methods are developed to predict the location of secondary structure elements in proteins for complementing or creating insights into experimental results. Cluster analysis is used as data mining model to retrieve the results.

III. Methodology

Advances in molecular biology in the last few decades, and the availability of equipment in this field have allowed the increasingly rapid sequencing of considerable genomes of several species. The most fundamental tasks in bioinformatics include the analysis of sequence information which involves the following the prediction of the 2D structure of a protein using algorithms that have been derived from the knowledge of physics, chemistry and from the analysis of other proteins with similar amino acid sequences as in Figure 2. One of the methods used for predicting secondary structure is GOR method.

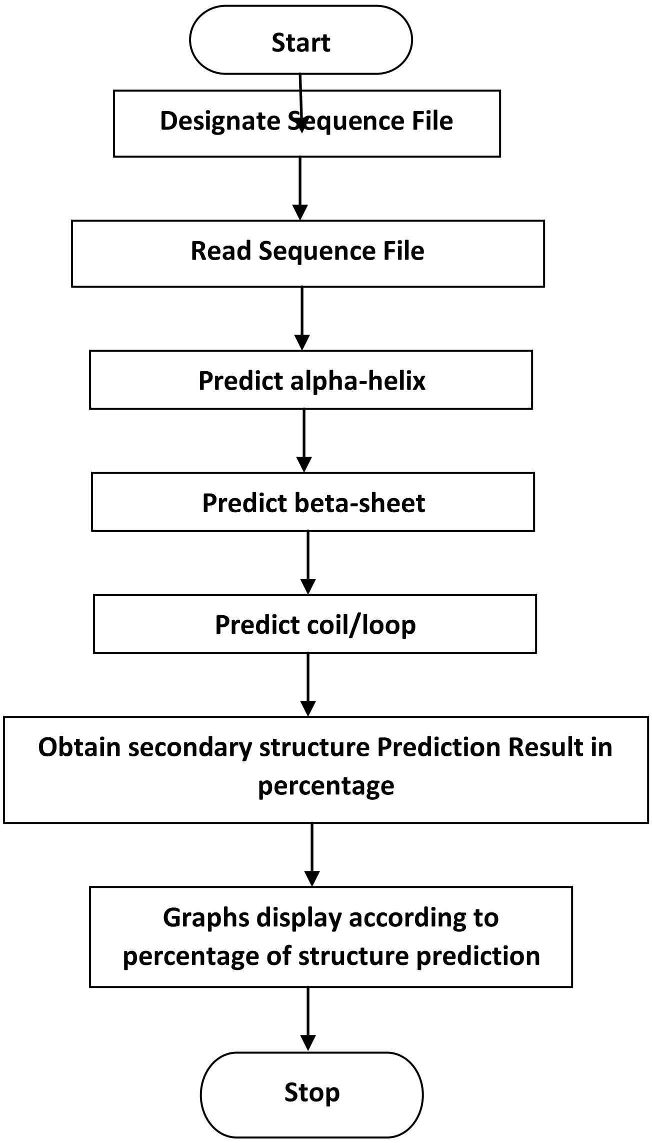

Start
Designate Sequence File
Read Sequence File
Predict alpha-helix
Predict beta-sheet
Predict coil/loop
Obtain secondary structure Prediction Result in percentage
Graphs display according to percentage of structure prediction
Stop

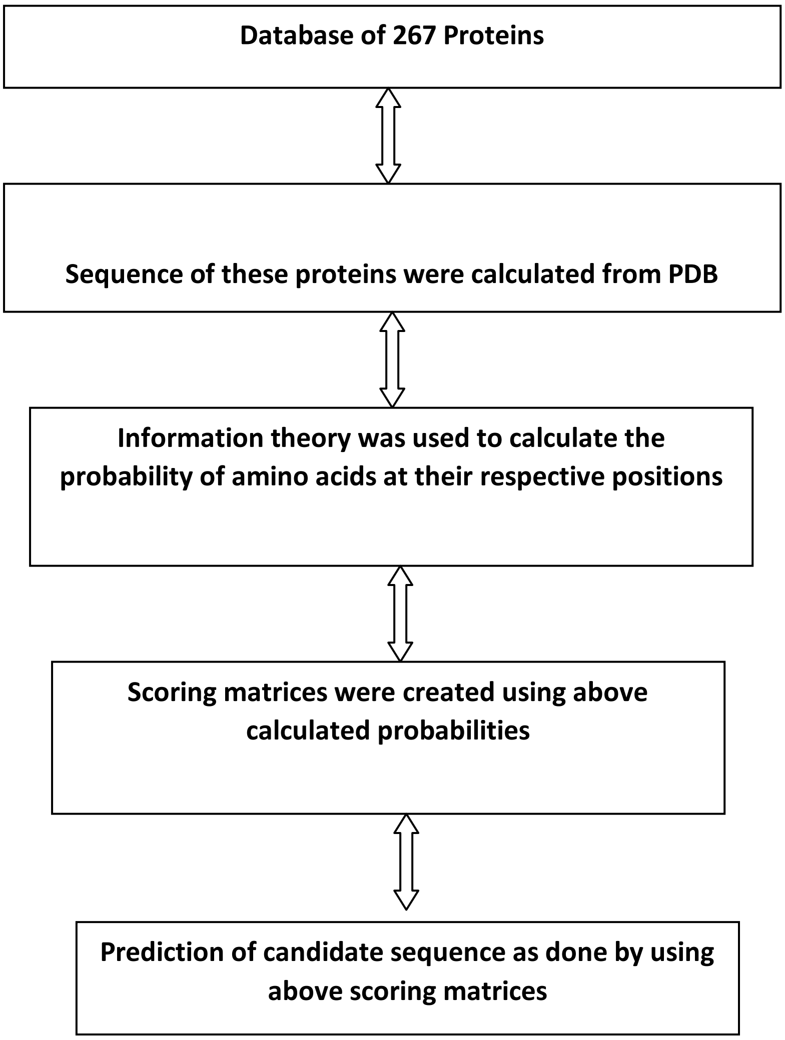
Figure 2: General Framework for Protein Secondary Structure Prediction (PSSP) Method

In prediction method for secondary structure of protein determines the accuracy in terms of present percentage of helix, sheet and coil. Formation of α-helix, β- sheet and coils are predicted with respect to each amino acid residue present in a sequence of amino acids residues. Result of the prediction of all secondary structure elements are combined to obtain the result of prediction of secondary structure of protein as in Figure 3.

Implementation of GOR method:

The GOR method is named after its authors Garnier-Osguthorpe-Robson. The GOR method uses both information theory and Bayesian statistics for predicting the secondary structure of proteins. The GOR method consist the information about a slightly longer segment of the polypeptide chain.

Work Flow Diagram of GOR method


Database of 267 Proteins
Sequence of these proteins were calculated from PDB
Information theory was used to calculate the probability of amino acids at their respective positions
Scoring matrices were created using above calculated probabilities
Prediction of candidate sequence as done by using above scoring matrices

Code relating the amino acid sequence and secondary structure of proteins established by the Robson using the combination of information theory and Bayesian statistics. Database of 267 protein standards are prepared with the help of visual studio. Three tables are prepared from this data of helix, beta sheet and coil. The information theory approach used by researchers to analyze the contributions of several traditional amino acid residues using mutual information concept. GOR method uses an information function as shown in Figure 3. To determine the structure for a given amino acid position j, the GOR method looks at a window of 8 amino acids before and 8 after the position of interest as in Equation1. Suppose a_j is the amino acid that we are trying to determine. GOR looks at the residues in Equation

$$a_{j-7}\ a_{j-8} \dots\dots..a_j \dots\dots\dots.a_{j+7}a_{j+8} \qquad \dots (1)$$

Intuitively, it assigns a structure based on probabilities it has calculated from protein databases. These probabilities are of the form as shown in Equatoins 2 and 3:

Pr[amino acid j is α |$\mathbf{a_{j-7}\ a_{j-8} \dots\dots..a_j \dots\dots\dots.a_{j+7}a_{j+8}}$] … (2)

Pr[amino acid j is β|$\mathbf{a_{j-7}\ a_{j-8} \dots\dots..a_j \dots\dots\dots.a_{j+7}a_{j+8}}$] … (3)

In GOR method, three scoring matrices, and each column consist the probability of finding each amino acid at one of the 17 positions, are prepared. Information theory forms on the basis of information function I(S, R)which will be fully represented in mathematical notation together with other functions and formula.

Shannon tells us the concept of mutual information and how the information is conveyed. The information function is described in terms of logarithm ratio of the conditional probability P (S|R)of observing conformation S.

The information available as to the joint occurrence of secondary structural conformation Sand amino acid R is given by Equation4:

$$\mathbf{I(S;R) = \log\frac{P(S|R)}{P(S)}} \qquad \dots (4)$$

whereP(S | R) is the conditional probability of conformation Sgiven residue R, and P(S) is the probability of conformation S. The information function is defined as the logarithmic ratio of the conditional probability P (S|R). Where S is observed conformation which can be one of three states*:* helix (H), extended (E), or coil (C)- for residue R, where R is one of the 20 possible amino acids and the probability P(S) of the occurrence of conformation S.
By Bayes' rule, the probability of conformation Sgiven amino acid R,(S | R) is given by Equation5:

$$\mathrm{P(S|R)} = \frac{\mathrm{P(S,R)}}{\mathrm{P(R)}} \qquad \dots (5)$$

whereP(S, R) is the joint probability of Sand Rand P (R) is the probability of R. These probabilities can be predicted from the frequency of each amino acid found in each structure and the frequency of each amino acid in the structural database. Given these frequencies in Equation6:

$$\mathrm{I(S;R)} = \frac{\log \mathrm{f_{S,R}}}{\mathrm{f_S}} \dots (6)$$

where$\mathrm{f_{S,R}}$is the frequency of amino acid Rin conformation Sand $\mathrm{f_S}$is the frequency of all amino acid residues found to be in conformation S.

IV. Results and Discussion:

In present work, we dealt with amino acid residues to determine the secondary structure of sequences. GOR method is used to predict the structure of amino acids. Combination of amino acids results in formation of protein through peptide bond.GOR method is window structure based experiment. It useu the 17*20 window size. It predicts the percentage of three conformational states according to the presence of it in protein sequence. The DSSP code is used in GOR method that will reduce the 8 classes code to 3 classes. Different types of sequence formats are used in present work as input. Each format has its own format and significance. Classification trees are generated from root node down to leaf node. It will check the values of one predictor or variable. MATLAB platform is used to done the present work.

V. Conclusion and future scope:

Author studied the GOR method based on information theory and Bayesian Statistics is quite successful in its accuracy of secondary structure prediction. Probabilities of three conformational states are predicted for each residue in the sequence with the help of GOR method and this information can be used for further analysis. These are results are achieved when predictions are made on single sequence. The developed method is highly stable and consistent when tested against the different DSSP secondary structure reduction methods conducted in this research. Information regarding the secondary structure elements such as helix, sheet and coil that form for a particular sequence of amino acid is distributed across whole window. This information is retrieved from database of 267 proteins. Different types of input formats of sequences are used to determine the accuracy of secondary structure prediction GOR method. Classification trees predict responses to data. To predict a response, follow the decisions in the tree from the root (beginning) node down to a leaf node. The leaf node contains the response. Classification trees give responses that are nominal, such as 'true' or 'false'. Each step in a prediction involves checking the value of one predictor (variable). Variety of different sequences formats can be introduced for further analysis.Varieties of Bioinformatics tools are available which can be used to incorporate new research in Bioinformatics field.Present GOR method is based on single sequence but in future it can be incorporated to multiple sequence alignment to achieve different results.Since the research in bioinformatics field increasing rapidly. So our requirement to achieve optimal result in less time.

REFERENCES

[1]. An, B., Zhai, Y. and Zhang, M. (2009) "Accuracy of Protein Secondary Structure Prediction Continues to Rise" International Conference on MASS' 09, pp.1-4.

[2]. Akitomi, J. (2007) "Method for predicting Secondary Structure of RNA, an apparatus for predicting and a predicting program" US Patent 0235155.

[3]. Balaban, D.J. and Aggarwal, A. (2005) "Method and apparatus for providing a Bioinformatics Database" US Patent 7215804.

[4]. Benner, S.A. (1997) "Application of Protein Structure Prediction" US Patent 6377893.

[5]. Chang, J. and Zhu, X. (2010) "Bioinformatics Database: Intellectual Property Protection Strategy" Journal of Intellectual property Rights Vol 15, pp.447-454.

[6]. Chen, X., Wang, M. and Zhang, H. (2011), "The use for classification trees for bioinformatics", John Wiley & Sons, Inc. WIREs Data Mining KnowlDiscov,pp 55–63.

[7]. Deris, S.B., Illias, R.B.M., Senafi, S.B., Abdalla S.D. and Arjunan S.N.V." Protein Secondary Structure Prediction From Amino Acid Sequence Using Artificial Intelligence Technique" pp. 1-245, 2007.

[8]. Exarchos, K.P., Exarchos, T.P., Papaloukas, C., Troganis, A.N. and Fotiadis, D.I. (2007) "Predicting peptide bond conformation using feature selection and the Naive Bayes approach" IEEE EMBS 2007, pp.5009-5012.

[9]. Fallahi, H. and Yarani, R. (2010) "Positional preferences by 20 amino acids in beta sheets" IEEE BIBMW 201, pp.806-807.

[10]. Gerhart, J. and Sacan, A. (2011) "BioDB: Integration of biological knowledgebases" IEEE BIBMW 2011, pp. 899.

[11]. Greene, L.A. (2011) "Polypeptide Structural Motifs Associated With Cell Signaling Activity" US Patent 0004185.

[12]. Gardner, S. (2003) "Modular Bioinformatics Platform" US Patent 0177143.

[13]. Garnier, J., Gibrat, J.F. and Robson B. 1996 “GOR method for predicting Protein Secondary Structure from Amino Acid Sequence” Methods in Enzomology, vol 266, pp. 540-553.

[14]. Haan, J.R.D. and Leunissen, J.A.M. (2005) “Protein Secondary Structure Prediction Comparison of Ten Common Prediction Algorithms Using a Neural Network” pp. 1-13.

[15]. Hategan, A. and Tabus, I. (2007) “Jointly Encoding Protein Sequences and their Secondary Structure Information” GENSIPS’07, pp. 1-4.

[16]. Ismail, W.M. and Chowdhury, S. (2010) “Preference of Amino Acids in Different Protein Structural Classes: A Database Analysis” ICBBE 2010, p. 1-5.

[17]. Kumar, B. and Jani, N.N. (2010) “Prediction of Protein Secondary Structure based on GOR Algorithm Integrating with Multiple Sequences Alignment” International Journal of Advanced Engineering and Applications, pp.177-182

[18]. Krane, D. and Raymer, M. (2006), “Fundamental concepts of bioinformatics”, Pearson Education Publishers.

[19]. Kloczkowski, A., Ting, K.L., Jernigan, R.L. and Garnier, J. 2002 “ Protein Secondary Structure Prediction based on GOR method incorporating with Multiple Sequence Alignment” pp.441-449.

[20]. Kloczkowski, A., Ting, K.L., Jernigan, R.L. and Garnier, J. 2002 “Combining the GORV Algorithm With Evolutionary Information for Protein Secondary Structure Prediction From Amino Acid Sequence” pp. 154-166.

[21]. Lin, C. and Lui, M. (2011) , “Adsorption Behaviors of Basic Amino Acids on Spherical Cellulose Adsorbent” ICBBE’011, pp. 1-7.

[22]. Ledda, F.G. (2011), “Protein Secondary Structure Prediction: Novel Methods and Software Architectures” University of Cagliari, pp.1-199.

[23]. Miyazawa, S. (2011), “Advantages of a Mechanistic Codon Substitution Model for Evolutionary Analysis of Protein-Coding Sequences”, PLoS ONE, vol. 6, 12, pp. 54-69.

[24]. Mount, D.W. (2004), “Bioinformatics: Sequence and Genome Analysis”, 2nd ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.. ISBN 0-87969-608-7.

[25]. Noguchi, T. (1996) “Prediction Method and Apparatus for a Secondary Structure of Protein” US Patent 5842151.

[26]. Pevsner, J. (2009), “Bioinformatics and Functional Genomics”, A john wiley& sons, Inc. Publication, pp 215-221.

[27]. Rastogi, S.C., Mendiratta, N., Rastogi, P. (2007), “Allignment of Multiple Sequences and Phylogenetic Analysi-Bioinformatics Methods and Applications”, 3rd edition, PHI publication, pp. 5-120.

[28]. Singh, R., Deol, S.K. and Sandhu, P.S. (2010) “Chou-Fasman Method for Protein Structure Prediction using Cluster Analysis” World Academy of Science, Engineering and Technology 72 2010,pp. 982-987.

[29]. Shi, O., Yang, H., Cai, C. , Yang, J. and Tian, X. (2008) “Disulfide Bond Prediction using Neural Network and Secondary Structure Information” ICBBE2008, pp. 656-659.

[30]. Schneiderbauer, S. (2008) “RNA Secndary Structure Prediction” Cognitive Science, University of Osnabruck, pp.3-42.

[31]. Singh, M., Sandhu, P.S. and Kaur, R.K. (2008) “Protein Secondary Structure Prediction” World Academy of Science, Engineering and Technology, pp. 458-461.

[32]. Senekal, F.P. (2008) “Protein Secondary Structure Prediction Using Amino Acids Regularities”, University of Pretoria, pp.1-178.

[33]. Sen Z.T., Jernigan L.R., Garnier J. and Kloczkowski A.(2005) “GOR V server for protein secondary structure prediction”vol 21., no. 11, pp 2787-2788.

[34]. Singh, M. “Predicting Protein Secondary and Super Secondary Structure” CRC Press, pp. 29.1-29.30, 2001.

[35]. Singh, M. “COS551 Introduction to Computational Biology” Carl Kingsford, pp. 1-15, 2000.

[36]. Wu, L., Dai, Q., Han, B., Zhu, L. and Li, L. (2010) “Prediction of protein structural class using a combined representation of protein-sequence information and support vector machine” IEEE BIBMW 2010, pp.101-106.

[37]. Xiong J., (2006), “Essential Bioinformatics”, United States Of America, Cambridge University Press, New York.

[38]. Zamani, M. and Kremer, S.C. (2011) “Amino acid encoding schemes for machine learning methods”IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), pp. 327-333.

[39]. Zimek, A., Buckwald, F., Frank, E. and Kramer, S. (2010), “A Study of Hierarchical and Flat Classification of Proteins”, IEEE/ACM Transactions on computational biology and bioinformatics, vol. 7, 3, pp. 563-571.