International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 7
July 2017

In-Silico Approach for Hypothetical Protein Function Prediction

Shabanam Khatoon
Department of Computer Science, Faculty of Natural Sciences
Jamia Millia Islamia, New Delhi
Suraiya Jabin
Department of Computer Science, Faculty of Natural Sciences
Jamia Millia Islamia, New Delhi
Punit Kaur
Department of Biophysics
All India Institute of Medical Sciences (AIIMS), New Delhi

ABSTRACT

Hypothetical proteins are the proteins whose existence has been predicted from nucleic acid sequences, but for which there is lack of experimental evidence that it is expressed in vivo. Function of these hypothetical proteins sometimes may be related to Cancer and many other therapeutic uses. The cellular function involves the process of interaction with other proteins and studying these complex interactions is a challenge in proteomics. Hypothetical proteins are characterized by low identity to known, annotated protein. Experimental methods can be time consuming and costly approach for function prediction of hypothetical proteins. It can be computationally predicted by homology modelling in few cases. In homology modeling, we align hypothetical protein with known protein sequence. It can also be predicted by determination its 3-D structure. We may also predict the function of HPs by understanding the prosthetic group and fold similarity with other proteins of known function. In recent years, protein-protein interaction (PPI) is an important research field. It deals with understanding of protein domain interaction. The function of protein may be predicted by PPI and by analyzing its interaction with protein of known function. The STRING database is used for protein-protein interaction. Majority of researchers have tried it by using machine learning methods such as ensemble methods, Genetic Algorithm, SVM, HMM, Artificial Neural Network, Neuro-Genetic hybrid algorithm and Fuzzy ANN etc. There are various types of Protein databases which are helpful in predicting the function of proteins. In this paper, we present survey of computational approaches used for prediction of hypothetical proteins with the help of bioinformatics tools and databases.

KEYWORDS: Hypothetical Proteins (HPs), Motifs, Domain, Database, Machine learning techniques

INTRODUCTION

The human genome sequence was completed in April 2003 some two years ahead of schedule [2]. The number of genes present in the genome was estimated at around 30,000 although recent estimates are lower standing at around 25,000 [3]. Ofran et al [4] estimated that of 2,000,000 known sequences, less than 25% were annotated to completion. Currently, there are approximately 2000 human protein sequences for which very little is known. For proteins with well characterised close relatives, it is trivial to infer function. Orphan proteins without discernible sequence relatives present a greater challenge. Here the task of experimental characterisation is blind and becomes unwieldy. It is highly unlikely that all known proteins will ever be completely experimentally characterised [5]. Thus there is an emergent need to develop fast and accurate computational approaches to fulfil this requirement.

Most new proteins come from genome sequencing projects, e.g.) Mycoplasma genitalium (484 proteins), Escherichia coli (4,288 proteins), S. Cerevisiae-yeast (5,932 proteins), C. Elegans-worm (~ 19,000 proteins), Homo sapiens (~ 40,000 proteins) and have unknown functions. Best annotated protein sequence databases

are SwissProt, Protein Information Resource (PIR-1) now part of UniProt–unified protein knowledgebase (http://pir.georgetown.edu).

Experimental characterization of protein's cellular function can be prohibitively expensive and take years to complete. Characterization of new proteins with unknown function through computational approaches is one of the most challenging problems in in-silico biology, which has attracted world-wide interests and great efforts. In this paper, we present a survey of different in-silico methods used in literature for automated function prediction of hypothetical proteins.

DEFINITION: HYPOTHETICAL PROTEIN (HP)

A hypothetical protein is a protein whose existence has been predicted but there is a lack of experimental evidence that it is expressed in vivo. A hypothetical protein function is predicted by domain homology searches. Conserved domains are present in the HP which need to be compared with the known family domains by which HP could be classified into particular protein families. The function of hypothetical protein could also be predicted by homology modelling. Proteins play a vital role in various cellular, biological and molecular functions. A protein database contains a large amount of computational nucleic acid sequence and protein sequence (amino acid sequence) of the living organisms. Prosite is a protein database. It contains biological information which describes protein families, protein domains and functional sites. This database helps to identify the possible functions of a new sequence from the existing sequence. If the new sequence is not closely related to those existing proteins, though a complete alignment cannot be found, one can identify the existence of pattern or motif from the database.

Not all of the hypothetical proteins are totally uninformative. Some of them have entries describing the domains or functions predicted from similarity search. But others are completely uninformative. About 50% of proteins in the RefSeq Protein database are actually hypothetical proteins.

PROTEIN EXISTENCE

In UniProtKB there are five types of evidence for the existence of a protein:

- **Experimental evidence at protein level**: The value 'Experimental evidence at protein level' indicates that there is clear experimental evidence for the existence of the protein.
- **Experimental evidence at transcript level:** The value 'Experimental evidence at transcript level' indicates that the existence of a protein has not been strictly proven but that expression data (such as existence of cDNAs) indicate the existence of a transcript.
- **Protein inferred from homology:** The value 'Protein inferred by homology' indicates that the existence of a protein is probable because clear orthologs exist in closely related species.
- **Protein predicted:** The value 'Protein predicted' is used for entries without evidence at protein, transcript, or homology levels.
- **Protein uncertain:** The value 'Protein uncertain' indicates that the existence of the protein is unsure.

Now we define different terminologies relevant for protein function prediction:

- **Identity:** *Identity* defines the percentage of amino acids (or nucleotides) with a direct match in the alignment.
- **Similarity:** Similarities are sequence similarities or other types of similarities. Similar pair of residues are structurally or functionally related including conservative substitutions. Amino acids which are similar in properties are acidic amino acids (Asp D, Glu E), basic amino acids (Lys K, Arg R, His H), hydrophobic amino acids (Trp W, Phe F, Tyr Y, Leu L, Ile I, Val V, Met M, Ala A) etc. When one amino acid is mutated to a similar residue such that the physiochemical properties are preserved, a *conservative substitution* is said to have occurred. For example, a change from aspartate to glutamate maintains the -1 negative charge (both are acidic amino acid). This is far more likely to be acceptable since the two residues are similar in property and won't compromise the translated protein. Thus, *percent similarity* of two sequences is the sum of both

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 7
July 2017

identical and similar matches (residues that have undergone conservative substitution). Similarity measurements are dependent on the criteria of how two amino acid residues are to each other.

- **Homologous:** Proteins are called homologous if they posses common ancestor**.** Homology among proteins or DNA is typically inferred from their sequence similarity. Significant similarity is strong evidence that two sequences are related by divergent evolution of a common ancestor. Alignments of multiple sequences are used to indicate which regions of each sequence are homologous.
- **Analogous:** Proteins that have no common ancestors but possess structural similarity are called analogs (Dokholyan and Shakhnovich, 2001).
- **Orthologs:** Orthologs arise from speciation, the same gene in different organisms. Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes.
- **Paralogs:** Paralogs may be derived from duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.
- **Equivalogs:** Equivalogs are proteins with equivalent functions.
- **Domain:** A domain can be defined as the independently folded parts of sequence or structure in a protein. All protein has a specific domain with a specific function. For example, most of the proteins involved in intracellular signaling contains PH domain (Pleckstrin homology domain). This domain can bind to certain biological molecules inside the membrane and recruit them to the membrane. A domain is structurally compact, independently folding unit forming a stable three –dimensional structure. Typically, a conserved domain contains one or more motifs (Koonin and Galperin, 2003).
- **Motifs:** A motif is a set of conserved amino acid residues that are important for function. Proteins can be characterized by more than one motif and it can be classified using certain specific motifs. Some of examples for motifs are:
- **Helix-turn-helix:** It is made up of of two α – helices connected with few amino acids. One is on the N-terminus end and the other at the C-terminus. Helix-turn-helix has important role in DNA recognizing and DNA binding
- **Helix-loop-helix:** It is composed of two α- helices that is joined by a loop. A loop is the area between two secondary structure elements. It describes the qualities of transcription factors. Transcription factor is a DNA binding protein which controls the transcription process of DNA.
- **Omega loop:** It is a loop shaped polypeptide chain. The motif contains a large number of hydrogen bond inside. Due to this it has an important role in protein stability and folding.

COMPUTATIONAL METHODS USED FOR PROTEIN FUNCTION PREDICTION

In this section we briefly describe most popular bioinformatics tools and databases used for facilitating protein function prediction.

- Similarity search: First start with Blastp, if our sequence is less than 40% identity go for PSI- Blast etc.
- Domain search: Do domain search using Interproscan, Pfam or CDART etc.
- Search for siganl peptide and TM: search for signal peptide using signp and TM using TMHMM, phobias.
- Comparative modelling: Do homology modelling using swiss model, if our sequence less than 40% identity from blast result go for abintio modelling using I-Tasser.
- Functional association prediction: Try searching sequence using STRING database.

Sequence similarity search: Basic local alignment tool (BLAST) Used for finding similar sequences in proteins.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 7
July 2017

Physiochemical properties prediction: ExPASy-Protparam tool Used for computation of various physical and chemical parameters like molecular weight, isoelectricpoint(Pi), aminoacid composition, atomic composition, extinction co-efficient, instability index, aliphatic index, and grand average of hydropathy(GRAVY).

Sub-cellularlocalization: signal Predicts signal peptide cleavage sites.

TMHMM used to authenticate whether the protein is a membrane protein or not.

HMMTOP Predict transmembrane topology.

I-TASSER Suite (http://zhanglab.ccmb.med.umich.edu/I-TASSER/download/), a stand-alone software package for protein structure and function modeling.

Domain analysis: Pfam Collection of multiple protein sequence alignments.

SUPERFAMILY is a database of structural and functional annotation for all proteins and genomes.

PANTHER (Protein analysis through evolutionary relationships) Identification and annotation of protein domains.

CDART (The conserved domain architecture comprehensively organized database of protein families and retrieval tool) sub-families, their evolutionary relationships in the form of phylogenetic trees.

SVMprot SVM (Support vector machine based classification of proteins)

Motif Analysis InterPro Scan Searches interPro for motif discovery. It is the integration of several large protein signature databases.

MOTIF used for Motif discovery.

MEME suite Database searching for assigning function to the discovered motifs.

Protein–Protein interaction STRING used for predicting protein–protein interactions.

Gene Ontology (GO) provides a hierarchical machine controlled vocabulary split into 3 categories: Molecular function, Cellular component and Biological process. **Molecular function**: The biochemical functions performed by a protein, such as ligand binding, catalysis of biochemical reactions and conformational changes. **Cellular function**: Many proteins come together to perform complex physiological functions, such as operation of metabolic pathways and signal transduction, to keep the various components of the organism working well. **Biological function**: The integration of the physiological subsystems, consisting of various proteins performing their cellular functions and the interaction of this integrated system with environment

MACHINE LEARNING APPROACHES FOR FUNCTION PREDICTION

The most commonly used machine learning techniques used for function prediction are Support Vector Machine, Artificial Neural Networks, K-Nearest Neighbour and Decision Tree. Machine learning techniques are widely used techniques for hypothetical protein function prediction. Huanget et.al proposed a novel scoring card method to estimate solubility scores of dipeptides and amino acid residues for predicting solubility of proteins and analyzing the tendency of physicochemical properties [10]. Instead of direct comparison or clustering of sequences, SVM classification is based on the analysis of physicochemical properties of a protein generated from its sequence .Samples of proteins known to be in a functional class (positive samples) and those not in the class (negative samples) are used to train a SVM system to recognize specific features and classify proteins into either the functional class or outside of the class. Such an approach may be applied to functional prediction for both distantly-related and closely-related proteins [12].

CONCLUSION

When we attempt to predict protein function, we need to first determine the cellular process in which protein participates or its physiological role or its enzymatic activity. The cellular function involve the process of interaction and understanding these complex interaction is a challenge in proteomics. Different researches

have indicated that integration of heterogeneous biological data plays key role in predicting function. This heterogeneous data mainly includes: Amino acid sequences, Protein structure, Genome sequences, Phylogenetic data, Microarray expression data, Protein interaction networks and protein complexes, biomedical literature, and combination of multiple data types. This data offers different types of insights into a protein's function and related concepts. For instance, protein interaction data shows which proteins come together to perform a particular function, while the three-dimensional structure of a protein determines the precise sites to which the interacting protein binds itself. An attempt has been made to introduce a very challenging problem in bioinformatics of hypothetical protein function prediction.

REFERENCES:

[1] Punta, M., & Ofran, Y. (2008). The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS computational biology, 4(10), e1000160.
[2] Pennisi E., (2003), Human genome. Reaching their goal early, sequencing labs celebrate. Science, 300(5618):409.
[3] Pennisi E., (2003), Human genome. A low number wins the GeneSweep Pool. Science, 300(5625):1484.
[4] Ofran Y, Punta M, Schneider R, and Rost B., (2005), Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. Drug Discov Today, 10(21):1475–1482.
[5] Baumgartner Jr. WA, Cohen KB, Fox LM, Acquaah-Mensah G, and Hunter L., (2007), Manual curation is not sufficient for annotation of genomic databases. Bioinformatics, 23(13):i41–48.
[6] M. P. S. Brown, W.N.G Rundy, D.Lin, N. Cristianini, C.W.S Ugnet, T.S. Furey, JR. M. Ares, and D. Haussler, (2000), Knowledge-based analysis of microarray gene expression data using support vector machines. Proc. Natl Acad. Sci. USA 97: 262–7.
[7] Noble, W., & Ben-Hur, A. (2007). Integrating information for protein function prediction. Bioinformatics-From Genomes to Therapies, 3, 1297-1314.
[8] Gabaldon, T; M.A. Huynen (2004). "Prediction of protein function and pathways in the genome era". Cellular and Molecular Life Sciences 61: 930–944. doi:10.1007/s00018-003-3387-y. PMID 15095013.
[9] Plessis, L.; N. Skunca; C. Dessimoz (2011). "The what, where, how and why of gene ontology--a primer for bioinformaticians". Brief Bioinform 12 (6): 723–735. doi:10.1093/bib/bbr002. PMID 21330331.
[10] Protein Structure and Function Prediction Using Machine Learning Methods Hemalatha N.1, SiddhantNaik2, Jeason Rinton Saldanha3 .
[12] SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence C.Z. Cai1,2, L.Y. Han1, Z.L. Ji1, X. Chen1 and Y.Z. Chen1,*