Enhancement of Accuracy of ANN Data Mining Algorithm for Protein Classification Based on Architecture Optimization

[1]Nandika Salwan, [2]Jasmine Kaur
[1]Dept. of CSE, RBCENTW, India
[2]Dept. of CSE, RIMT-IET, India

Abstract
Protein Classification is a very important aspect of Bioinformatics and no one has yet optimized the Architecture of Artificial Network Based Technique for protein classification. So there is a need for changing the architecture of ANN to increase the effectiveness of ANN for protein classification.

Keywords
Proteins, Data Mining, Artificial Neural Network

I. Introduction

Proteins are large, complex molecules that play many critical roles in the body. They do most of the work in cells and are required for the structure, function, and regulation of the body's tissues and organs. Proteins are made up of hundreds or thousands of smaller units called amino acids, which are attached to one another in long chains. There are 20 different types of amino acids that can be combined to make a protein. The sequence of amino acids determines each protein's unique 3-dimensional structure and its specific function.

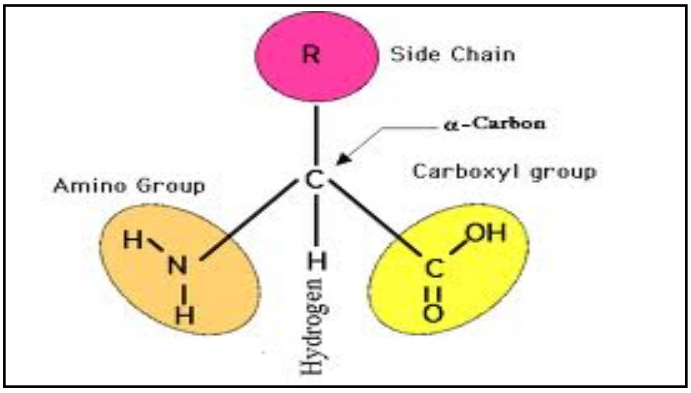

Fig. 1: Structure of Protein

The amino group (-NH_2) of one amino acid is linked with the carboxylic group (- COOH) of the adjacent amino acid to form peptide bond.

II. Classification of Proteins

Enzymes - catalyze chemical and biochemical reactions within living cell and outside. Sucrase and Trypsin are examples of Enzymes.
Hormones - responsible for the regulation of many processes in organisms. Insulin and Growth hormones are examples of Hormones.
Transport Proteins - transport or store some other chemical compounds and ions. Hemoglobin and Lipoproteins are examples of Transport Proteins.
Antibodies/Protection Proteins - involved into immune response of the organism to neutralize large foreign molecules, which can be a part of an infection. Immunoglobulin is an example of Protection Proteins.
Structural proteins- responsible to maintain structures of other biological components, like cells and tissues. Collagen and Keratin are examples of Structural Proteins.
Motor proteins- convert chemical energy into mechanical energy.
Receptors - responsible for signal detection and translation into other type of signal.
Signaling proteins - involved into signaling translation process.
Storage proteins - contain energy, which can be released during metabolism processes in the organism. Casein and Ferritin are examples of Storage Proteins.

III. Structures of Proteins

A. Primary Structure of Protein

The primary structure of proteins refers to the linear number and order of the amino acids present. The convention for the designation of the order of amino acids is that it contains:
N-terminal end that is the end bearing the residue with the free α-amino group and is to the left.
C-terminal end that is the end with the residue containing a free α-carboxyl group and is to the right.

B. Secondary Structure of Protein

The ordered array of amino acids in a protein confer regular conformational forms upon that protein. These conformations constitute the secondary structures of a protein. Proteins fold into two broad classes of structure termed as follows:
Globular Proteins are compactly folded and coiled.
Fibrous Proteins are more filamentous or elongated.

C. Tertiary Structure of Protein

Tertiary structure refers to the complete three-dimensional structure of the polypeptide units of a given protein. It is the spatial relationship of different secondary structures to one another within a polypeptide chain. The interactions of different domains is governed by several forces. These includes:
Hydrogen Bonding
Hydrophobic Interactions
Electrostatic Interactions
Van Der Waals Forces.

D. Quaternary Structures of Protein

The structure formed by monomer-monomer interaction in an oligomeric protein is known as quaternary structure. Oligomeric Protein is composed of multiple identical polypeptide chains or multiple distinct polypeptide chains.
For Example: Hemoglobin

IV. Data Mining

Data mining (sometimes called data or knowledge discovery)

is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining technique is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries.

Data mining consists of five major elements:

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

V. Techniques of Data Mining

A. Classification

Is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines.

B. Clustering

Is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

C. Regression

Attempts to find a function which models the data with the least error.

D. Artificial Neural Networks

An Artificial Neural Network (ANN), usually called neural network (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks.

E. Association Rules

Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

VI. Artificial Neural Network

An Artificial Neural Network (ANN), usually called Neural Network (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data.

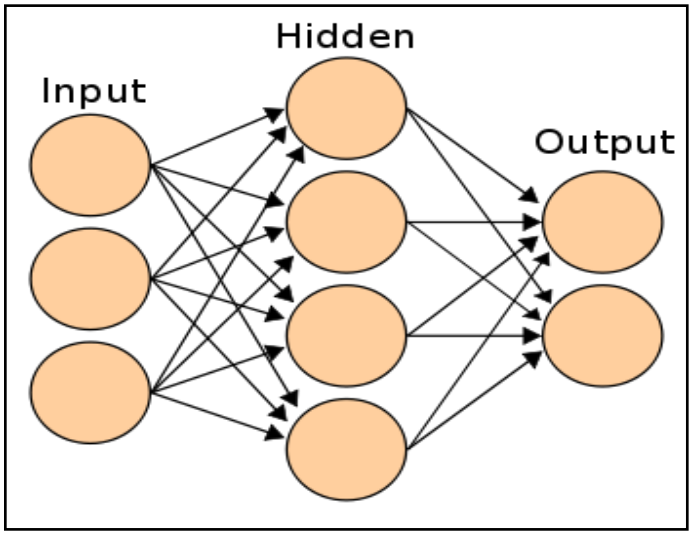

Fig. 2: Working of Neural Network

VII. Objectives

The objectives are:

1. To study the protein classification process in the area of bioinformatics.
2. To Study the metrics required for protein classification two parameters are used i.e speed and accuracy.
3. To develop ANN based system for protein classification.
4. To conduct iterations to improve the architecture of ANN based system for protein classification to improve the accuracy and reduce the time taken for protein classification.

VIII. Conclusion and Future Scope

A. Conclusion

Detection of disease at early stage is more important so that accurate result could be shown. With comparison of different architectures of Neural Network, it is found that we are getting accurate results for curing the disease. As per previous results for curing the disease our results are more accurate and detection is done within short period.

B. Future Scope

The work can be extended in following directions:

1. This work can be extended to solve other diseases.
2. Any other method can also be discovered so that accurate results can be produced.
3. Further investigation can be done so that diseases can be detected within short period.

References

[1] Dianhui Wang; Nung Kion Lee; Dillon, T.S.; Hoogenraad, N.J.,"Protein sequences classification using Radial Basis Function (RBF) neural networks", Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference on, pp. 764- 768, Vol. 2, 18-22 Nov. 2002

[2] Wu, C.H.; Sheng Zhao; Simmons, K.; Shivakumar, S., "Motif neural network design for large-scale protein family identification", Neural Networks,1997. International Conference on, pp. 86-89, Vol.1, 9-12 Jun 1997.

[3] Dianhui Wang; Nung Kion Lee; Dillon, T.S.,"Data mining for building neural protein sequence classification systems with improved performance", Neural Networks, 2003. Proceedings of the International Joint Conference on, pp. 1746- 1751, Vol. 3, 20-24 July 2003.

[4] Sharma, S.; Kumar, V.; Sobha Rani, T.; Durga Bhavani, S.; Bapi Raju, S.,"Application of neural networks for protein sequence classification", Intelligent Sensing and Information Processing, 2004. Proceedings of International Conference on, pp. 325- 328, 2004.

[5] Vipsita, S.; Shee, B.K.; Rath, S.K,"An efficient technique for protein classification using feature extraction by artificial neural networks", India Conference (INDICON), 2010 Annual IEEE, pp. 1-5, 17-19 Dec. 2010

[6] Rossi, A.L.D.; De Oliveira Camargo-Brunetto, M.A.,"Protein Classification Using Artificial Neural Networks with Different Protein Encoding Methods", Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on, pp.169-176, 20-24 Oct. 2007

[7] Sasagawa, F.; Tajima, K.,"Application of a neural network with a modular architecture to prediction of protein secondary structures-overlearning effects on predictions", Neural Networks, 1993. IJCNN '93-Nagoya. Proceedings of 1993 International Joint Conference on, Vol. 1, pp. 1007- 1010 Vol. 1, 25-29 Oct. 1993

[8] Cerri, R.; Barros, R.C.; de Carvalho, A.C.P.L.F.,"Hierarchical multi-label classification for protein function prediction: A local approach based on neural networks", Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on, pp. 337-343, 22-24 Nov. 2011

[9] Wei Zhong; Altun, G.; Hu, H.-J.; Harrison, R.; Tai, P.C.; Yi Pan, "Factoring tertiary classification into binary classification improves neural network for protein secondary structure prediction", Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB '04. Proceedings of the 2004 IEEE Symposium on, pp. 175- 181, 7-8 Oct. 2004

[10] Daugherity, W.C.,"A neural-fuzzy system for the protein folding problem", Industrial Fuzzy Control and Intelligent Systems, 1993., IFIS '93. Third International Conference on, pp. 47-49, 1-3 Dec 1993

[11] Hashemi, H.B.; Shakery, A.; Naeini, M.P.,"Protein Fold Pattern Recognition Using Bayesian Ensemble of RBF Neural Networks," Soft Computing and Pattern Recognition, 2009. SOCPAR '09. International Conference of, pp. 436-441, 4-7 Dec. 2009

[12] Vipsita, S.; Shee, B.K.; Rath, S.K.,"Protein superfamily classification using Kernel Principal Component Analysis and Probabilistic Neural Networks", India Conference (INDICON), 2011 Annual IEEE, pp.1-6, 16-18 Dec. 2011

[13] Vipsita, S.; Rath, S.,"An evolutionary approach for protein classification using feature extraction by artificial neural network", Computer and Communication Technology (ICCCT), 2010 International Conference on, pp. 516-520, 17-19 Sept. 2010.

[14] Yuehui Chen; Xueqin Zhang; Yang, M.Q.; Yang, J.Y.;, "Ensemble of Probabilistic Neural Networks for Protein Fold Recognition", Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on, pp. 66-70, 14-17 Oct. 2007

[15] Zeng, Hanglin; Zhou, Ling; Li, Linjiang Li; Wu, Yongqiang, "An improved prediction of protein secondary structures based on a multi-mold integrated neural network", Natural Computation (ICNC), 2012 Eighth International Conference on, pp. 376-379, 29-31 May 2012.