Chapter 54
Open-Source Essential Protein Prediction Model by Integrating Chi-Square and Support Vector Machine

S. R. Mani Sekhar
REVA University, India & Ramaiah Institute of Technology, India

Siddesh G. M.
Ramaiah Institute of Technology, India

Sunilkumar S. Manvi
REVA University, India

ABSTRACT

Identification and analysis of protein play a vital role in drug design and disease prediction. There are several open-source applications that have been developed for identifying essential proteins which are based on biological or topological features. These techniques infer the possibility of proteins to be essential by using the network topology and feature selection, which can ignore some of the features to reduce the complexity and, subsequently, results in less accuracy. In the paper, the authors have used selenium driver to scrap the dataset. Later, the authors integrated the chi-square method with support vector machine for the prediction of essential proteins in baker yeast. Here, chi-square is a test of dissimilarity used for altering the record, and afterward, the support vector machine is used to classify the test dataset. The results show that the proposed model Chi-SVM model achieves an accuracy of 99.56%, whereas BC and CC achieved an accuracy of 84.0% and 86.0%. Finally, the proposed model is validated using Statistical performance measures such as PPA, NPA, SA, and STA.

DOI: 10.4018/979-8-3693-3026-5.ch054

1. INTRODUCTION

In the current computational domain, proteomics act as a sub-branch of bioinformatics, which studies the structure and functions of proteins. It also helps in identifying the procedure and encoding order used in the analysis of the large scale data set. Free and open-source software (FOSS) allows the developer to use, change, and distribute software without significant limitations. However, it also provides a complete software package to users freely. Now a day's requirement of open source software in bioinformatics and chem-informatics etc is increasing rapidly. cPath (Cerami, Bader, Gross, & Sander, 2006) an open-source software contains a web interface for gathering, storing, and querying bio-informatics data. It also supports the build-in mapping service and helps to link to external resources (Breitwieser & Colinge, 2013). Developed a tool to analysis proteins localization and processing of Tandem Mass Tags peptides. In addition it also supports statistical computation with user friendly reports generation. For study of protein function and sequence (Deng, Yuan, Huang, & Wang, 2013) created a SFAPS module by using spectrum function. This function helps in converting numerical sequences into the characteristic frequency of the protein interface. Subsequently helps in analyze sequence of proteins. ATGme (Daniel et al., 2015) designed an open-source web-based tool for optimizing the protein sequence by classifies organisms into rare and very high rare codons. After that, help the user to generate an individual custom optimized DNA sequence.

Proteins are the best adaptable macromolecules in living creatures and also play important roles in biological procedures. They are made of individual or several chains of amino acid residues. An amino acid is made up of an organic composite that has an essential carbon atom, called an alpha carbon (Cα). The bond is consists of four valences with a link to the carboxylic group, hydrogen atom, an amino group, and an adjacent chain. The amino acids are the strong structured chunks of protein. This sequence of amino acids in a given polymer helps in identifying protein structure and protein function. It can also be called as a polypeptide, though a small chain of amino acids. This chained bond starts with the amino cluster and ends with the carboxyl cluster. According to Bioinformatics researcher DNA, RNA, Protein, and other molecules do not work separately. They can be linked together to perform a specific biological action. Interaction is a process when two molecules linked to perform a certain function. These type of interaction are mainly divided into four groups by considering their molecule types as shown below

- *Protein-protein interactions*: collaboration between proteins to derive biological procedures.
- *Gene regulatory interactions*: interaction of genetic data to normalize protein expression level.
- *Metabolic interactions*: create support between enzyme proteins and helps in changing a substrate cell into product cell through numerous catalysis responses.
- *RNA-DNA interactions:* provide collaboration among RNA-RNA or RNA-DNA connections and also plays a vital role in critical diseases.

PPI performance numerous functions in a different cell, include metabolic pathways, DNA, replication, transduction, and transcription etc. In a PPI network, a vertical position signifies the proteins, and the linking line shows the communication link between proteins. This network can be noticed as a graph. Thus the closely connected protein clusters can be computed efficiently.

Essential Proteins are made up of a set of minimal genome, which can support cell survival with basic requirements. A living creature cannot breed and live without them (Kamath et al., 2003; Wang, X Peng, W Peng, & Wu, 2014). Proteins are mostly classified into two types Essential proteins and non-essential