Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 132 (2018) 1654–1662

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

Predicting the protein structure using random forest approach

Charu Kathuria[a]*, Deepti Mehrotra[a], Navnit Kumar Misra[b]

[a] *Amity University, Noida,Uttar Pradesh,201313, India*
[b] *Brahmanand College, The Mall, Kanpur, 208004 India*

Abstract

Predicting the secondary structure of proteins is a challenging task. A large variety approaches exist that include observation using equipment's and theoretical evaluation, in which the optimal structure is determined. The secondary structure determines 3D tertiary structure of protein, on which features and functionalities of protein depend. This paper use classification technique, Random Forest to build a model which is able to determine structure of unknown proteins. The dataset included the amide frequencies of proteins whose structure is known. Machine learning model is developed that can predict the structure of protein that still need to be exploited. The accuracy of the model is determined using ROC curve. The results confirm the performance of the model constructed using amides dataset.

*Keywords:*Protein secondary structure prediction; amide frequencies; random forest; ROC;area under the curve

* Corresponding author. Tel.: +91 9999964785.
E-mail address: charu.kathuria05@gmail.com

10.1016/j.procs.2018.05.134

1. Introduction

Protein is an important molecule in all living organisms and is responsible for any changes in organism's body. Keeping track on protein functions can help in early disease diagnosis and various drugs design. These functions are correlated with protein structure which consists of sequence of 20 different amino acids. Each of these amino acids has some distinctive features and properties [1]. Integrating different sequences of amino acids can generate infinite number of proteins with unique features. So, an extreme measure of protein sequences is available in various datasets, however the number of structures known to human is very less near about 0.2% of the whole [2-5]. To determine structure of other proteins various computational and experimental methods have been generated. Experimental methods like X-ray crystallography, multidimensional magnetic resonance, etc. produces accuracy in results but are very costly, time consuming and apart from this it is not applicable to all proteins. Instead from few decades various computational methods are widely used in this field of bioinformatics and include the characteristics of simple, low cost and fast speed, this helps in overcoming the disadvantages faced by experimental methods [2, 4, 6]. The principle followed by these methods is to analyze the structure and sequence of known proteins for predicting the unknown proteins [7]. This situation is analyzed as a classification problem in which class of a given instance is then predicted on account of its properties and features. Several approaches of classificationhave been applied such as neural network [8,9,10], hidden markov model [11], support vector machine [12, 13, 14], decision tree [13] and so on.

In this paper a model is generated to solve the above addressed problemusing one of the efficient mining techniques named Random Forestand its accuracy is validated using ROC curve. In further sections an introduction to protein structure is given followed by the methodology and experimental part concluded with results and conclusion.

2. Protein Structure concepts and techniques

Hierarchy of protein structure: primary, secondary, tertiary and quaternary are the main four levels. In this tertiary and quaternary structures forms a 3D structure of proteins which further decides its functional properties. Determining this 3D structure is a difficult task, so all the research is focused to secondary structure prediction of proteins which forms a bridge between primary and tertiary structure. Secondary structure of protein thus forms a crucial point in protein science and an essential step in 3D structure studies which helps to figure out the relation between functions and the primary structure of proteins [15]. Secondary structure is primary fold of polypeptide chain and the basis of spatial structure for a protein. It broadly occurs in three different shapes: alpha helical structure, beta structure and others. These structures can be determined with the help of amide frequencies calculated and observed with the help of IR Spectroscopy or Raman Spectroscopy using the tool FTIR. These amide frequencies occur in different categories as Amide (I – VII, A).

Mining techniques used for analyzing and predicting the data can be categorized into Supervised and Unsupervised learning [16]. In supervised learning input and output data is predefined and the mapping of input to output result is predicted. This mapping function is further used to predict the new input data whose outputs need to be known. In unsupervised learning only the input data is provided with no output information. This input is further analyzed on the basis of its features and structure to predict its detailed information. Classification, regression, association, etc are part of supervised learning, used in different areas for generating a predictive model which can be used on test data to predict the desired results.

In this work Random Forest (RF) classification technique is considered. RF was developed by Breiman in 2001 as a classification and regression tree (CART) algorithm which can construct a large number of decision trees from the given training data [17]. The name is derived from the tree like structure, having random characteristics as the split of nodes is actually performed by randomly selecting the best features from the given list of attributes. Random forest has provided the magnificent performance on a large number of problems which acts as an

enhancement of single tree classifiers. Fundamentally it is used to improve the accuracy with its low classificationerror in respect to other classification algorithms. It also overcomes the problem of over fitting and the number of trees generated can be used for further references [18]. So, this is the one of the major reason of considering this technique in the below model.

3. Methodology

This paper includes the following methodology in which dataset is decomposed in two portions: training data and test data. The training data is basically used in generating the model with a classification technique and the test data is further used to predict the accuracy of this model. The model generated is then validated using ROC curve to demonstrate the desired results.

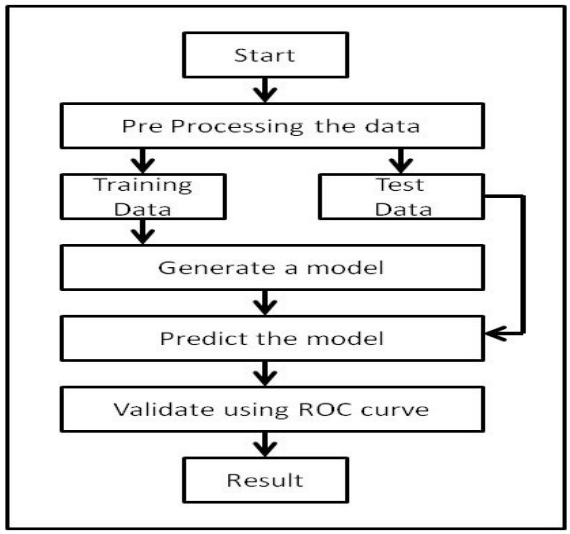

Fig. 1. Flowchart of the model constructed

4. Experiment

The model is generated using R studio which is a free R editor. It's open source software that provides an IDE (Integrated Development Environment) to R language by including functionalities and features. R programming language is basically used by data miners for data analytics, statistical computations, graphics and so on.

4.1 Dataset

The dataset consists of amide frequencies of known proteins which actually help in determining the structure that is related to its functional behavior and features. These are the observed frequencies which are collected using Infrared Spectroscopy [19, 20]. The dataset used is broadly classified in alpha and non-alpha structures as during preprocessing all alpha structure polypeptides are taken as alpha in their structure attribute and rest all other structure of polypeptides are considered as non alpha. Some of the vibrational frequencies of polypeptides having alpha structure are taken from [21-24] and non-alpha are taken from [25-30]. It's a classical method for determining the structure of bio molecules [31]. Thus more than 50 polypeptides are considered with their amide frequencies ranging from Amide (A, I-VI).

4.2 Model Construction

For convergence Random Forest (RF) technique of classification is considered in R programming language by including its respective package named randomForest.The dataset is imported in Rstudio and constructed a model using the methodology discussed with RF classification technique. The model created using training data is predicted on test data to calculate its accuracy. The following screenshots shows the implementation of the model designed in Rstudio.

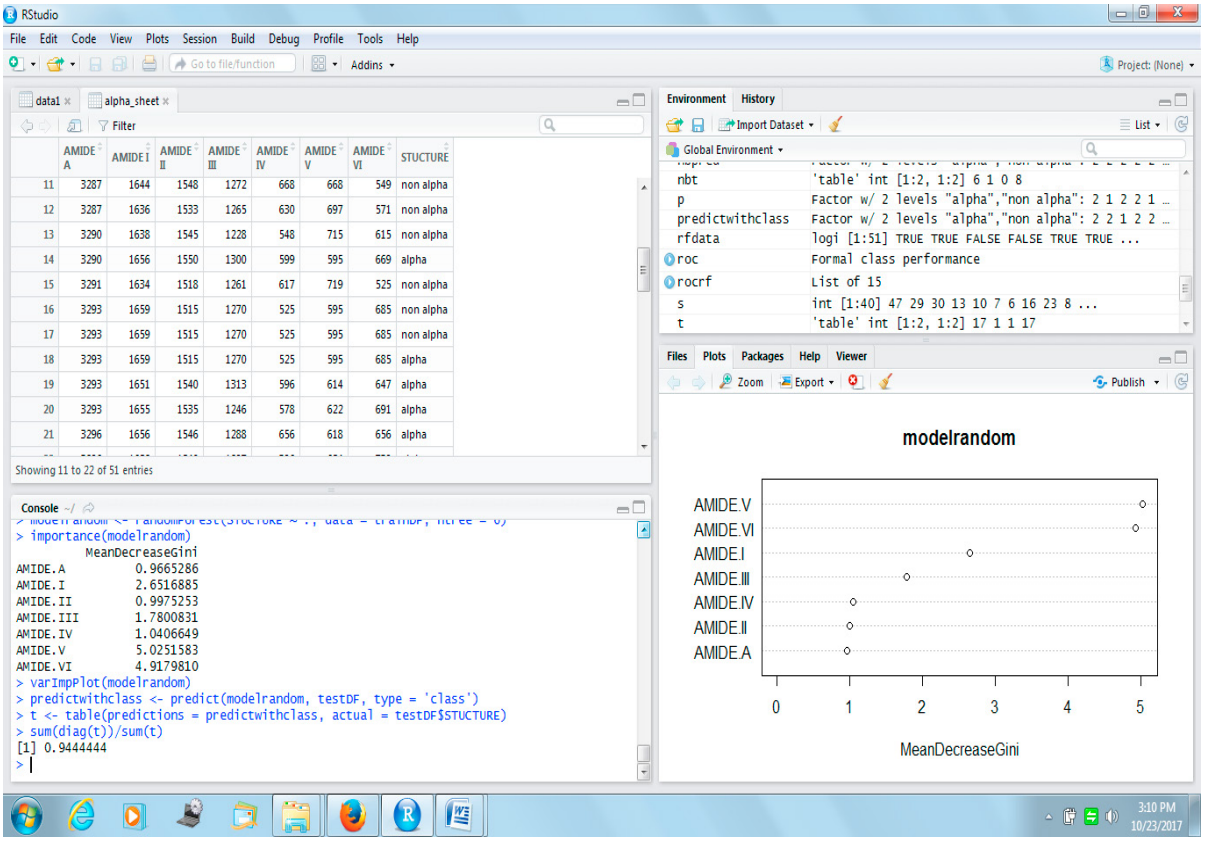

Fig. 2 Screenshot of the model generated

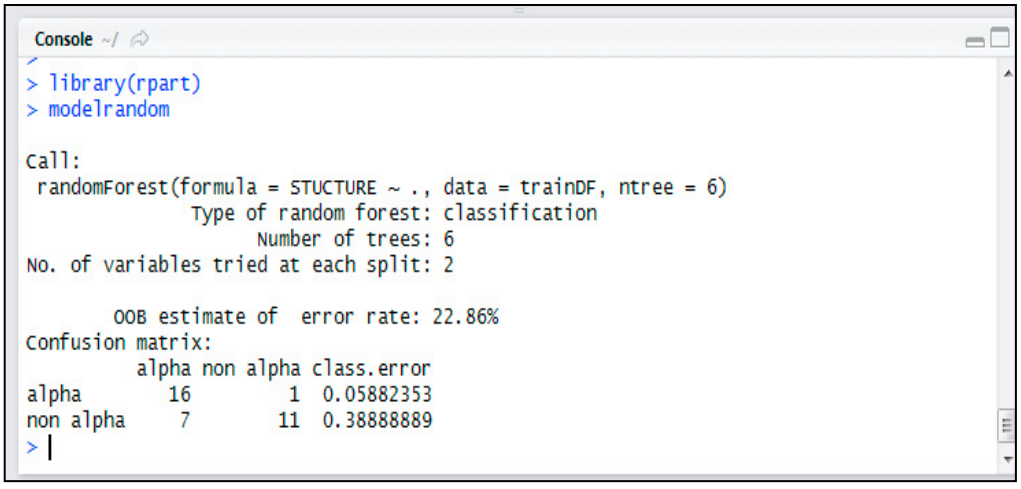

Fig. 3. Screenshot of the confusion matrix

4.3 Validation

The model constructed above is validated using Receiver Operator Curve (ROC) having True Positive Rate (TPR) as mentioned in equation (1) also named as sensitivity defines the ratio of favourable tuples that are accurately identified.

$$TPR = TP / (TP + FN) \quad (1)$$

It also includes False Positive Rate (FPR) as mentioned in equation (2) or specificity that defines the ratio of unfavourable tuples that are not accurately identified as favourable.

$$FPR = FP / (FP + TN) \quad (2)$$

Equation (1) and (2) consists of four parameters where TP defines True Positive, FP as False Positive, TN as True Negative and FN as False Negative [32]. ROC curve is a perfect visualization tool that depicts the relationship between TPR and FPR in x and y axis respectively. Hence accuracy of the model can be determined using a measure named area under the curve (AUC) whose value as 1.0 demonstrates the accurate model.

5. Results and Conclusion

In this paper we have proposed a model using Random Forest classifier to predict the secondary structure of proteins using their amide frequencies. The model accuracy is validated with ROC curve and area under the curve has been calculated having value 0.963. The following screenshot shows the ROC curve formed along with the measured area under the curve.

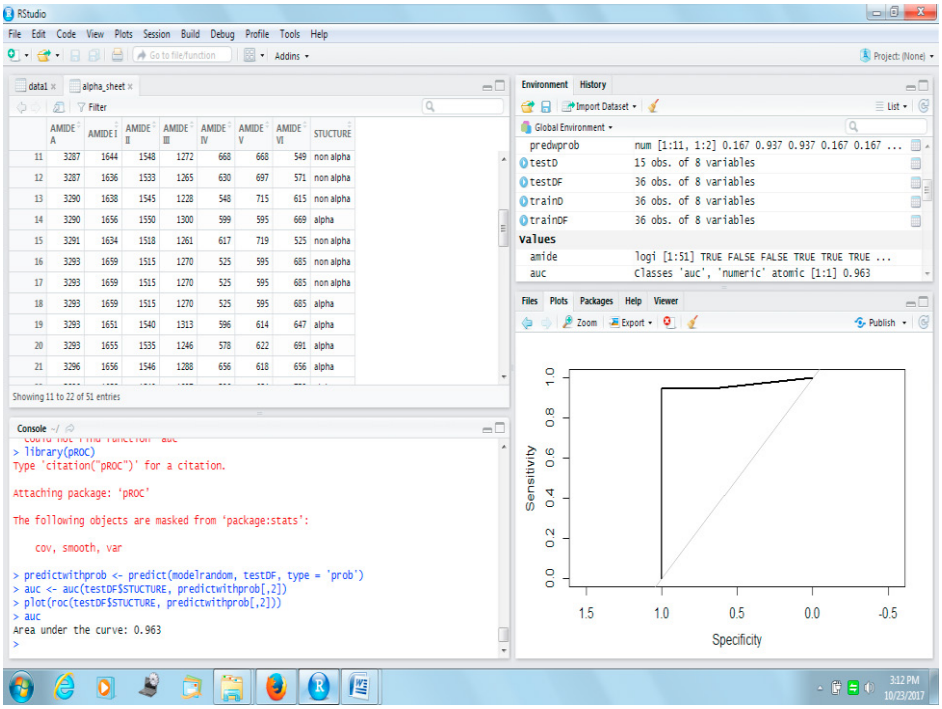

Fig. 4. Screenshot of the ROC curve with Area under the curve using Random Forest algorithm.

As previously discussed that the functionality of protein depends on its structure and its prediction is a challenging task. The α– helical structure also termed as alpha structure is one of the common protein secondary structures. In this paper a model is designed that can predict whether a protein with unknown structure having its amide

frequencies bare an alpha helical structure or not by using a data mining classification technique.
The reason of considering Random forest in the above model is its results which are far better than Decision Tree classifier. The figure below shows the result of Decision Tree classifier when used in the model.

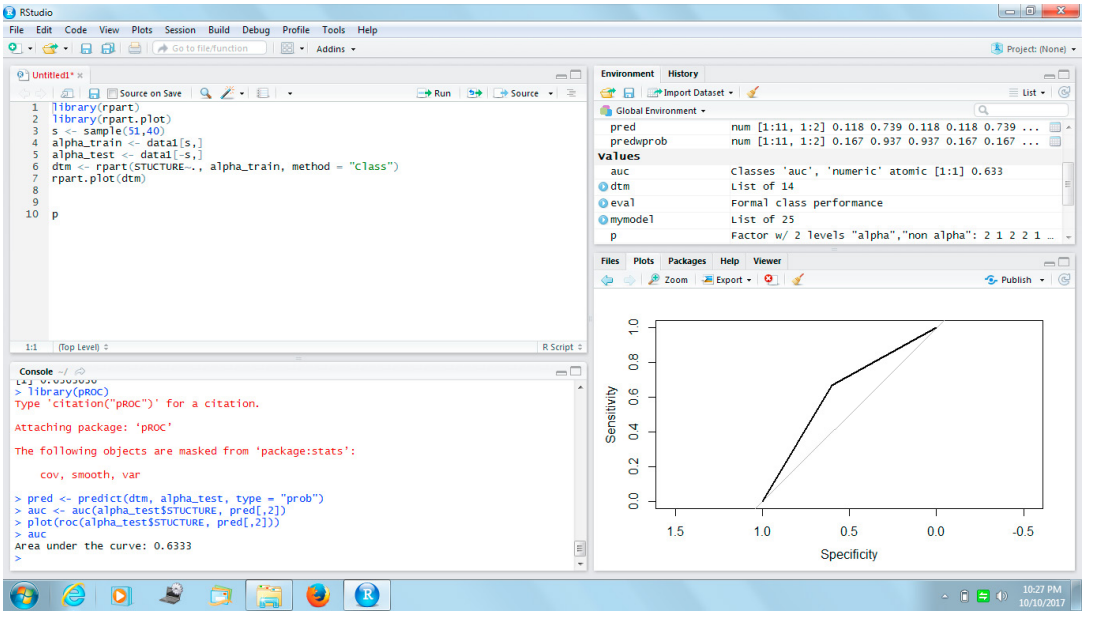

Fig 5. Screenshot of ROC curve with Area under the curve using Decision Tree Classifier.

Thus the above results of two algorithms Random forest and Decision Tree concludes that the outcome of Random forest (AUC = 0.963) is much better than the Decision Tree (AUC = 0.6333) algorithm. The ROC curve suggests that the model prepared using Random Forest has given the accurate prediction. In this work binary classification approach is adopted, but a large variant of protein structure exist. The work can be extended to classify these protein structures using multiclass classification approach.

Acknowledgement

We Acknowledge the Macromolecular Laboratory, Physics Department, University of Lucknow, Lucknow 226007, for sharing the data for this research work.

References

[1] Buxbaum, Engelbert (2007) "*Fundamentals of protein structure and function.*" Vol. 31. New York: Springer.

[2] Li, Dapeng, Tonghua Li, Peisheng Cong, WenweiXiong, and Jiangming Sun. (2011) "A novel structural position-specific scoring matrix for the prediction of protein secondary structures." *Bioinformatics* 28, no. 1: 32-39.

[3] Błażewicz, Jacek, Piotr Łukasiak, and Szymon Wilk. (2007) "New machine learning methods for prediction of protein secondary structures." *Control and Cybernetics* 36: 183-201.

[4] Ho, Hui Kian, Lei Zhang, KotagiriRamamohanarao, and Shawn Martin. (2012) "A survey of machine learning methods for secondary and supersecondary protein structure prediction." In *Protein Supersecondary Structures*, pp. 87-106. Humana Press, Totowa, NJ.

[5] Jin, Xin, RencanNie, Dongming Zhou, Shaowen Yao, Yanyan Chen, Jiefu Yu, and Quan Wang. (2016) "A novel DNA sequence similarity calculation based on simplified pulse-coupled neural network and Huffman coding." *Physica A: Statistical Mechanics and its Applications* 461: 325-338.

[6] Meng, Fanchi, and Lukasz Kurgan. (2016) "Computational Prediction of Protein Secondary Structure from Sequence." *Current protocols in protein science*: 2-3.

[7] Feng, Yonge, Hao Lin, and Liaofu Luo. (2014) "Prediction of protein secondary structure using feature selection and analysis approach." *Acta biotheoretica* 62, no. 1: 1-14.

[8] Rashid, Shamima, Saras Saraswathi, Andrzej Kloczkowski, Suresh Sundaram, and Andrzej Kolinski. (2016) "Protein secondary structure prediction using a small training set (compact model) combined with a Complex valued neural network approach." *BMC bioinformatics* 17, no. 1: 362.

[9] McGuffin, Liam J., Kevin Bryson, and David T. Jones. (2000) "The PSIPRED protein structure prediction server." *Bioinformatics* 16, no. 4: 404-405.

[10] Wang, Sheng, Jian Peng, Jianzhu Ma, and Jinbo Xu. (2016) "Protein secondary structure prediction using deep convolutional neural fields." *Scientific reports* 6: 18962.

[11] Karplus, Kevin, Rachel Karchin, Jenny Draper, Jonathan Casper, Yael Mandel-Gutfreund, Mark Diekhans, and Richard Hughey. (2003) "Combining local-structure, fold-recognition, and new fold methods for protein structure prediction." *Proteins: Structure, Function, and Bioinformatics* 53, no. S6: 491-496.

[12]Hua, Sujun, and Zhirong Sun. (2001) "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach1." *Journal of molecular biology* 308, no. 2: 397 407.

[13] He, Jieyue, Hae-Jin Hu, Robert Harrison, Phang C. Tai, and Yi Pan. (2006) "Rule generation for protein secondary structure prediction with support vector machines and decision tree." *IEEE Transactions on nanobioscience* 5, no. 1: 46-53

[14]Hu, Hae-Jin, Yi Pan, Robert Harrison, and Phang C. Tai. (2004) "Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier." *IEEE Transactions on NanoBioscience* 3, no. 4: 265-271.

[15] Yoo, Paul D., Bing B. Zhou, and Albert Y. Zomaya. (2008) "Machine learning techniques for protein secondary structure prediction: an overview and evaluation." *Current Bioinformatics* 3, no. 2: 74-86.

[16] Han, J., Kamber, M., & Pei, J (2006).: Data mining, southeast asia edition: Concepts and techniques. Morgan Kaufmann.

[17] Breiman, Leo. (2001) "Random forests." *Machine learning* 45, no. 1: 5-32.

[18] Biau, GÃŠrard, Luc Devroye, and GÃĄbor Lugosi.(2008) "Consistency of random forests and other averaging classifiers." *Journal of Machine Learning Research* 9, no. Sep(2008): 2015-2033.

[19]Bandekar, Jagdeesh. (1992) "Amide modes and protein conformation." *Biochimica et Biophysica Acta (BBA) Protein Structure and Molecular Enzymology* 1120, no. 2: 123-143.

[20]Bandekar, Jagdeesh, and S. Krimm. (1979) "Vibrational analysis of peptides, polypeptides, and proteins: Characteristic amide bands of β-turns." *Proceedings of the National Academy of Sciences* 76, no. 2: 774-777.

[21]Kapoor, Deepti, Navnit K. Misra, Poonam Tandon, and V. D. Gupta. (1998) "Phonon dispersion and heat capacity of poly (l-aspartic acid)." *European polymer journal* 34, no. 12: 1781-1791.

[22]Misra, Navnit Kumar, Deepti Kapoor, Poonam Tandon, and VishwambharDayal Gupta. (1997) "Vibrational dynamics and heat capacity of poly (L-lysine)." *Polymer journal* 29, no. 11: 914.

[23]Misra, N. K., D. Kapoor, P. Tandon, and V. D. Gupta (2000)"Phonon dispersion and heat capacity in cross-β form of poly (O-acetyl, l-serine)." *Polymer* 41, no. 6: 2095-2104.

[24]Rabolt, J. F., W. H. Moore, and S. Krimm. (1977) "Vibrational analysis of peptides, polypeptides, and proteins. 3. α-Poly (l-alanine)." *Macromolecules* 10, no. 5: 1065-1074.

[25]Baran, Enrique J., Inés Viera, and María H. Torre. (2007) "Vibrational spectra of the Cu (II) complexes of L asparagine and L-glutamine." *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 66, no. 1: 114-117.

[26]Moore, Willie H., and Samuel Krimm. (1976) "Vibrational analysis of peptides, polypeptides, and proteins. II. β-Poly (L-alanine) and β-poly (L-alanylglycine)." *Biopolymers* 15, no. 12: 2465-2483.

[27] Dwivedi, Anil M., and S. Krimm. (1982) "Vibrational analysis of peptides, polypeptides, and proteins. X. Poly (glycine I) and its isotopic derivatives." *Macromolecules* 15, no. 1: 177-185.

[28] Sharma, Poonam, Navnit K. Misra, Poonam Tandon, and V. D. Gupta. (2002) "Phonon dispersion in poly (L arginine)." *Journal of Macromolecular Science, Part B* 41, no. 2: 319-340.

[29] Pande, Sapna, Poonam Tandon, and V. D. Gupta. (2002) "Vibrational dynamics and heat capacity of poly (l

ornithine)." *Journal of Macromolecular Science, Part B* 41, no. 1: 117-136.
[30] Jain, Ashika, Radha Mohan Misra, Poonam Tandon, and VishwambharDayal Gupta. (2006) "Vibrational dynamics and heat capacity of syndiotactic poly (methyl methacrylate)." *Journal of Macromolecular Science, Part B: Physics* 45, no. 2: 263-284.
[31]Barth, Andreas. (2007) "Infrared spectroscopy of proteins." *Biochimica et Biophysica Acta (BBA) Bioenergetics* 1767, no. 9: 1073-1101.
[32] Fawcett, Tom. (2006) "An introduction to ROC analysis." *Pattern recognition letters* 27, no. 8: 861-874.