BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(7)

An improved hypergeometric probability method for identification of functionally linked proteins using phylogenetic profiles

Appala Raju Kotaru[1], Khader Shameer[2], Pandurangan Sundaramurthy[3, 4]*, Ramesh Chandra Joshi[1]*

[1]Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee, 247667, Roorkee, India; [2]Division of Biomedical Statistics and Informatics & Division of Cardiovascular Diseases, Mayo Clinic, Rochester 55905, USA; [3]Department of Mathematics, Indian Institute of Technology Roorkee, 247667, Roorkee, India; [4]School of Advanced Sciences, VIT University, Vellore - 632014, Tamil Nadu, India; Ramesh Chandra Joshi – Email: rcjosfec@gmail.com; Pandurangan Sundaramurthy – Email: sundaramurthy.p@vit.ac.in; *Corresponding authors

Received January 16, 2013; Accepted March 06, 2013; Published April 13, 2013

Abstract:
Predicting functions of proteins and alternatively spliced isoforms encoded in a genome is one of the important applications of bioinformatics in the post-genome era. Due to the practical limitation of experimental characterization of all proteins encoded in a genome using biochemical studies, bioinformatics methods provide powerful tools for function annotation and prediction. These methods also help minimize the growing sequence-to-function gap. Phylogenetic profiling is a bioinformatics approach to identify the influence of a trait across species and can be employed to infer the evolutionary history of proteins encoded in genomes. Here we propose an improved phylogenetic profile-based method which considers the co-evolution of the reference genome to derive the basic similarity measure, the background phylogeny of target genomes for profile generation and assigning weights to target genomes. The ordering of genomes and the runs of consecutive matches between the proteins were used to define phylogenetic relationships in the approach. We used *Escherichia coli K12* genome as the reference genome and its 4195 proteins were used in the current analysis. We compared our approach with two existing methods and our initial results show that the predictions have outperformed two of the existing approaches. In addition, we have validated our method using a targeted protein-protein interaction network derived from protein-protein interaction database STRING. Our preliminary results indicates that improvement in function prediction can be attained by using coevolution-based similarity measures and the runs on to the same scale instead of computing them in different scales. Our method can be applied at the whole-genome level for annotating hypothetical proteins from prokaryotic genomes.

Keywords: Protein function prediction, phylogenetic profiles, functional annotation, functional similarity.

Background:
Predicting the functions of uncharacterized proteins from their sequence is one of the major goals of bioinformatics. Large-scale genome projects and high-throughput experiments are generating enormous amounts of data. The central challenge of bioinformatics however, is to derive biologically valid information to understand the function of proteins. The concept of protein function is highly context-sensitive and typically acts as an umbrella term for all types of activities that a protein is involved in, be it cellular, molecular, metabolic, structural or physiological mechanisms. Proteins play a crucial role in mediating function in different contexts by interacting with other biological macromolecules or small molecules **[1-3]**. The functions of proteins in different cellular or pathological contexts were previously deciphered through biochemical experiments. However, irrespective of the validated functional

data, these approaches have low throughput because of the experimental effort required in analyzing a single gene or protein. Due to such limitation of experimental methods for functional characterization, a large fraction of proteins in the protein sequence database remains un-annotated. Approximately 20%, 7%, 10% and 1% of annotated proteins in the *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Caenorhabiditis elegans* genomes respectively, have been experimentally characterized **[4].** The delay in the experimental characterization of the biochemical role of proteins resulted in a continually expanding sequence-function gap **[3, 5].** This may further hinder the systematic understanding of the biological processes or molecular mechanisms mediated by them. To minimize such sequence-to-function gap various bioinformatics procedures have been proposed to predict the functional mechanisms associated with a protein and its role in mediation complex diseases including cancer **[6-8]**. In a typical function association experiment, researchers have been using nucleotide sequences in case of genes, or amino acid sequences in case of proteins to determine the function of genes or the corresponding proteins **[9]**. This approach relies on the fact that sets of genes that have sufficiently similar sequences may also perform the same function. The explosive growth of the amount of sequence information available in public databases has made such an approach particularly accurate. Phylogenomics approaches **[10, 11]** can be considered as a useful tool in functional genomics, personalized medicine and genomic medicine **[12, 13]** due to its predictive power. Several database and webservers that utilize the phylogenomic information for functional inference of related proteins are available. Multiple approaches are currently available to investigate the mechanisms leading to the accomplishment of a protein's function and its therapeutic role. These procedures have generated a wide variety of useful data such as gene expression data sets **[9, 14-17],** phylogenetic data **[6, 18-27]** and protein interaction networks **[7].** These data sets offer various insights into a protein's function and related concepts. Recently, protein structural data was also analyzed in the perspective of phylogeny and the pathogenic role of functional variants were also assessed from evolutionary perspective to understand the role of non-synonymous variations and disease phenotypes **[28]**. In this study, we considered phylogenetic profiles **[1-3]** that come under phylogenetic data category for predicting the function of gene products **[10, 11, 29]**. The phylogenetic profile of a protein is a binary vector whose length is the number of available genomes. The vector contains a 1 in the i^{th} position if the i^{th} genome contains a homologue of the corresponding gene, and 0 otherwise. Phylogenetic profiles were used as a quantitative method to annotate proteins, derive evolutionary relationships **[25, 30],** examine functional transfer between whole genomes, predict functionally linked proteins **[21, 27, 29, 31, 32],** correlate genotype and phenotype **[30],** predict metabolic activities **[23]** and conduct network analysis including network organization **[22, 31, 33].**

In this paper we propose a new method using the weighted hypergeometric probabilistic approach to incorporate two important aspects of functional relatedness. We considered the co-evolution of the reference genome that gives the basic similarity measure, which is the background phylogeny of target genomes for profiles generation and the ordering of genomes was used to derive phylogeny. We compared our results with existing approaches such as weighted hypergeometric probability without runs, weighted hypergeometric probability with runs. Our work focuses on the *Escherichia coli K12* genome where we considered 305 genomes for phylogenetic profiles generation. Functional interactions (protein-protein interactions in this context) were extracted from the STRING database for the validation of generated functional interactions **[34]**. Our experimental results show the accuracy achieved by incorporating the phylogeny in addition to the similarity measure for identification of functionally linked proteins that could include the same pathways or pathophysiological mechanisms. The concept of phylogenetic profiles was introduced and extensively used in function association in the seminal work by Pelligrini *et al* **[6, 18]**. Earlier work in the analysis of phylogenetic profiles can be briefed into three categories based on the measure used to compare pairs of proteins. During the initial stage of application of phylogenetic profiles in functional correlation studies, more emphasis was on various ways of comparing two phylogenetic profiles. These methods ignored the underlying phylogeny of the organisms. The underlying hypothesis was that proteins, which function together in a pathway or a protein complex, are likely to have a similar evolutionary path **[6, 18]**. Enault **[35]** proposed an approach for relaxing phylogenetic profiles particularly for the annotation of bacterial genomes. The modification suggested here is to use the normalized BLAST score **[36]** denoting the best match for a protein in a genome, instead of using a 0 or 1. Wu *et al.* advocated the use of more general measures of similarity for pairs of phylogenetic profiles **[20]**. Three popularly used measures of similarity, namely the Hamming distance (D), Pearson's Correlation Coefficient (r) and mutual information (MI) were evaluated for the task. The second category of work used the underlying phylogeny of organisms and also the relative positions of the genomes while generating phylogenetic profiles. Vert and others proposed the use of support vector machines (SVM) for learning protein functions from their phylogenetic profiles **[19, 37, 38].** However, instead of the common kernel functions such as linear kernels used in SVMs, a tree kernel is proposed to calculate the similarity of the profiles in higher dimensional space used by SVM. Narra *et.al* used the extended real-valued profiles to the above approach **[39].** Here, all the internal nodes of the phylogenetic tree were also assigned scores equal to the average scores at their child nodes. An extended profile can be constructed for each protein by a post-order traversal of the tree. Recently, Kotaru and Joshi proposed a method for classification of phylogenetic profiles using supervised machine learning method which supports vector machine classification along with radial basis function as kernel for identifying functionally linked proteins **[40].** In evaluation using three-fold cross validation on the same data, performance of the radial basis kernel is similar to polynomial kernel. In case of some functional classes application of both kernels together were better than linear and tree kernel **[19]**, and over all radial basis kernel have shown to outperform the polynomial kernel **[39]**, linear kernel and tree kernel **[19].** The third category of work is an approach that considers only an ordering of genomes and not a full phylogenetic tree. Cokus *et al*. proposed a method based on similar kind of metric, which considers ordering of genomes and clustering optimization using swiveling technique **[41, 42].** They showed that such an approach superior to the first class of metric that considered only co-evolution because the current method considers both

co-evolution and phylogeny. One drawback such method was while calculating the runs probability in the conditional probability, rather than considering the runs of the proteins, the similarity was taken as the right side of conditional probability.

Methodology:

The phylogenetic profile of a protein can be described as a string that encodes the presence or absence of the protein from the reference genome in every sequenced target genome. It is a binary vector whose length is the number of sequenced target genomes. The vector contains 1 in the i^{th} position if the i^{th} genome contains a homologue of the corresponding gene, else a zero **[6, 18, 41].** The homologue of the genes is obtained using the BLASTP (protein-protein Basic Local Alignment Search Tool) algorithm **[36].** Phylogenetic profiles were generated using 305 prokaryotic genomes using proteome sequences downloaded from NCBI database http://www.ncbi.nlm.nih.gov/ **[43].** Profiles were computed for each target organism using BLASTP searches (using an e-value 0.01) **[36]** to define the presence and absence of homologs across the genomes. All the 4,195 genes encoded in the genome of *Escherichia coli K12* were used as query for sequence searches they have the most comprehensive annotations and therefore allow us to accurately assess the performance of methods. We believe similar to previous methods that used phylogenetic profiles for gathering functionally similar proteins in a high-throughput manner from sequence approach, the proposed approach can be applied to other fully sequenced genomes.

A modified weighted hypergeometric probability method

The proposed method considers both similarity and background phylogeny. Briefly the whole methodology is as follows: identifying the order of the genomes using the hierarchal clustering and optimal leaf ordering algorithm; then calculating the two probabilities of the similarity and runs between a given pair of proteins; and finally calculating the total score which gives the functional relatedness between the pair of proteins using the above two probabilities. This is similar to the model Cokus *et al* **[41].** The proposed method was formulated based on two basic hypotheses. The first basic hypothesis is based on the similarity between the two given proteins. The second hypothesis is based on the runs of consecutive matches both the proteins span. Here, a run was defined as the maximal non-empty string of consecutive occupancy matches between two phylogenetic profiles. For calculating runs the ordering of genomes is important because the number of runs generally changes as target genomes were permuted. The ordering of genomes is established such that the order reflects the evolutionary relationships among the target genomes **[44].** Further for hierarchal clustering, we used the target genomes' phylogenetic profiles. For calculating the distance matrix, we used Jaccard dissimilarity to measure the distance between two genomes. Complete linkage was used here to define the pairs using the largest distance between objects in the two clusters. Here, we used the ordering of the genomes that are the leaves of the tree generated from clustering. We have used the complete dendrogram obtained in the above step to infer the co-clustering pattern. Hierarchal clustering is generally topological in nature and there is an ambiguity about the ordering of genomes, we used optimized swiveling approach to handle such ambiguities. . A detailed explanation of the concept of runs, ordering of genome and optimal swiveling is available elsewhere (See Cokus *et al* **[41]).** In short, the two basic hypotheses of our proposed method are: 1) The greater the similarity, the more the proteins that are functionally related **[6, 18]**; 2) The profiles with more runs are more likely to involve functionally related proteins than profiles in which all the matches are concentrated in one interval of the tree **[41].**

The weighted hypergeometric similarity probability was defined as the probability of two phylogenetic profiles having a certain number of matches using an extension of the hypergeometric distribution that accounts for number of proteins in each genome. The basic assumption was based on biologically plausible hypothesis that protein pairs with more matches in their profiles are more likely to co-evolve. We used the same similarity probability defined by Cokus *et al.* **[14].** The similarity probability for a pair of genes (Gene 1 and Gene 2) is defined as the number of genomes that have the first gene (Gene 1) is a >= 0, the number of genomes that have the second gene (Gene 2) is b >= 0 and the number of genomes that have both genes (Gene 1 and Gene 2) is c >= 0. The similarity *P*-value, the number of genomes with both genes, is at least as large as c, given that a and b are defined using equation (1).

$$P(c >= \text{observed} \mid a, b) = P(c >= \text{observed}, a, b) / P(a, b)$$

Where a = number of genomes with Gene 1; b = number of genomes with Gene 2; c = number of genomes with Gene 1 and Gene 2.

The weighted hypergeometric runs probability was defined as the probability of two profiles having a certain number of runs using an extension of the hypergeometric distribution that accounts for the number of proteins in each genome.

The runs probability for a pair of genes (Gene 1 and Gene 2) was defined as the number of runs that have the first gene (Gene 1) in some number r >= 0, the number of runs that have the second gene (Gene 2) is s >= 0 and the number of runs that have both genes t >= 0 is the value of the unique entry of P that is P[r+1, s+1, t+1]. The runs *P*-value, then the number of genomes with both genes (Gene 1 and Gene 2) is at least as large as c given a and b is defined using equation (2).

$$P(t >= \text{observed} \mid r, s) = P(t >= \text{observed}, r, s) / P(r, s)$$

Let k take values 0, 1…n and random variables R_k, S_k and T_k take values in 0…k and R_k be the number of runs that have the Gene 1, S_k be the number of runs that have Gene 2 and T_k be the be the number of runs that have both genes (Gene 1 and Gene 2), restriction to genomes 0…k. To obtain the conditional distribution of T_n given R_n and S_n it is sufficient to calculate the joint distribution of R_n, S_n and T_n. P` represent a 3-dimensional table with three variables, and the runs *P*-value was derived using equation (3)

$$P(t >= \text{observed} \mid r, s) = {}^{n/2}\Sigma_{t=t} P`[r+1, s+1, t`+1] \; / {}^{n/2}\Sigma_{t=0} P`[r+1, s+1, t`+1]$$

We further expanded the approach to score pairs of proteins using the following assumptions. If H is the weighted hypergeometric similarity *P*-value for a given pair of proteins encoded by Gene 1 and Gene 2 and R was the modified

weighted runs *P*-value for the same pair of proteins (Protein 1 and Protein 2), then we scored the pair of proteins as H*R or, on a logarithmic scale score, defined it using equation (4). The lesser the score of a given pair of proteins, the more the pairs were considered as functionally related. The score was derived using equation 4.

$$\text{Score} = \log_{10}H + \log_{10}R$$

The proposed method was implemented in MATLAB (MathWorks, Massachusetts, USA) and validated using a benchmark data set; we have also applied the method to a specific example to illustrate the application. Source code in MATLAB is available from the corresponding authors upon request. An outline of the methodology is provided in (**Figure 1)**.

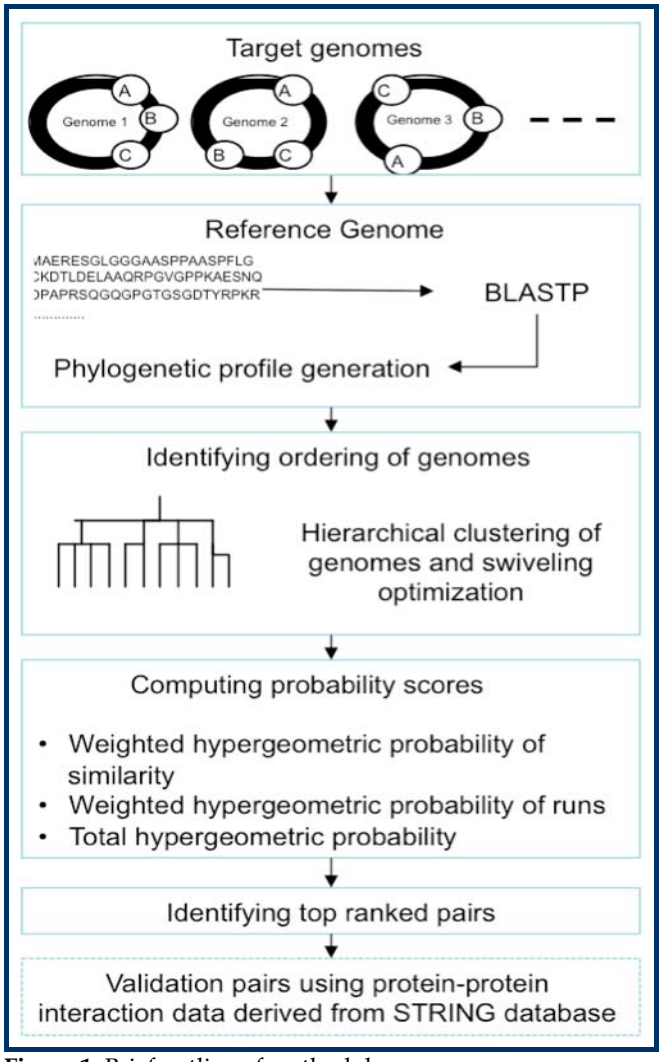

Figure 1: Brief outline of methodology

Results:

All pairs of proteins with a probability score > 0.5 were considered for evaluation in our study. A total of 100, 000 protein pairs were evaluated and provided as Supplementary Material (Data submitted to Dryad (URL: http://www.datadryad.org/); Supplementary file (xls) doi:10.5061/dryad.m6t4j). The phylogenetic profiling approach that uses the weighted hypergeometric probability with runs proposed by Cokus *et al.* outperformed all the previous methods **[41].** We compared our method with the weighted hypergeometric probability with runs, with a hypothesis that, if the results are comparable or better than these methods then it indicates that the proposed method could perform better than all the above methods. All the three methods were benchmarked against pairs derived from the STRING **[8]** database.

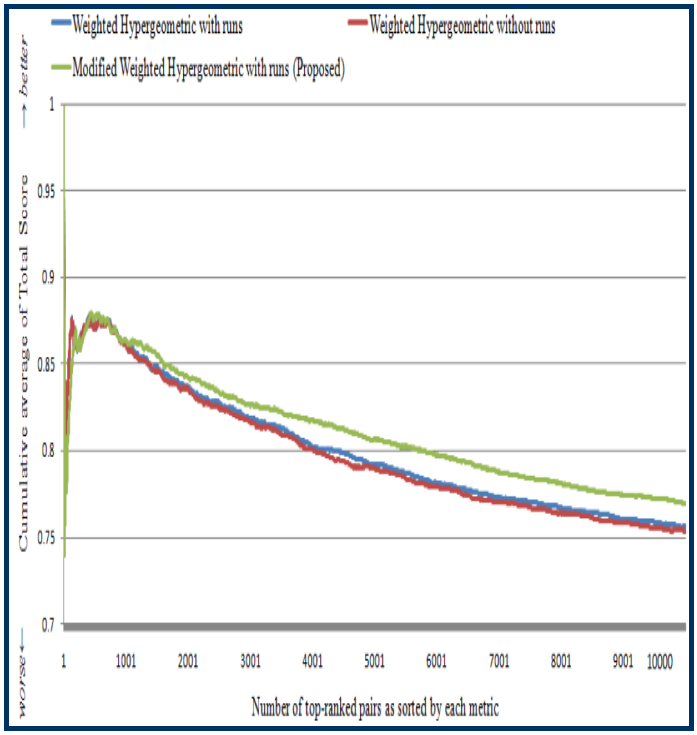

Figure 2: Pairwise comparison of weighted hypergenometric probability method with runs (blue), weighted hypergeometric probability method with runs (red) and modified weighted hypergeometric method runs (proposed) (green).

Validation of method based on comparison of benchmark pairs

Figure 2 compares three methods, considering one method at a time. Each method assigns a *P*-value to every pair of genes (100,000 pairs). Gene pairs are then sorted in ascending order by the *P*-value. The graph in (**Figure 2)** was plotted as given x-axis value x, y was plotted as the mean (total score) of first x gene pairs after sorting based on *P*-value. Here, the total score is the score obtained from the STRING database. Data used to generate the plot is provided in the supplemental file. STRING is an integrated database of known and predicted protein-protein interactions. A given interaction in the STRING database was derived from one or a combination of association methods: gene fusion, neighborhood, co-occurrence, experiments, databases, text mining and homology. Detailed explanation of scoring of protein-protein interaction reported in STRING database can be found elsewhere **[34]**. The score ranges from value 0 to 1. The greater value of the score indicates the strength of the functional relatedness between the proteins. From the graph it shows that the pairs obtained from the proposed - method modified weighted hypergeometric probability with runs (green line) – shown comparable or better performance than the other two methods weighted hypergeometric probability with runs (blue line) **[41]** and weighted hypergeometric probability without runs (red line)

[45]. The cumulative average considering the 10,000 pairs for the proposed method was 0.769, whereas the values of weighted hypergeometric probability with runs was 0.756 and for weighted hypergeometric probability without runs was 0.753 **Table 1 (see supplementary material); (Figure 2).** The margin observed here between the two methods was 0.003; though it is a numerically low value, since it is an average on 10,000 pairs. We noted that our approach (weighted hypergeometric probability with runs) is complementary or in the context of the analyzed genomes when compared to the existing method by a value of 0.013.

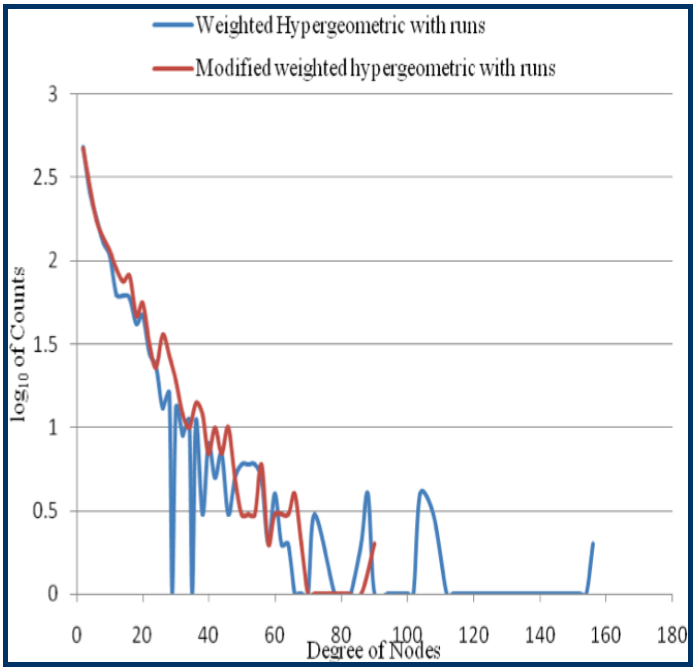

Figure 3: Network degree distributions derived using two different methods

Comparative analysis of degree distribution of interaction networks

A network (an undirected graph with no multiple edges and no self-edges) was obtained from a computational method by ranking gene pairs by the *P*-values from that method and then collecting the top ranked 10,000 pairs. The nodes are the genes mentioned in the kept gene pairs and an edge was placed between two different genes if and only if the gene pair consisting of the two genes was among the kept gene pairs. The degree of a node was defined as the number of edges incident with that node. **Figure 3** shows two histograms (with a logarithmic scale for frequency) of node degree, one (blue) for the network from the weighted hypergeometric with runs computational method and the other (red) for the modified weighted hypergeometric with runs computational method. We can observe that the proposed weighted hypergeometric with runs network (blue) contains many nodes with 90 or more edges, while the modified weighted with runs (red) have almost none. From the above graph, we can say that the network formed by the proposed method show sub-clusters clusters when compared to the pure weighted hypergeometric probability with runs. Large clusters of protein pairs are often not very useful for functional studies since they bring together proteins with a broad range of diverse functions. In contrast, small clusters can contain proteins with well-defined functional relationships and can be tested using targeted functional genomics or protein-protein interaction experiments.

Application of modified hypergeometric probability method using nitrate reductase

To illustrate an example using our method, we considered the subunits of nitrate reductase enzyme **[46-48]** from *Escherichia coli K12*. (**Figure 4a)** shows all the interactions of the six subunits of nitrate reductase, which are narY, narH, narZ, narV, narJ and narG, those that are present in the STRING database. **Figure 4b** shows the network containing all the interactions of the six subunits of nitrate reductase that are observed using our proposed methodology modified weighted hypergeometric probability with runs. Visualization was generated using Cytoscape **[49]**. These proteins belong together as they are subunits of a protein complex that catalyzes the reduction of nitrate to ammonia. In the network generated using our methodology, the interactions missing are narG-narZ, narZ-narJ and narJ-narG and these links had less scores implying less significant edges. The network obtained from the approach was similar to the existing interactions observed in the STRING database (version 8).

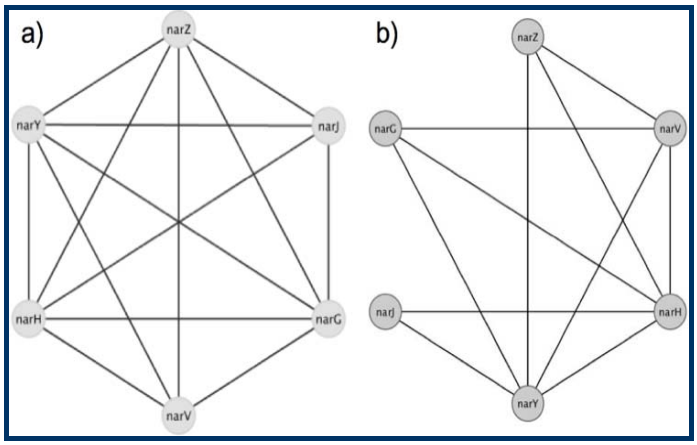

Figure 4: a) shows the interactions of the six subunits of nitrate reductase, mediated by narY, narH, narZ, narV, narJ and narG derived from protein-protein interaction database STRING**; b)** shows the network based on protein pairs mediated by six subunits of nitrate reductase that were identified using our proposed methodology.

Discussion:

Rapidly lowering costs of next-generation sequencing methodologies are increasing the repertoire of gene and protein sequences in molecular databases. Various bioinformatics methods based on similarity measures, phylogenetic approaches, machine learning, protein-protein interaction, gene expression and integrative approaches are available for the prediction of functions of proteins encoded in a genome. In this manuscript we explored the possibility of using phylogenetic profiles to find the functional similarity of genes; a new probability based function association method using phylogenetic profiles was proposed and validated. We proposed a new approach using functional protein association network for identifying functionally linked proteins. We used the probabilistic approach to incorporate two important aspects of functional relatedness, which are similarity measure and the number of runs the profiles span given the ordering of

genomes. We tested the method using the 4195 phylogenetic profiles of *Escherichia coli K12* generated using 305 genomes. The functional links obtained by our proposed method are validated using the STRING database functional links. We compared our results with hypergeometric probability with runs and hypergeometric probability without runs. The cumulative average of STRING score considering the top ranked 10,000 pairs for the proposed method was 0.769; whereas the values of weighted hypergeometric probability with runs was 0.756 and for weighted hypergeometric probability without runs was 0.753. Our proposed method weighted hypergeometric probability with runs has shown comparable or better results than existing method in our evaluation at the same time the method may need additional testing with independent datasets and statistical validation to estimate errors and homology bias that could be introduced during sequence search in generating phylogenetic profiles. For example we considered the top hit from the BLAST search in our analysis – including additional hits and filtering orthologs or paralogs may have further influence the performance of the method. Further the proposed method can be extended to an increasing number of target genomes.

Conclusion:

In the current era of rapidly increasing number of genomic and transcriptomic sequencing projects, assigning functions to individual gene products, fusion transcripts and novel protein isoforms and novel protein products will remain as a primary challenge in bioinformatics. We envision that bioinformatics approaches including the application of phylogenetic profile based methods could enhance function assignment in such scenarios.

References:

[1] Pazos F *et al. EMBO J*. 2008 **27**: 2648 [PMID: 18818697]
[2] Rentzsch R *et al. Trends Biotechnol*. 2009 **27**: 210 [PMID: 19251332]
[3] Rost B *et al. Cell Mol Life Sci*. 2003 **60**: 2637 [PMID: 14685688]
[4] Lee D *et al. Nat Rev Mol Cell Biol*. 2007 **8:** 995 [PMID: 18037900]
[5] Friedberg I, *Brief Bioinform*. 2006 **7**: 225 [PMID: 16772267]
[6] Marcotte EM *et al Science*. 1999 **285**: 751 [PMID: 10427000]
[7] Sharan R *et al. Mol Syst Biol*. 2007 **3**: 88 [PMID: 17353930]
[8] Hu P *et al. Nat Rev Cancer*. 2007 **7:** 23 [PMID: 17167517]
[9] Ben-Dor A *et al. J Comput Biol*. 1999 **6**: 281 [PMID: 10582567]
[10] Eisen JA, *Genome Res*. 1998 **8**: 163 [PMID: 9521918]
[11] Sjolander K, *Bioinformatics*. 2004 **20**: 170 [PMID: 14734307]
[12] Fernald GH *et al. Bioinformatics*. 2011 **27**: 1741 [PMID: 21596790]
[13] Kumar S *et al. Trends Genet*. 2011 **27**: 377 [PMID: 21764165]
[14] Mehta SR *et al. J Infect Dis*. 2012 **205**: 1529 [PMID: 22448013]
[15] Walker MG *et al. Genome Res.* 1999 **9**: 1198 [PMID: 10613842]
[16] Luo F *et al. BMC Bioinformatics*. 2007 **8**: 299 [PMID: 17697349]
[17] Van Noort V *et al. Trends Genet*. 2003 **19**: 238 [PMID: 12711213]
[18] Pellegrini M *et al. Proc Natl Acad Sci U S A*. 1999 **96**: 4285 [PMID: 10200254]
[19] Vert JP, *Bioinformatics*. 2002 **18**: S276 [PMID: 12169557]
[20] Wu J *et al. Bioinformatics*. 2003 **19**: 1524 [PMID: 12912833]
[21] Mikkelsen TS *et al. Bioinformatics.* 2005 **21**: 464 [PMID: 15374867]
[22] Wu J *et al. Genome Inform*. 2005 **16**: 142 [PMID: 16362916]
[23] Chen L & Vitkup D, *Genome Biol*. 2006 **7**: R17 [PMID: 16507154]
[24] Jothi R *et al. BMC Bioinformatics*. 2007 **8:** 173 [PMID: 17521444]
[25] Vitulo N *et al. BMC Bioinformatics*. 2007 **8**: S23 [PMID: 17430568]
[26] Gonzalez O & Zimmer R, *Bioinformatics*. 2008 **24**: 1257 [PMID: 18381403]
[27] Jiang Z, Crit *Rev Biotechnol*. 2008 **28**: 233 [PMID: 19051102]
[28] Sjolander K, *PLoS Comput Biol*. 2010 **6**: e1000621 [PMID: 20126522]
[29] Engelhardt BE *et al. PLoS Comput Biol*. 2005 **1**: e45 [PMID: 16217548]
[30] Snitkin ES *et al. BMC Bioinformatics*. 2006 **7**: 420 [PMID: 17005048]
[31] Wu J *et al. BMC Bioinformatics*. 2006 **7**: 80 [PMID: 16503966]
[32] Ranea JA *et al. PLoS Comput Biol*. 2007 **3**: e237 [PMID: 18052542]
[33] Antonov AV *et al. Comput Biol Chem*. 2008 **32**: 412 [PMID: 18753010]
[34] Jensen LJ *et al. Nucleic Acids Res*. 2009 **37:** D412 [PMID: 18940858]
[35] Enault F *et al. Bioinformatics*. 2003 **19**: i105 [PMID: 12855445]
[36] Altschul SF *et al. Nucleic Acids Res*. 1997 **25**: 3389 [PMID: 9254694]
[37] Wang X *et al. IEEE Computer Society*. 2008 **59**: 62
[38] Roger C *et al. Machine Learning and Applications, Proceedings of the Fourth International Conference.* 2005
[39] Narra K *et al. Communications in Computer and Information Science Volume*. 2009 **40**: 510
[40] Kotaru AR *et al Contemporary Computing*. 2009 **40**: 510
[41] Cokus S *et al. BMC Bioinformatics*. 2007 **8**: S7 [PMID: 17570150]
[42] Bar-Joseph Z *et al. Bioinformatics*. 2003 **19:** 1070 [PMID: 12801867]
[43] Slonim N *et al. Mol Syst Biol*. 2006 **2**: 2006 [PMID: 16732191]
[44] Fitz-Gibbon ST & House CH, *Nucleic Acids Res*. 1999 **27** : 4218 [PMID: 10518613]
[45] Kharchenko P *et al. BMC Bioinformatics*. 2006 **7**: 177 [PMID: 16571130]
[46] Hackett CS & MacGregor, *J Bacteriol*. 1981 **146**: 352 [PMID: 7012121]
[47] Stewart V, *Mol Microbiol*. 1993 **9**: 425 [PMID: 8412692]
[48] Bonnefoy V *et al. Antonie Van Leeuwenhoek*. 1994 **66**: 47 [PMID: 7747940]
[49] Smoot ME *et al. Bioinformatics*. 2011 **27**: 431 [PMID: 21149340]

Edited by P Kangueane

Citation: **Kotaru *et al.*** Bioinformation 9(7): 368-374 (2013)

BIOINFORMATION

open access

Supplementary material:

Table 1: Cumulative Average of Total Score for top 10,000 pairs

Method	Score
Modified Hypergeometric Probability method with runs (Proposed)	0.769
Hypergeometric Probability with runs	0.756
Hypergeometric Probability without runs	0.753