Comparative Study of Various Genomic Data Sets for Protein Function Prediction and Enhancements Using Association Analysis [*][†]

Rohit Gupta [‡][§] Tushar Garg [‡] Gaurav Pandey [‡] Michael Steinbach [‡]
Vipin Kumar [‡]

Abstract

The prediction of protein function is a key task in bioinformatics and a variety of techniques and data sets have been employed for that purpose. Using the popular keyword recovery measure, which is based on standard keyword annotations of the SwissProt database, this paper presents a comparative study of the information provided for protein function prediction by different types of data sets: phylogenetic profiles, protein interaction networks, and gene expression data. The technique employed is to evaluate the average keyword recovery achieved when the top (most strongly connected or similar) pairs of proteins are taken from each data set. The results show that protein interaction data contains the most information, then gene expression data, and finally, phylogenetic profiles. In addition, the average keyword recovery is also computed for the top pairs derived from the raw protein interaction data using a measure, h-confidence, which comes from the data mining area of association analysis. This approach gives improved results over raw protein interaction data and even better results when applied to protein complexes that were computationally generated using the raw protein complex data. The paper also briefly discusses the fact that the different data types appear to be complementary.

1 Introduction.

Proteins are the most essential and versatile macromolecules of life, and the knowledge of their functions is a crucial link in the development of new drugs, better crops, and even the development of synthetic biochemicals such as biofuels. Experimental procedures for protein function prediction are inherently low throughput and are thus unable to annotate a non-trivial fraction of proteins that are becoming available due to rapid advances in genome sequencing technology. This has motivated the development of computational techniques that utilize a variety of high-throughput experimental data for protein function prediction, such as protein and genome sequences, gene expression data, protein interaction networks and phylogenetic profiles. (See [7] for a survey of several hundred articles on this topic.) This paper contributes to efforts in the computational prediction of protein function by presenting a comparative study of the information provided for protein function prediction by different types of data sets: phylogenetic profiles, protein interaction networks, and gene expression data.

The technique employed here is to identify pairs of proteins that are strongly connected or similar using the information from a data set and then to evaluate the ability of the top (most strongly connected or similar) pairs of proteins from each data set to predict protein function. Given the strength of various pairs of proteins, the top pairs of proteins for a data set are evaluated with respect to the popular keyword recovery measure, which is based on standard keyword annotations of the SwissProt database [11]. The higher the average keyword recovery the more information relevant to protein function prediction is present.

Using average keyword recovery metric [6], it is possible to compare the performance of the various data sets and determine the relative levels of information. Our results show that protein interaction data contains the most information, then gene expression data, and finally, in a result that is somewhat surprising, phylogenetic profiles. This result is surprising because previously published results have shown phylogenetic profiles to have more information than gene expression data with respect to protein function prediction [6].

Besides the functional links that can be directly derived, it is also possible to identify strongly interacting pairs using a more indirect approach. Specifically, the average keyword recovery was also computed for the top pairs derived from the raw protein interaction data using a measure, h-confidence, which comes from the data

[*] This paper appeared in the Data Mining for Biomedical Informatics Workshop which was held in conjunction with the 7th SIAM International Conference on Data Mining on April 28, 2007 in Minneapolis, MN.

[†] This work was supported by NSF grants #CRI-0551551, #IIS-0308264, and #ITR-0325949

[‡] Department of Computer Science and Engineering, University of Minnesota {rohit, garg, gaurav, steinbac, kumar}@cs.umn.edu

[§] Corresponding Author

mining area of association analysis [8]. This approach gives improved results over raw protein interaction data and even better results when applied to protein complexes that were computationally generated using the raw protein complex data [3]. This result also emphasizes the importance of the processing of the data to extracting the maximum amount of information.

We believe that our work has the following significant contributions:

1. Owing to the importance of the knowledge of protein function, several approaches have been proposed in the literature for predicting protein function from a variety of biological data types, such as protein interaction networks, gene expression data and phylogenetic profiles, as well as a combination thereof [7]. Past studies have indicated that phylogenetic and expression profiles, which provide information at the level of individual proteins, generally have comparable power for predicting protein function, with the former slightly outperforming the latter [6, 9]. On the other hand, interaction networks provide information at a more global level, so that the direct and indirect interactions between proteins can be used to infer functional knowledge about proteins more accurately [6, 5]. Thus, although these studies provide an implicit understanding of the relative capabilities of the different data sets and their sources, there have not been any studies that compare these capabilities for function prediction explicitly. In this paper, we attempt to fill this gap by explicitly comparing the potential of a diverse variety of biological data sets, namely microarray data, phylogenetic profiles, protein interaction networks and protein complexes for the function prediction task.

2. Traditional methods for the inference of function from protein interaction networks involve the representation of the network as a graph consisting of proteins represented as nodes and interactions as edges, and the subsequent application of one or more of neighborhood-, global optimization- and clustering-based techniques [7]. However, a new successful category of approaches involves the representation of the set of interactions and complexes as a binary matrix and applications of techniques from the field of association analysis to extract dominant groups of proteins, which are hypothesized to represent functional modules [12, 4]. In particular, the concept of the hyperclique pattern [13] has been shown to produce very pure functional modules with respect to the Gene Ontology [12]. In this study, we show that the simple application of just pairwise hypercliques leads to significantly more accurate inference of protein function as compared to those obtained using some simple graph-based approaches. Thus, we illustrate the potential of hypercliques in particular, and association analysis in general, for the analysis of biological data.

2 Data Sources.

We use the following genomic data for the task of protein function prediction on a global scale.

Protein-protein interaction data

High-throughput protein-protein interaction data of yeast Saccharomyces cerevisiae is taken from [3]. This comprises of 6228 unique interactions among 2372 proteins.

Protein complex data

The protein complex data is taken from [3]. These protein complexes are computationally obtained from raw protein interaction data using the TAP-MS (tandem affinity purification with mass spectrometry) approach. Even though individual protein complexes may not provide explicit information about physical interactions, they provide information about functional relationships.

Microarray gene expression data

We considered two gene expression data sets for this study. [10] has time series of 4 experimental conditions for all the genes in budding yeast, making it effectively a matrix of (6178×74). We call this data GE:Spellman. Another expression data that we have used in this study is constructed by the combination of several well-known data sets available for yeast. This microarray data set, called GE:Barkai in this study, consists of a total of 1011 experimental conditions for 6206 genes [1].

Phylogenetic profile data

The phylogenetic profile of a protein is essentially a bit vector that encodes for the presence or absence of that protein in various organisms. For this study, we used the phylogenetic profile data for all the proteins of yeast constructed using the genomes of 59 different organisms [2].

SwissProt keywords

For evaluation purposes, we used keyword annotations of all the proteins from the SwissProt database [11]. In all, there are 368 keywords assigned to yeast proteins. We omitted one of the keywords 'Complete Proteome' as this was present in all the proteins.

3 Methods.

Our general approach involves constructing weighted graphs in which nodes represent proteins and the weight between any two proteins in the constructed graph is the value of the similarity (according to the measure used) between the corresponding two proteins. We then generate several subgraphs from this weighted graph by preserving only those links whose weight is more than specified similarity threshold. Generally speaking, for a higher threshold, the constructed subgraph has stronger links and fewer proteins and vice-versa. If the threshold were fully relaxed, the constructed subgraph would be same as the original weighted graph. Below we discuss how we constructed weighted graphs for the different data sets we used in this study.

Gene Expression Data

To build a graph using all possible co-expressed protein pairs based on gene expression data, we used the Pearson correlation coefficient to measure the association of each possible protein pair.

Phylogenetic Profiles

To build a graph using all possible protein pairs that co-evolved together based on the phylogenetic profiles, we used the mutual information measure, as suggested by [2].

Weighted Protein-Protein Interaction Data

We build a graph using the known interaction weights between each pair of proteins [3].

Derived Graphs Using Protein Interaction and Protein Complex Data

We derive two more graphs from protein interaction data and protein complex data using an indirect measures of the strength of the association of a pair of proteins, namely, h-confidence [13]. This measure was used for finding functional modules in protein complex data [12], where functional modules are groups of proteins that are likely to have similar functions. Full details of h-confidence and its associated hyperclique pattern are provided in [12] and [13]. However, we provide a brief explanation here.

Consider a set of protein complexes, which can be regarded as sets of proteins. This data can be represented as a binary matrix, where the rows represent complexes and the columns represent proteins. Specifically, each row contains a 1 in precisely those columns representing the proteins that belong to the complex corresponding to that row. For a pair of items, the h-confidence of a pair of proteins is just the number of times they appear together divided by the maximum number of times that one of the proteins occurs by itself. The h-confidence of a set of items is between 0 and 1. It is close to 1 if proteins mostly occur together and close to 0 if they often occur without one another. Although the starting binary matrix can be a set of protein complexes, we can also use the raw protein interaction graph as well. In both cases, we build subgraphs using all pairs with an h-confidence above a certain level. Different subgraphs are generated by using different h-confidence levels.

Evaluation

The subgraph generated by each data set was evaluated using keyword recovery, based on standard keyword annotations of the SwissProt database [11]. More specifically, the widely used keyword recovery metric was applied. This metric was introduced by [6] and is defined by the following equation:

$$KeywordRecovery = \frac{1}{A}\sum_{i=1}^{A}\sum_{j=1}^{x}\frac{n_j}{N}$$

where A is the total number annotated protein, x is the number of keywords associated with the ith protein, N is the total number of keywords associated with the neighbors of the ith protein, and n_j is the number of keywords from neighboring proteins match the jth keyword of the ith protein.

Informally, the keyword recovery of a specified protein is just the fraction of the keywords of its neighbors (not eliminating duplicates) that are also keywords of the specified protein, and the overall keyword recovery is just the average of the individual keyword recoveries. The maximum possible value is 1, and the minimum is 0. Use of this measure for evaluation requires eliminating proteins without any SwissProt keyword.

The motivation for the keyword recovery metric is the following: If protein function for an unlabeled protein is predicted as a set of the keywords of its neighbors, then keyword recovery serves as a measure of the quality of that approach. Of course, more sophisticated approaches are possible, but the keyword recovery measure is simple and provides an easily understood metric of the amount of information related to function prediction that can be found in a data set or a portion of a data set.

The keyword recovery measure is applied to the weighted graphs defined above. The weights on the edges of these graphs are not used in the computation of the keyword recovery, but can be used to select a subgraph, e.g., by eliminating edges with low weights. Indeed, as is shown below, those graphs with higher edge weight typically yield higher keyword recovery scores.

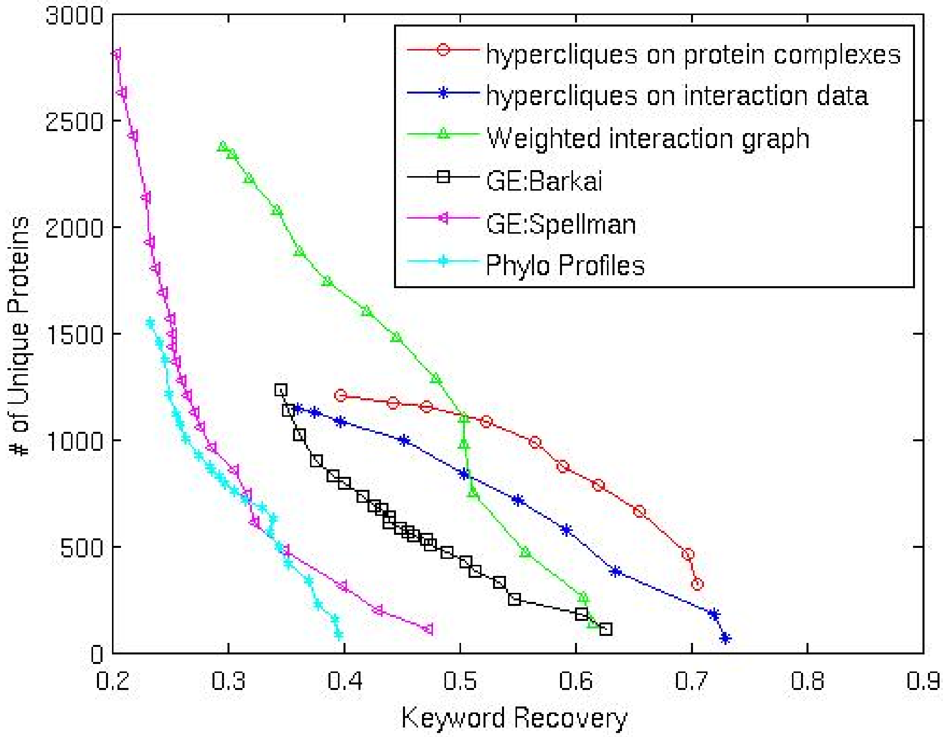

Figure 1: Performance of different biological data sets for yeast protein function prediction

4 Results and Discussion.

Figure 1 shows the performance of the different data sets for the prediction of protein function evaluated using the SwissProt keyword recovery methodology described earlier. This figure plots the average keyword recovery against the corresponding number of proteins for which predictions can be made at that value of recovery, for each of the graphs generated from the original data sets. However, before we discuss these results in more detail, it is important to note that graphs for the protein interaction and protein complex data sets are limited by the number of proteins (2372 in protein interaction data and 2450 proteins in protein complex data), while gene expression and phylogenetic profiles are available for nearly all the yeast proteins.

It is easy to see that it is better to have a plot closer to the top-right corner in Figure 1, because this indicates that a large number of proteins can be annotated at any level of keyword recovery. More specifically, following important conclusions can be made from these plots:

1. Phylogenetic profiles seem to perform the worst among all data sets in terms of keyword recovery. In particular, both the gene expression data sets outperform the phylogenetic profile data set. Even though this result does not match those reported by earlier studies [6, 9], it should be noted that the expression data sets we used are richer than those used earlier. Also, using all SwissProt keywords, we were able to reproduce almost exactly the results of [6] for phylogenetic profiles, which indicate that these profiles do not contain abnormally spurious information.

2. Even for the same type of data, results may vary widely for different data sets. This can be observed in our results for gene expression data, where the 'GE:Barkai' data set clearly outperforms the 'GE:Spellman' data set in function prediction. This difference may explain why different studies on gene expression are targeted towards understanding proteins that belong to different sets of functional classes. In such a case, the protein function prediction accuracy over all functional classes may not provide a meaningful conclusion. This factor may have led to the above difference in performance since the 'GE:Spellman' data set is not rep-

resentative of all the functional classes of yeast. On the other hand, these results show merit of using a comprehensive gene expression data set, such as 'GE:Barkai' for predicting protein function.

3. Weighted protein interaction graphs substantially outperform both gene expression data and phylogenetic profiles in recovering a larger number of proteins at a substantially high keyword recovery. For instance, at a reasonable recovery threshold of 0.5, predictions could be made only for 428 proteins using the 'GE:Barkai' data set, while the corresponding number for interaction data was 1102, which is significantly higher. This result confirms the greater power of interaction data for protein function prediction.

4. Interestingly enough, keyword recovery performance on the graph derived from protein interaction data using h-confidence measure is much higher than that obtained from the raw weighted interaction graph. Moreover, the sets of protein pairs identified using the simple protein interaction network and the h-confidence measure are significantly complementary (more than 75%) to each other.

5. Finally as shown by the plot, functionally linked protein pairs derived from protein complex data using the h-confidence measure have the best performance for a keyword recovery threshold of about 50% or more. This may be attributed to the rich functional information in the protein complex data, which is missing in the raw protein interaction network. Assuming that all the proteins in a complex are functionally linked to each other, protein complexes have a keyword recovery of only 0.20 is obtained from all the complexes. This result indicates that although protein complexes by themselves are sufficiently stable, it is important to identify functionally coherent groups of proteins within complexes. Moreover, the set of functional links obtained from these two approaches are also substantially complementary to each other.

Finally, it is also important to check that the protein pairs that we have identified are not found to be functionally coherent just by random chance. For this, we performed a random trial with weighted protein-protein interaction data. We randomize the interactions while maintaining the total number of proteins and interactions among them in the protein-protein interaction matrix. Using the same keyword recovery analysis, we obtain a recovery of 0.11, which is much lower than the performance of all other data sets. This indicates the potential of the protein pairs that we have identified either by using the weights on the interaction edge or by using the hyperclique patterns in inferring protein function.

5 Conclusion and Future Work.

In this study, we systematically compared the performance of various biological data sets, namely protein complexes, protein-protein interaction, gene expression, and phylogenetic profiles, with respect to the task of predicting functions of yeast (Saccharomyces cerevisiae) proteins. Using a keyword recovery approach based on SwissProt keyword annotations of yeast proteins, our results suggest that the hypercliques derived from the protein complex data set consistently outperforms other data sets in protein function prediction. The relatively poor performance of phylogenetic profiles, especially with respect to expression profiles, was somewhat unexpected, but to our knowledge other such comparisons have not utilized such a comprehensive expression data set.

Moreover, results on protein-protein interaction and protein complex data suggest that pairwise hyperclique patterns not only provide complementary information to the raw physical protein interactions but also improve the accuracy by which a protein function can be inferred. It is noteworthy that the best results are obtained with pairwise hyperclique patterns derived from protein complexes, which were in turn derived from the raw protein interactions.

There are several important tasks for future work. One task is to extend our analyses to more data sets. In particular, the two gene expression data sets used show markedly different performance, and we plan to investigate different types of expression data. Also, it is important to investigate the effect of different similarity measures on the results. There may be different measures (and in some cases preprocessing schemes) that will yield even better results. The use of evaluation measures other than keyword recovery may also impact results and we propose to extensively evaluate this issue. Furthermore, there are a number of different ways to combine the results of different data sets and we plan to investigate various strategies for that task. Finally, we will investigate whether the approach based on defining the similarity of pairs of proteins in terms of h-confidence (or some other association measure) will also give improved results for phylogenetic and expression data.

References

[1] Bergmann S., Ihmels J. and Barkai N., 2003. Iterative

signature algorithm for the analysis of large-scale gene expression data, Phys. Rev. E 67, 031902.

[2] Date, S. V. and Marcotte, E. M. 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat Biotechnology. 21, 9, 1055-1062.

[3] Gavin et al. 2006, Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631-636.

[4] Hu, H., Yan, X., Huan, Y., Han, J., and Zhou, X. J. 2005. Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics 21, Suppl. 1, i213-i221.

[5] Lanckriet, G. R. G., Bie, T. D., Cristianini, N., Jordan, M. I., and Noble, W. S. 2004. A statistical framework for genomic data fusion. Bioinformatics 20, 16, 2626-2635.

[6] Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. 1999. A combined algorithm for genome-wide prediction of protein function. Nature 402, 6757, 83-86.

[7] Pandey, G., Kumar, V. and Steinbach M., Computational Approaches for Protein Function Prediction: A Survey, TR 06-028, Department of Computer Science and Engineering, University of Minnesota, Twin Cities.

[8] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining. Pearson Addison-Wesley, 2006.

[9] Pavlidis, P., Weston, J., Cai, J.,and Grundy, W. N. 2002. Learning gene functional classifications from multiple data types. J Comput Biol. 9, 2, 401-411.

[10] Spellman, P. T., Sherlock, G., Zhang, D. M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, b. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9, 12, 3273-3297.

[11] The Universal Protein Resource (UniProt), Nucleic Acids Research Advance Access published on January 12, 2007, Nucl. Acids Res. 35: D193-D197.

[12] H. Xiong, X. He, C. Ding, Y. Zhang, V. Kumar, S. R. Holbrook. Identification of Functional Modules in Protein Complexes via Hyperclique Pattern Discovery. Proc. of the Pacific Symposium on Biocomputing, pp. 221-232, 2005.

[13] Xiong, H., Tan, P., and Kumar, V. 2006. Hyperclique pattern discovery. Data Mining Knowledge Discovery. 13, 2 (Sep. 2006), 219-242.