# Combining Multi-Source Information through Functional Annotation based Weighting: Gene Function Prediction in Yeast

Shubhra Sankar Ray, Sanghamitra Bandyopadhyay, *Senior Member, IEEE,* and Sankar K. Pal, *Fellow, IEEE,*

***Abstract*—*Motivation:* One of the important goals of biological investigation is to predict the function of unclassified gene. Although there is a rich literature on multi data source integration for gene function prediction, there is hardly any similar work in the framework of data source weighting using functional annotations of classified genes. In this investigation we propose a new scoring framework, called *Biological Score (BS)* and incorporating data source weighting, for predicting the function of some of the unclassified Yeast genes.**

***Methods:* The *BS* is computed by first evaluating the similarities between genes, arising from different data sources, in a common framework, and then integrating them in a linear combination style through weights. The relative weight of each data source is determined adaptively by utilizing the information on Yeast GO-Slim process annotations of classified genes, available from Saccharomyces Genome Database (SGD). Genes are clustered by a method called *K-BS*, where, for each gene, a cluster comprising that gene and its K nearest neighbors is computed using the proposed score (*BS*). The performances of *BS* and *K-BS* are evaluated with gene annotations available from Munich Information Center for Protein Sequences (MIPS).**

***Results:* We predict the functional categories of 417 classified genes from 417 clusters with 98.20 positive predictive value using *K-BS*. The functional categories of 12 unclassified Yeast genes are also predicted.**

***Conclusions:* Our experimental results indicate that considering multiple data sources and estimating their weights with annotations of classified genes can considerably enhance the performance of *BS*. It has been found that even a small proportion of annotated genes can provide improvements in finding true positive gene pairs using *BS*.**

***Index Terms*—gene expression, protein sequence, transitive homology, phenotypic profile, combinatorial optimization, bioinformatics.**

## I. Background

Increasing quantities of high-throughput biological data have become available in recent years. Many of these, such as phenotypic profiles [1], gene expression microarrays [2], protein sequences [3], KEGG pathway [4], protein-protein interaction data [5], [6], protein phylogenetic profiles [7] and Rosetta Stone sequence [8] assess functional relationships between genes on a large scale. These high-throughput data can be the key to assign accurate functional annotation to a significant number of unclassified genes [9] but they alone often lack the degree of specificity needed for accurate gene function prediction. This improvement in specificity can be achieved through the incorporation of heterogeneous functional data in an integrated analysis [9].

The value of combining informations, obtained from different methods, for gene function predictions, has been illustrated by several studies [3], [9]. Marcotte et al. [3] predicted many potential protein functions for Saccharomyces cerevisiae based on a heuristic combination of different types of data. Von Mering et al. [10] first developed quantitative methods to measure functional relationship among genes from three different sources of information. In [9], heterogeneous data sources are integrated in Bayesian network approach and functional modules are predicted by using a clustering algorithm based on the principle of *KNN* algorithm. Lee et al. [11] compared different classes of data and integrated them by using Bayesian Score.

While there are many works regarding data source integration, the relevance of integrating information from data sources in a linear combination style through functional annotation based adaptive weights, is still unexplored. Moreover, all the pre-mentioned works do not incorporate transitive nature of protein homology and KEGG pathway similarity extraction excluding Yeast genes. In this investigation, we present a new computational framework, using functional annotation based weighting of data sources, for the prediction of gene function in yeast. The novelty of our method lies in the way of estimating the weights in a linear combination style, using gene annotations in Eq. 4 (described in Section II-C). In the related work of Lee et al. [11], all available log likelihood scores derived from the various data sets are added with a rank-order dependent weighting scheme. They used a single free parameter for estimating weights.

## II. Methods

We mainly focus on integrating phenotypic profiles, microarray gene expression, KEGG pathway related protein database in Protein Information Resource (PIR) [12], protein sequence similarity by transitive homology, and protein-protein interaction information as data sources. The main steps of our methodology for predicting gene functions can be summarized as:

S1) extract pairwise similarity of genes, obtained from different data sources (see Section II-A);

S. S. Ray and S. K. Pal are with the Center for Soft Computing Research, Indian Statistical Institute, Calcutta 700108, India (e-mail: shubhra_r@isical.ac.in; sankar@isical.ac.in).

S. Bandyopadhyay is with the Machine Intelligence Unit, Indian Statistical Institute, Calcutta 700108, India (e-mail: sanghami@isical.ac.in).

S2) separately re-score the similarities in a common framework of Yeast GO-Slim: Process annotations (see Section II-B);

S3) integrate the re-scored similarities from different data sources through the proposed scoring framework (see Section II-C) and calculate the combined score;

S4) for each gene $g$, form a cluster comprising that gene and its K nearest neighbors using the proposed score and predict the function of $g$ by noting the functional enrichment of the cluster using MIPS [13] annotation (see Section II-D).

Each of the above steps are discussed in detail in the following subsections.

### A. Data Sources and Similarity Extraction Techniques

Here we describe the different data sources and their respective similarity extraction techniques.

*1) Phenotypic Profile:* Recently, Brown et al. [1] presented a method for the analysis of the function of genes in budding yeast. The method is based on hierarchical clustering of the quantitative sensitivity profiles of the 4756 strains with individual homozygous deletion of all nonessential genes. They showed the method to be superior than other global methods for identifying various interrogated functions. The detailed procedure of generating and normalizing the data is available in Brown et al. [1]. The normalized data is downloaded from the supplimentary material and we use Pearson correlation for phenotypic profile similarity extraction. The genes with more than 50% missing values are first eliminated from the dataset. For the remaining genes missing values are estimated using LSimpute_adaptive method available in LSimpute [14] software.

*2) Gene Expression:* We use the All Yeast [2], [15] data for gene expression similarity extraction. Brown et al. [1] have shown that even with 30 distinct biological conditions for gene expression, GO term ribosome biogenesis (GO:0007046) tends to dominate gene pairs implicated by coexpression. As we have already used phenotypic profiles, which implicate gene relationships over a broad range of biological processes, here we only use the widely studied All yeast data and use centered Pearson correlation for extracting gene expression similarity. The All Yeast dataset is downloaded from Stanford Microarray Database [16] with default normalization parameters, as suggested by the experts. The missing values in the dataset are estimated using LSimpute_adaptive [14] in a similar fashion to phenotypic profiles, mentioned in Section II-A1.

*3) KEGG pathway:* The pathway information for genes in KEGG [4] can be utilized as a reference for functional reconstruction. All the protein sequences, except Yeast proteins, corresponding to each pathway (121 pathways in the second level) are downloaded from PIR [12]. Profile vector for each protein in Yeast is computed by comparing its sequence across 121 pathway databases, using BLAST [17]. The method is similar to phylogenetic profile [7] construction, where, each pathway database is replaced by all proteins within a species. The pathway profiles of genes, computed using KEGG pathway databases, are denoted as KEGG profiles. To find the similarity between two genes using KEGG profiles, we used the ratio of dot product value and OR value between two profiles. The similarity matrix has a highest similarity value of 1. Hence, the similarity values, obtained by all pairwise comparison, have a dynamic range from 0 to 1 and its normalization is unnecessary. Note that, the genes, whose protein sequences are not available, are assigned a pathway profile similarity value of 0 w.r.t. all other genes (proteins).

*4) Protein Sequence:* Comparing the protein sequences presents an alternative prominent approach for gene annotation and analysis. Intuitively one can assume that all the protein relations, arising from direct protein similarity search, are available in the literature and will not help in predicting functions for unclassified genes in a widely studied organism like Yeast. As compared to direct protein similarity search, the field of searching gene/protein similarity through phylogenetic profiles (PP) [7], Rosetta Stone sequence (RS) [8], and transitive homology [18] are relatively new methods. In this investigation, transitive homologues are used instead of PP and RS, for extracting protein similarity, as its performance is reported to be better than PP and RS in literature [19], [20].

Transitive homology detection method [18], [20] works by searching the query sequence against the database with a conservative threshold to find the closely homologous sequences and using these homologous sequences as seeds to search the database to find remotely homologous sequences with a less conservative threshold. The method has been shown to be close to the profile [7] based methods and better than a direct pairwise homology search [18]. To find the transitive homologues, homology comparisons are performed among target proteins and 37,66,477 proteins downloaded from UniProt [21], by using BLASTP in BLAST [17]. Before comparison all the yeast proteins are removed from the downloaded database. Let the similarity (E-values using BLAST) between two protein sequences A and B be $B_{A,B}$. The value $B_{A,B}$ is replaced by $B_{A,C} \times B_{C,B}$ if there exists a sequence C such that $B_{A,C} \times B_{C,B}$ is larger than the current value of $B_{A,B}$. This transformation takes advantage of the transitive homology of sequences A and B through the intermediate sequence C, assuming that sequences A and C and sequences B and C are independently homologous [20]. Instead of storing raw BLAST score as the similarity between two protein sequences, we use the metric of ProClust [22] where the metric value scales from 0 to 1. Here also the genes, whose protein sequences are not available, are assigned a transitive protein similarity value of 0 w.r.t. all other genes.

*5) Protein-Protein Interaction:* Protein-protein maps promise to reveal many aspects of the complex regulatory network underlying cellular function [10]. For this study, manually curated catalogues of known protein-protein interactions are downloaded from BioGRID [6]. For a given pair of genes/proteins the similarity value is 1 or 0, indicating a interaction present or absent, respectively. The BioGRID database/catalogue includes more than 90000 interactions by combining results obtained from different experiments. The related references of experiments are available in BioGRID.

### B. Scoring the Similarities in a Common Framework

Scoring the data sets by a single criterion allows us to directly measure the relative merit of each data set and then to integrate the data sets with weights that reflect this merit, even when the data sets are accompanied by their own intrinsic scoring schemes (such as Pearson Correlation for gene expression). In this regard, the similarities arising from various data sources are separately re-scored, based on the common framework of Yeast GO-Slim process annotations of genes in the SGD database [23]. The proportion of true positive (TP) gene pairs at a particular similarity value (computed from a data source) can be used as a single criterion for re-scoring the similarity values, 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) GO (Gene Ontology) term annotation. In [9] proportion of TP pairs (positive predictive value ($PPV$)) of a method is defined as

$$PPV = \frac{no.\ of\ predicted\ pairs\ with\ common\ GO\ term}{total\ no.\ of\ predicted\ pairs}. \tag{1}$$

The hierarchical nature of GO and multiple inheritance in the GO structure can lead to evaluation problems if we consider only the particular GO term with which a gene is annotated [9]. To alleviate this problem, we consider the SGD Yeast GO-Slim process annotations, where every gene is annotated in the same level without any tree based structure. For every gene $g$, that has undergone Yeast GO-Slim process annotation, a vector

$$V(g) = (v_1, v_2, \cdots, v_j) \tag{2}$$

is used to represent its category (Yeast GO-slim process) status, where $j$ is the number of categories. The value of $v_j$ is 1 if gene $g$ is in the $j$th category; otherwise is zero. Based on the information about categorization, the positive predictive value ($PPV$) at a given similarity value, can be defined as

$$PPV = \frac{\sum_{i=1}^{n}\sum_{m=1}^{j}(V(g_i)_m \times V(g_{ir})_m)}{n}, \tag{3}$$

where $\sum_{m=1}^{j}(V(g_i)_m \times V(g_{ir})_m)$ is set to 1 if $\sum_{m=1}^{j}(V(g_i)_m \times V(g_{ir})_m) \geq 1$, $g_i$ and $g_{ir}$ form a gene-pair, $n$ is the number of predicted gene pairs at a given similarity value, and $V(g_i)_m$ represents the $m^{th}$ entry of vector $V(g_i)$. Hence, if a gene pair, associated with any method or data-source, belongs to two or more GO categories then it contributes with a 1 at the numerator of the $PPV$. Its contribution to the denominator is 1 if both the genes in a gene pair belong to at least one GO category (i.e., both are classified). If any one of them is unclassified then contribution of that gene pair in the denominator of $PPV$ will be 0.

Figure 1 compares the similarity values obtained from different data sources in terms of their $PPV$. The $PPV$ for intermediate similarity values, that are not plotted in Fig. 1, are calculated from the slopes of the respective curves. The similarities extracted from protein-protein interactions are binary relations in our study. Therefore, $PPV$ 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.

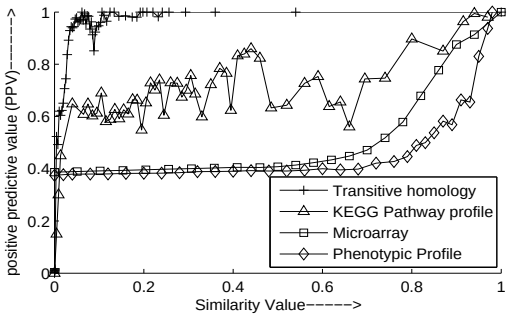


Fig. 1. Comparing the re-scored similarity values for different types of data sources to obtain equivalency in the common framework of Yeast GO-Slim process annotations. The positive predictive values ($PPV$) versus the similarity values are plotted for each data source.

### C. New Framework for Data Source Integration

As the similarities computed from different data sources are re-scored (see Section II-B) on a single criterion and common framework of Yeast GO-Slim process annotations, they are directly comparable and can be integrated even when the natures of experiments are distinct. The $PPV$ reflects the usefulness of a data source at a given similarity value, but do not provide any information about weight of one data source in presence of the other data sources, in predicting gene pairs. Consequently, it will be more appropriate and better if

1) $PPV$ of each data source, in presence of other data sources, is separately weighed by a factor and then integrated;
2) factors are dependent on the $PPV$ of the integrated $PPV$ of different data sources.

Such an attempt is made in this article with a new score where, $PPV$s computed from phenotypic similarity ($P$), gene expression similarity ($M$), KEGG pathway profile similarity ($K$), protein similarity through transitive homologue ($B$), and protein-protein interaction information ($I$) between two genes $X$ and $Y$ are integrated through weights $a$, $b$, $c$, $d$, and $e$ in a linear combination style. The weights of the $PPV$s, computed from different data sources, are determined by adaptively maximizing the $PPV$ of the new score using Yeast GO-Slim process annotations [23] of known genes. This score is referred to as $Biological\ Score$ ($BS$) and is defined as

$$\begin{aligned} BS_{X,Y} &= \frac{a \times P_{X,Y} + b \times M_{X,Y} + c \times K_{X,Y}}{a+b+c+d+e} \\ &\quad + \frac{d \times B_{X,Y} + e \times I_{X,Y}}{a+b+c+d+e} \end{aligned} \tag{4}$$

where $a$, $b$, $c$, $d$, and $e$ are varied within range 0 to $\alpha$ in steps of 1 to find a combination that maximizes the $PPV$ for a user defined number of top gene pairs. The weighting scheme enables all possible weighting (including equal and zero weighting). Note that, the weights $a$, $b$, $c$, $d$, and $e$ are assigned to the complete $PPV$ matrices calculated from individual data sources. The following can be stated about the score:

1) $0 \leq BS_{X,Y} \leq 1$
2) $BS_{X,Y} = BS_{Y,X}$ (symmetric).

The proposed scoring framework for data source integration, in Eq. 4, is based on data source weighting where the re-scored similarity spaces, available from different data sources, are adaptively transformed using a set of weighting coefficients. Intuitively, more important similarity spaces should be assigned larger weights than less important ones, while irrelevant ones should be assigned zero weight. Although the proposed framework has some common working principle with feature weighting (FW) [24], it cannot be categorized as FW because what is computed using $BS$ is the pair-wise gene similarities and not the set of features of any individual gene.

*Estimation of Weights for Maximization of* $PPV$: We maximize the $PPV$, using Yeast GO-Slim process annotations, for top gene pairs by varying the weights $a$, $b$, $c$, $d$, and $e$ in the $BS$ (Eq. 4). For each set of values of $a$, $b$, and $c$, the top gene pairs are identified with a gold standard cut-off value. Our gold standard cut-off value and gold standard of top gene pairs are determined from KEGG pathway profiles, which provides $26432$ gene pairs with similarity value 1 and constant $PPV$ of $.81$. These gene pairs are the most predictive of all, whereas the $PPV$ of other data sources, as well as gene pairs below top $26432$ for KEGG pathway profiles, vary considerably. We now use the following steps to estimate the weight factors $a$, $b$, $c$, $d$, and $e$ in the $Biological\ Score$:

S1) All the factors are assigned an initial value of 1.
S2) $BS$ values are calculated for all the gene pairs and sorted in descending order to identify the cut-off value above which the top $26432$ gene pairs are available.
S3) $PPV$ is calculated for the top $26432$ gene pairs.
S4) The weight factors are now varied in steps of .1 and the steps from 2 to 3 are repeated to find a combination of weights for which the $PPV$ is maximized.

Figure 2 shows how $PPV$, using Yeast GO-Slim process, varies for different values of weight factors ranging from 0 to 100, in steps of 1. The curves show instances where one weight factor is varied and the other weight factors are kept constant. Experiments are also conducted by excluding the KEGG pathway profile database and the corresponding curves are reffered to as $c = 0$.

### D. Gene Function Prediction

The biological function for each gene is predicted from a cluster comprising that gene and its top K nearest neighbors [3], [9] by selecting a gold standard $BS$ cut-off value obtained from KEGG pathway profiles using MIPS October 2005 classification. This gene clustering method using $BS$ is denoted as *K-BS*, where each gene is considered once for its function prediction and allows its neighbor genes to be a member of multiple gene clusters. As Yeast GO-Slim process annotations was used for determining the weights of the data sources, 510 different MIPS (October 2005) functional categories are used to evaluate the biological significance of the clusters generated by our *K-BS*. One or several predominant functions are then assigned to each cluster and the target gene by calculating the P-values for different functional categories.

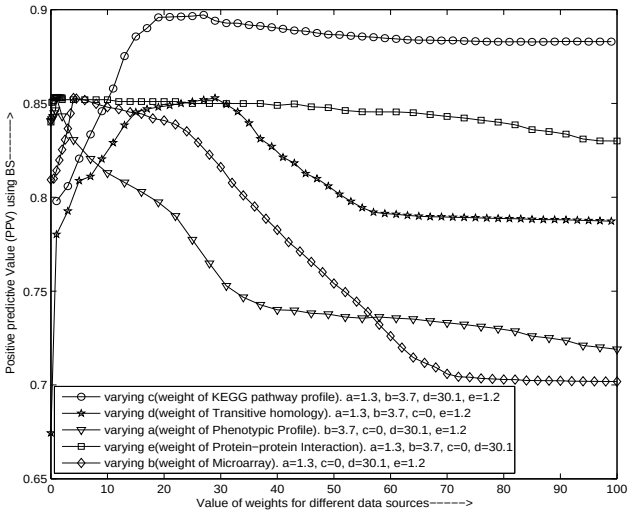


Fig. 2. Comparing the values of $PPV$ using $BS$, by varying weights of $PPV$ of different data sources for top $26432$ gene pairs. When a particular weight is varied the other weights are kept constant at the values shown in the figure. The curves obtained with c=0 indicate that KEGG pathway profile is excluded in the integration process.

## III. Results

As Yeast GO-Slim process was used for determining the weights of the data sources, MIPS annotation is now used to evaluate the performance of *BS*. Genes/proteins that could not be mapped to their MIPS identifier are eliminated. Our gold standard $PPV$ of top gene pairs is now changed and determined from KEGG pathway profiles, which provides $26432$ gene pairs with constant $PPV$ of $.8874$, using top level classification of MIPS annotation. In this section, first we discuss about the various paremeters involved in the clustering method and the biological significance of some clusters in Section III-A. Influence of number of classified genes on the proposed scoring framework is demonstrated in Section III-B. In Section III-C, we present the comparisons of our method with Lee et al.'s [11] probabilistic network and individual data sources. Finally, the performance of *BS* and some comparisons based on independent training (estimating weight factors) and test set with null intersection are presented in Section III-D.

### A. Gene Function Prediction based on Clustering Results

Genes are considered to be linked if they are among the 10 closest neighbors within a given distance or similarity cut-off [3]. The biological function for each gene is predicted from the cluster consisting the top 10 neighbors of that gene by selecting $K$ to be at most $10$ and $BS$ cut-off value of 0.77. Above this cut-off value the gold standard $PPV$ of 0.8874 is achieved for 36033 gene pairs using the MIPS October 2005 classification. We found several clusters to be significantly enriched with genes of a similar function. Clusters with P-values greater than $10^{-5}$ are not reported.

To predict a genes function from it's neighbor genes we use the following steps:

S1) 2507 clusters are identified with at-least three or more members by selecting $K = 10$ and with $BS$ gold standard cut-off value 0.77.
S2) Out of these clusters, 1915 clusters are identified with functional enrichment in one or more categories and P-values less than $10^{-5}$.
S3) From functionally enriched clusters we predict the functions of 1855 classified and 60 unclassified genes by assigning the function related with the smallest $P$-value. This in practice resulted in more accurate predictions than if multiple functions are allowed per cluster.

The functions of 1855 classified genes are predicted with 95.16 $PPV$. The functional enrichments for clusters intended for 60 unclassified yeast genes are available in tabular form (tab delimited file) at http://www.isical.ac.in/˜scc/Bioinformatics/AdS/unclassifiedpr-ediction.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, function related no. of genes in the cluster, and the function related no. of genes in the genome, respectively. The table also includes all the genes within each cluster, the $PPV$ arising from various data sources, and the $BS$ values. A table with similar format, containing the predicted functions of 1855 classified yeast genes is available at http://www.isical.ac.in/˜scc/Bioinformatics/AdS/classifiedpredi-ction.xls.

Out of 60 unclassified genes, YEL041W and YDR459C are now classified in MIPS, and our function predictions for these two genes are in agreement with MIPS. YEL041w and its four neighbors YJR049C, YPL188W, YDR226W, and YER170W form a cluster. From the functional enrichment of the cluster we correctly predict that YEL041w is related with the category 'phosphate metabolism' (with p-value $1.42\times10^{-6}$) as the four remaining genes belong to this category. The cluster containing gene YDR459C and its ten neighbor genes shows functional enrichment in categories 'protein modification' (8 out of 11, $P$-value $1.16 \times 10^{-6}$), 'modification with fatty acids' (4 out of 11, $P$-value $2.3 \times 10^{-7}$) and 'modification by acetylation, deacetylation' (4 out of 11, $P$-value $4.4\times10^{-6}$). We correctly predict that YDR459C is related to 'modification with fatty acids'.

Our top predictions consist the function of 12 unclassified (MIPS 2007) and 417 classified genes at $BS$ cut-off value $0.77$, and *P*-value cut-off $1 \times 10^{-13}$. At these cut-off values, the functions of the classified genes are predicted with 98.20 $PPV$. Table I summarizes the top 12 predicted functions for 12 unclassified genes. Each of the clusters contain 11 genes and they are available in the table representing 60 clusters for function prediction of unclassified genes. Since four of the 12 clusters show functional enrichment in a single category of 'DNA topology', we analyze these clusters manually. We observe that 15 classified, 4 unclassified and 2 recently deleted (YEL076C and YPR203W) genes are distributed in these clusters with 80% genes in common. On examination of the literature for 4 unclassified genes (YHR218W, YHR219W, YBL112C, and YLR464W), we find that their involvement in DNA processing and DNA topology is likely due to their relation to helicase-proteins [23], [25].

### *B. Influence of Number of Classified Genes on Functional Annotation based Weighting*

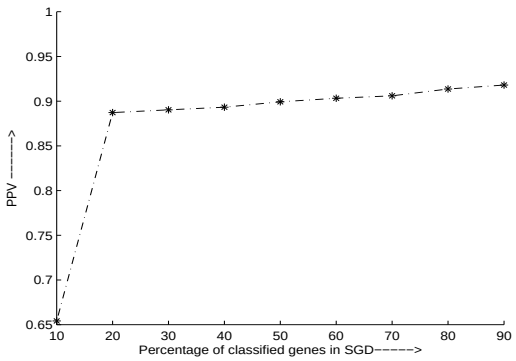


Fig. 3. Variation of $PPV$, using *BS*, with nine different percentages of classified genes.

Here we study how the increase in the number of classified genes in Yeast GO-Slim affects the $PPV$ for the classified genes in MIPS for top $26432$ gene pairs using *BS*. We found that even with 20% of classified genes the estimated values of $a$, $b$, $c$, $d$, and $e$, in maximizing $PPV$, differs by an amount of 1 than the estimated values with 90% of classified genes. Hence, the value of $PPV$ also varies by an amount of 0.02 to 0.03 with classified genes ranging from 20% to 90%. Fig. 3 shows that the percentage of classified genes clearly has a limited contribution to the $PPV$ of the *BS*. Thus *BS* may also be successfully used for organisms where the number of classified genes is as low as 20%.

### *C. Comparative Performance of Methods and Data Sources*

In order to demonstrate the power of data source integration, we compare the $PPV$ of gene pairs identified by the $BS$ with those identified by the individual data sources. Since $BS$ uses GO annotations for adapting its weights, it is not used for performing the comparisons. Rather, the MIPS annotation of classified genes is used (see Fig. 4). We sorted the similarity values computed from $BS$, phenotypic profiles, gene expression, KEGG profiles, and protein similarity from transitive homology in descending order, and drew a curve for top gene pairs verses $PPV$ from the sorted data for each form of data source. In contrast, $PPV$ for protein-protein interactions has a constant value of 0.69 and not shown in Fig.4. We found that the curve of $BS$ is above the other curves. Moreover, the top $26432$ gene pairs has an $PPV$ greater than the gold standard KEGG pathway profiles. The gene pairs are also reasonably distinct from gene pairs of KEGG pathway profiles. Figure 4 also compares the performance of $BS$ and 'final log likelihood

TABLE I
TOP 12 FUNCTION PREDICTIONS OF UNCLASSIFIED GENE AT $BS$ CUT-OFF VALUE OF 0.77

| Unclassified Gene | Functional category | *P*-value | Genes within cluster | Genes within category |
|---|---|---|---|---|
| YIL080W | ABC transporters | 2.2204e-16 | 8 | 28 |
| YLR057W | modification with sugar residues | 2.2871e-14 | 8 | 67 |
| YHR218W | DNA topology | 0 | 9 | 52 |
| YHR219W | DNA topology | 0 | 10 | 52 |
| YIL170W | C-compound and carbohydrate transport | 1.3656e-14 | 8 | 63 |
| YDR441C | purin nucleotide/nucleoside/nucleobase metabolism | 6.7724e-15 | 8 | 58 |
| YCL074W | TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS | 3.3307e-16 | 8 | 34 |
| YBL112C | DNA topology | 0 | 10 | 52 |
| YLR464W | DNA topology | 2.6645e-15 | 8 | 52 |
| YMR010W | modification with sugar residues (e.g. glycosylation, deglycosylation) | 0 | 9 | 67 |
| YIL067C | vesicle fusion | 2.2204e-16 | 9 | 32 |
| YHR049W | metabolism of secondary products derived from glycine, L-serine and L-alanine | 3.3307e-16 | 7 | 19 |

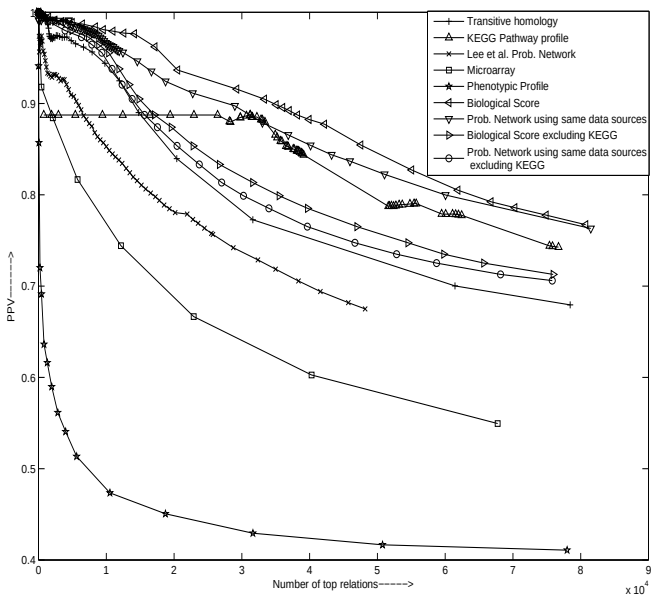


Fig. 4. Comparison between the $Biological\ Score$ ($BS$), Lee et al.'s Probabilistic Network, and individual data source in terms of $PPV$ versus the number of top gene pairs.

scores' of Lee et al.'s probabilistic network (downloaded from the website mentioned in [26]) in terms of $PPV$ with MIPS annotation. The curve of Lee et al.'s probabilistic network is drawn from top 34,000 gene pairs, as mentioned in [11]. For a direct comparison between our method and the probabilistic network, we implemented the probabilistic network as described in Lee et al. using the same datasources as in $BS$ and plotted the respective curve in Fig. 4. From the figure it is clear that the top gene pairs identified in this investigation is better than any other existing network or data sources. The above statement is true not only for gold standard 24632 gene pairs but also for top $80000$ gene pairs which can be used further for gene function prediction. The top $1,00,000$ gene pairs predicted by our method with $PPV$ above 0.755 (not shown in the data) are available in http://www.isical.ac.in/˜scc/Bioinformatics/AdS/toprelation.txt in tabular (tab delimited) form. The $PPV$ computed from individual data source are also shown in the file. The KEGG pathway profile information may be a bit redundant with functional annotations available in MIPS and Yeast GO-Slim process. In this regard experimensts are also conducted by excluding the KEGG pathway profile dataset from the datasource integration procedure in $BS$ and our implemented version of Lee et al.'s probabilistic network while, all other aspects are kept unchanged. The two corresponding curves are also shown in Fig. 4.

### *D. Evaluation Based on Independent Training and Test Sets*

To perform a fair evaluation of the methdos, the training and test set should be independent with null intersection and in this regard we also experimented with a method based on cross-validation. The KEGG pathway profile dataset also remains excluded from the integration procedure to avoid any redundancy in KEGG pathway information and annotations available in MIPS and Yeast GO-Slim process.

In this study, we randomly picked 3036 genes with Yeast GO-Slim process annotations (using Eq. 2), to train the weights in $BS$ and then evaluated the performance with the remaining 3036 genes with MIPS annotations. All links among the genes within the same training subset and the same test subset are calculated, with neither links nor genes shared between the training and test sets. Because data are integrated using weights derived only from a part of genes with Yeast GO-Slim process annotations, the performances measured on remaining genes with MIPS annotations are expected to be free from circular logic and memorization of the annotation set during the training procedure. All other steps prior to the final assessment of *BS* are performed using only the training set. The final assessment is performed on the independent test set. The cross-validation procedure is repeated 10 times and the performance of *BS* is evaluated.

Fig. 5 shows the curves comparing *BS*, Lee et al.'s Probabilistic Network and individual data sources in terms of $PPV$ for top gene pairs, in one of the cross-validation procedures. Similar curves are obtained when the cross-validation procedure is repeated. The curves show that *BS* performs better than Lee et al.'s Probabilistic Network and individual data source.

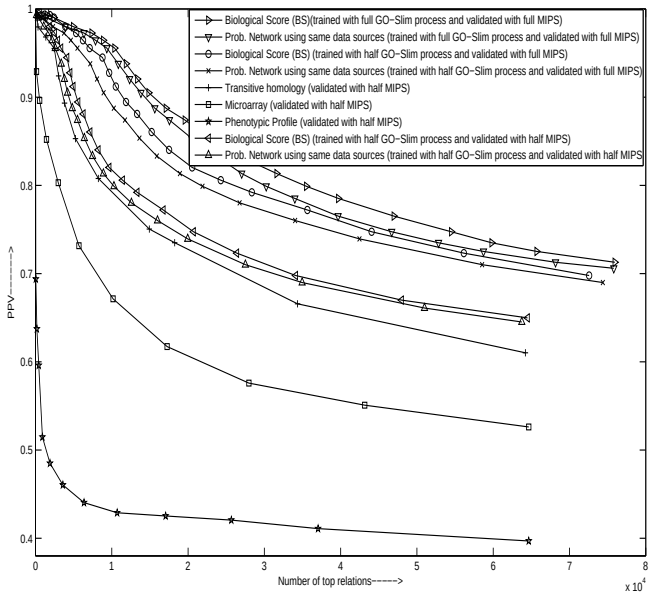


Fig. 5. Comparison between the $Biological\ Score$ ($BS$), Lee et al.'s Probabilistic Network, and individual data source in terms of $PPV$ versus the number of top gene pairs. The cross-validation results of $BS$ and Lee et al.'s Probabilistic Network (with two different sets of 3036 genes) are shown in curve 5 and 6, respectively.

In order to compare the performance of cross-validation results with the results reported in Section III-C for *BS* and Lee et al.'s Probabilistic Network without KEGG as a datasource, the top two curves are provided in Fig. 5 (these curves also appear in Fig. 4, and are provided here for convenience). These two curves are at the top of Fig. 5 and are superior to the two other curves of *BS* and Lee et al.'s Probabilistic Network where, half of the genes with Yeast GO-Slim process annotations are used to train weights and half of the genes with MIPS annotations are used for evaluation. Experiments are also conducted by randomly picking half of the genes with Yeast GO-Slim process annotations to train weights and all the genes with MIPS annotations for evaluation of *BS* and Lee et al.'s Probabilistic Network. The corresponding curves are shown in Fig. 5 for the purpose of illustration.

In clustering solutions, using *K-BS* and repeteting cross-validation procedure, on average 642 clusters are identified with functional enrichment in one or more categories by selecting $K \leq 10$, $BS \geq 0.77$, and $P - values < 10^{-5}$. From functionally enriched clusters, on average we predict the functions of 405 classified genes with 94.9 $PPV$ and 237 unclassified genes by assigning the function related with the smallest $P$-value. In one of the cross-validation process (out of 10 repetitions), functions of 454 classified yeast genes are predicted with 96.4 $PPV$ from 454 clusters. The predicted functions of 454 classified yeast genes are available at http://www.isical.ac.in/˜scc/Bioinformatics/AdS/classifiedpredi-ctionreview.xls.

## IV. Conclusion

In this study we proposed a framework for data source integration, through functional annotation based weighting, to predict gene function for yeast. Five data sources, namely, phenotypic profiles, gene expression data, KEGG profiles, protein-protein interaction and protein sequence similarity through transitive homologues are used. Functional categories of 60 unclassified (MIPS October 2005) Yeast genes and 1855 classified genes are predicted with 95.16 $PPV$. Evaluation on the predicted gene pairs confirmed the validity and potential value of the proposed framework for gene function prediction.

Although a neighbor based clustering method needs a user defined neighbor number, from this investigation we find that *K-BS* is a highly accurate and efficient gene function annotation tool. The system integrates heterogeneous biological information in a functional annotation based weighting framework, leading to more biologically accurate gene groupings, which can be used for gene function prediction. The flexibility of the system also allows for easy inclusion of other data sources. Furthermore, we plan to examine our proposed framework on a larger test-bed by including similarities arising from gene-fusion and gene-order conservation based methods.

## Acknowledgement

We would like to acknowledge Dr. Maria C. Costanzo, Curator of SGD, for mapping the gene names to ORFs and anonymous reviewers for their suggestions in improving the quality of research. Support of the Dept. of Science and Technology, Govt. of India to the Center for Soft Computing Research through its IRHPA scheme is acknowledged. The support was provided when one of the authors, S. K. Pal, was a J. C. Bose Fellow of the Govt. of India.

## References

[1] J. A. Brown, G. Sherlock, C. L. Myers, N. M. Burrows, C. Deng, H. I. Wu, K. E. McCann, O. G. Troyanskaya, and J. M. Brown, "Global analysis of gene function in yeast by quantitative phenotypic profiling," *Molecular System Biology*, vol. 2, no. 2006.0001, pp. 1–9, 2006.

[2] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 14863–14867, 1998.

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

[4] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, "From genomics to chemical genomics: new developments in kegg," *Nucleic Acids Res.*, vol. 34, pp. D354–D357, 2006.

[5] L Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit J. U. Bowie, and D. Eisenberg, "The database of interacting proteins," *Neuclic Acid Research*, vol. 32, pp. 449451, 2004.

[6] T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, G. C. Hon, C. L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O. G. Troyanskya, T. Ideker, K. Dolinski, N. N. Batada, and M. Tyers, "Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae," *Journal of Biology*, vol. 5, no. 4, pp. 1–28, 2006.

[7] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 4285–4288, 1999.

[8] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, "Detecting protein function and protein-protein interactions from genome sequences," *Science*, vol. 285, pp. 751–753, 1999.

[9] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, "A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae)," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 14, pp. 8348–8353, 2003.
[10] C. V. Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, pp. 399–403, 2002.
[11] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte, "A probabilistic functional network of yeast genes," *Science*, vol. 306, pp. 1555–1558, 2004.
[12] W. C. Barker et al., "The protein information resource (pir)," *Nucleic Acids Research*, vol. 28, no. 1, pp. 41–44, 2000.
[13] Munich Information for Protein Sequences, "http://www.mips.com," .
[14] T. H. B, B. Dysvik, and I. Jonassen, "Lsimpute: accurate estimation of missing values in microarray data with least squares methods," *Nucleic Acids Research*, vol. 32, no. 3: e34, pp. online, 2004.
[15] Website, "http://rana.lbl.gov/EisenData.htm," .
[16] G. Sherlock et al., "The stanford microarray database," *Nucleic Acids Research*, vol. 29, no. 1, pp. 152–155, 2001.
[17] S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.
[18] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia, "Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods," *J Mol Biol*, vol. 284, pp. 1201–1210, 1998.
[19] H. Xie, A. Wasserman, Z. Levine, A. Novik, V. Grebinskiy, Avi Shoshan, and Liat Mintz, "Large-scale protein annotation through gene ontology," *Genome Research*, vol. 12, pp. 785–794, 2002.
[20] Q. Ma, G. W. Chirn, R. Cai, J. D. Szustakowski, and N. Nirmala, "Clustering protein sequences with a novel metric transformed from sequence similarity scores and sequence alignments with neural networks," *BMC Bioinformatics*, vol. 6, no. 242, 2005.
[21] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh, "The universal protein resource (uniprot)," *Nucleic Acids Research*, vol. 33, pp. 154–159, 2005.
[22] P. Pipenbacher, A. Schliep, S. Schneckener, A. Schonhuth, D. Schomburg, and R. Schrader, "Proclust: improved clustering of protein sequences with an extended graph-based approach," *Bioinformatics*, vol. 18, no. 2, pp. S182S191, 2002.
[23] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, and J. M. Cherry, "Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go)," *Nucleic Acids Research*, vol. 30, no. 1, pp. 69–72, 2002.
[24] D. Wettschereck, D. W. Aha, and T. Mohori, "A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms," *Artificial Intelligence Review*, vol. 11, no. 1-5, pp. 273–314, 1997.
[25] A. Shiratori, T. Shibata, M. Arisawa, F. Hanaoka, Y. Murakami, and T. Eki, "Systematic identification, classification, and characterization of the open reading frames which encode novel helicase-related proteins in saccharomyces cerevisiae by gene disruption and northern analysis," *Yeast*, vol. 15, no. 3, pp. 219–253, 1999.
[26] I. Lee, R. Narayanaswamy, and E. M. Marcotte, *Yeast Gene Analysis*, chapter Bioinformatic prediction of yeast gene function, Elsevier Press, Amsterdam, 2006.

**Shubhra Sankar Ray** is a Visiting Research Fellow at the Center for Soft Computing Research: A National Facility, Indian Statistical Institute, Kolkata, India. He received the M.Sc. in Electronic Science and M.Tech in Radiophysics & Electronics from University of Calcutta, Kolkata, India, in 2000 and 2002, respectively. Till March 2006, he had been a Senior Research Fellow of the Council of Scientific and Industrial Research (CSIR), New Delhi, India, working at Machine Intelligence Unit, Indian Statistical Institute, India. His research interests include bioinformatics, evolutionary computation, neural networks, and data mining.

**Sanghamitra Bandyopadhyay** did her BS, MS and Ph. D. in Computer Science in 1991, 1993 and 1998 respectively. Currently she is an Associate Professor at Indian Statistical Institute, India. She has worked in Los Alamos National Laboratory, Los Alamos, USA in 1997, University of New South Wales, Sydney, Australia, during 1999, University of Texas at Arlington, USA in 2001, University of Maryland at Baltimore, USA in 2004, Fraunhofer Institute, Germany in 2005, and Tsinghua University, China in 2006 and 2007. Dr. Bandyopadhyay is the first recipient of Dr. Shanker Dayal Sharma Gold Medal and also the Institute Silver Medal for being adjudged the best all round post-graduate performer in IIT, Kharagpur, India, in 1994. She has also received the Young Scientist Awards of the Indian National Science Academy (INSA), the Indian Science Congress Association (ISCA) in 2000 and the Indian National Academy of Engineers (INAE) in 2002. She also received the 2006-2007 Swarnajayanti Fellowship in Engineering Sciences from the Govt. of India. She is a senior member of IEEE. Dr. Bandyopadhyay has co-authored more than one hundred and twenty five technical articles in international journals, book chapters and conference/workshop proceedings. She has delivered many invited talks and tutorials around the world. She was the Program Co-Chair of the First International Conference on Pattern Recognition and Machine Intelligence, (PReMI'05) held in Kolkata, India, during December 18-22, 2005. Her research interests include Bioinformatics, Soft and Evolutionary Computation, Image Processing, Pattern Recognition and Data Mining.

**Sankar K. Pal** is the *Director* and *Distinguished Scientist* of the Indian Statistical Institute. He has founded the Machine Intelligence Unit in 1993, and the Center for Soft Computing Research: A National Facility in 2004 at the Institute in Calcutta. He received a Ph.D. in Radio Physics and Electronics from the University of Calcutta in 1979, and another Ph.D. in Electrical Engineering along with DIC from Imperial College, University of London in 1982.

He worked at the University of California, Berkeley and the University of Maryland, College Park in 1986-87; the NASA Johnson Space Center, Houston, Texas in 1990-92 & 1994; and in US Naval Research Laboratory, Washington DC in 2004. Since 1997 he has been serving as a *Distinguished Visitor* of IEEE Computer Society (USA) for the Asia-Pacific Region, and held several visiting positions in Hong Kong and Australian universities.

Prof. Pal is a *Fellow* of the IEEE, USA, Third World Academy of Sciences, Italy, International Association for Pattern recognition, USA, and all the four National Academies for Science/Engineering in India. He is a co-author of thirteen books and about three hundred research publications in the areas of Pattern Recognition and Machine Learning, Image Processing, Data Mining and Web Intelligence, Soft Computing, Bioinformatics, Neural Nets, Genetic Algorithms, Fuzzy Sets, and Rough Sets.

He has received the 1990 *S.S. Bhatnagar Prize* (which is the most coveted award for a scientist in India), and many prestigious awards in India and abroad including the 1999 *G.D. Birla Award*, 1998 *Om Bhasin Award*, 1993 *Jawaharlal Nehru Fellowship*, 2000 *Khwarizmi International Award* from the *Islamic Republic of Iran*, 2000-2001 *FICCI Award*, 1993 *Vikram Sarabhai Research Award*, 1993 *NASA Tech Brief Award (USA)*, 1994 *IEEE Trans. Neural Networks Outstanding Paper Award (USA)*, 1995 *NASA Patent Application Award (USA)*, 1997 *IETE-R.L. Wadhwa Gold Medal*, and the 2001 *INSA-S.H. Zaheer Medal*.

Prof. Pal is an *Associate Editor* of IEEE Trans. Pattern Analysis and Machine Intelligence, IEEE Trans. Neural Networks, Pattern Recognition Letters, Neurocomputing, Applied Intelligence, Information Sciences, Fuzzy Sets and Systems, Fundamenta Informaticae and Int. J. Computational Intelligence and Applications; *a Member, Executive Advisory Editorial Board*, IEEE Trans. Fuzzy Systems, Int. Journal on Image and Graphics, and Int. Journal of Approximate Reasoning; and a *Guest Editor* of IEEE Computer.