Predicting Gene Function in Yeast through Adaptive Weighting of Multi-Source Information

Shubhra Sankar Ray[1,*], Sanghamitra Bandyopadhyay[2], Sankar K. Pal[1]

1. Center for Soft Computing Research, Indian Statistical Institute, Kolkata, India
2. Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

*E-mail: shubhra_r@isical.ac.in

Background

The value of combining informations obtained from different methods, for gene function predictions, has been illustrated by several studies [1, 2, 3, 4]. We propose a new scoring framework, called *Adaptive Score* (*AdS*), for predicting the function of a few unclassified Yeast genes. We mainly focus on phenotypic profiles [5], microarray gene expression (All Yeast data [6]), KEGG pathway database [7], protein sequence similarity through transitive homologues, and protein-protein interactions from BioGRID [8] as data-sources. We use the Pearson correlation for similarity extraction from phenotypic profile and gene expression data. All the protein sequences, except Yeast proteins, corresponding to each KEGG pathway (121 pathways in the second level) are downloaded from PIR to extract profile similarity between two Yeast proteins. Profile vector for each protein in Yeast is computed by comparing its sequence across 121 pathway databases, using BLAST. The method is similar to phylogenetic profile [9] construction. To find the similarity between two genes using KEGG profiles, we used the ratio of dot product value and OR value between two profiles. To detect similarity between two proteins sequences through transitive homologues, 37,66,477 protein sequences are downloaded from UniProt and compared with target proteins by using BLAST [10], the metric of ProClust [11], and the method described in [12]. For protein-protein interaction study, manually curated catalogues of known interactions are downloaded from BioGRID [8] and binary interactions are used as the common unit of analysis.

Benchmarking

The similarities arising from various data-sources are separately benchmarked, based on the super GO-Slim process annotations of genes in the Saccharomyces Genome Database (SGD). The proportion of true positives (TP) gene-pairs at a particular similarity value (computed from a data-source) can be used as a benchmarking method [1], where TP gene-pairs are defined as pairs of genes i and j, such that genes i and j have an overlapping (explicit or implicit) super GO-Slim process term annotation. Figure 1.a compares the similarity values obtained from different data-sources in terms of their $proportionTP$. The $proportionTP$ values for intermediate similarity values are calculated from the slopes of the respective curves. The $proportionTP$ for protein-protein interactions has a constant value 0.69 at a similarity value of 1 and hence it is not shown in Fig. 1.a.

New Scoring Framework

The $proportionTP$ values reflect the accuracy of similarity values, but do not provide any information about importance/weight of one data-source in presence of the other data-sources,

in predicting gene-pairs. Consequently, it will be more appropriate if *proportionTP* values of each data-source, in presence of other data-sources, is weighed by a different factor and then integrated; and the factors are dependent on the *proportionTP* of the integrated *proportionTP* values of different data-sources. In this investigation we propose a new score where, *proportionTP* values (computed from different data-sources) between two genes X and Y are added through weights a, b, c, d, and e in a linear combination style. This score is referred to as *Adaptive Score* (*AdS*) and is defined as

$$AdS_{X,Y} = \frac{a \times Pp_{X,Y} + b \times Pm_{X,Y} + c \times Kp_{X,Y} + d \times B_{X,Y} + e \times I_{X,Y}}{a + b + c + d + e} \quad (1)$$

where a, b, c, d, and e are varied within range 0 to α in steps of 1 to find a combination that maximizes the *proportionTP* (using super GO-Slim process) for a user defined cut-off of top gene-pairs. The weights a, b, c, d, and e are assigned to the complete *proportionTP* matrices calculated from individual data-sources. Our gold standard cut-off (user defined) of top gene-pairs is determined from KEGG pathway profiles, which provides 26432 gene-pairs with similarity value 1 and gold standard constant *proportionTP* value of .81. These gene-pairs are the most accurate of all, whereas the accuracy (*proportionTP*) of other data-sources, as well as gene-pairs below top 26432 for KEGG pathway profiles, vary considerably.

Evaluation

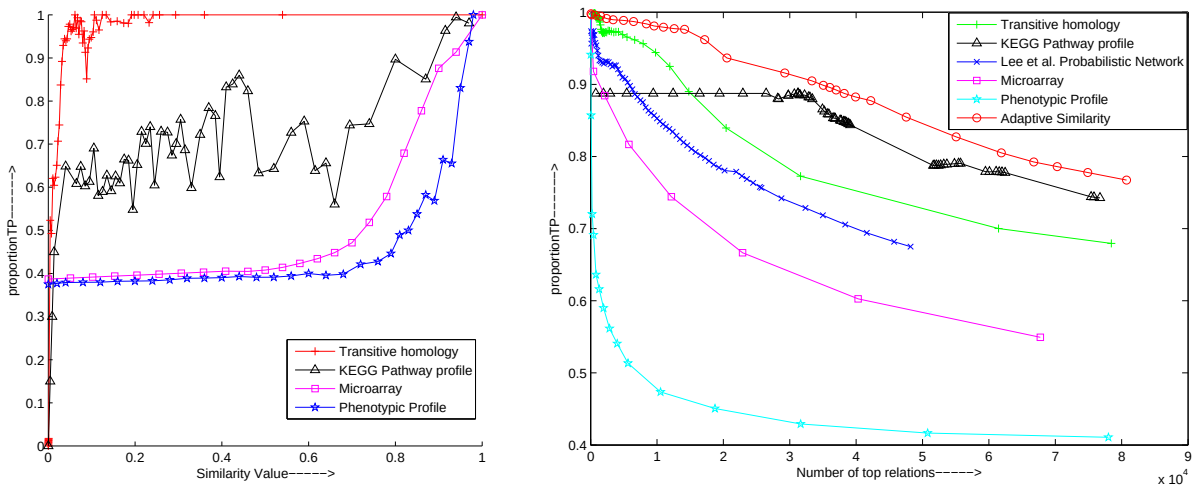
As super GO-Slim process was used for determining the weights of the data-sources, top level classification of MIPS October 2005 annotation is now used to evaluate the performance of *AdS*. We sorted the similarity values computed from different data-sources in descending order, and drew a curve for top gene-pairs verses *proportionTP* from the sorted data for each form of data-source (Fig. 1.b). In contrast, *proportionTP* for protein-protein interactions has a constant value of 0.69 and not shown in Fig.1.b. Figure 1.b also compares the performance of *AdS* and 'final log likelihood scores' of Lee et al.'s [3] probabilistic network (downloaded from the website mentioned in [13]) in terms of *proportionTP* with MIPS annotation. From the figure it is clear that the gene-pairs identified in this investigation is better than any other existing network or data-sources. The top $1,00,000$ gene-pairs predicted by our method with *proportionTP* values above 0.755 are available at http://www.isical.ac.in/~scc/Bioinformatics/AdS/toprelation.txt in tabular form. The *proportionTP* values computed from individual data-source are also shown in the file.

Results and Discussions

Genes are considered to be linked if they are among the 10 closest neighbors within a given distance or similarity cut-off [2]. Genes are clustered with a method based on the K Nearest Neighbors (KNN) algorithm [14] by selecting $K = 10$ and using *AdS* with gold standard cut-off value 0.77. The method is denoted as *KNN-AdS*. The biological significance of the clusters generated by our *KNN-AdS* is evaluated with 400 different MIPS functional categories. Clusters with P-values greater than 10^{-5} are not reported. 2507 clusters are identified with at-least three or more members. Out of these clusters, 1915 clusters are identified with functional enrichment in one or more categories. From functionally enriched clusters we pre-

dict the functions of 1855 classified genes (with 95.16% accuracy) and 60 unclassified genes by assigning the function related with the smallest P-value. The functional enrichment, in one or more categories, for clusters intended for 60 unclassified yeast genes are available in tabular form at http://www.isical.ac.in/~scc/Bioinformatics/AdS/unclassifiedprediction.xls. The function with the smallest P-value in the table represents the predicted function for the unclassified gene, and the three values in the parenthesis denote the function related P-value, no. of genes in the cluster, no. of genes in the genome, respectively. The table also includes all the genes within each cluster, the *proportionTP* values arising from various data-sources, and the *AdS* values. A table containing the predicted functions of 1855 classified yeast genes is available at http://www.isical.ac.in/~scc/Bioinformatics/AdS/classifiedprediction.xls. Out of 60 unclassified genes, YEL041W and YDR459C are now (April 2007) classified in MIPS, and our function predictions for these two genes are in agreement with present MIPS classification. YEL041w is related with the category 'phosphate metabolism' (4 out of 5 genes, p-value 1.42×10^{-6}). YDR459C is related with category 'modification with fatty acids' (4 out of 11, P-value 2.3×10^{-7}). Our top predictions consist the function of 12 unclassified (MIPS 2007) and 417 classified genes at P-value cut-off 1×10^{-13} and with 98.20% accuracy. The related Table for top 12 predictions and discussions are available at http://www.isical.ac.in/~scc/Bioinformatics/AdS/top12predictions.pdf.

Figures



a) b)

Figure 1: a) *proportionTP* Vs. similarity values for different types of data-sources. b) Comparing *Adaptive Score* and individual data-source in terms of *proportionTP* versus the number of top gene-pairs.

References

At http://www.isical.ac.in/~scc/Bioinformatics/AdS/top12predictions.pdf the references are available.