Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Functional annotation of hypothetical proteins – A review

Selvarajan Sivashankari[1] and Piramanayagam Shanmughavel[2*]

[1]Department of Bioinformatics, Kongunadu arts and science college, Coimbatore - 641029, India; [2]Computational Biology Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore - 641046, India;
Piramanayagan Shanmughavel* - Email: shanvel_99@yahoo.com; * Corresponding author

received August 15, 2006; revised October 20, 2006; accepted November 1, 2006; published online December 29, 2006

Abstract:

The complete human genome sequences in the public database provide ways to understand the blue print of life. As of June 29, 2006, 27 archaeal, 326 bacterial and 21 eukaryotes is complete genomes are available and the sequencing for 316 bacterial, 24 archaeal, 126 eukaryotic genomes are in progress. The traditional biochemical/molecular experiments can assign accurate functions for genes in these genomes. However, the process is time-consuming and costly. Despite several efforts, only 50-60 % of genes have been annotated in most completely sequenced genomes. Automated genome sequence analysis and annotation may provide ways to understand genomes. Thus, determination of protein function is one of the challenging problems of the post-genome era. This demands bioinformatics to predict functions of un-annotated protein sequences by developing efficient tools. Here, we discuss some of the recent and popular approaches developed in Bioinformatics to predict functions for hypothetical proteins.

Keywords: hypothetical; protein function; functional annotation; prediction; assignments

Background:

Genome research started 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 is under way. Inconsistency in the accuracy of genome annotation that was the subject of many heated discussions at the beginning of the genome era. **[1]** 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. **[2]** The remaining genes are either (i) homologous to genes of unknown function, and are typically referred to as "conserved hypothetical" genes, or (ii) do not have any known homologs termed "hypothetical" or "non characterized" or "unknown" because it is unclear whether they encode actual proteins. Since it is often unclear whether they encode actual proteins, the latter genes are commonly referred to as "hypothetical", "uncharacterized", or "unknown" proteins. As of April 25, 2006, the NCBI protein database contained 19,85,480 protein sequences from ~373 completely sequenced genomes; one out of three proteins had no assigned function and one out of ten proteins was annotated as "conserved hypothetical". Even for *Escherichia coli* strain K-12, the best studied of all organisms, there are still ~2000 genes that have never been experimentally characterized, almost half of all proteins encoded in its genome. At the current rate of experimental characterization of new *E. coli* genes, 20 ± 30 per year, it will take many decades before the biological function of all these proteins is established. **[3]**

Several approaches have been developed for predicting protein function using the information derived from sequence similarity, phylogenetic profiles, protein-protein interactions, protein complexes and gene expression profiles. The classical way to infer function is based on sequence similarity using sequence database searching programs such as FASTA **[4]** and PSI-BLAST. **[5]** Lack of sequence similarity in the database to the protein of interest creates difficulties for functional predictions. However, examples of dissimilar function for similar proteins are also available. Thus, approaches to predict protein fucntion by *in silico* methods are discussed below.

Methodology:

Methods based on protein-protein interactions

Proteins often interact with one another in a mutually dependent way to perform a common function. As an example, the transcription factors interact among themselves to bring about transcription. It is therefore possible to infer the functions of proteins based on their interaction partners. The Rosetta-Stone approach **[6]** is a method to predict function based on protein fusion events. Two polypeptides A and B in one organism are likely to interact if their homologs are expressed as a single polypeptide AB in another organism. The latter polypeptide (AB) is called a Rosetta stone protein, as it contains information about both A and B. This method can be effective because a biochemical function in many cases depends on the action of a multi-meric complex demonstrating a correlation between co-interacting proteins and their functions. Although Rosetta protein approach seems approved, Rosetta protein may not be a proof for protein-protein interactions. **[7]**

Methods based on comparative genomics

Comparative genomics is the study of relationships between genomes of different species. This method is based on the assumption that proteins that function together

either in a metabolic pathway or in structural complex are expected to evolve together. During evolution, all such functionally linked proteins tend to be either preserved or eliminated in a new species. Proteins within these groups are defined as functionally linked. For example, two proteins are functionally linked if they have homologs in a group of organisms. Phylogenetic Profiling can detect such functionally linked Proteins. **[8]** To represent a group of organisms that contain a homolog a phylogenetic profile for each protein is created. Phylogenetic profile is a string with one bit and 'n' entries, where n is the number of genomes under consideration. If the n^{th} genome contains a homolog for the protein then the nth entry is represented as unity in the phylogenetic profile. These profiles are clustered to determine which proteins have common profiles. Proteins with identical or similar profiles are functionally linked. This method can identify functionally linked proteins with no amino acid sequence similarity so that the function of the hypothetical protein can be known. This method was tested using three proteins the ribosome protein RL7 and the flagellar structural protein FlgL, as well as a protein known to participate in a metabolic pathway, the histidine biosynthetic protein HIS5. **[8]** The comparisons of phylogenetic profiles for flagellar proteins have revealed that proteins with similar profiles are likely to be functionally linked. Thus, phylogenetic profiling can aid in predicting functions of several other proteins with the same profiles and no assigned function (Hypothetical Proteins).

Function assignment based on 3D structures

Structures **[9]** of hypothetical proteins may provide a hint for their biochemical or biophysical functions. 3D structure can aid the assignment of function for uncharacterized proteins. During evolution, the folding patterns of proteins are often preserved and hence structure based comparisons can identify homologs where the sequence based comparisons become futile. As an example, the crystal structure of a hypothetical protein, MJ0577, from a hyper-thermophile, *Methanococcus jannaschii*, at 1.7 Å resolutions contains a bound ATP, suggesting MJ0577 is an ATPase. The structure also shows different ATP binding motifs that are shared among many homologous hypothetical proteins in this family. **[10]** Thus, structure-based assignment of molecular function is a viable approach for large-scale biochemical assignment of proteins and for discovering new motifs. Nevertheless, prediction of protein function from sequence and structure is a difficult problem, because homologous proteins do different functions in several cases. Many methods of function prediction rely on identifying similarity in sequence and/or structure between a protein of unknown function and one or more well understood proteins. **[11]**

Clustering approaches

Clustering is the process of grouping on the basis that genes of the same cluster are involved in similar function. Hence, the protein that is coded by this gene will also have the same function. Clustering of genes is done by several approaches. According to Overbeek and colleagues **[12]** clusters of genes is based on the definition that a set of genes occurring on a prokaryotic chromosome will be called a ''run'' if and only if they all occur on the same strand and the gaps between adjacent genes are 300 bp or less. It should be noted that any pair of genes occurring within a single run is called ''close.'' Given two genes Xa and Xb from two genomes Ga and Gb, Xa and Xb are called a ''bidirectional best hit (BBH)'' only if recognizable similarity exists between them and there is no gene Zb in Gb that is more similar than Xb is to Xa, and there is no gene Za in Ga that is more similar than Xa is to Xb. Genes (Xa, Ya) from Ga and genes (Xb, Yb) from Gb form a ''pair of close bidirectional best hits (PCBBH)'' if and only if Xa and Ya are close, Xb and Yb are close, Xa and Xb are a BBH, and Ya and Yb are a BBH. 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 phosphoglycerate mutase, EC 5.4.2.1), *eno* (enolase, EC 4.2.1.11) and homologous to a hypothetical transcriptional regulator of *Bacillus megaterium*. This conveys functional coupling within members of a gene clusters which has led to the development of database for COG (clusters of orthologous groups). **[13]** COG includes proteins that are orthologs. This also involves one-to-many and many-to-many relationships. Howver, it should be noted that the COG database has a large set of "uncharacterized proteins".

Genome context methods

New methods are designed to detect alleged functional constraints on genome evolution, and are called 'genomic context' approaches. They predict functional associations between protein coding genes by analyzing gene fusion events, the conservation of gene neighborhood, or the significant co-occurrence of genes across different species. Unlike homology-based annotation, genomic context methods predict functional associations between proteins, such as physical interactions, or co-membership in pathways, regulators or other cellular processes. Characterizing protein function in this manner is intuitive and generally applicable, but it should be noted that it does not provide information about the exact biochemical or enzymatic function of a protein. Genomic context methods have been successfully used to study protein associations, either individually or in combination with other methods or data sets. Various new methods have been proposed to predict functional interactions between proteins based on the genomic context of their genes. The types of genomic context that they use are (1) fusion genes; (2) conservation of gene-order or co-occurrence of genes in potential operons; and (3) co-occurrence of genes across genomes

(phylogenetic profiles). Despite these efforts more than 35% of genes in prokaryotes are still annotated as 'function unknown'. **[14]** With this approach new functional features of *M. genitalium* proteins were detected. Hence, there is a correlation between the spatial proximity of genes on the genome and the directness of the interaction between proteins they encode.

Knowing the importance of context information the database STRING **[15]** was developed which is a pre-computed global resource for the exploration and analysis of these associations. Since the three types of evidence differ conceptually, and the number of predicted interactions is very large, it is essential to be able to assess and compare the significance of individual predictions. Thus, STRING contains a unique scoring-framework based on benchmarks of the different types of associations against a common reference set, integrated in a single confidence score per prediction. The graphical representation of network inferred, weighted protein interactions provides a high-level view of functional linkage, facilitating the analysis of modularity in biological processes. STRING is updated continuously, and currently it contains 261,033 orthologs in 89 fully sequenced genomes. The database predicts functional interactions at an expected level of accuracy of at least 80% for more than half of the genes.

Other approaches:

Other function prediction methods using high throughput data include machine learning and data mining approaches **[16]** and Markov random fields. **[17]** Instead of searching for a simple consensus among the functions of interacting partners, Deng and colleagues **[18]** used the Bayesian approach to assign a probability for a hypothetical protein to have the annotated function. Another Bayesian approach for combining heterogeneous data in yeast for function assignment has been applied by Troyanskaya and colleagues. **[19]** Cluster analysis of gene-expression profiles is a common approach used to predict function based on the assumption that genes with similar functions are likely to be co-expressed. **[20-22]** Using protein-protein interaction data to assign function to novel proteins is yet another approach. Schwikowski and colleagues (2000) applied neighbor-counting method in predicting function. **[23]** They assigned function to an unknown protein based on the frequencies of its neighbors having certain functions. The method was improved by Hishigaki and colleagues (2001), who used c2 statistics. **[24]** Both the approaches give equal significance to all the functions contributed by the neighbors of the protein.

Conclusion:

The abundance of hypothetical proteins makes their study a formidable task. There is a clear need for rational criteria that would allow sorting 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 advent of comparative genomics, wide (better yet, universal) phylogenetic distribution and indispensability for cell growth have been taken into consideration by some researchers when choosing uncharacterized genes for experimental study. Significant positive correlation between the phyletic spread of a gene and the likelihood that it is essential for cell growth has been demonstrated. On a number of occasions, experiments with proteins that met one or both of these criteria led to major discoveries. For example, the three-dimensional (3D) structure solved by Thomas I. Zarembinski and colleagues **[25]** led to subsequent functional characterization of *Methanococcus jannaschii* protein MJ0226, a member of the widely distributed HAM1 protein family, whose only known function until then had been modulation of sensitivity to 6-N-hydroxylaminopurine mutagenesis in yeast. Characterization of this protein as a XTP and ITP specific pyro-phosphatase, explained its role in mutagenesis control. Moreover, this protein perfectly fits the description of ITP pyrophosphatase (ITPase) from human erythrocytes (EC 3.6.1.19) that had been first reported in 1964, purified and extensively characterized five years later, but had never been identified with a gene. Based on the sequence of MJ0226 protein, its human homolog has been characterized and shown to account for the ITPase activity in humans. Furthermore, although mutations in the ITPase gene did not seem to have a clear disease phenotype, ITPase deficiency has been associated with adverse reactions to purine analog azathioprine, which is used as immunosuppressant in the treatment of cancer and inflammatory bowel disease. This example shows how a supposedly arcane study of an archaeal "conserved hypothetical" protein can have immediate consequences for understanding human physiology and might be relevant for human health.

The recent identification of NAD kinase (EC 2.7.1.23) follows a similar pattern. **[26]** Also, the enzyme has been experimentally characterized many years ago, both in avian tissues and in yeast, but the associated gene remained unknown. Again, the characterization of a bacterial enzyme **[27]** allowed assigning this function to a family of previously uncharacterized 'conserved hypothetical' proteins. Finally, studies on the bacterial enzyme **[28]** paved the way for the identification of an orthologous enzyme in humans and in yeast. Of course, it would be wrong to assume that functional characterization of a "conserved hypothetical" protein would always turn up a previously described enzymatic activity. **[29]** In many cases, the underlying biology and/or biochemistry could be unknown or at least not properly appreciated. Thus, the case of identifying the *E. coli* product of hemK gene **[30]** as a glutamine N5-methyltransferase of peptide release factors pointed out the importance of this post-translational modification that had been previously overlooked. The orthologs of HemK in humans **[31]** and other eukaryotes are annotated (without experimental support) as DNA

methyltransferases. Nonetheless, the glutamine methylation in eukaryotic proteins remains to be investigated. **[32]** Similarly, the recent recognition of the roles of suf genes **[33]** in the assembly of iron - sulfur clusters in bacteria has important implications for understanding the functioning of chloroplasts, where these processes seem to be similar. As in the case of HemK, the original annotation of SufC as 'ABC-type transporter ATPase' turned out to be less than precise: although SufC is certainly an ATPase of the ABC-type ATPase family, it does not seem to participate in transport.

Thus, sequencing of multiple genomes from all walks of life and the concomitant development of computational approaches of comparative genomics create an opportunity for biology that was hardly imaginable 10 years ago: a directed, systematic effort aimed at producing a complete catalog of biochemical activities, biological functions and the responsible genes, at least for simpler, prokaryotic life forms. A co-ordinated program on elucidation of the functions of conserved hypothetical proteins has the potential of taking us a long way on the road to this lofty goal. It is worth emphasizing that the number of conserved hypothetical proteins that are widely represented among diverse life forms is not huge, a few thousand on the outside. However incomplete the current collection of genomes turns out to be, genes from new genomes increasingly fall within already established orthologous gene sets. Thus, although a truly comprehensive gene catalog might belong in the distant future, a concise dictionary of the main functions and the corresponding genes is likely to be well within reach of the current generation of researchers, provided the development of new and newer algorithms for functional annotateion.

References:

[01] S. E. Brenner, *Trends Genet.,* 15:132 (1999) [PMID:10203816]
[02] P. Bork, *Genome Res.,* 10:398 (2000) [PMID:10779480]
[03] E. Kolker, *et al., Nucleic Acids Research,* 32:2353 (2004) [PMID:15121896]
[04] W. R. Pearson, *et al., Proc Natl Acad Sci.,* 85:2444 (1998) [PMID:3162770]
[05] S. F. Altschul & E. V. Koonin, *Trends Biochem Sci.,* 23:444 (1998) [PMID:9852764]
[06] E. M. Marcotte, *et al., Science,* 285:751 (1999) [PMID:10427000]
[07] R. A. Veitia, *Genome Biology,* 3:interactions1001 (2002) [PMID:11864366]
[08] M. Pellegrini, *et al., Proc. Natl. Acad. Sci.,* 96:4285 (1999) [PMID:10200254]
[09] F. C. Bernstein, *et al., J. Mol. Biol.,* 112:535 (1977) [PMID:875032]
[10] T. I. Zarembinski, *et al., Proc. Natl. Acad. Sci.,* 95:15189 (1998) [PMID:9860944]
[11] J. C. Whisstock & A. M. Lesk, *Q Rev Biophys.,* 36:307 (2003) [PMID:15029827]
[12] R. Overbeek, *et al., Proc. Natl. Acad. Sci.,* 96:2896 (1999) [PMID:10077608]
[13] R. L. Tatusov, *et al., Nucleic Acids Research,* 29 (2001) [PMID:11125040]
[14] M. Huynen, *et al., Genome Res.,* 10:1204 (2000) [PMID:10958638]
[15] C. Von Mering, *et al., Nucleic Acids Research,* 31 (2003) [PMID:12519996]
[16] S. Clarke, *Proc. Natl Acad. Sci.,* 99:1104 (2002) [PMID:11830650]
[17] M. Deng, *et al., Proc IEEE Comput Soc Bioinform Conf.,* 1:197 (2002) [PMID:15838136]
[18] M. Deng, *et al., Journal of Computational Biology,* 10:947 (2003) [PMID:14751964]
[19] O. G. Troyanskaya, *et al., Proc Natl Acad Sci.,* 100:8348 (2003) [PMID:12826619]
[20] M. B. Eisen, *et al., Proc Natl Acad Sci.,* 95:14863 (1998) [PMID:9843981]
[21] P. O. Brown, *et al., Bioinformatics,* 1:S49 (2001) [PMID:11472992]
[22] P. Pavliditis, *et al., Genome Res.,* 14:1085 (2004) [PMID :15173114]
[23] B. SchwiKowski, *et al., Nat Biotechnology,* 18:1242 (2000) [PMID:11101803]
[24] H. Hishigaki, *et al., Yeast,* 18:523 (2001) [PMID :11284008]
[25] T. I. Zarembinski, *et al., Proc. Natl. Acad. Sci.,* 95:15189 (1998) [PMID:9860944]
[26] M. Y. Galperin & E. V. Koonin, *Nucleic Acids Research,* 32:5452 (2004) [PMID:15479782]
[27] D. K. Apps, *et al., Eur. J. Biochem.,* 55:475 (1975) [PMID:239]
[28] Y. M. Tseng, *et al., Biochim. Biophys., Acta,* 568:205 (1979) [PMID:132429]
[29] S. Kawai., *et al., Eur. J. Biochem.,* 268:4359 (2001) [PMID:11488932]
[30] K. Nakahigashi, *et al., Proc. Natl Acad. Sci.,* 99:1473 (2002) [PMID:11805295]
[31] V. Heurgue-Hamard, *et al., EMBO J.,* 21:769 (2002) [PMID:11847124]
[32] S. Clarke, *Proc. Natl Acad. Sci.,* 99:1104 (2002) [PMID:11830650]
[33] L. Loiseau, *et al., J. Biol. Chem.,* 278:38352 (2003) [PMID:12876288]

Edited by R. Sowdhamini

Citation: Sivashankari & Shanmughavel, Bioinformation 1(8): 335-338 (2006)