Study of Protein Structure Using Data Mining - A Survey

[I]Pankaj Kumar Dwivedi, [II]Sachin Dube
[I,II]School of CSE, Galgotias University, Greater Noida, U.P, India

Abstract

Bioinformatics or procedural biology is that field of science which merges or combines the biology, computer science and information technology as a single process. Now-a-days in computation molecular biology, secondary structure prediction of a protein is a major downside. Secondary structure prediction depends on its amino acid sequence. Recent studies likes machine learning techniques for classification and regression task. The researchers used varied data processing and machine learning tool for protein structure prediction. Our work centered on to determine the protein structure of critical diseases using Support Vector Machine and Neural Network that provides us a sample data, helpful in determining the prediction of protein function.

Keywords

Proteins, Macro-molecule Protein Structure, Protein Function Prediction, Support vector machine, Neural Network.

I. Introduction

A. Protein

Proteins are the foremost and important building blocks of everyone's life. They are accountable for catalyzing and governing all biochemical reactions, moving molecules. Thus, form the core structure for skin, hair and tendon. Collection of twenty amino acids shaped the structure of proteins and function is nearly related to its structure.[2] Several functions of proteins includes advocating the body from germs, providing support, movement of muscles etc. One of the main function of protein is that of Enzymes, which conciliate the chemical reactions occurring in the body. They do not change the reaction rather they speed up them. Proteins are composed of several structures named as Primary Structure, Secondary Structure, Tertiary Structure and Quaternary Structure. The Primary Structure defines the arrangement of amino acids sequence, the secondary structure describes the native folding conformities maintained by H-bond and is classified within three classes as followed – alpha-helicies, beta sheets or strands and coil. [3]. To have the knowledge about the function of proteins, one should know about its structure first. Today Protein structure prediction (PSP) from chain of amino acid series is one of the most important research area of bioinformatics. This is because of the fact that the biological function of the protein is decided by its three dimensional view. Amino acids are differentiated into four groups Polar, Non- polar, Basic, Acidic. Non-Polar are again grouped under Hydrophobic (attracted towards water) and Hydrophilic (repelled by water) [1, 2].

Proteins may also be classified into 3 main classes, that are related with the typical 3D view structure of protein: Spherical Proteins or Ball Shaped Proteins, Fibrous Proteins and Membrane Proteins or Thin Layer Proteins. Approximately entire ball-shaped proteins are able to dissolve in a liquid and plenty of are biological catalyst. Fibrous Proteins are responsible for structural functions in body. The cells of skin, tendons, hairs etc depends on the molecules of this protein. Membrane proteins gain interactions with the cell or biological membranes. They functions as the receptors and proposed channels for charged molecules so that they can easily pass through cell membranes. Proteins don't seem to be perfectly strong molecules. Moreover to the levels of structure of proteins ,it may shift within some connected structures, where they accomplish the functions. Within the context of these useful shifts the tertiary and quaternary structures are referred to conformations or agreements and transformations within them are known as conformational or agreemental changes.

B. Protein Structure

Proteins or polypeptides are large, organic molecules and are among the foremost necessary elements within the cells of living organisms. They're a lot of diverse in structure and function than the other kind of molecule. It will act as Biological catalyst, anti-bodies, Hormones, Transport Molecules, hair, skin, muscle, cartilage, claws, nails, horns, feathers etc. Super-molecule structure features primarily four levels of category: Primary, secondary, tertiary and Quaternary protein structure. Fig. 1 shows different levels of protein structure.

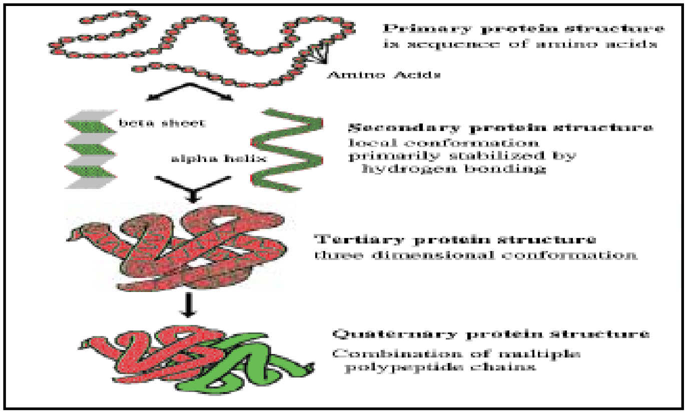
Primary protein structure
is sequence of amino acids
Amino Acids
beta sheet
alpha helix
Secondary protein structure
local conformation
primarily stabilized by
hydrogen bonding
Tertiary protein structure
three dimensional conformation
Quaternary protein structure
Combination of multiple
polypeptide chains

Fig.1 :

Protein development passes through diverse levels of structures [4]. The basic or primary structure of a protein is the ordered series, of the amino acid residues. Each and every alpha amino acids embodies a backbone part which is present in all types of amino acids, and a side chain that is matchless to each and every residue. The secondary macromolecule structure containing the H-bond is most important for understanding and foreshadowing the further tertiary structure when structure breakthrough is without series similarity in datasets. In the tertiary structure, alpha-helix and beta sheets or strands from a secondary structure are gathered into a stabilize globule. In contrariety to the second structure of protein, the third structure is firmed by hydrophobic interactions within the non-polar aspect chains.[5] Several proteins are formed by combining more than one polypeptide chain, hence forming the complex quaternary structure. Hemoglobin is one of the example of this complex structure.[5]

C. Protein Function Prediction

Protein also known as biological macromolecules are accountable for various tasks carried out in all living organisms. These macromolecules play a vital role in the function and structure of building blocks of life i.e. cells.[9]

The necessary function a protein may carry out involves the structural support, protecting body against germs and diseases, speeding up the biochemical reactions, squeezing of muscles etc. Protein react to their functions, according to the environment in which they are in.[9] Proteins are helpful in many ways at various levels. They can also useful in identifying the diseases which occurred due to their modification and mutations. Protein performance prediction is of great importance from different viewpoint of biology.[9] There are variety of machine and biological challenges that build super molecule of protein function prediction troublesome. If we talk in the reference of biology, then function is ascertain in the reference of organism and it barely certified by a single experiment. Some experiments do not able to reflect the activity of proteins. Thus we came across many challenges in prediction of protein function as well as its structure. The function prediction holds a significant place in all the research area of protein . It not only proved beneficial to biological areas but also have remarkable impact on other fields such as computational and statistical areas. To understand the function of protein one should also know about its structure because proteins are capable to accomplish their tasks only by winding up their amino acids sequences.[7]

II. Dataset

A key to grasp the function of biological
Macromolecules, e.g., proteins, is the determination or prediction of its structure. Large-scale gene-sequencing projects accumulate a vast variety of putative protein sequences [6]. However, information regarding three-dimensional structures is available for only little fractions of notable proteins. So experimental structure prediction has improved. This creates a necessity for extracting structural information from sequence databases. To facilitate the requirement various protein databases are available online. Following is the details of three freelance and identical protein databases that are used for our study. Protein Data Bank (PDB) – Protein data bank is basis of Structural Classification of Protein (SCOP) database, which is a publicly accessible over the web. All the chains available from PDB are compared with each other using the Basic Local Alignment Search Tool (BLAST) algorithm as enforced within the National Centre for Biotechnology Information (NCBI) toolkit library.[7]

III. Methodology

A. Support Vector Machine (SVM)

SVM is about novel learning techniques which is acquaint in the environment of Structural Risk Minimization (also called as SRM) inductive principle and in the theory of Vapnik Chervonenski (known as VP). The kernel function creation is one of the meaningful step in SVM classification system. When there is a linearly separable data, linear kernel is the best choice one can made. It does not advocate any need to map data instances into a high dimensional area. On the other hand, non-separable data categorization uses the polynomial kernel. SVM includes number of properties which can be seen as – effectively avoiding the over-fitting, features of high dimensional areas can be easily handled etc.[7] [10] It is also seen as a group of supervised learning procedure which can be used for many purposes such as regression, classification and outliers detection. The principle of SVM works as follows, it takes a collection of input information and then forecast the possible classes for that input information. This makes SVM a non-probabilistic linear classifier. A SVM algorithm builds a prototype which assigns values to each and every class of the input data (fig.2). With the help of this separation of classes, there will be a transparent gap between the possible category and due to this , mapping of new training examples can also be done with ease.[10]

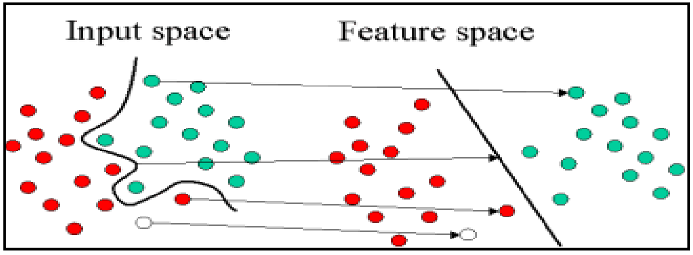

Fig. 2 :

Vapnik, Burgus in 1998 developed the SVM to solve the various problems of the classification. Their development is based on the statistical learning theory. SVM can be seen a new toolbox of data mining approach. It gains attraction from many areas due to its speculative functions. It shows remarkable work in protein structure prediction also.The traditional machine learning approach were not able to perform many functions of bioinformatics applications, thus the introduction of SVM proved beneficial to various areas such as pattern recognition, protein prediction interactions etc.

B. Artificial Neural Network

Neural Networks can be defined at various levels. At one level we can say that class of mathematical algorithm constitutes the neural networks, because network can sometimes be seen as graphical notations. At another level, it can be seen as moulded or synthetic biological networks. Nowadays, we have very few knowledge of these neural networks in organisms. Therefore, here the most suitable approach to understand the neural network is one with the algorithms.[11]

Many researchers , practioners, engineers etc have keen interest in the structure of neural networks to come across new innovative ideas.[11]

Artificial Neural Networks(ANN) is a term of biology but it can be seen in the close relation of real neural system. The architecture of neural and artificial neural networks shows a huge difference. To have the information about neural networks , we should have the knowledge of actual brain functions but this knowledge is very limited so far. We do not have proper models which defines the successful functions of brains same as it works. So, brain is still a metaphor for almost all neural networks.[11]

There is a very little correspondence between the natural and artificial neural networks or systems, but it is necessary to first understand about biological neurons. Human brains incorporates 1011 (approx.) computing elements which is known as "Neurons". The communication between neurons is carried by a connected network of axons and synapses.[11]

IV. Conclusion

In this paper we proposed a method to predict the disease by using the protein structure. In this method we basically dependent on sample data, which is collected from protein data bank (PDB) and using support vector machine (SVM) and artificial neural network, we can prediction of the protein function.

References

[1] *C.N. Rokde and .M. K shirsagar, "Bioinformatics: Protein structure prediction," 4th International Conference on Computing Communication and Networking Technologies". Nagpur, Maharashtra, July2013.*

[2] *M. Taufer, C. An, A. Kerstens, and C.L .Brooks III, "Predictor@ home: A 'Protein structure prediction supercomputer' based on Public-resource computing," Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium., California, US, IEEE, 2005, pp 1530-2070.*

[3] *Arvind kumar Tiwari and R.B Mishra, "Protein Function Prediction Using Support Vector Machine," International Journal of Computational Bioinformatics and In Silico Modeling., Vol. 2, No. 5 (2013): 239-244.*

[4] *Marco Vassura, Luciano Margara, Pietro Di Lena, Filippo Medri, Piero Fariselli, and Rita Casadio," Reconstruction of 3D Structures from Protein Contact Maps", VOL. 5, NO. 3, JULY-SEPTEMBER 2008.*

[5] *Chandrayani N.Rokde and Dr. Manali Kshirsagar, "Bioinformatics: Protein Structure Prediction," International Conference on Computing, Communications and Networking Technologies., Tiruchengode, India, July 4-6 2013.*

[6] *Jung-Ying Wang, "Application of Support Vector Machines in Bioinformatics,"2002.*

[7] *Protein Secondary Structure Prediction Using Support Vector Machines (SVMs) 2013 International Conference on Machine Intelligence Research and Advancement, 2013*

[8] *Predrag Radivojac, Indiana University, A (not so) Quick Introduction to Protein Function Prediction, September 27, 2013*

[9] *Jieyue He, Hae-Jin Hu , Robert Harrison, Phang C. Tai and Yi Pan, "Transmembrane segments prediction and understanding using support vector machine and decision tree," Expert Systems with Applications, pp 64-72, 2006.*

[10] *Introduction to Artificial Neural Systems, Jacek M. Zurada, pp 25-26.*