Combining Homolog and Motif Similarity Data with Gene Ontology Relationships for Protein Function Prediction

Hafeez ur Rehman, Alfredo Benso, Stefano Di Carlo,
Gianfranco Politano, and Alessandro Savino
Department of Control and Computer Engineering
Politecnico di Torino, I-10129, Torino, Italy
Email:<firstname.lastname>@polito.it

Prashanth Suravajhala
Bioclues Organization, ICICI Knowledge Park,
Picket, Secunderabad 500011 AP, India
Email: prash@bioclues.org

***Abstract*—Uncharacterized proteins pose a challenge not just to functional genomics, but also to biology in general. The knowledge of biochemical functions of such proteins is very critical for designing efficient therapeutic techniques. The bottleneck in hypothetical proteins annotation is the difficulty in collecting and aggregating enough biological information about the protein itself. In this paper, we propose and evaluate a protein annotation technique that aggregates different biological information conserved across many hypothetical proteins. To enhance the performance and to increase the prediction accuracy, we incorporate term specific relationships based on Gene Ontology (GO). Our method combines PPI (Protein Protein Interactions) data, protein motifs information, protein sequence similarity and protein homology data, with a context similarity measure based on Gene Ontology, to accurately infer functional information for unannotated proteins. We apply our method on *Saccharomyces Cerevisiae* species proteins. The aggregation of different sources of evidence with GO relationships increases the precision and accuracy of prediction compared to other methods reported in literature. We predicted with a precision and accuracy of 100% for more than half proteins of the input set and with an overall 81.35% precision and 80.04% accuracy.**

***Index Terms*—Function Prediction, Gene Ontology, Protein Interaction Network, Protein motifs**

I. Introduction

As long as there are hundreds of conserved proteins with unknown function even in model organisms, such as *Escherichia coli, Drosophila melanogaster* or *Saccharomyces cerevisiae*, the possibility of a 'complete' understanding of these organisms as biological systems remains a challenge. Complete comprehension of protein function is a prerequisite for rational development of antibacterial compounds, drugs, and vaccines. Hypothetical proteins on the other hand cannot be taken into account as potential targets in a drug or vaccine manufacturing process since their role is poorly defined in the metabolic pathways. To make drugs more efficient and to widen the set of their possible targets, it is necessary to devise effective computational techniques for the precise annotation of uncharacterized proteins.

Until recently, several approaches have been developed for predicting protein function using high throughput datasets. These techniques utilize information derived from sequence similarity, phylogenetic profiles, protein 3D structure, protein-protein interactions, protein complexes, gene expression profiles etc., [1]. The most recent and prominent set of techniques uses protein-protein interactions data in a variety of ways to infer protein function [2], [3], [4], [5]. These methods are based on the idea that interacting proteins share common functions; therefore, these methods tend to directly assign functions to an unannotated protein based on the functions of its neighbors.

Direct annotation of protein functions lacks both in terms of precision and accuracy. For precise and accurate function prediction, the context information of protein functions must be incorporated in the methodology by utilizing the relationships between them. Some researchers, e.g., [6], tried to incorporate protein-protein interactions, and protein homology with Gene Ontology (GO) [7], structural relationships to predict protein functions. This method operates on a fixed size ontology structure for GO term relationships. However, protein function annotations vary from protein to protein and may not fit into a fixed ontology size. The limitation of such methods is increasing complexity, for larger ontology sizes. Hence, only a subset of the functions can be taken into consideration. Incorporating all functions of a protein and their diverse level of annotation in GO gives a detailed view of protein's cellular activity. Therefore, a detailed function coverage based on GO will improve the predictive power.

Approaches that try to aggregate different types of biological data, each focusing on a different aspect of cellular activity, demonstrated to produce good results as shown in [8], [9], [10]. Unfortunately, for most uncharacterized proteins we rarely find enough biological information in their own networks, which could be used for their functional association with other proteins. For such proteins, it is important to target biological data that could provide a functional link to annotated proteins. Many hypothetical proteins are found with no edges in their own network, but are connected to homologs of other species network. For example in Figure 1, the protein *YKL033W-A* of *Saccharomyces cerevisiae* species is not con-

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

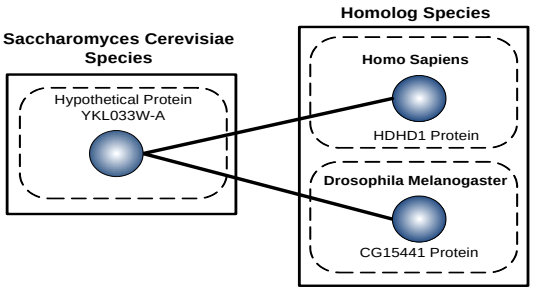

Figure 1. An example of *Saccharomyces Cerevisiae* hypothetical protein connected with homolog proteins of the other species networks.

nected to any protein in its own network but is connected with high homolog similarity edges to protein *HDHD1* of *Homo Sapiens* species and to protein *CG15441* of *Drosophila Melanogaster* species networks. Another type of biological information that could be used for uncharacterized proteins are the motifs conserved in those proteins that associate them to a particular molecular activity.

In this paper we propose a novel method for protein function prediction that integrates different biological information present across many uncharacterized proteins. The conceptual innovation of our method is to integrate functionally related information and to grab full functional coverage of related proteins in the GO structure by selecting flexible ontology sizes that increase the prediction accuracy. In the combined model, we show how this two way integration improves both the precision and accuracy of prediction compared with direct annotation transfer.

The rest of the paper is structured as follows: In section II we present the proposed computational model that evaluates two measures: the functional potential, which is calculated by integrating heterogeneous biological information, and the context based similarity measure, which is based on GO relationships. Section III details the effectiveness of the proposed method when applied to *Saccharomyces cerevisiae* species proteins. We present the results by relating them with state of the art methods of protein function prediction. In section IV we conclude the paper with some future developments.

II. Methods

Our scheme exploits the fact that interacting proteins are likely to collaborate on a common purpose, thus the function of an unannotated protein can be deduced when the function of its binding partners is known. Along with this, we also combine heterogeneous information that is conserved across many proteins and serves as functional evidence, by calculating functional potential scores for interacting proteins. After having a vivid idea about the function from potential interactors, we calculate a similarity measure based upon Gene Ontology. Functional terms with high similarity value are the target annotations. We divide this strategy into three major steps namely, A) Interacting protein selection, B) Filtering based upon similarity scores, and C) Context similarity score based on Gene Ontology.

A. Interacting Protein Selection

We obtain our protein dataset from UniProt [11] database. For unannotated proteins we consider related protein-protein network information which is passed as input to the proposed technique. We select protein-protein interaction data from two databases: IntAct [12] and DIP (Database of Interacting Proteins) [13]. We only consider interactions for which there is an experimental evidence.

B. Filtering based upon Similarity Scores

To increase the predictive power of our automated annotation system we calculate similarity scores among interacting proteins by integrating heterogeneous sources of data. This is particularly important as each type of data typically captures distinct aspects of cellular activity. We name this overall score as Functional Potential (FP). In the second step we compute Functional Potential measure $FP_{(i,j)}$ to filter proteins which have high potential of being functionally similar to unannotated protein. The functional potential measure $FP_{(i,j)}$ is based upon three functional indicators: (1) Motif Similarity Score (2) Homolog Similarity Score, and (3) Sequence Similarity Score.

1) Motif Similarity Score: Patterns of evolutionarily conserved motifs in a protein-sequence reflect the tendency of biochemical functions of an annotated protein. Motif information can also be conserved in unannotated proteins, so the number of common motifs conserved in two connected proteins can be a strong functional clue for functionally unknown proteins. Based on this fact we incorporate motif information from the ProSite database [14], and introduce a similarity measure. This measure is normalized to $M_{i,j}$ and is calculated for same number of common motifs between two interacting proteins P_i and P_j as follow,

$$M_{i,j} = \frac{Common_{Motif}(P_i, P_j)}{Min_{Motif}(P_i, P_j)} \quad (1)$$

Where $Common_{Motif}(P_i, P_j)$ is the number of common motifs conserved between two interacting proteins and $Min_{Motif}(P_i, P_j)$ is the minimum number of motifs conserved in one of the two proteins.

2) Homolog Similarity Score: The second measure that contributes to increase the functional potential of a protein is the homologs similarity between two proteins P_i and P_j of different species. Evolutionary relationships between species suggest that orthologous proteins of different species, which share high sequence similarity and whose functions have been established before speciation, are likely to share similar protein classifications. To capture homolog similarity based upon orthologs, we define a homolog sequence similarity score between protein P_i and P_j as $H_{(i,j)}$ a normalized pairwise BLAST score [15]. We use normalized BLAST scores, defined as the BLAST score (homolog) divided by self score of query (which is BLAST score of the protein against itself), as defined in equation 2. We only consider scores above 0.5 threshold value, which is a strong similarity indicator as described by [6].

$$Blast\_Score(P_i, P_j) = \frac{BLAST(P_i, P_j)}{BLAST(P_i)} \quad (2)$$

3) Sequence Similarity Score: The third measure that increases the functional potential of a protein is the sequence similarity between two proteins P_i and P_j of the same species. Sequence similarity of proteins by itself is also a strong hint for functional relevance. Proteins with highly similar sequences are found to have been involved in similar functional activities. We define a sequence similarity measure between protein P_i and P_j as $S_{(i,j)}$ a normalized pairwise BLAST score. We calculate this score in the same way as in equation 2, only the proteins in this case are from the same species.

All the similarity scores values lie between 0 and 1 and the overall functional linkage potential $FP_{(i,j)}$ between interacting protein P_i and its neighbor P_j is calculated as follows,

$$\mathbf{FP_{(i,j)} = M_{i,j} + H_{i,j} + S_{i,j}} \quad (3)$$

The interacting nodes with high value of $FP_{(i,j)}$ are more likely to participate in common functions. After this step, we have a set of potential interactors for unannotated protein.

C. Context Similarity Score based on Gene Ontology

From the set of annotated potential interactors, we obtain the annotation set for our protein under test. Each annotation is represented in GO with a node label. Nodes (classes or labels) are connected to other nodes through parent-child edges, which impose hierarchical inter-relationships between them. Thus, it is possible to compute the similarity between two GO nodes, referred to as context similarity, on the basis of their relative positioning in the hierarchy. We use Gene Ontology structural data, downloaded from the Gene Ontology database [7], for molecular function class hierarchies.

Potential interactors P_j and P_k of protein P_i are annotated with a number of functions, we map those functions on Gene Ontology to obtain related term dependencies. For proteins with multiple functions we define the functional context terms $F_1, F_2, ..., F_n$ as the top most annotations of the Gene Ontology. For protein annotations under the same functional context, we define a functional similarity $Sim(T_{P_j}, T_{P_k})$ between two terms T_{P_j} and T_{P_k} of protein P_j and P_k as follows,

$$\mathbf{Sim(T_{P_j}, T_{P_k}) = \frac{Sim_{TO}(T_{P_j}, T_{P_k})}{Min(T_{P_j}, T_{P_k})}} \quad (4)$$

Where $Sim_{\mathbf{TO}}(T_{P_j}, T_{P_k})$ is the number of terms overlapping between the GO hierarchies of T_{P_j} and T_{P_k} terms, under the same context term. The $Min(T_{P_j}, T_{P_k})$ is the minimum length (number of terms) between the two hierarchies of T_{P_j} and T_{P_k} terms.

For annotating protein P_i, we need to calculate similarity (defined in equation. 4) among all terms of its interactors. We calculate similarity scores for all annotations of the interacting proteins and the protein annotations under each functional context crossing the defined similarity threshold are considered as potential functions for the unannotated protein.

III. Experimental Setup and Results

We applied our methodology to *Saccharomyces cerevisiae* species proteins, one of the most complete and extensively studied data sets. To calculate the prediction performance and effectiveness of our method we use cross validation approach. For evaluation of our methodology, we computed several performance measures, such as: precision, recall, accuracy and F1 as in [6]. With a $FP_{(i,j)}$ threshold of 0.5 and different $Sim(T_{P_i}, T_{P_j})$ similarity threshold values, we report the prediction results calculated for above measures in the following subsections.

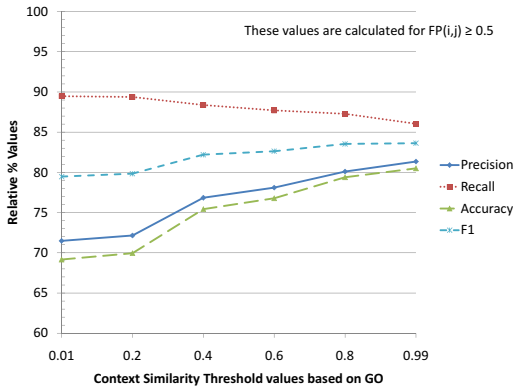

Figure 2. Effect of different $Sim(T_{P_i}, T_{P_j})$ threshold values on precision, recall, accuracy, F1 when applied to *Saccharomyces cerevisiae* proteins.

A. Effect of GO Similarity on Precision, Recall, and Accuracy

To observe the effect of similarity based on Gene Ontology we calculated the precision, recall, accuracy and F1 measure for different threshold values of $Sim(T_{P_i}, T_{P_j})$ by fixing the $FP_{(i,j)}$ potential to 0.5. The complete plot of the results is shown in Figure 2. Except recall, all other measures show an increasing trend. It can be clearly seen that as we raise the similarity threshold values, precision, accuracy and F1 measure are continuously increased. This is due to the fact that a higher similarity threshold selects annotations which are highly related from a functional point of view, and thus part of the same molecular activity. Hence, it can be seen that using GO term specific similarities values improved the precision, accuracy and F1 to 10%, 12%, and 5% respectively, as compared to direct annotation transfer. Another important observation of using GO classification is the decrease in FPR (False Positive Rate) with increasing similarity values. The FPR is decreased from 71% to 27% as shown in Figure 3, which means the predictions are more centered towards semantically related annotations.

B. Comparison with other approaches

In this section, we compare our method to the most widely used group of techniques for function prediction that integrate multiple information sources. One of such techniques is presented by Nariai et al. [8], which is based on Bayesian probabilistic approach. Since methods based upon these approaches

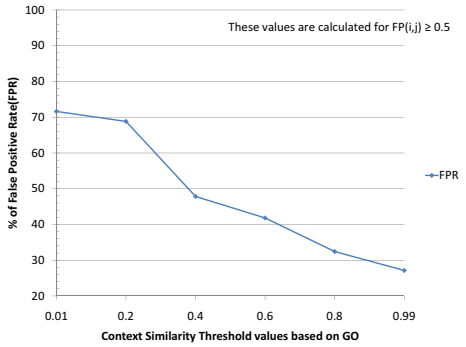

Figure 3. Effect of different $Sim(T_{P_i}, T_{P_j})$ threshold values on false positive rate, when applied to *Saccharomyces cerevisiae* proteins.

are the widely used and accepted in the paradigm of protein function prediction, therefore we compare our results with most recent and established of these computational techniques. We compare Narai's best accepted prediction results i.e., with the optimum values of precision and accuracy, with our results. We report the precision, recall, accuracy and F1 measure values for both methods in Figure 4. Our method outperforms Narai's method with respect to all reported measures. The accuracy of our method is higher due to large number of true negatives and less number of false negatives. Overall our method shows higher values of precision, recall, accuracy and F1 which improves the overall prediction confidence.

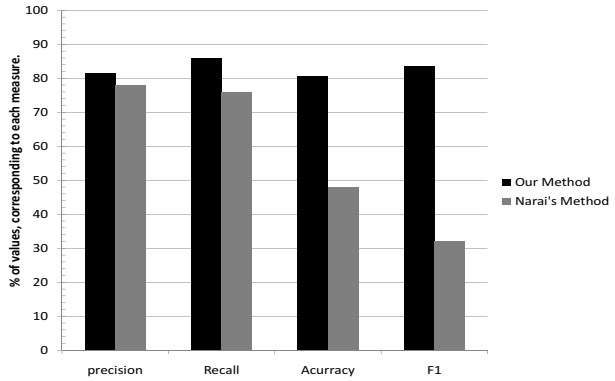

Figure 4. Comparison of precision, recall, accuracy and F1 of our method (black) with Narai's [8] method (grey).

IV. Conclusion

In this work, we presented a new method that uses existing biological data with Gene Ontology relationships to infer function of uncharacterized proteins. We combined different sources of information that are present across proteins of unknown function. Along with this, term specific relationships are utilized for defining functional contexts for activities of interacting proteins, which improves the prediction accuracy by involving only related functions. This approach may be easily extended by integrating more sources of biological information to further improve the function prediction confidence.

References

[1] Y. Galperin and V. Koonin, "A survey-conserved hypothetical proteins: prioritization of targets for experimental study," *Nucleic Acids Research*, vol. 38, no. 18, pp. 5452–5463, 2004.

[2] S. Letovsky and S. Kasif, "Predicting protein function from protein-protein interaction data: A probabilistic approach," *Bioinformatics*, vol. 19, no. 1, pp. i197–i204, 2003.

[3] U. Karaoz, T. M. Murali, and et al., "Whole-genome annotation by using evidence integration in functional-linkage networks," *Proc. Nat'l Academy of Sciences USA*, vol. 101, pp. 2888–2893, 2004.

[4] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein-protein interactions in yeast," *Nature Biotechnology*, vol. 18, pp. 1257–1261, 2000.

[5] N. Yosef, R. Sharan, and N. Stafford, "Improved network-based identification of protein orthologs," *Bioinformatics*, vol. 24 no. 16, pp. i200–i206, 2008.

[6] A. Mitrofanova, V. Pavlovic, and B. Mishra, "Prediction of protein functions with gene ontology and interspecies protein homology data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8 no. 3, pp. 775–784, 2011.

[7] "The gene ontology database." Jan 2012. [Online]. Available: http://www.geneontology.org/

[8] N. Nariai, E. Kolaczyk, and S. Kasif, "Probabilistic protein function prediction from heterogeneous genome-wide data," *PLoS ONE*, vol. 2, no. 3, p. e337, 2007.

[9] S. Carroll and V. Pavlovic, "Protein classification using probabilistic chain graphs and the gene ontology structure," *Bioinformatics*, vol. 22, no. 15, pp. 1871–1878, 2006.

[10] A. Mitrofanova and et al., "Integrative protein function transfer using factor graphs and heterogeneous data sources," *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 314–318, IEEE Computer Society, 2008.

[11] "The uniprot consortium: Reorganizing the protein space at the universal protein resource (uniprot)," *Nucleic Acids Res. 40: D71-D75 (2012).*

[12] S. Kerrien and et al., "The intact molecular interaction database in 2012. [pmid: 22121220]," *Nucl. Acids Res.*, doi: 10.1093/nar/gkr1088. [Online]. Available: http://www.ebi.ac.uk/intact

[13] L. Salwinski and et al., "The database of interacting proteins," *Nucl. Acids Res.*, pp. 449–51, 2004. [Online]. Available: http://dip.doe-mbi.ucla.edu

[14] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, and et al., "The prosite database," *Nucl. Acids Res.*, pp. D227–230, 2006. [Online]. Available: http://prosite.expasy.org/

[15] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol. 215:*, pp. 403–410, 1990. [Online]. Available: http://blast.ncbi.nlm.nih.gov/Blast.cgi