JOURNAL OF TROPICAL LIFE SCIENCE
2020, Vol. 10, No. 2, 141 – 148
http://dx.doi.org/10.11594/jtls.10.02.06

Research Article

A Comparative Genomics Pipeline for *In Silico* Characterization and Functional Annotation of Short Hypothetical Proteins

Soumyajit Guha [1,2]*, Shuvam Das [2], Sayak Ganguli [3]

[1] Department of Microbiology, Vijaygarh Jyotish Ray College, Kolkata 700032, India
[2] The Biome Research Facility, AD60 - Salt Lake City, Kolkata 700064, India
[3] Department of Biotechnology, St. Xaviers College, Kolkata 700016, India

Article history:
Submission September 2019
Revised January 2020
Accepted February 2020

**Corresponding author:*
E-mail:
soumyajitguhas42@gmail.com

ABSTRACT

Hypothetical proteins are the proteins whose existence has been anticipated, but for which there are certain scarcities of experimental evidences about its structure, function or linkage to any known genes. Sequencing of several genomes has resulted in numerous predicted open reading frames to which structure or function(s) cannot be readily assigned and sometimes they can make up a significant portion of a genome. In this study, we designed a pipeline for the study and efficient functional annotation of short hypothetical proteins (only which were < 400 amino acids) comparing two case studies, using amino acid sequence informations retrieved from two different protein databases. The investigation and *in-silico* analysis of likely functional aspects of hypothetical proteins were performed employing various computational methods and tools based on sequence similarity, identification of targeting signals, presence of known protein domains, physicochemical characterization, etc. Our annotation pipeline was able to annotate 90 hypothetical proteins out of 100 compared to evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) databases' annotation of 82 proteins, which is about 8% more compared to eggNOG for case study 1 and 78 hypothetical proteins out of 96 compared to eggNOG's annotation of 58 proteins, which is about 20.83% more compared to eggNOG for case study 2. It was also seen that some hypothetical proteins had a high aliphatic index, indicating higher thermostability in extreme environments. From this study subcellular localization involving cytoplasmic proteins and membrane proteins were also predicted with higher accuracies than other proteins. Hypothetical proteins can provide an insight of different unknown structures and functions of proteins and can be an important area for further research.

Keywords: *Annotation, database, hypothetical protein, in silico, protein sequences, subcellular*

Introduction

Research on genome had started as early in 1995 with the sequencing of the first complete genome of a cellular life form: the 1.8 Mb genome of *Haemophilus influenzae* strain Rd KW20. Eight years later, the genomes of over 100 organisms have been sequenced, and sequencing of many more are underway. Inconsistency in the accuracy of genome annotation was a subject of many heated discussions at the beginning of the genome era. Still, the so-called "70% hurdle" holds, as functions of only ~ 50 ±, 70% of the genes in any given genome can be predicted with reasonable confidence. The remaining genes are either homologous to the genes of unknown function, and are typically referred to as "conserved hypothetical" genes, or do not have any known homologs termed "hypothetical" or "unknown" or "non characterized" as it is unclear whether they encode actual proteins. Many times, it is doubtful whether they code for actual proteins, the latter genes are

How to cite:
Guha S, Das S, Ganguli S (2020) A Comparative Genomics Pipeline for *In Silico* Characterization and Functional Annotation of Short Hypothetical Proteins. Journal of Tropical Life Science 10 (2): 141 – 148. doi: 10.11594/jtls.10.02.06.

generally referred to as 'hypothetical', 'uncharacterized', or 'unknown' proteins. Gene products of many completely sequenced organisms fall under the 'hypothetical' category and they cannot be linked with any previously known and characterized proteins and as such their function is completely unknown [1]. Up to December 2019, NCBI (National Center for Biotechnology Information) database contains 166,841,351 protein sequences that are still unknown. Thus, the determination of protein function is one of the challenging problems of the post-genome era [2]. Hypothetical proteins can be an important thing; as they can offer the presentation of new structures and novel functions. Furthermore, new hypothetical proteins may serve as markers and pharmacological targets. Here lies the application of bioinformatics to predict functions of un-annotated protein sequences. The choice of dataset was based on the testing the pipeline on two counts - the first choice was an ecological niche area which would actually represent a conglomeration of multiple organism types thus presenting us with a wide variety of annotation challenges. Furthermore, the deep-sea (at a water depth of > 2,000 m) constitutes the largest biome on earth. Still, relatively much remains unknown about its microbial community's structure, function, and adaptation to the cold and deep biosphere. Marine sediment is an extensive unexplored source of enzymes with unique properties that may be useful for various industrial and biotechnological purposes. However, since many microbes are unculturable in the laboratory, a cultivation-independent metagenomic approach would be advantageous for the identification of novel enzymes [3].

The choice of the second dataset stems from the fact that the pipeline needed to be tested based on an organism specific dataset and hence we chose the Gram-positive bacteria *Streptococcus pyogenes*, which is a serious threat to humans and are responsible for a wide variety of infections, some of which can be life-threatening. These infections affect about 700 million people worldwide with a mortality rate of only 0.1%. Out of these 700 million people, 650,000 people suffer from severe or invasive infection(s) of *S. pyogenes* with a mortality rate reaching nearly 25 % [4]. Conventional antibiotics are now becoming less effective due to the emergence of drug-resistant strains of the bacteria that are reported to be resistant against antibiotics like penicillin [5], erythromycin [6], etc. and it is now becoming a serious challenge to cope up with them. So, there is a need to find new targets against which the bacteria have not yet grown any resistance. Many hypothetical proteins are now also used as targets for the therapeutic purpose [7]. Therefore, with the objective of *in-silico* analysis of novel proteins and enzymes, and understanding of their functions, localization, etc. a pipeline was designed, common for both the case studies; which is begun by using the GenBank database of NCBI-UniProt to search for hypothetical proteins (amino acid sequences length < 400).

Material and Methods

Sequence retrieval and selection of the hypothetical proteins

Primary amino acid sequences (each sequence consisting not more than 400 amino acids) of the hypothetical proteins, on a first come first serve basis were retrieved in FASTA format from the NCBI protein database, for case study 1 and the UniProt database, for case study 2. These sequences were downloaded and saved as text files for further operations. This was a computational approach where various databases and online based servers were used consequently to collect important information about the hypothetical proteins. Protein sequences retrieved from the NCBI database were then run in CDD-search suite for clustering.

Sequence analysis

The FASTA sequences of the hypothetical proteins were used as query sequences for various analyses involving physicochemical properties, conserved domains, signal peptides, subcellular localization, etc. either individually or as a collection in a .txt file and uploaded. Online tools and interfaces such as the CDD-BLAST (Conserved Domain Database - Basic Local Alignment Seach Tool), Pfam [7], InterProScan [8] etc. that can automatically annotate protein families and domains providing insights into sequence or structure or function relationships.

Physicochemical characterization of the hypothetical proteins

The hypothetical proteins in raw sequence format were evaluated for physicochemical proper-

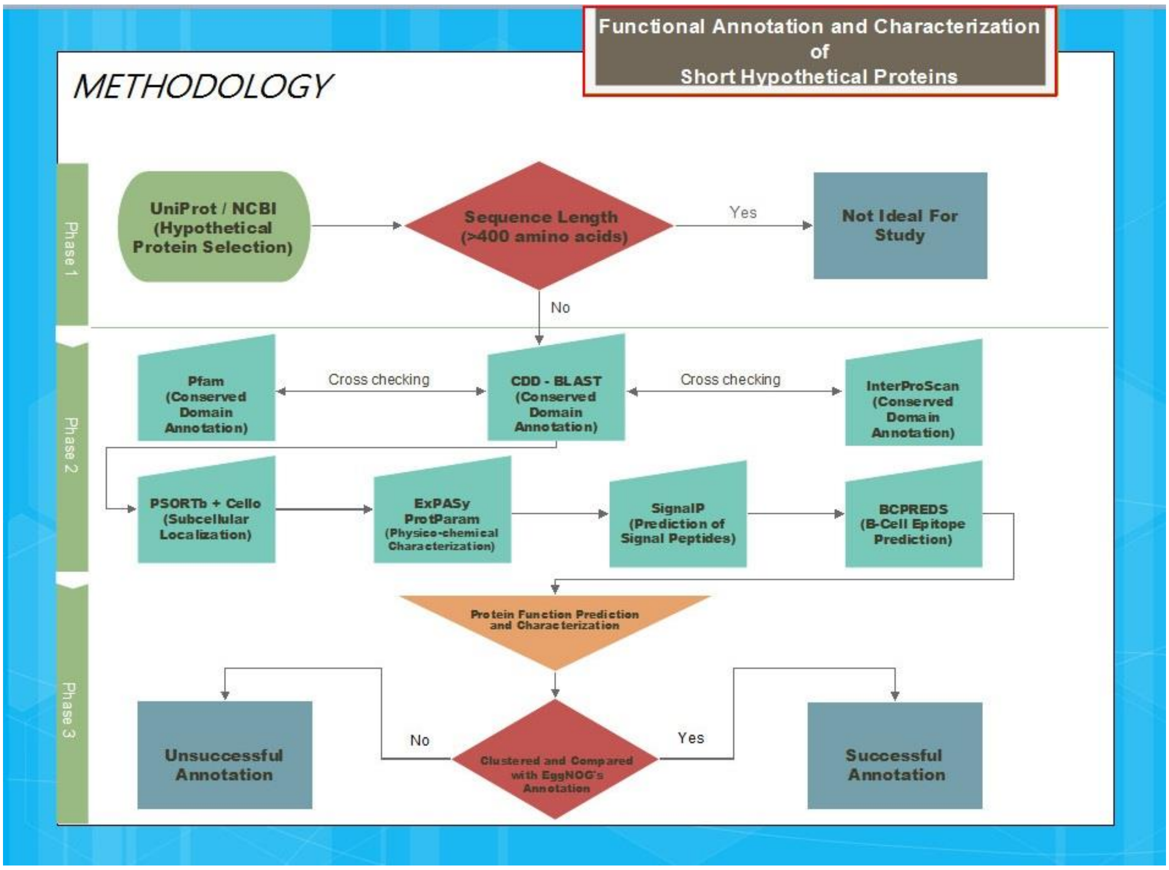
Functional Annotation and Characterization of Short Hypothetical Proteins
METHODOLOGY
Phase 1
Phase 2
Phase 3
UniProt / NCBI (Hypothetical Protein Selection)
Sequence Length (>400 amino acids)
Yes
Not Ideal For Study
No
Pfam (Conserved Domain Annotation)
Cross checking
CDD - BLAST (Conserved Domain Annotation)
Cross checking
InterProScan (Conserved Domain Annotation)
PSORTb + Cello (Subcellular Localization)
ExPASy ProtParam (Physico-chemical Characterization)
SignalP (Prediction of Signal Peptides)
BCPREDS (B-Cell Epitope Prediction)
Protein Function Prediction and Characterization
Unsuccessful Annotation
No
Clustered and Compared with EggNOG's Annotation
Yes
Successful Annotation

Figure 1. Flowchart of methodology for annotation of short hypothetical proteins

ties using the ProtParam tool of the ExPASy (Expert Protein Analysis System) server. The parameters that are calculated by the program and reported include the molecular weight, theoretical pI (isoelectric point), amino acid composition, total number of positive and negative residues, instability index, aliphatic index, extinction coefficient, and GRAVY (grand average of hydropathicity). The instability index gives an estimate of the protein stability in a test tube. An instability index value of less than 40 is predicted to be stable, and above that to be unstable. The aliphatic index of a protein is defined as the relative volume held by its aliphatic side chain amino acids. The extinction coefficient stipulates the amount of light a protein absorbs at a certain wavelength. By dividing the sum of hydropathy values of all of the amino acids by the number of residues in the sequence, the GRAVY value for a peptide or protein is obtained [10].

Functional classification of hypothetical proteins

Updated databases such as Pfam, CDD and InterProScan were used to identify the characteristic functional domains by utilizing the sequence of these hypothetical proteins. CDD is an open source database present in the NCBI for the functional characterization of protein sequences by using annotation of conserved domain footprints. In addition, the motif detection was performed using InterProScan and to predict the functional signature in the sequences. The function of the hypothetical proteins was predicted using InterProScan, a tool that integrates various protein signature recognition methods and databases.

Sub-cellular localization analysis

As there is a positive correlation between the subcellular location and its biological function of a protein, the knowledge about the subcellular location can help to characterize the protein functions. Therefore, PSORTb tool at ExPASy server was utilized to identify the subcellular localization of nonhomologous essential protein sequences. The proteins can be present in feasible subcellular locations namely, cytoplasm, plasma membrane, nucleus, periplasm, outer membrane, and extracellular. In the absence of experimental information, subcellular localization prediction tools such as PSORTb and CELLO can be used. In the current study, locations of the short-listed proteins were identified using PSORTb 3.0.2 [9] and CELLO 2.5 [10]. PSORTb sorts proteins by means of various modules like SVM, S-TMHMM, and SCL-BLAST using about 11692 proteins of known localization from bacteria (Gram-positive and negative) and archaea as the training set. SignalP is another powerful tool for predicting secretory proteins. The developers of SignalP have recently assessed the performance of the SignalP 3.0 method on the current data set, by categorizing all proteins as secretory except cytoplasmic proteins and reported an overall accuracy of 95% [11]. Though the performance of SignalP has been found to be better than PSLpred, one cannot compare PSLpred and SignalP, as the number of classes predicted by these two methods is different**.**

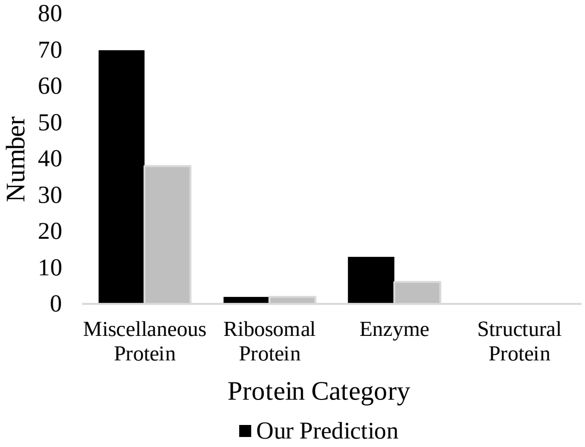

Figure 2. Categorization of the hypothetical proteins (Case study 1: 90 Hypothetical Proteins and Case study 2: 85 Hypothetical Proteins)

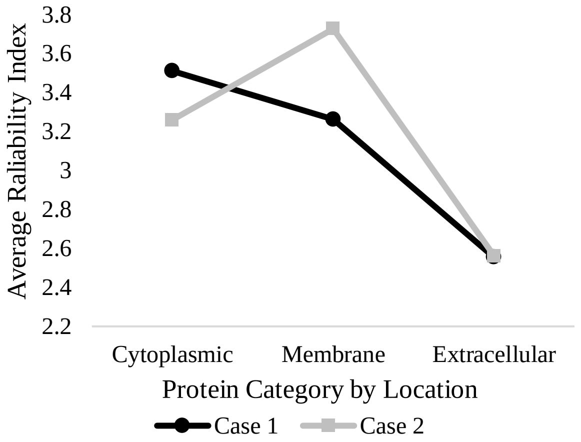

Figure 3. Average reliability index of hypothetical proteins (Case study 1: 100 Hypothetical Proteins and Case study 2: 96 Hypothetical Proteins)

Automatic annotation and then clustering of protein functions

For most comparative genomics study, the identification of orthologous genes is a necessity. The maximum portion of the functionally annotated genes in genomes or metagenomes are usually retrieved by comparative analyses and inferences from existing functional knowledge via homology. The eggNOG database is a biological information database presented by the EMBL (European Molecular Biology Laboratory). It was created in 2007 and later updated to version 4.5 in 2015. An important characteristic of eggNOG is that it supplies functional annotations for the orthologous groups. These annotations are produced by a pipeline, which summarizes the available functional information on the proteins in each cluster: the textual annotation for these proteins, their annotated GO (Gene Ontology) terms, their membership to KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways and the presence of protein domains from SMART (Simple Modular Architecture Research Tool) and Pfam. The textual descriptions usually allow for the most detailed annotation of the protein functions, so it first uses the Ukkonen's algorithm to retrieve the LCS (Longest Common Subsequence) amongst the description lines of any two given proteins within a cluster. Then it scores each LCS based on the number of protein descriptions matched within the cluster, the number of occurrences of each word of the LCS in these descriptions, and the presence of words such as ‘hypothetical’, ‘putative’, or ‘unknown’. These scores are finally normalized against a score distribution based on randomized clusters of the same size, and the highest scoring LCS is chosen, provided that it scores above a threshold. EggNOG's orthologous groups contain 1,241,751 genes and render at least a detailed functional description for 77% of them [12]. For each orthologous group, the pipeline also searches for overrepresented GO terms, KEGG pathways or protein domains. To find terms that are sufficiently specific and at the same time are likely to describe the entire orthologous group, it is provided with a scoring function that considers back

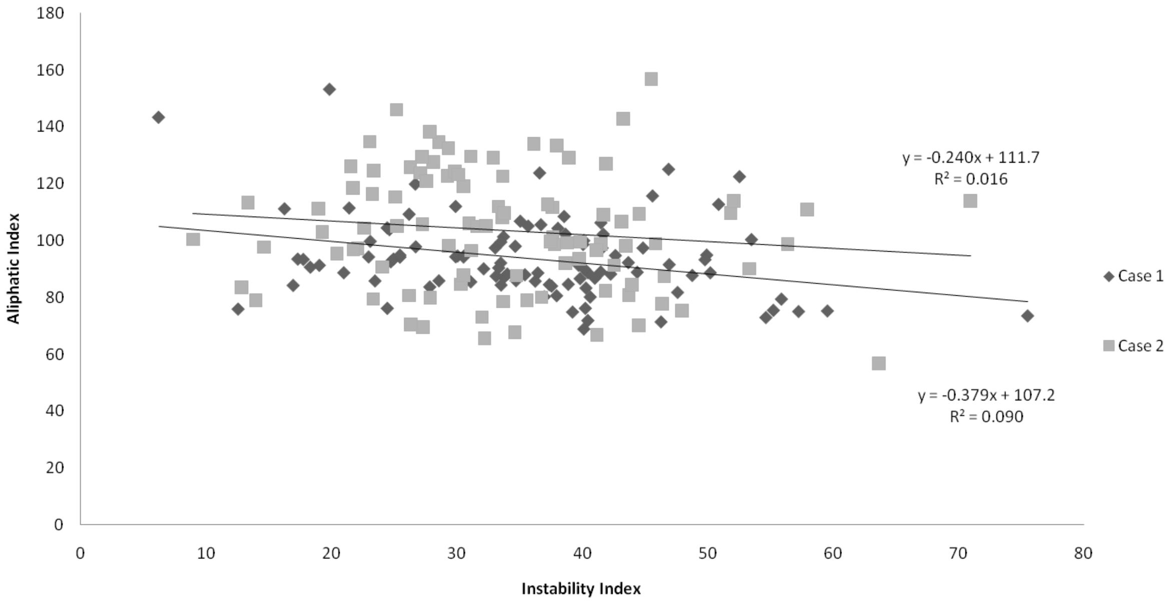

Figure 4. Correlation between instability and aliphatic Index (Sample size: Case study 1 = 100; Case study 2 = 96)

Table 1. Hypothetical protein annotation comparison

Case	Total number of proteins studied	eggNOG's annotation	Our pipeline's annotation	Number of proteins annotated more by our pipeline	Total number of proteins studied
1	100 (100%)	82 (82%)	90 (90%)	8 (8%)	100 (100%)
2	96 (100%)	58 (60.42%)	78 (81.25%)	20 (20.83%)	96 (100%)

ground frequency, term frequency within the group, and the ratio of the two (i.e. the fold over-representation). In case no satisfactory LCS was found, a description line is constructed based on the highest scoring GO term or KEGG pathway. As a single domain may not properly reflect the function of a complete protein, description lines are constructed based on overrepresented domains only if all other options have been exhausted. Thus, to determine our pipeline's efficiency, we compared our annotations with that of the automated annotation of eggNOG. A flowchart of the methodology followed is shown in Figure 1.

Results and Discussions

We studied a total of 196 hypothetical proteins (100 in case 1 and 96 in case 2 respectively) and categorized them as miscellaneous, ribosomal, enzyme, and structural proteins. Out of 100 hypothetical proteins, of case study 1, we were able to classify 90 hypothetical proteins on the above-said category and in case of case study 2; out of 96 hypothetical proteins, 85 were categorized (Figure 2). Hypothetical proteins can participate in wide range of cellular activities like cell structure and morphology, signaling, stimulus and stress responses, catalysis, ribo-nuclear organization and processing, biogenesis, etc. Relatively high number of Miscellaneous proteins in both cases of our study suggest that they may take part in different cellular activities and can also be responsible for functions that are still unknown. We also determined subcellular locations of all the hypothetical proteins using CELLO 3.0 and classified them into three groups i.e. Cytoplasmic protein, Membrane protein and Extracellular Protein. Information about subcellular location of a hypothetical protein is very important as it provides an insight into the probable function of the protein and it is also very useful from the point of view of computatio-

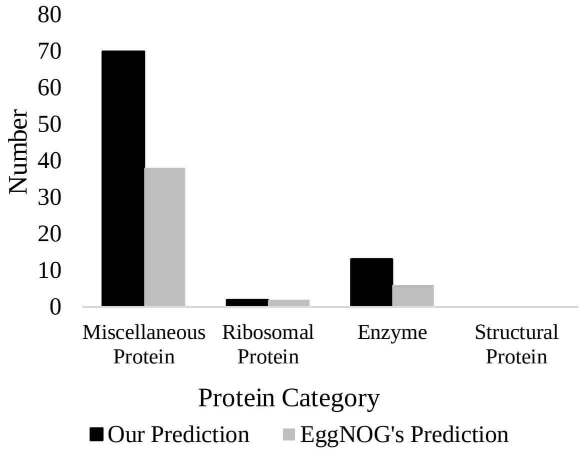

Figure 5. Comparison of annotation between EggNOG and ours. (Case study 1; Total no of proteins 100)

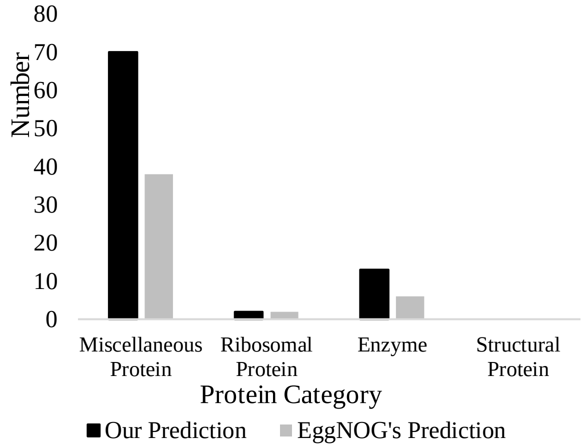

Figure 6. Comparison of annotation between EggNOG and ours. (Case study 2; Total no of proteins 96)

nal biology as many whole genome sequencing projects have been finished and many resulting protein sequences are still lacking detailed functional information. Along with this, we determined the reliability index of each proteins in both cases. Reliability index is used to determine the level of certainty in the prediction of a particular sequence. After determining the reliability index of each hypothetical protein, we calculated their mean reliability index after separating them on the basis of their subcellular location. The data obtained for mean reliability index of case study 1 is 3.511; 3.263; 2.555 for cytoplasmic, membrane, and extracellular proteins respectively and for case study 2 it is 3.259; 3.727; 2.563 for cytoplasmic, membrane, and extracellular proteins respectively (Figure 3).

Hence, from the data, we can infer that cytoplasmic and membrane proteins are predicted at higher accuracy than Extracellular Proteins. We also determined the instability and aliphatic index of the hypothetical proteins in both cases and calculated correlation between them. The aliphatic index is directly related to the mole fraction of Ala, Ile, Leu, and Val in the protein. The aliphatic index of proteins from thermophilic bacteria was observed to be remarkably higher than that from ordinary proteins and hence, it can serve as a measure of thermostability of proteins. Higher aliphatic index of a protein suggests an increase in the thermostability of the protein which in turn might favor an increase in its solubility and that thermostability and solubility on over expression have a positive correlation [13]. On the other hand, instability index is a measure of the stability of a protein, thus the instability index can be used as a measure of *in vivo* half-life of a protein [14]. Proteins having an *in vivo* half-life of less than 5 hour have been shown to have an instability index of more than 40, whereas those with an *in vivo* half-life of more than 16 hour [15] have an instability index of less than 40, thus, a protein whose instability index is smaller than 40 is predicted as stable, a value above 40 speculates that the protein may be unstable. The correlation between *in vivo* half-life of a protein and solubility could be rationalized by the role played by longer-lived partially folded intermediates of a protein. These long-lived intermediates can interact with a greater chance with other partially folded intermediates and also exhaust the limited *in vivo* supply of chaperones [16], thus contributing to inclusion body formation. In this study for case study 1 we found a negative correlation between instability and aliphatic index which is statistically significant and for case study 2 the correlation is statistically insignificant (Figure 4). Finally we compared our annotated data with eggNOG server in both cases to compare our annotation's success with eggNOG, in case study 1, out of four categorized proteins, i.e. miscellaneous, ribosomal, enzyme and structural protein; eggNOG successfully predicted 39, 23, 15 and 5 respectively and our pipeline's successful prediction is 40, 30, 15 and 5 respectively (Figure 5); on the other hand, in case study 2, out of four above said proteins, eggNOG's successful prediction is 38, 2, 6 and 0 and our pipeline's successful prediction is 70, 2, 13 and 0 (Figure 6).

Also we have found that in case study 1, out of 100 hypothetical proteins, 82 proteins were successfully annotated by eggNOG and our pipeline of annotation successfully annotated 90 hypothetical proteins which is 8% more compared to eggNOG server and in case study 2, out of 96 hypothetical proteins, 58 proteins were successfully annotated by eggNOG and our pipeline of annotation successfully annotated 78 hypothetical proteins which is 20.83% more compared to eggNOG server (Table 1).

Conclusion

The profusion of hypothetical proteins makes their study a formidable task. There is a clear requirement for a rational criterion that would allow classifying these protein families and selecting the most important ones, i.e. prioritizing the targets for experimental studies; two obvious criteria are the number of proteins in the family and its phyletic spread. Since the emergence of comparative genomics, wide phylogenetic distribution and indispensability for cell growth have been taken into consideration by some researchers when choosing uncharacterized genes for experimental study. Notable positive correlation between the phyletic spread of a gene and the likelihood that it is essential for cell growth has been demonstrated. On many occasions, experiments with proteins that met one or both of these criteria led to major discoveries. Hypothetical proteins are considered as orphaned proteins because of their unknown structure and function. In this study, different bioinformatics tools are applied to the various hypothetical proteins to find their localizations, functions etc. It was found that selected hypothetical proteins had a high aliphatic index thus indicating higher thermostability that would help the proteins to survive in extreme temperature conditions. Our study also revealed that subcellular localization involving cytoplasmic proteins and membrane proteins are predicted with higher accuracies than other cell proteins. This study can also help designing of specific PCR (Polymerase Chain Reaction) primers. Hypothetical proteins can be a treasure trove for modern science as it can provide new and interesting scientific data; this study provides brief information about the probable function(s) of the hypothetical proteins of Streptococcus pyogenes along with their cellular location and available epitopes which can be used as a potential target for drug design. This is a stepping stone for the study of hypothetical proteins, further studies can be done to reveal the detailed structure and function of the hypothetical proteins, and they can also be used as effective target for more efficient drug delivery. The most important aspect of this study is the establishment of a superior annotation pipeline of hypothetical proteins than existing automated annotation platforms like eggNOG. Our annotation pipeline was able to annotate completely and accurately 90 (90%) hypothetical proteins out of 100 (100%) compared to eggNOG's annotation of 82 (82%) proteins, which is 8 (8%) more compared to eggNOG for case study 1 and in case study 2 our annotation pipeline was able to annotate completely and accurately 78 (81.25%) hypothetical proteins out of 96 (100%) compared to eggNOG's annotation of 58 (60.42%) proteins, which is 20 (20.83%) more compared to eggNOG.

Acknowledgment

Authors wish to acknowledge their sincere thanks to The Biome Research Facility, Kolkata for their support in this research work. No external grant has been used for this work.

References

1. Eisenstein E, Gilliland GL, Herzberg O et al. (2000) Biological function made crystal clear - Annotation of hypothetical proteins via structural genomics. Current Opinion in Biotechnology 11 (1): 25–30. doi: 10.1016/S0958-1669(99)00063-4.
2. Sivashankari S, Shanmughavel P (2006) Functional annotation of hypothetical proteins – A review. Bioinformation 1 (8): 335–338. doi: 10.6026/97320630001335.
3. Naveed M, Matloob M, Aziz U et al. (2016) Structural and Functional Characterization of a Hypothetical protein of Streptococcus Pyrogenes: An In-Silico Approach.
4. Kim KS, Kaplan EL (1985) Association of penicillin tolerance with failure to eradicate group A streptococci from patients with pharyngitis. The Journal of Pediatrics 107 (5): 681–684. doi: 10.1016/S0022-3476(85)80392-9.
5. Lamagni TL, Darenberg J, Luca-Harari B et al. (2008) Epidemiology of severe Streptococcus pyogenes disease in Europe. Journal of Clinical Microbiology 46 (7): 2359–2367. doi: 10.1128/JCM.00422-08.
6. Barragan-Osorio L, Giraldo G, J. Almeciga-Diaz C et al. (2015) Computational Analysis and Functional Prediction of Ubiquitin Hypothetical Protein: A Possible Tar-

get in Parkinson Disease. Central Nervous System Agents in Medicinal Chemistry 16 (1): 4–11. doi: 10.2174/1871524915666150722120605.

7. Finn RD, Mistry J, Schuster-Böckler B et al. (2006) Pfam: clans, web tools and services. Nucleic acids research 34 (Database issue): D247-51. doi: 10.1093/nar/gkj149.
8. Quevillon E, Silventoinen V, Pillai S et al. (2005) InterProScan: protein domains identifier. Nucleic Acids Research 33 (Web Server): W116–W120. doi: 10.1093/nar/gki442.
9. Yu NY, Wagner JR, Laird MR et al. (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes | Bioinformatics | Oxford Academic. Bioinformatics 26 (13): 1608–1615.
10. Yu CS, Chen YC, Lu CH, Hwang JK (2006) Prediction of protein subcellular localization. Proteins: Structure, Function, and Bioinformatics 64 (3): 643–651. doi: 10.1002/prot.21018.
11. Bendtsen JD, Jensen LJ, Blom N et al. (2004) Feature-based prediction of non-classical and leaderless protein secretion. Protein Engineering, Design and Selection 17 (4): 349–356. doi: 10.1093/protein/gzh037.
12. Jensen LJ, Julien P, Kuhn M et al. (2007) eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Research 36 (Database): D250–D254. doi: 10.1093/nar/gkm796.
13. Ikai A (1980) Thermostability and Aliphatic Index of Globular Proteins. The Journal of Biochemistry 88 1895–1898. doi: 10.1093/oxfordjournals.jbchem.a133168.
14. Guruprasad K, Reddy BVB, Pandit MW (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Engineering, Design and Selection 4 (2): 155–161. doi: 10.1093/protein/4.2.155.
15. Rogers S, Wells R, Rechsteiner M (1986) Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. Science 224 (4655): 1343–1346. doi: 10.1126/science.6374895.
16. Fink AL (1998) Protein aggregation: Folding aggregates, inclusion bodies and amyloid. Folding and Design 3 (1): R9–R23. doi: 10.1016/S1359-0278(98)00002-9.