Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS
ISSN 2454-6348

Original Review Article DOI: 10.26479/2018.0404.22

A SURVEY ON PROTEIN FUNCTION PREDICTION: COMPUTATIONAL METHODS AND TOOLS

R. Ranjani Rani, D. Ramyachitra*

Department of Computer Science, Bharathiar University, Coimbatore, India.

ABSTRACT: In computational biology, protein function prediction is considered to be one of the painstaking challenges. Rapid development in the high-throughput protein data leads to large amount of uncharacterized function of a protein. These overwhelming bulk of protein data is considered to be a difficult task for identifying the function of a protein using an experimental study. Thus, the function of a protein has been determined using the computational intelligence techniques. The prediction of a protein functionality can be identified using various ways such as homology sequence, structure, interaction networks, phylogenetics and gene expression data. This paper summarizes about the review of recent computational techniques along with optimistic datasets and tools for predicting protein function with their performance measure. Also, compared the performance of heterogenous dataset for predicting protein function.

KEYWORDS*:* Protein Function Prediction; Sequence; Structure; interaction networks; gene expression; phylogenetics.

Corresponding Author: Dr. D. Ramyachitra* Ph.D.

Department of Computer Science, Bharathiar University, Coimbatore, India.

Email Address: jaichitra1@yahoo.co.in

1. INTRODUCTION

The most important class of biomolecule is represented as protein in living organisms. Proteins are composed of amino acids. It is a macromolecule that helps in functioning of cells and it performs specific function in the body. Every single protein is virtually involved in metabolism, body movement and structural support. It plays a major role in biotechnology as well as in medicine with respect to the development of new drugs, crops and biochemical's such as bio fuels [1]. The biological or biochemical role of a protein can be assigned using protein function prediction. The function of a protein can be identified using the sequence and structure of a protein. The prediction can be done based on fold similarity, three dimensional templates, gene sequence data, etc. Heterogeneous data has also been used in computational method to predict function of a protein [2]. Due to numerous high throughput empirical measures, a huge volume of protein sequences, structures, gene expressions, protein interaction network are created. The difference between the unannotated and annotated protein functions are huge in volume. The prediction of protein function in laboratory needs an enormous labor effort and time to analyze a solitary protein or gene. So, to eradicate this disadvantage, many computational tools and techniques have been established to predict for protein function from various heterogeneous data like protein sequence, structure etc. In this paper, the review of various computational approaches for predicting protein function has been discussed. Also, the various heterogenous benchmark datasets and computational tools used for predicting the function of a protein has been discussed. The remaining sections of a paper are ordered as follows: Section 2 defines the introduction about protein function prediction. Section 3 discusses about the various heterogenous databases used for predicting the protein function. Section 4 lists the various computational tools and methods for predicting the function of a protein. Section 5 evaluates and compares the top 5 computational tools for protein function prediction and finally the concluding observations of this review are projected.

2. PROTEIN FUNCTION PREDICTION

In common, the protein function prediction is defined as the prediction of assigning a biochemical or biological character to the indigently studied proteins. They are predicted based on the previously studied heterogenous data such as protein sequence, structure, interaction network, gene expression, text mining, genomic data etc. The protein function can be described from various degrees such as phenotypical and physiological degrees. To obtain all different features of degrees, the Gene Ontology Consortium delivers three different classification of functions such as cellular component, biological process and molecular function. The cellular component defines the position of the structural component in which the gene activates. The biological process acquires the functional definition of protein function and permits to specify the processes of the gene in a cell. The molecular function describes the gene product that involved in the cell. There are diverse categories of data that can make predictions of protein function such as protein sequence, structure, interaction network, genomic data, phylogenomic data, gene expression data and text mining in literature data. The classification of heterogenous data employed for protein function prediction is given in Fig 1.

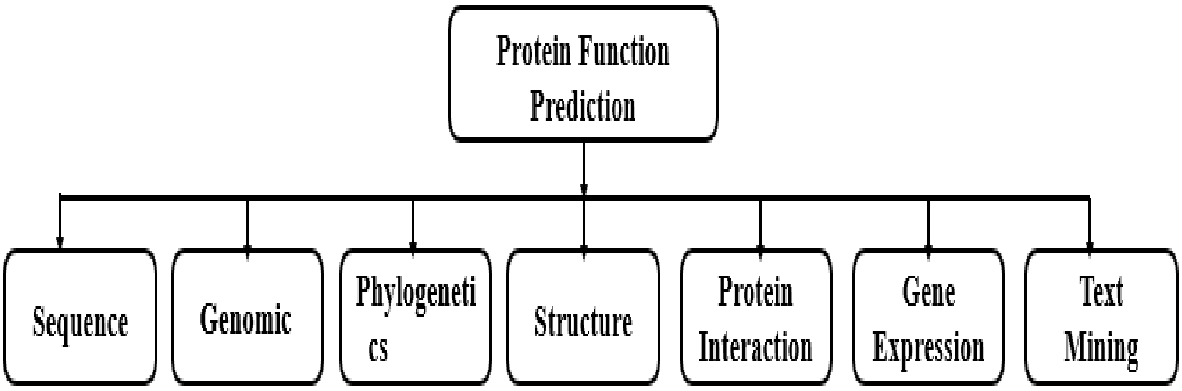

Fig.1. Classification of Protein Function Prediction Techniques

a) Function Prediction using Protein Sequence

The protein function can be predicted based on homology detection using the protein sequences. Initially, a huge number of unannotated protein sequences are existing in huge volume of data. Protein sequences are the collection of amino acids in which they predict the function of a protein by identifying the common residues with same function. The sub classification of sequence-based approaches is depicted in Fig 2.

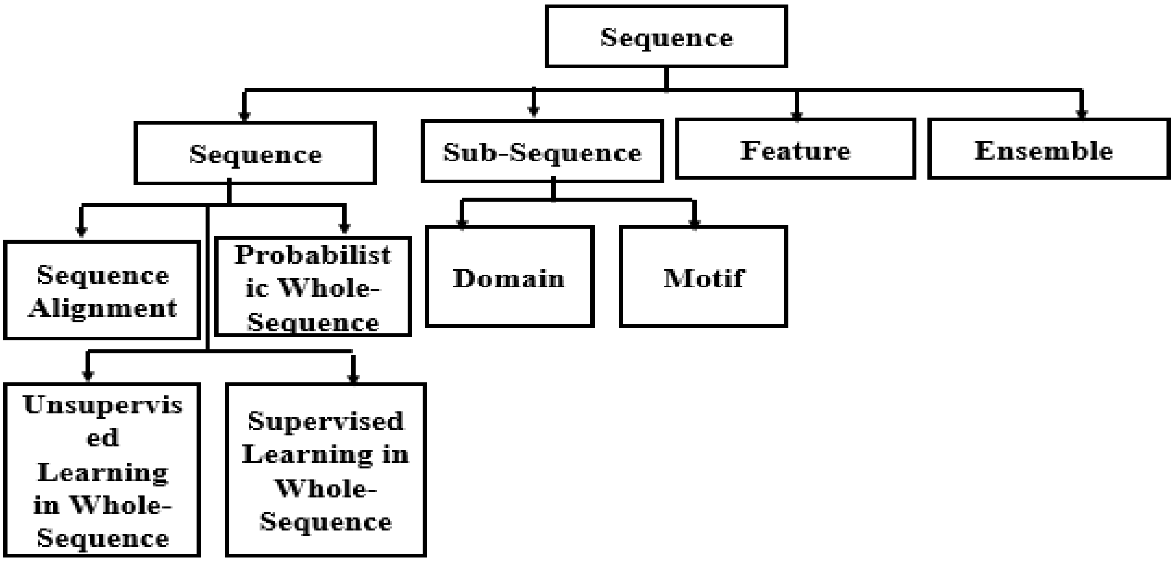
Sequence
Sequence
Sub-Sequence
Feature
Ensemble
Sequence Alignment
Probabilistic Whole-Sequence
Domain
Motif
Unsupervised Learning in Whole-Sequence
Supervised Learning in Whole-Sequence

Fig.2. Classification of Protein Function Prediction using Sequences

There are large number of computational methods are available to predict the function of a protein using the primary sequences. Some of the recent techniques of function prediction depending on the sequences are GOFDR [3] that detects the functional discriminating residues, ProLanGO [4] translates the function issue to language conversion issue by converting the “ProLan” to “GOLan” by employing neural networks. DeepGO [5] employs deep learning approach to study features from sequences and interaction network.

b) Function Prediction using Genomic Context

The growing amount of whole sequenced genomes has allowed prediction of gene function by use of complete genome assessment. The whole genome data provides the better understanding of the biological process of the protein function. Not only they provide the gene function but also the gene fusion, neighborhood of gene and the co-expression of genes across whole genome. They are also used to identify any undiscovered pathways and function of a protein. The classification of genomic context approach is depicted in Fig 3.

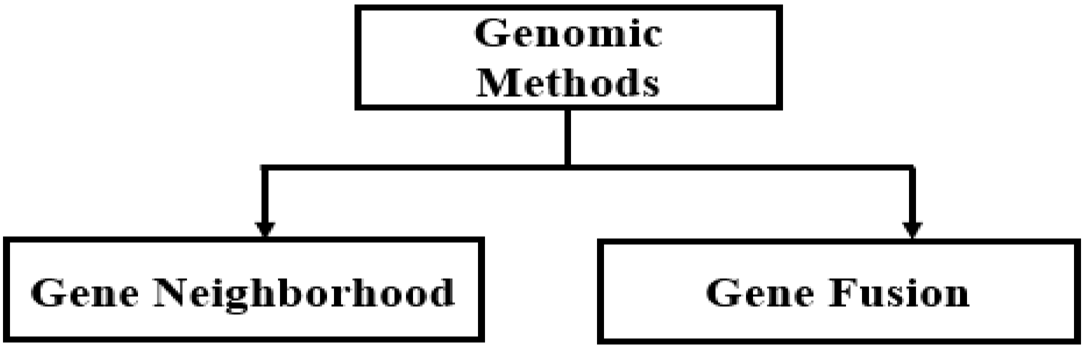

Fig. 3. Classification of Protein Function Prediction using Genomic Context

Some of the available literature reviews for predicting function of a protein using genomic context are SynFPS [6] that collects the large weakly associated genomes into various clusters depending on their gene distribution. Then from every cluster the data is extracted and given into the SVM technique that predicts the protein function. VISTA [7] was employed to align the large sequences and from that the protein functional annotations are identified. Various qualitative and quantitative evaluation and corresponding interferences are obtained [8]. Numerous unidentified gene neighbourhood and function of a genome are predicted. Various other available approaches in the field of genomic context is also provided [9].

c) Function Prediction using Phylogenomics

In common, the homology-based function prediction approaches are defined by employing the similarity of the sequences. But, some of the non-homologous proteins may have similar functional annotations. Thus, the Phylogenomics approach is used for predicting the functional annotations among genes of non-homologous by comparing their phylogenetic distribution. The function of a protein can be found using phylogenetic tree, profile and also by combining them together. The classification of Phylogenomics to predict the protein function is given in Fig 4.

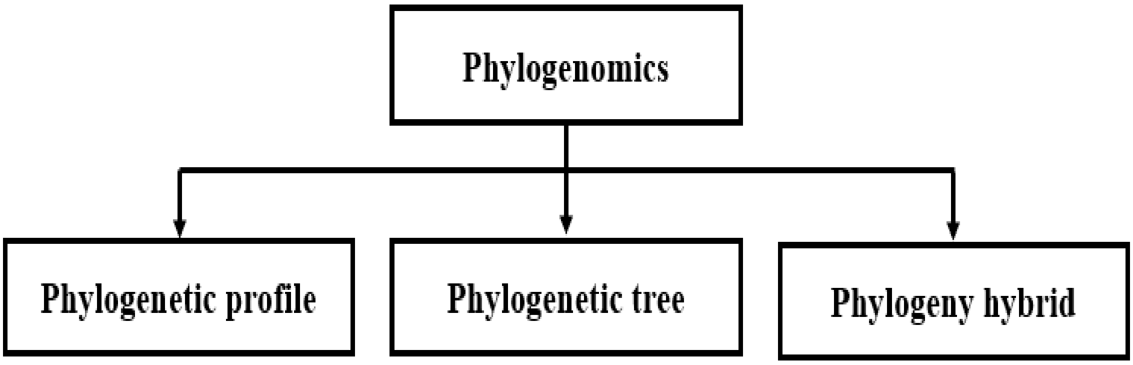

Fig.4. Classification of Protein Function Prediction using Phylogenomics

Some of the existing literature reviews on function prediction using Phylogenomics are SVD-Phy [10] employs the truncated singular score decomposition to resolve the issue of unproductive profiles by solving the false positive predictions. Gene3D phylo tuner [11] approach is implemented on eukaryotic genomes whose phylogenetic distribution information is very weak. In that they detected many novel functional annotations of genomes that was not predicted by any other methods. SIFTER (Statistical Inference of Function Through Evolutionary Relationships) [12] method employs the statistical graph model to calculate the probabilities of unannotated proteins in Phylogenomics.

d) Function Prediction using Protein Structure

Due to the reason of structural genomics, the many protein structures have been evolved without knowing its functional annotations. Many of the proteins that have similar protein functions does not seems to be homologous in nature. Thus, the prediction of protein function using sequence homology-based approaches is not sufficient. Hence, the protein structure is employed to predict the accurate function of a protein using its three-dimensional spatial coordinates of its residues. The classification of solving function prediction using protein structure is given in Fig 5.

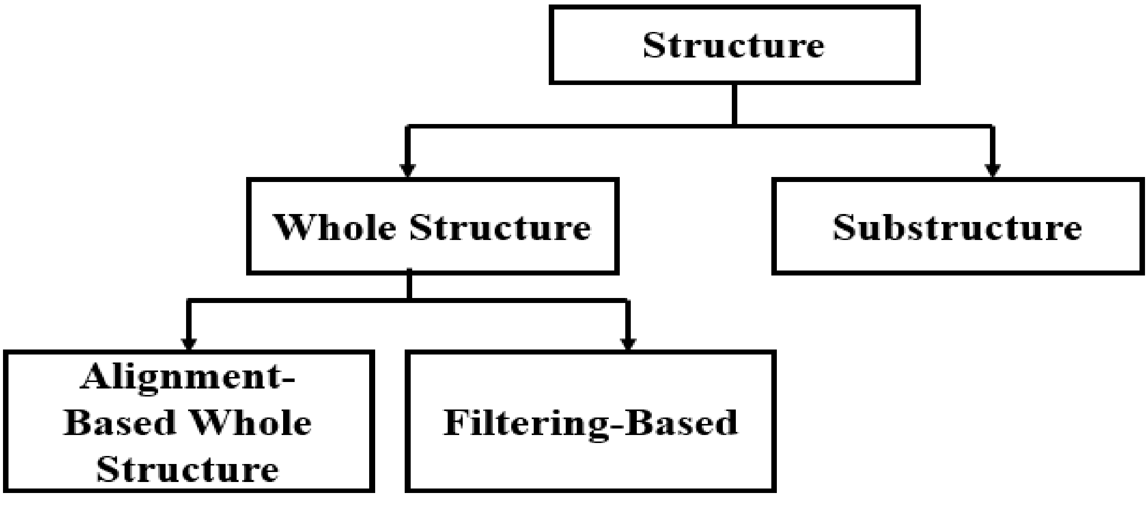

Fig. 5. Classification of Protein Function Prediction using Protein Structure

The various existing methodologies employed for protein function prediction by employing protein structure are COFACTOR [13] structurally includes the low-resolution structural models using BioLIP library. A novel protein structure illustration was employed [14] to predict the accurate function of a protein and classified the functions based on KNN, Random Forest and Naive Bayes approaches. A neural network approach has been employed [15] to classify the protein function by using the structural features and parameters.

e) Function Prediction using Protein Interction Networks

Almost in many situations an isolated protein cannot accomplish its function. Generally, in order achieve the function of a protein it may interact with other neighbor proteins and it is denoted as protein-protein interaction (PPI) network. The general method to imagine the PPI network is as an undirected graph in which the proteins are denoted as nodes and interaction between proteins are denoted as edges of the graph. The methods to predict the protein function using PPI networks are classified into four groups and they are given below Fig 6.

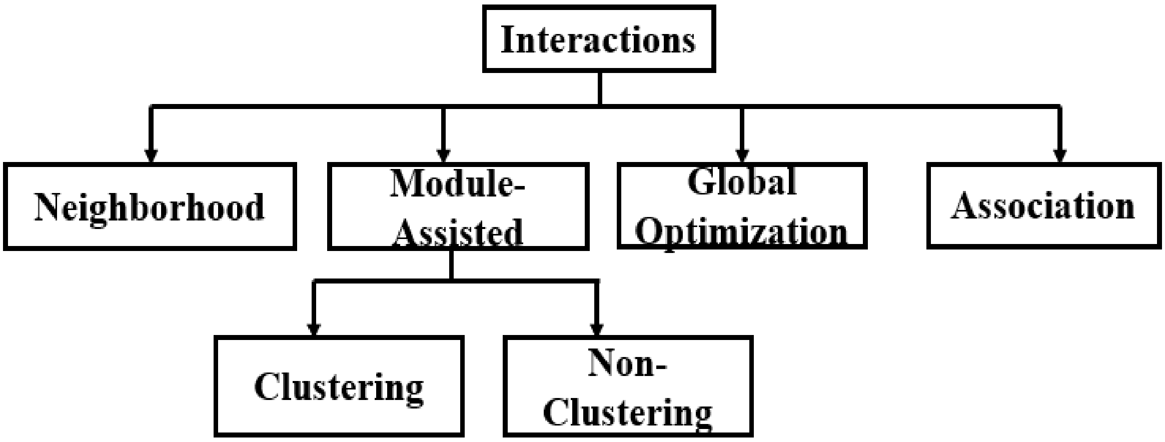

Fig. 6. Classification of Protein Function Prediction using Protein Interaction Networks.

Some of the recent literatures for predicting protein function based on PPI networks are given here. In [16], the function is predicted by using the global information of PPI networks and it studies up to 5 functional levels. In [17] the authors projected the two approaches to detect protein function are using protein domain and protein complexes. From the results, it is proved that using protein complexes provides better results. In [18], the protein function is predicted without using the similarity measures instead they employ Guilty by the association technique. To reduce the number of false rate, the dynamic PPI network has been constructed by using the static PPI and gene expression data for protein function prediction [19]. To improve the biological significance of protein function the multilayer protein networks (FP-MPN) is built by combining many interaction networks and they reduce the noise from interaction networks [2].

f) Function Prediction using Gene Expression

The beginning of microarray technologies has allowed the differential expressions of thousands of genes under numerous empirical situations. Due to this high throughput microarray experiments, many researches are taken place at the system level instead of molecular level thus leads to the accurate prediction of functional annotations. The prediction of gene functional using the gene expression data leads to the accurate prediction. The functional annotations can be predicted using the clustering, classification and temporal analysis and it is depicted in following Fig 7.

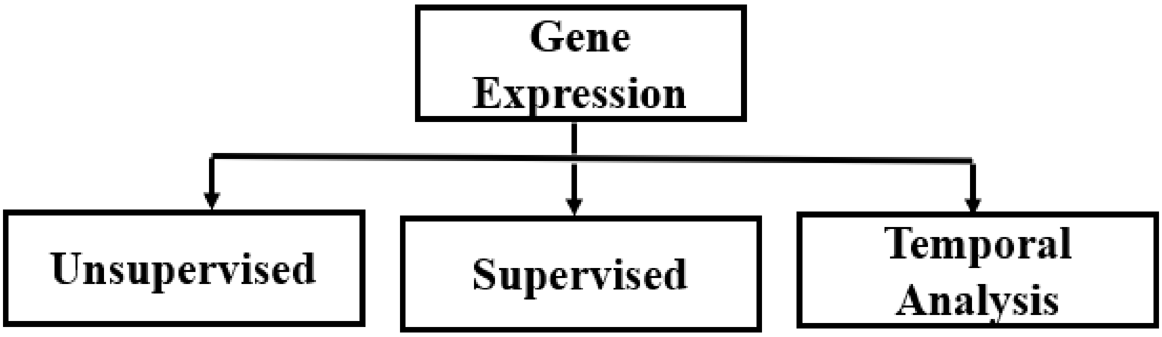

Fig.7. Classification of Protein Function Prediction using Gene Expression

Some of the existing literature review of the protein function prediction using gene expression are FNC [20] is termed as fuzzy nearest clusters. It has two phases, initially the hierarchical clustering is done to detect the homogenous expressed genes and in the second phase, the classification is done to predict the biological role depending on their similarities. A hypergraph is constructed using gene expression data [21], and Laplacian based semi-supervised learning approaches are employed on the hypergraph to predict the function of a protein. VIRGO [22] (ViRtual Gene Ontology) approach built a functional linkage network by employing the gene expression and interaction data, then it names the genes in the constructed network with their functional annotations in the GO and finally scientifically broadcasts these names across the network to exactly guess the functions of unnamed genes.

g) Function Prediction using Text Mining

In recent times, many biological information is identified and published in the reputed scientific journals. The prediction of the characteristics of biological data automatically from the literature using text mining becomes an interesting research area in bioinformatics. Text based parameters are also becoming successful for predicting the protein function prediction. Few of the recent literature reviews for function prediction using text mining are SMISS [23] that projected the statistical multiple integrative scoring system (SMISS). This score integrated the 3 various probabilistic values of protein sequence, functions and interaction networks. LEAP-FS [24] (Literature Enhanced Automated Prediction of Functional Sites) approach enhances the prediction using the text mining and structural analysis techniques. The text-based approach provides the exact residue of functional annotation which is same in the functional site of a protein.

3. LIST OF COMMON DATABASES USED FOR PROTEIN FUNCTION PREDICTION

Numerous protein sequence, structure, interaction and genome databases have developed from a universal effort to curate the evidence on protein information. For managing huge volume of protein data in bioinformatics, a detailed understanding of the design and information of these databases is mandatory. The list of popular and common databases of protein sequences, structures, interaction network and genome used for predicting protein function are given below. The various heterogeneous databases are displayed at its specific type of data in depicted in Table 1.

Table 1: List of Common Databases used for Protein Function Prediction

Sr.No	Data Type	Database Name
1	Sequence	UNIPROT (Universal Protein)
		NCBI (National Centre for Biotechnology Information) all
		SWISS-PROT
		Pfam (Protein Families)
		TrEMBL (TRanslations of European Molecular Biology Laboratory)
2	Interaction Network	BioGrid (Biological General Repository for Interaction Datasets)
		DIP (Database of Interacting Proteins)
		MINT (Molecular INTeraction database)
		BIND (Biomolecular Interaction Network Database)
		STRING
3	Structure	SCOP (Structural Classification of Proteins)
		CATH
		PDB (Protein Data Bank)
4	Genome	Gene Ontology Consortium
		Reactome
		KEGG Pathway (Kyoto Encyclopedia of Genes and Genomes)

4. LIST OF COMMON COMPUTATIONAL TOOLS FOR PROTEIN FUNCTION PREDICTION

The prediction of protein function in laboratory needs an enormous labor effort and time to analyze a solitary protein or gene. So, to eradicate this disadvantage, many computational tools and techniques have been developed to predict the function of a protein from various heterogeneous data like protein sequence, structure etc. The list of various common computational tools is given in Table 2.

Table 2: List of common Computational Tools for Protein Function Prediction

Sr.NO	Tools
1	SVM-Prot
2	InterPro
3.	FFAS-3D (Fold and Function Assignment System)
4	COFACTOR
5	PANNZER (Protein ANNotation with Z-score)
6	EFICAz (Enzyme Functional Inference by a Combined Approach)
7	SIFTER (Statistical Inference of Function Through Evolutionary Relationships)
8	PFP/ESG (Protein Function Prediction/ Extended Similarity Group)
9	DeepGO
10	PANTHER (Protein Analysis THrough Evolutionary Relationships)
11	GOStruct
12	CombFunc
13	MS-kNN (Multi-Source k-Nearest Neighbor)
14	HHblits (HMM-HMM–based lightning-fast iterative sequence search)
15	GPCRpred

5. COMPARISON OF VARIOUS HETEROGENEOUS DATASET FOR PROTEIN FUNCTION PREDICTION USING VARIOUS COMPUTATIONAL TOOLS

The goal of this review is to know the better accurate results of predicting protein function from various heterogenous data. It is not easy to conclude which method can provide better performance and the results are based on the system configuration and dataset structure. Here, the saccharomyces cerevisiae organism is taken as a common input from sequence, structure, interaction network, genome, phylogenetic, gene expression and text mining database. After that the saccharomyces cerevisiae organism is given as an input for heterogenous computational tools such as GOFDR (Sequence) [3], COFACTOR (Structure) [13], FP-MPN (interaction network) [2], SynFPS (genome) [6], SVD-Phy (phylogenetic) [10], FNC (Gene Expression) [20] and SMISS (Text mining) [23]. From the experiment it is observed that the FP-MPN approach has highest accuracy values when compared to others. Also, it is observed that when the protein interaction is given as an input as they are interacting with other similar proteins, the possibility of achieving the better protein function prediction accuracy is more. The Comparison of prediction accuracy with heterogeneous dataset is depicted in Fig 8.

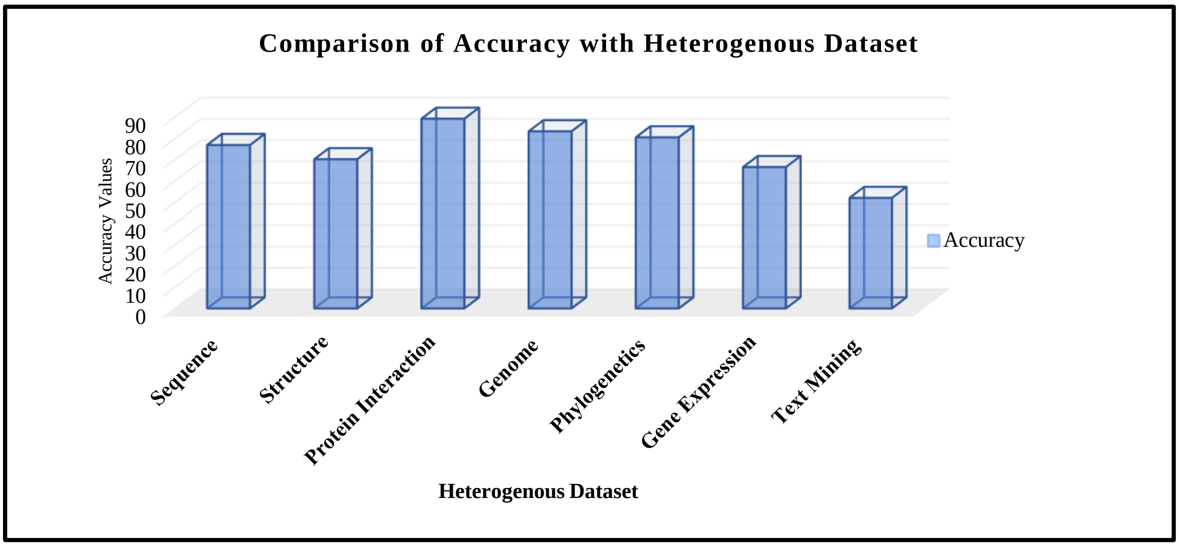

Fig. 8. Comparison of Accuracy of Various Computational Tools with Heterogeneous Data

6. CONCLUSION

The prediction of protein function is a vital task in bioinformatics that provides the molecular function, biological process and cellular component of an unannotated protein by gene ontology consortium. The goal of this review is to analyze the various classification of heterogenous data to predict the function of a protein. The various types of databases and the popular computational tools used for function prediction is listed out. Also, the computational analyses of various heterogenous approaches for function prediction is carried out on saccharomyces cerevisiae organism. From the obtained results, the significance performance was achieved by the interaction network data. Due to various approaches for protein function prediction, it is not easy to accomplish, whether certain approaches are better than remaining approaches from the literatures as they are assessed on different parameter setting. Thus, the performance of function prediction of a protein is based on the type of dataset and the type of parameters used. This review is useful for the beginners who study about the computational approach of protein function prediction.

CONFLICT OF INTEREST

There are no conflicts of interest.

REFERENCES

1. Pandey G, Kumar V, Steinbach M. Computational Approaches for Protein Function Prediction: A Survey. In Proceedings of Computational AF.TR 06-028, Department of Computer Science and Engineering, University of Minnesota, Twin Cities. 2007; 3–105.
2. Zhao B, Hu S, Li X, Zhang F, Tian Q, Ni W. An Efficient Method For Protein Function Annotation Based On Multilayer Protein Networks. Hum Gen. 2016.10(33).
3. Gonga Q, Ninga W, Tian W. GoFDR: A sequence alignment based method for predicting protein functions. Methods. 2016. 93; 3-14.

4. Cao R, Freitas C, Chan L, Sun M, Jiang H, Chen Z. ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. Molecules. 2017. 22(1732).
5. Kulmanov M, Khan MA, Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2017. 34(4); 660-668.
6. Li J, Halgamuge SK, Kells CI, Tang SL. Gene function prediction based on genomic context clustering and discriminative learning: an application to bacteriophages. BMC Bioinformatics. 2007. 8.
7. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids, Res. 2004. 32; W273-W279.
8. Huynen M, Snel B, Lathe W III, Bork P. Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. Gen Res. 2000. 10 (8); 1204-1210.
9. Gabaldon T, Huynen MA. Prediction of protein function and pathways in the genome era. Cellular and Molecular life sciences. 2004. 61(78); 930-944.
10. Franceschini A, Lin J, Mering CV, Jensen LJ. SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. Bioinformatics. 2016. 32(7); 1085-1087.
11. Ranea JAG, Yeats C, Grant A, Orengo CA. Predicting Protein Function with Hierarchical Phylogenetic Profiles: The Gene3D Phylo-Tuner Method Applied to Eukaryotic Genomes. Plos Computtaional Biology. 2007. 3(11).
12. Sahraeian SM, Luo KR, Brenner SE. SIFTER search: a web server for accurate phylogeny-based protein function prediction. Nucl Acids Res. 2015. 43; W141-W147.
13. Zhang C, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. Nucleic Acids Res. 2017. 45; W291-W299.
14. Maghawry HA, Mostafa MG, Gharib TF. A New Protein Structure Representation for Efficient Protein Function Prediction. Jnl of Comp Biol. 2014. 21(12).
15. Stawiski EW, Baucom AE, Lohr SC, Gregoret LM. Predicting protein function from structure: Unique structural features of proteases. Proc of the nat aca of sci of the USA. 2000. 97(8); 3954-3958.
16. Rahmani H, Blockeel H, Bender A. Predicting the functions of proteins in Protein-Protein Interaction networks from global information. Proc of the third Int Workshop on Mac Lear in Sys Biol. 2009. 8; 82-97.
17. Peng W, Wang J, Cai J, Chen L, Li M, Wu FX. Improving protein function prediction using domain and protein complexes in PPI networks. BMC Sys Biol. 2014. 8(35).

18. Piovesan D, Giollo M, Ferrari C, Tosatto SCE. Protein function prediction using guilty by association from interaction networks. Amino Acids. 2015. 47; 2583-2592.
19. Zhao B, Wang J, Wu FX, Pan Y. Predicting Protein Functions Based on Dynamic Protein Interaction Networks”. In: Harrison R., Li Y., Măndoiu I. (eds) Bioinformatics Research and Applications. Lecture Notes in Computer Science.2015. 9096.
20. Li XL, Tan YC, Ng SK. Systematic gene function prediction from gene expression data by using a fuzzy nearest-cluster method. BMC Bioinformatics.2006. 7.
21. Tran L. Hypergraph and protein function prediction with gene expression data. *CoRR* abs/1212.0388: n. pag.,2012.
22. Massjouni N, Rivera CG, Murali TM. VIRGO: computational prediction of gene functions. Nucleic Acids Res. 2006. 34 (2); W340-W344.
23. Cao R, Cheng J. Integrated protein function prediction by mining function associations, sequences, and protein–protein and gene–gene interaction networks. Methods. 2016. 15(93); 84-91.
24. Verspoor KM, Cohn JD, Ravikumar KE, Wall ME. Text Mining improves prediction of protein functional sites. PLOS One. 2012.