# Review Article: Necessity of User Friendly Sequence Analysis Tools & Servers and Different Analysis Through Computational Biology Approaches in Modern era.

Gaurav Kumar Srivastava[1], Dr. Santosh Kumar[2], Dr. Himanshu Pandey[3]

[1]*Research Scholar, Maharishi University of Information technology, lucknow*

[2]*Assosiate Professor, Maharishi University of Information technology, lucknow*

[3]*Assistant Professor, BBDNIIT, lucknow*

***Abstract-*** The Indian subcontinent is a vast repository of medicinal plants that are used in traditional medical treatments. In India around 20,000 medicinal plants have been recorded.But very few plants are in use for curing different diseases. The medicinal plants are listed in various indigenous systems such as Siddha (600), Ayurveda (700) and Amchi (600), Unani (700), Allopathy which 30 plant species for ailments. It has stood the test of time for centuries in protecting the human health and vigor against diseases. This database will contain all genome for different Ayurvedic plants having medicinal applications. High-throughput biology technologies have resulted the complete sequences and functional genomics data for several organisms but still, up to 50% of genes within a genome are unknown and labeled “unknown”, “uncharacterized” or “hypothetical”. Hypothetical proteins are the proteins that have been predicted to be expressed from an open reading frame. Function prediction of uncharacterized proteins in structural biology is a great challenge. Genome encodes thousands of sequences that play significant role in diverse biological process. As the traditional molecular or biochemical experiments for function prediction of genes, genomes are time consuming and costly, hence, it raises the demands of bioinformatics to predict function of protein sequences by developing new tools that would be user friendly. In this review, while focusing on proteins and Gene function, discussion has been made on some of the recent sequence based approaches for function prediction of uncharacterized /hypothetical proteins.

***Keywords-*** *High-throughput, functional genomics, Uncharacterized proteins, hypothetical proteins.*

## I. INTRODUCTION

With the development of biological sciences, huge amount of data including the primary data, such as, genomic sequences along with functional genomic data from high throughput experiments are available globally in the form of various storage devices.But there is a deficiency in the functional annotation of newly sequenced data. Hence, one of the major tasks in this post genomic is to assign function to gene products based on amino acid sequences and genome annotation **[Mazandu*et al.*, 2011]**. The Gene Ontology Consortium classified protein function into three main categories: molecular function, biological process and cellular components **[Ashburner*et. al*., 2000].** Structural Genomics is used to determine the three dimensional structure of a given protein by different experimental methods such as X-ray crystallography, NMR spectroscopy and also by the mean of computational approaches like molecular modeling. On the other hand functional genomics is used to predict the function and interactions of protein and gene using data produced by genome sequencing. Hence, both functional and structural genomics aim is to discover biological function of genes and proteins. There is a challenge in structural genomics in prediction of the function of uncharacterized proteins. When proteins cannot be related to other proteins of known activity, identification of function based on sequence or structural homology is not possible and in such cases it would be useful to assess structurally conserved binding sites in connection with the protein function. Computational proteomics plays a crucial role in the annotation of newly sequenced genomes [**Galperin*et.al*., 2000.]**and also in the interpretation of high throughput experimental data such as gene expression patterns by microarray or protein-protein interaction data. **[Andrade*et. al.*1999].**The function prediction of protein is based on various methodology such as homology based methods, sequence motif based methods, structure based methods, genomic context based methods and network based methods **[Gabaldon*et. al.*, 2004].**

Proteins are involved in many cellular processes such as signal transduction, enzyme catalysis and gene expression, they also interact with other proteins to form multi-protein complexes. Currently, some approaches that were currently used to understand protein interactions such as, using the yeast two-hybrid system or tandem-affinity-purification mass spectroscopy, but these methods have some limitations in explaining how the proteins may interact with each other. Although many protein crystal structures are available in the Protein Data Bank (PDB), the problem is that there is only a small population of solved structures for protein-protein complexes, since the dynamics of complex formation complicates crystallization [**Moreiraet. al.2010]**. There are many proteins whose existence has been predicted through wet lab experiments but their functions are not known. Such kinds of protein are known as hypothetical/uncharacterized protein. Basically, hypothetical proteins are created by gene prediction software during genome analysis. When the bioinformatics tool used for the gene identification, finds a large open reading frame without a characterized homologue in the protein database, it returns "hypothetical protein" as an annotation remark. Through high throughput technologies sequencing of several genomes has resulted in numerous predicted open reading frames but their functions cannot be readily assigned. These proteins are either orphan or conserved hypothetical proteins, the quantum of which is 20-40%of proteins encoded in each newly sequenced genome.

Now –a- days the sequencing genomes of numerous organisms have been worked out and it helps in getting large amount of information about cellular biology. Today, it is a biggest challenge for bioinformatics to use this information in discovering the function of proteins. The functional assignments of genes come mostly from various biochemical experimentations, which could be further extended by matching recently sequenced proteins to those that have already been characterized **[Bork *et.al.*, 1999].** As, the characterization of protein function remains a fundamental challenge in functional genomics research, in this paper we uses of computational techniques to predict the function of uncharacterized protein as also further description of advanced understanding of their structural function will be discussed. There are many computational servers available for predicting protein functions.

## II. CURRENT STATUS

As on April 11 2016, NCBI reported about 146,925 uncharacterized proteins in Animals, [68,783] in Plants, [32,583] in Fungi, [7,144] in Protists, [18, 768] in Archaea, and [1,830] in viruses. **[http://www.ncbi.nlm.nih.gov/protein/?term=uncharacterized0+ proteins)**. The mRNA profiling and gene expression analysis provide information for further study of genes using NGS technologies. Figure 1.shows the functional annotation of uncharacterized proteins.

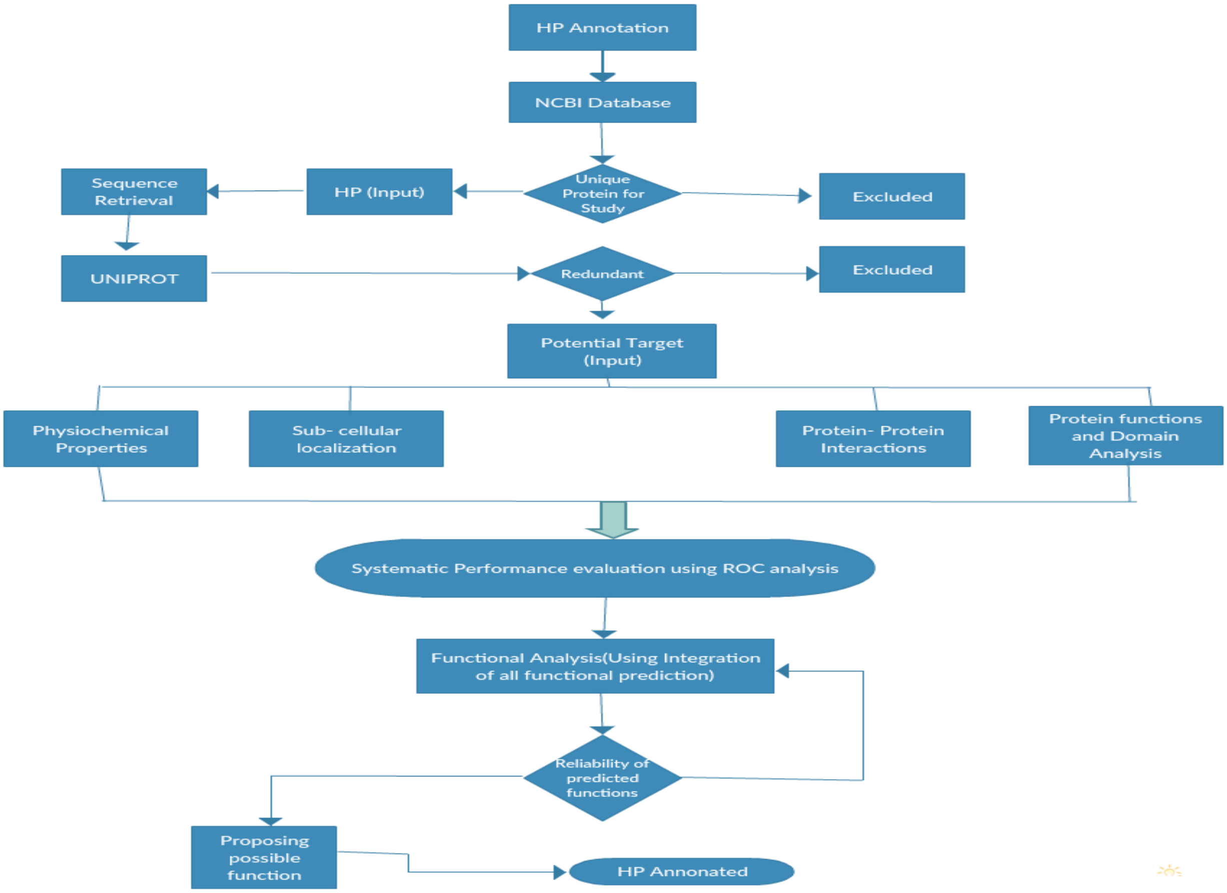
HP Annotation
NCBI Database
Unique Protein for Study
Excluded
HP (Input)
Sequence Retrieval
UNIPROT
Redundant
Excluded
Potential Target (Input)
Physiochemical Properties
Sub- cellular localization
Protein- Protein Interactions
Protein functions and Domain Analysis
Systematic Performance evaluation using ROC analysis
Functional Analysis(Using Integration of all functional prediction)
Reliability of predicted functions
Proposing possible function
HP Annotated

Fig.1: Flowchart showing the computational framework used for annotating function of hypothetical proteins (HPs).Adopted from Shahbaaz*et al.* (2013).

## III. APPROACHES TO ANALYZE FUNCTION OF UNCHARACTERIZED PROTEINS

Functional analyses of proteins through bioinformatics analysis can provide understanding to the work in assigning functions to proteins, especially in a high-throughput setting. There are different approaches through which the function of uncharacterized proteins can be analyzed. By the use of different types of software, it is easy analyze protein sequences and structures for fold and motif similarities.

## IV. SEQUENCE ALIGNMENT METHODS

The amino acid substitution matrix and an alignment algorithm are needed in Protein function analysis to determine the homologous sequences, and the sequences that have been descended from a common ancestral protein sequence.Different substitution matrices have been developed such as BLOSUM **[Henikoff*et. al.*, 1992],** PAM **[Schwartz *et. al.*, 1978]** and Gonnet**[Gonnet*et. al.*, 1994.]** each one measuring these probabilities using different sets of starting alignments that have been manually created.

### A. SEQUENCE BASED FUNCTION PREDICTION OF UNCHARACTERIZED PROTEINS

BLAST and Dali are examples of the classical approaches that rely on sequence or fold similarity searching utilized for function inference. Other sequence level tools include: PSI-BLAST **[Altschul*et.al.*, 1997]** as well as sequence motif search functions in databases like PROSITE **[Sigrist*et.al.*, 2010]**, Pfam**[ Punta *et.al.*, 2012]**, and InterPro**[ Hunter *et.al.*, 2011].**The structural level information like 3D motifs and patterns can be analyzed using programs like Profunc**[Laskowski*et.al.*, 2005]**, SPRITE **[Nadzirin*et.al.*, 2012]**, and PINTS **[Stark *et.al.*, 2003]**.

### B. ANNOTATION DATABASES

Both protein and pattern databases are referred as annotation databases only the difference is that protein databases are based on pairwise sequence similarity determined by BLAST or PSI-BLAST whereas, pattern databases are based on multiple sequence similarity. There are mainly two protein databases that are widely used for annotation: UniProt**[The UniProt Consortium, 2007]**, and SwissProt**[Barioch*et.al.*, 1996]**. Ten different pattern databases can be used for the annotation suing their respective search tools: Clusters of Orthologous Groups (COG) **[Atwood, 2000]**, Protein Clusters (PRK) **[O'Neill *et al.*, 2007; Tatusov*et al.*, 2000],** The InterPro**[Apweiler*et al.*, 2001],**Pfam**[Sonnhammer*et al.*, 1997],**PIRSFScan**[Wu *et al.*, 2003]**, SMART **[Schultz *et al.*, 1998],**and TIGRFAM **[Haft *et al.*, 2003].**BLOCKS **[Henikoff*et al.*, 1999]**, Conserved Domain Database (CDD) **[Marchler-Bauer *et al.*, 2002],**SuperFamily**[Gough *et al.*, 2001].**

### C. CLUSTERING APPROACHES

Clustering is the process of grouping of protein sequences of same functions. Clustering of genes is done by several approaches. By gene clustering method, the function of a hypothetical protein from *E. coli* was predicted to be transcription regulation because it belonged to a cluster containing *tpi*(triose phosphate isomerase, EC 5.3.1.1), *gap* (glyceraldehyde 3- phosphate dehydrogenase, EC 1.2.1.12), *pgk*(phosphoglycerate kinase, EC 2.7.2.3), *pgm*(2,3-bisphosphoglycerate independent phosphoglyceratemutase, EC 5.4.2.1), *eno*(enolase, EC 4.2.1.11) and homologous to hypothetical transcriptional regulator of *Bacillus megaterium*.[**Selvarajan*et. al.*, 2006].** COG (clusters of orthologs) is a databasethat has a large set of "uncharacterized proteins", which includes proteins that are orthologs**. [Tatusov, *et al.*2001]**

### D. STRUCTURE BASED FUNCTION PREDICTION OF UNCHARACTERIZED PROTEINS

The function of a protein is inherently linked to its structure and it is crucial for protein sequences lacking both experimentally determined functions and structures. Methods for structure prediction are increasingly more abundant and accurate including homology modeling **[Fiser*et.al.*, 2003**], *ab initio* modeling [**Skolnick *et.al.*, 2003]** and threading **[Kihara*et.al.*, 2004**] methods, which thread a query sequence through a library of known protein folds. Proteins that share 30% sequence similarity are generally recognized as having similar folds, **[Rost*et.al.*, 1999]** and it has been assumed that during evolution the global folds tend to be more conserved than amino acid sequence **[Wilson *et al.*, 2000].** Protein structure classification databases, such as SCOP (Structural Classification of Proteins)**[Murzin*et.al.*, 1995]**and CATH **[Orengo*et.al.*, 1997]**are the useful resources for predicting protein function.

### E. ACCURACY AND LIMITATIONS OF SEQUENCE BASED FUNCTION PREDICTION METHODS

Sequence similarity searches are generally considered to be simple, accurate and reliable methods of function annotation **[Table. 1].** Within these, FASTA is slightly more accurate then BLAST.52 PSI-BLAST should be used to find more distant homologies.

| *Annotation database* | *Agree* | *Disagree* | *Indeter.* | *Accuracy* | *Cond_Acc* |
|---|---|---|---|---|---|
| TIGRFAM(7.0) | *20* | 7 | 3 | 66.7% | *74.1%* |
| CDD (2.12) | *19* | 5 | 6 | 63.3% | *79.2%* |
| SwissProt (54.4) (B) | *19* | 2 | 9 | 63.3% | *90.5%* |

| | | | | | |
|---|---|---|---|---|---|
| InterPro (16.1) | *18* | 9 | 3 | 60.0% | *66.7%* |
| PRK (1.0) | *18* | 1 | 11 | 60.0% | *94.7%* |
| UniProt (37.4) (PB) | *18* | 9 | 3 | 60.0% | *66.7%* |
| Pfam (22.0) | *15* | 10 | 5 | 50.0% | *60.0%* |
| COG (1.0) | *14* | 9 | 7 | 46.7% | *60.9%* |
| SMART (5.1) | *14* | 11 | 5 | 46.7% | *56.0%* |
| InterPro (16.1)(-Pfam) | *13* | 6 | 11 | 43.3% | *68.4%* |
| TIGRFAM (7.0)(-Pfam) | *12* | 0 | 18 | 40.0% | *100.0%* |
| SuperFamily (1.69) | *10* | 15 | 5 | 33.3% | *40.0%* |
| BLOCKS (14.3) | *7* | 10 | 13 | 23.3% | *41.2%* |
| PIRSF | *5* | 0 | 25 | 16.7% | *100.0%* |

TABLE 1. Table showing accuracy of annotation databases

"InterPro(-Pfam)" and "TIGRFAM (-Pfam)" are results from InterPro and TIGRFAM without Pfam results. "(B)" and "(PB)" stand for BLAST and PSI-BLAST searches of the associated database, respectively. "Cond_Acc" is "Conditional Accuracy." Database versions, in parenthesis, are supplied as well. **[Table adapted from LOUIE *et. al.* 2008]**

The top six databases in the above table have overall accuracy rates of at least 60%: TIGRFAM, CDD, SwissProt, InterPro, PRK, and UniProt, where TIGERFAM and PIRSF shows Conditional accuracy : 100%. The annotations produced by these databases were very likely to agree with the benchmark. Thislevel of accuracy can only be achieved by using top-hits from multiple databases. The top hits from TIGRFAM, Pfam, PRK, and COG can achieve this level of accuracy. Pfam and SuperFamily, search for very distant relationships between proteins.

### F. ACCURACY AND LIMITATIONS OF STRUCTURE BASED FUNCTION PREDICTION METHODS

Structural similarity is a very accurate method of predicting protein function as theglobal fold of a protein determines the shape and the location of active and binding sites, whereas the local structural environment determines the catalytic mechanisms of enzymes. The table 2 shows the complete set of bioinformatics tools and databases used for function annotation of uncharacterized proteins.

| S.No. | Software | Function |
|---|---|---|
| A | **A Sequence similarity search** | |
| 1. | Basic local alignment tool (BLAST) | Used for finding similar sequences in protein databases |
| B | **Physiochemical characterization** | |
| 2. | ExPASy – Protparam tool | Protparam tool Used for computation of various physical and chemical parameters like molecular weight, isoelectric point (Pi), amino acid composition, atomic composition, extinction co-efficient, instability index, aliphatic index, and grand average of hydropathy (GRAVY) |
| C | **Sub-cellular localization** | |
| 3. | signalP | Predicts signal peptide cleavage sites. |
| 4. | secretomeP | Used for identifying proteins involved in non-classical secretory pathway. |
| 5. | PSORT B | Predicts subcellular localization of bacterial proteins |
| 6. | PSLpred | Predicts subcellular localization of proteins from Gram-negative bacteria**.** |

| | | |
|---|---|---|
| 7. | CELLO | Assign localization to both prokaryotic and eukaryotic proteins |
| 8. | TMHMM | used to authenticate whether the protein is a membrane protein or not |
| 9. | HMMTOP | Predict transmembrane topology |
| **D.** | **Domain analysis** | |
| 10. | Pfam | Collection of multiple protein sequence alignments |
| 11. | SVMprot | SVM (Support vector machine based classification of proteins) |
| 12. | SYSTERS | For grouping of proteins on the basis of their functions. |
| 13. | SUPERFAMILY | Hierarchical domain classification of PDB structures. NCBI Entrez protein database search of domain architecture |
| 14. | CATH (Class, Architecture, Topology, Homology) | Used for finding protein similarities across evolutionary distances based on domain architecture. |
| 15. | CDART (The conserved domain architecture and retrieval tool) | Classification based on HMM–HMM search. PANTHER is a comprehensively organized database of protein families sub-families, their evolutionary relationships in the form of phylogenetic trees |
| 16. | PANTHER (Protein analysis through evolutionary relationships) | Identification and annotation of protein domains. |
| 17. | SMART | Automatic hierarchical clustering of the protein sequences |
| **E** | **Motif Analysis** | |
| 19. | InterProScan | InterProScan Searches for motif discovery. It is the integration of several large protein signature databases. |
| 20. | Motif | Used for motif discovery |
| 21. | MEME Suite | Database searching for assigning function to the discovered motifs. |
| **F** | **Protein–Protein interaction** | |
| 22. | STRING | Used for predicting protein–protein interactions. |

| | | |
|---|---|---|

Table 2. List of bioinformatics tools and databases used for sequence based function annotation of uncharacterized proteins

## V. CONCLUSION

The large quantity of uncharacterized proteins makes the study of proteins important for their structural and functional information. The introduction of whole-genome sequences and mRNA profiling, has created new opportunities for computational biologists. The information from comparative genome analysis is used to reconstruct a protein's evolution, and also helps in finding the functions of uncharacterized proteins. Hence, the ability to analyze the expression levels of every gene within a genome is also developing our ability to understand the function of co-regulated proteins and its transcriptional regulation. Identification of the uncharacterized proteins are important for the functional interpretation of fully sequenced genomes and further understanding of the diverse functions of its structures. Development of computational approaches and programs create an opportunity for biologists to produce a complete record of their biological functions and the genes and proteins involved. That's why there is necessity of user friendly sequence analysis tools & servers and different analysis through computational biology approaches in modern era.

In the future, these discernments will be used by computational biologists to model cellular pathways. It is already possible to begin to model developmental pathways signal transduction pathways and metabolic pathways and also compare the predictions of these models to experimental results. In the next few years there will undoubtedly be new approaches that combine genome wide experimental measurements with complex mathematical modeling, to gain an exceptional understanding of protein function of uncharacterized proteins and cellular biology.

## VI. REFERENCES

[1]. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402.

[2]. Andrade MA, Brown NP, Leroy C, et. al. (1999) automated genome sequence analysis and annotation, Bioinformatics 15(5):391-412, 1999.

[3]. Apweiler, R., Attwood, T., Barioch, A., Bateman, A., and Birney, E. (2001). The InterPro database, and integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29, 37–40.

[4]. Attwood, T.K. (2000). The quest to deduce protein function from sequence: the role of pattern databases. Int J Biochem Cell Biol 32, 139–155.

[5]. Barioch, A., and Apweiler, R. (1996) The Swiss-Prot protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res 24, 21–25.

[6]. Brenton Louie, Peter Tarczy-Hornoch, Roger Higdon, and Eugene Kolker1 (2008).Validating Annotations for Uncharacterized Proteins in Shewanellaoneidensis. OMICS A Journal of Integrative Biology Volume 12, Number 3, 2008 © Mary Ann Liebert, Inc.DOI: 10.1089/omi.2008.0051

[7]. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. J Mol Biol. 1998;283:707–725.[PubMed].

[8]. Edwards JS, Palsson (2000) BO: The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc. Natl. Acad. Sci. USA 2000, 97:5528-33.

[9]. Enault, F.; Suhre, K.; Claverie, J. M. Phydbac (2005) "Gene function Predictor": A Gene annotation tool based on genomics context analysis. BMC Bioinforma. 2005,6.

[10]. Pandey. H and Kumar. S, "Proposed Methodology for crowdsourcing and agile development", International Journal of Advanced Research in Engineering and Technology (UGC approved Journal), Volume-9, Issue-2, PP. 68-76, ISSN: 0976-6499(Online), Mar–April. 2018.

[11]. Fiser A, Sali A, (2003) Modeller: Generation and refinement of homology-based protein structure models, Methods Enzymol374:461–491.

[12]. Gabaldon, T; M.A. Huynen (2004)."Prediction of protein function and pathways in the genome era". Cellular and Molecular Life Sciences 61 (7–8): 930–944. doi:10.1007/s00018-003-3387-y. PMID 15095013.

[13]. Galperin MY, Konin EV, (2000)Who's your neighbor? New computational approaches for functional genomics, Nat Biotechnol 18(6): 609-613, 2000.

[14]. Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001).Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J MolBiol 4, 903–919.

[15]. Gonnet G.H., Cohen M.A., Benner S.A. (1994): Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix.";Biochem. Biophys. Res. Commun. 1994, 199:489-496.

[16]. Haft, D.H., Selengut, J.D., and White, O. (2003).The TIGRFAMs database of protein families. Nucleic Acids Res 31, 371–373.

[17]. Pandey. H and Kumar. S, "Enhancing Efficiency of agile software engineering by using crowdsourcing", International Organization of Scientific Research: Journal of Engineering (UGC approved Journal), Volume-8, Issue-4, PP. 71-77, ISSN: 2250-3021 (Online), Apr. 2018.

[18]. Henikoff S and Henikoff JG (1992).Amino acid substitution matrices from protein blocks. Proc. Natl Acad. Sci. USA 1992, 89:10915–10919

[19]. Henikoff, S., Henikoff, J., and Pietrokovski, S. (1999). Blocks_: a non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics 15, 471–479.

[20]. Hunter S., Jones P., Mitchell A., Apweiler R., Attwood T.K., Bateman A., Bernard T., Binns D., Bork P., Burge S., et al.

InterPro (2011): New developments in the family and domain prediction database.Nucleic Acids Res. 2012;40:D306–D312.

[21]. Pandey. H and Kumar. S, “Crowdsourcing Rules in Agile Software Engineering to Improve Efficiency using Ontological Framework”, American International Journal of Research in Science, Technology, Engineering & Mathematics (UGC approved Journal), Volume-1, Issue-22, PP. 24-30, ISSN: 2328-3580(Online), Mar-May. 2018.

[22]. Kihara D, Skolnick J, (2004). Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR Q, Proteins 55(2):464–473.

[23]. Laskowski R.A., Watson J.D., Thornton J.M. (2005) ProFunc: A server for predicting protein function from 3D structure. Nucleic Acids Res. 2005;33:W89–W93.

[24]. Marchler-Bauer, A., Panchenko, A., Shoemaker, B., Thiessen, P.,Geer, L., and Bryant, S. (2002). CDD: a database of conserveddomain alignments with links to domain three-dimensionalstructure. Nucleic Acids Res 30, 281–283.

[25]. Pandey. H and Kumar. S, “A Study of the pertinence of Crowdsourcing in Agile Software Development”, International Journal of Research in Electronics and Computer Engineering (UGC approved Journal), Volume-5, Issue-4, PP. 539-543, ISSN: 2348-2281 (Online), Oct.-Dec. 2017.

[26]. Mazandu, G.K; Mulder, N.J. (2011) Scoring protein relationships in functional interaction networks predicted from sequence data. PLoS One 6, doi:10.1371/journal.phone.0018607.

[27]. Michael Ashburner, et. al. (2000).The Gene Ontology Consortium (2000). "Gene ontology: tool for the unification of biology". Nature Genetics 25 (1):25–29. doi:10.1038/75556. PMC 3037419.PMID 10802651.

[28]. Moreira, I.S., Fernandes, P.A., and Ramos, M.J. (2010).Protein–protein docking dealing with the unknown. Journal of Computational Chemistry 31, 317-342

[29]. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures, J MolBiol247(4):536–540.

[30]. Nadzirin N., Gardiner E.J., Willett P., Artymiuk P.J., Firdaus-Raih M. (2012) SPRITE and ASSAM: Web servers for side chain 3D-motif searching in protein structures. Nucleic Acids Res. 2012;40:W380–W386.

[31]. O’Neill, K., Klimke, W., and Tatusova, T. (2007). Protein Clusters:A Collection of Proteins Grouped by Sequence Similarity and Function.(NCBI, Bethesda, MD).

[32]. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997). CATH–a hierarchic classification of protein domain structures, Structure 5(8):1093–1108.

[33]. Punta M., Coggill P.C., Eberhardt R.Y., Mistry J., Tate J., Boursnell C., Pang N., Forslund K., Ceric G., Clements J., et al. (2012) The Pfam protein families database. Nucleic Acids Res. 2012;40:D290–D301.

[34]. Gupta. B and Pandey. H, “Web Mining for Personalization: A Survey in the Fuzzy Framework”, Asian Journal of Computer and Information Systems, Volume 04, Issue 01, PP. 1-7, ISSN: 2321 – 5658, Feb. 2016.

[35]. Pandey. H and Kumar. S, “Anatomize the suitability of crowdsourcing in agile software development”, Global Journal of Engineering Science and Researches, (UGC approved Journal), ISSN 2348 – 8034, Feb.2018.

[36]. R. L. Tatusov, et al., Nucleic Acids Research, 29 (2001) [PMID:11125040]

[37]. Rost B, (1999). Twilight zone of protein sequence alignments, Protein Eng12(2):85–94.

[38]. Schwartz R.M., Dayhoff M.O. (1978): Matrices for detecting distant relationships. (In) Atlas of Protein Sequence and Structure, 5 suppl. 3:353-358, Nat. Biomed. Res. Found., Washington D.C

[39]. Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. ProcNatlAcadSci USA 95, 5857–5864.

[40]. Schoeberl B, Eichler-Jonsson C, Gilles ED, Muller G. (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nat Biotechnol. 2002, 20(4):370-5.

[41]. SelvarajanSivashankari and PiramanayagamShanmughavel (2006).Functional annotation of hypothetical proteins – A review.Bioinformation 1(8): 335 -338 (2006)

[42]. Shahbaaz,M.,Hassan,M.I.,andAhmad,F.(2013).Functionalannotation of conservedhypotheticalproteinsfrom Haemophilusinfluen-zaeRd KW20. PLoSONE8:e84263.doi:10.1371/journal.pone. 0084263

[43]. Sigrist C.J., Cerutti L., de Castro E., Langendijk-Genevaux P.S., Bulliard V., Bairoch A., Hulo N. (2010). PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res.2010;38:D161–D166.

[44]. Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki M, Szilagyi A, Kihara D, (2003) TOUCHSTONE: A unified approach to protein structure prediction, Proteins (53 Suppl 6):469–479.

[45]. Stark A., Sunyaev S., Russell R.B. (2003) A model for statistical significance of local similarities in structure. J. Mol. Biol. 2003;326:1307–1316.

[46]. Sonnhammer, E., Eddy, S., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28, 405–420.

[47]. Tatusov, R., Galperin, M., Natale, D., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28, 33–36.

[48]. Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y et. al.: E-Cell: software environment for whole-cell simulation. Bioinformatics 1999, 15:72-84.

[49]. Von Dassow G, Meir E, Munro EM, Odell GM: The segment polarity network is a robust developmental module. Nature 2000, 406:188-92.

[50]. Wilson CA, Kreychman J, Gerstein M (2000). Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores, J MolBiol297 (1):233–249.

[51]. Wu, C., Huang, H., Yeh, L., and Barker, W. (2003).Protein family classification and functional annotation.CompBiolChem 27, 37–47.