A Weighted Power Framework for Integrating Multisource Information: Gene Function Prediction in Yeast

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

***Abstract*—Predicting the functions of unannotated genes is one of the major challenges of biological investigation. In this study, we propose a weighted power scoring framework, called *weighted power biological score (WPBS)*, for combining different biological data sources and predicting the function of some of the unclassified yeast *Saccharomyces cerevisiae* genes. The relative power and weight coefficients of different data sources, in the proposed score, are estimated systematically by utilizing functional annotations [yeast Gene Ontology (GO)-Slim: Process] of classified genes, available from Saccharomyces Genome Database. Genes are then clustered by applying *k-medoids* algorithm on WPBS, and functional categories of 334 unclassified genes are predicted using a *P*-value cutoff 1×10^{-5}. The WPBS is available online at http://www.isical.ac.in/∼shubhra/WPBS/WPBS.html, where one can download WPBS, related files, and a MATLAB code to predict functions of unclassified genes.**

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

I. Background

THE availability of high-throughput biological data, such as, phenotypic profiles [1], gene expression microarrays [2], protein sequences [3], Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway [4], and protein–protein interaction data [5] has opened a new direction in genomic analysis and function prediction of unclassified genes by combining multisource as well as multiscale information from these biological data-sources. A single data source often lacks the degree of accuracy needed for accurate gene function prediction, and this can be improved by integrating different data sources in an efficient manner. Predicting functions of unclassified yeast genes is an important task in biological research as it is considered as a model eukaryotic organism. According to Saccharomyces Genome Database (SGD) [6] and Munich Information for Protein Sequences (MIPS) [7], there are 6069 and 6130 genes for yeast *Saccharomyces cerevisiae*, of which 4387 and 4737 genes, respectively, are classified into some biological process and the remaining genes are unclassified. Out of 1682 and 1393 unclassified genes in SGD and MIPS, 802 and 240 genes, respectively, are either pseudogenes or dubious open reading frames (ORFs). Hence, the number of unclassified genes, without pseudogenes or dubious open reading frames, is 880 and 1153 in SGD and MIPS, respectively. Functional prediction of these genes may also help in classifying human genes with unknown functions.

Mering *et al*. [8] first developed quantitative methods to measure and predict functional relationship among genes by first benchmarking, and then integrating information from different data sources. In [9], proteins are grouped by correlated evolution [10], correlated gene expression [2], and patterns of domain fusion [11] to determine functional relationships among the 6217 proteins of the yeast *Saccharomyces cerevisiae*. Troyanskaya *et al*. [12] integrated data sources in the Bayesian network approach and predicted functional modules by using a clustering algorithm based on the principle of K-nearest neighbor (KNN) algorithm. Interacting networks are predicted in [13] which, not only identifies highly interacting and functionally connected genes, but also those which are sparsely connected with others. Lee *et al*. [14] derived log likelihood scores from the various datasets, weighted them with a rank-order dependent weighting scheme and added them to find a combined similarity using the Bayesian Score. Our previous work in Ray *et al*., 2009, [15] focuses on integrating multiscale information from data sources in a linear combination style through multiple free parameters. Functional categories of 12 unclassified yeast genes are also predicted in this study.

However, the performance of our previous integration method [15] can be improved by incorporating additional free parameters to get estimate of relative powers of individual information, obtained from different datasources. In this regard, we present a new weighted power scoring framework, called *weighted power biological score* ($WPBS$), where besides the existing linear weights, we incorporate new different free parameters involving power estimates of positive predictive values (PPV), obtained from different data sources, namely, phenotypic profiles, cDNA microarray expression, KEGG pathway information, protein similarity through transitive homologues, and protein–protein interaction information.

II. Proposed Approach for Multiscale Data Integration

The main steps of our methodology involves: 1) extraction of pairwise similarity of yeast *Saccharomyces cerevisiae*

Manuscript received July 22, 2011; revised December 27, 2011; accepted January 18, 2012. Date of publication February 3, 2012; date of current version March 21, 2012. *Asterisk indicates corresponding author.*

*S. S. Ray is with the Center for Soft Computing Research: A National Facility, Indian Statistical Institute, Kolkata 700108, India and also with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India (e-mail: shubhra@isical.ac.in).

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

S. K. Pal is with the Center for Soft Computing Research: A National Facility, Indian Statistical Institute, Kolkata 700108, India (e-mail: sankar@isical.ac.in).

genes from different data sources having different observational scales; 2) separately rescoring the similarities, obtained from different data sources, in the common scale of PPV, using yeast GO-Slim: Process annotations; 3) power and weight coefficient estimation and then integration of the PPVs, computed from gene similarities for different data sources, through the proposed scoring framework, WPBS, by adaptively maximizing the PPV of the score itself using yeast GO-Slim process annotations [6] of known genes, and 4) predicting functions of classified as well as unclassified genes from clusters, obtained by applying *k-medoids* algorithm on the proposed score. The function of a gene is predicted by calculating the functional enrichment of the cluster using MIPS annotation.

A. Methods for Measuring Similarities Between Genes for Different Data Sources

Different data sources and their respective pair-wise gene similarity extraction techniques are described in this section.

1) Phenotypic Profile: Brown *et al.* [1] first used phenotypic profiles for functional analysis of genes in budding yeast by hierarchical clustering of the quantitative sensitivity profiles of the 4756 strains with individual homozygous deletion of all nonessential genes. From the clustering solutions, functional predictions related to DNA repair, damage checkpoint pathways, and other functions, are made for some unclassified genes. Analysis of the phenotypic profiles places a total of 860 genes of unknown function in clusters with genes of known function. These complete phenotypic profile data are downloaded from the supplementary material of [1] and used in our investigation as one of the data sources. Pearson correlation is used as a similarity extraction technique for phenotypic profile [1].

2) Gene Expression: To identify relationship among genes, involved in multiple biological functions or processes, many microarray experiments with different biological origins are conducted world wide. A key goal of microarray experiments is to extract the fundamental patterns of gene expression inherent in the data. In this investigation, the widely studied all yeast microarray data [2], [16], having 6221 genes and 80 time points, is used for microarray gene expression analysis. We use centered Pearson correlation for extracting gene expression similarity as mentioned in the previous section.

3) KEGG Pathway: The KEGG database [4] provides pathway information for genes involved in 221 pathways. This information can be used as a reference for functional reconstruction. For each of these 221 pathways, all the protein sequences except yeast proteins, are downloaded from Protein Information Resource (PIR) [3]. Profile vector for each protein in yeast is computed by comparing its sequence across 221 pathway databases, using BLAST [17]. 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. Hence, the similarity matrix has a highest similarity value of 1.

4) Protein Similarity Through Transitive Homology: Intuitively, one can assume that all the protein relations arising from direct protein similarity search is available in the literature and will not help in predicting functions for unclassified genes in a widely studied organism like yeast. On the contrary, the transitive homology detection method [18] 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. For example [18], consider the transitive homology between sequence a and sequence b through the third sequence c. The *E*-values between sequence a and sequence c, sequence c and sequence b, as well as sequence a and sequence b are 0.01, 0.005, and 20 respectively. The protein similarities $B_{a,c}$, $B_{c,b}$, and $B_{a,b}$ are 0.8, 0.9, and 0.2 respectively. The similarity between sequence a and sequence b cannot be detected with their direct *E*-value. However, the value of $B_{a,b}$ is assigned to $0.8 \times 0.9 = 0.72$ because of the transitive sequence similarity. In this investigation, homology comparisons are performed among target proteins and 3 766 477 proteins downloaded from PIR [3], by using BLASTP in BLAST [17].

5) Protein–Protein Interaction: Protein interactions assemble the molecular machines of the cell and represent the dynamics of virtually all cellular responses [19]. While, genetic interactions reveal functional relationships between and within regulatory modules, protein–protein maps reveal many aspects of the complex regulatory network underlying cellular function [8], [20]. Large datasets of protein and genetic interactions in a consistent, well-annotated format is essential for interrogation of gene function. In this regard, manually curated catalogues of 238 313 protein–protein interactions are downloaded from BioGRID [5] on February 2011 and binary interactions (1 or 0) are used for indicating an interaction is present or not.

B. Computing PPV From Different Data Sources

The five datasets, used in this investigation, are accompanied by their own way of extracting gene similarities (such as Pearson Correlation for phenotypic profile) and they have different observational scales. So, a unified observational scale is needed to integrate these datasets. This can be achieved by rescoring the similarities in the common scale of PPV and using yeast GO-Slim process annotations of genes in the SGD database [6]. As mentioned earlier, according to yeast GO-Slim process and MIPS, there are 6069 and 6130 annotated genes (ORFs) for yeast of which 4387 and 4737 genes, respectively, are classified into some biological or functional process and the remaining genes are unclassified. The PPV at a particular similarity value (computed from a data source), using gene annotations, is defined as [12]

$$\text{PPV} = \frac{\text{number of gene pairs sharing common annotations}}{\text{total number of gene pairs}} \tag{1}$$

where gene pairs sharing common annotations are pairs of genes having an overlapping yeast GO-Slim process annotation and the total no. of gene pairs is the available gene pairs at a particular similarity value for a particular data source. Fig. 1 compares the similarity values obtained from different data sources in terms

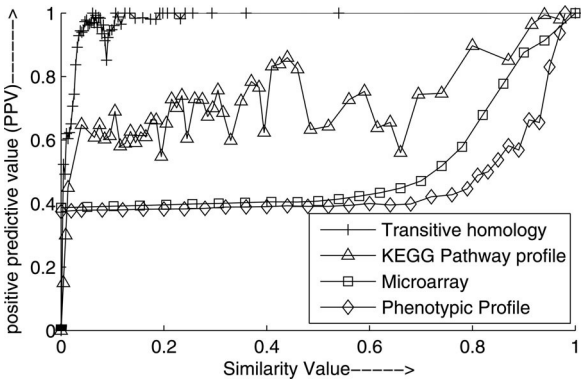

Fig. 1. Comparison among the data sources in terms of PPV versus the similarity values.

TABLE I
DISTANCE CORRELATIONS AMONG DATA SOURCES

Source	PhP	GE	KP	TH	PPI
PhP	1.00	0.09	0.11	0.10	0.00
GE	0.09	1.00	0.03	0.07	0.00
KP	0.11	0.03	1.00	0.19	0.00
TH	0.10	0.07	0.19	1.00	0.00
PPI	0.00	0.00	0.00	0.00	1.00

of their PPV. The protein–protein interactions (not shown in Fig. 1) are binary relations and have a constant PPV of 0.69 at a similarity value of 1.

C. Evaluation for Statistical Dependence Among Data Sources

A multisource integration framework is meaningful, when these sources are independent of each other. In this regard, we checked the statistical dependence of every data source with respect to other data sources, using distance correlation. Distance correlation is a measure of statistical dependence between two random variables or two random vectors of arbitrary, not necessarily equal dimension. Its important property is that this measure of dependence is zero if and only if the random variables are statistically independent. Its maximum value is 1, indicating an absolute dependence. The distance correlation between two variables, *X* and *Y*, is defined as

$$d\text{Cor}(X,Y) = \frac{d\text{Cov}(X,Y)}{\sqrt{d\text{Var}(X)d\text{Var}(Y)}} \tag{2}$$

where dCov(X,Y) is the distance covariance between X and Y and dVar(X) and dVar(Y) are the distance variances of X and Y, respectively.

Table I summarizes the results obtained by calculating the distance correlations among different data sources. Here, PhP, GE, KP, TH, and PPI indicates phenotypic profile, gene expression, KEGG pathway, transitive homology, and protein–protein interaction, respectively. From the results, it is observed that the off diagonal elements are close to 0.1 or 0.0, and indicate that the dependence is negligible among data sources.

D. New Framework for Data Source Integration

The scoring of different data sources, based on the unified observational scale of PPV's (see Section II-B), allows us to directly compare and integrate the different types of datsets. The PPV reflects the accuracy of datasets, but do not provide any information about relative power and weight estimate of one data source in presence of the other data sources. In this regard, a new score, where PPVs 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 and power estimates a_p, b_p, c_p, d_p, and e_p, respectively. This score is referred to as WPBS and is defined as

$$WPBS_{x,y} = \frac{a.P_{x,y}^{a_p} + b.M_{x,y}^{b_p} + c.K_{x,y}^{c_p} + d.B_{x,y}^{d_p} + e.I_{x,y}^{e_p}}{a+b+c+d+e} \tag{3}$$

where, a, b, c, d, e, a_p, b_p, c_p, d_p, and e_p, are varied within range 0 to α in steps of 1 to find a combination that maximizes the PPV of WPBS for a user defined number of top gene pairs. For each set of values of power and weighting coefficients, the top gene pairs are identified with a gold standard cutoff value, determined from KEGG pathway profiles, which provide 25 110 gene pairs with similarity value 1 (so sorting gene pairs, corresponding to their similarity value, is not possible and top 25 110 pairs are treated as a single group) and constant PPV of 0.87, using yeast GO slim process annotations. The final values of a, b, c, d, e, a_p, b_p, c_p, d_p, and e_p, are found to be 1.3, 3.7, 1.0, 30.1, 1.2, 8.1, 1.4, 3.5, 7 and 4, respectively. The following can be stated about the score:

1) $0 \leq \text{WPBS}_{x,y} \leq 1$
2) $\text{WPBS}_{x,y} = \text{WPBS}_{y,x}$ (symmetric).

In WPBS, the rescored similarity spaces, available from different data sources, are adaptively transformed using a set of power and weight coefficients. Intuitively, more important similarity spaces should be assigned larger weights and smaller powers (as PPVs are less than 1) than less important ones, while irrelevant ones should be assigned zero weight. Some typical instances in the exhaustive search process, showing the variation of PPV of WPBS for different sets of power and weight coefficients (assigned to PPVs of different data sources), ranging from 0 to 100, are shown in Fig. 2.

III. RESULTS

We now use MIPS October 2009 classification to evaluate the performance of WPBS as yeast GO-Slim process was used for determining the weights and power coefficients. First, we compare the gene pairs, obtained from various integration methods and data sources, and then we predict the functions of unclassified genes from clustering solutions. Evaluation of gene pairs, based on independent training and test set, is also performed.

A. Comparative Study

Here, we compare the PPV of gene pairs identified by the proposed framework, WPBS, with those identified by our previous

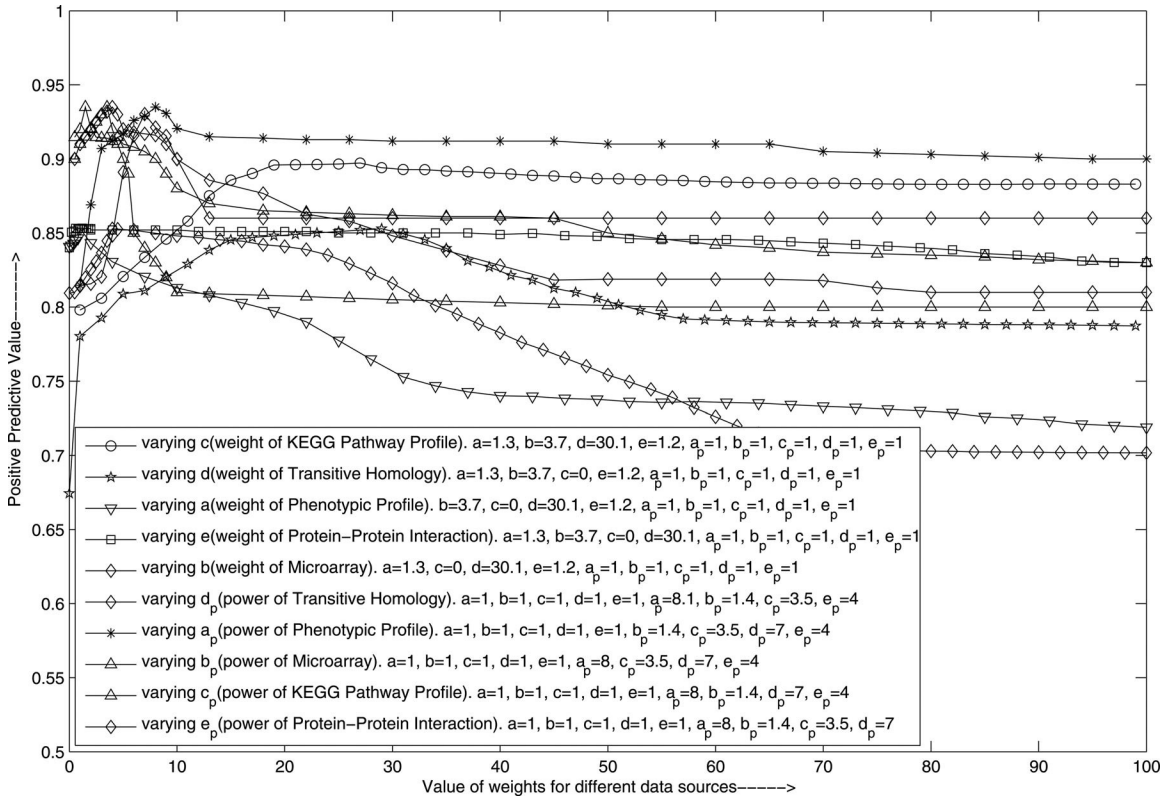

Fig. 2. Comparing the values of PPV using WPBS, by varying weight and power coefficients of PPV of different data sources for top 25 110 gene pairs. When a particular coefficient is varied the others are kept constant at the values shown in the figure.

"biological score" in Ray coworker, 2009, [15], Lee *et al.*'s probabilistic network (top 34 000 gene pairs downloaded from the website [21]), Lee *et al.*'s probabilistic network using the same datasources as in WPBS, and individual data sources, using MIPS annotations (see Fig. 3). We sorted the similarity values obtained from WPBS, other scores and datasources, separately, in descending order, and drew a curve for top gene pairs verses PPV from the sorted data for each form of data source. The PPVs for protein–protein interactions has a constant value of 0.69 and not shown in Fig. 3. It is evident from Fig. 3 that the curve of WPBS is above the other curves, achieved higher PPV, better than related methods and data sources, and the top 25 110 gene pairs obtained from WPBS has a PPV greater than the gold standard KEGG pathway profiles. Moreover, the performance of WPBS is also found superior to the other scoring frameworks for top 100 000 gene pairs that can be used further for any gene network or gene function prediction. The top 100 000 gene pairs predicted by WPBS with PPV above 0.78 are available in http://www.isical.ac.in/∼shubhra/WPBS/toprelationwpbs.txt in tabular (tab delimited) form.

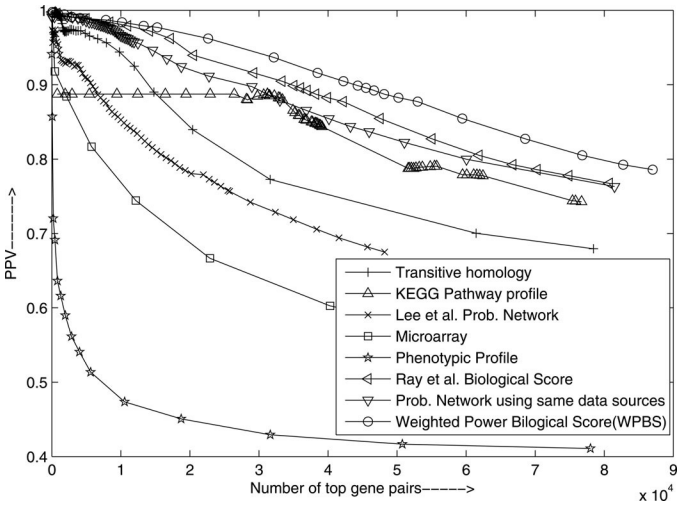

Fig. 3. Comparison between the WPBS, related methods, and individual data source in terms of PPV (using MIPS October 2009 annotations) versus the number of top gene pairs.

B. Gene Function Prediction Based on Clustering Results

In this investigation, genes are clustered by applying *k-medoids* algorithm on WPBS. The *k-medoids* clustering is one of the simplest partitive unsupervised algorithms that partitions the data into k clusters. The main idea is to define k medoids, one for each cluster, and chosen from the datapoints. The next step is to assign each data point to the nearest medoid. When all the points are so assigned then, k new medoids are recalculated from the data points of each cluster by minimizing the squared error, involving the distance between points, labeled to be in a cluster, and a point designated as the medoid of that cluster. The whole process is repeated until no more changes are observed in the locations of all the k medoids. For simplicity, the value of k in this investigation, is chosen as 510 as there are 510 functional categories in MIPS and the results for one of those instances,

TABLE II
TOP 39 FUNCTION PREDICTIONS OF UNCLASSIFIED GENE (ACCORDING TO MIPS AND SGD 2009 CLASSIFICATION) AT *P*-VALUE LESS THAN 10^{-16}

Unclassified Gene in MIPS & SGD	Predicted functional category	*P*-value	Function related genes within cluster	Function related genes within genome
YLR108C	ENERGY	1.8408e-18	16	348
YLR454W	ABC transporters	3.8091e-18	11	28
YDR493W YKR074W YMR098C	ribosomal proteins	1.8748e-17	34	221
YMR178W YML011C YPL168W	biosynthesis of vitamins, cofactors, and prosthetic groups	5.7559e-17	26	109
YDL156W YBR151W YIR016W	protein/peptide degradation	1.9301e-17	14	251
YIL041W	CELLULAR TRANSPORT	1.3967e-17	28	981
YMR251W YPR127W YIL165C YBR053C YMR041C YPR098C	C-compound and carbohydrate metabolism	5.3975e-20	21	496
YGR146C YLR257W YOR086C YCR070W YOR385W YML037C	modification by phosphorylation, dephosphorylation, autophosphorylation	2.6629e-17	27	184
YER189W YFL065C YPR202W YHR219W YPR203W YHL049C YBL110C YBL112C YLR464W YEL075C YEL076C	DNA topology	9.0061e-17	14	52
YDL144C	ribosome biogenesis	2.2567e-17	26	284
YBR262C YCL049C	transported compounds	7.3302e-17	27	555
YGR235C	protein/peptide degradation	7.7267e-17	17	251

having the highest PPV, are reported. To predict a genes function from the other genes in its cluster we use the following steps.

S1) 192 clusters are identified with functional enrichment in one or more categories by selecting *P*-value less than 10^{-5}, using 510 different MIPS functional categories, downloaded on October 2009. Clusters with *P*-values greater than 10^{-5} are not reported.

S2) From these functionally enriched clusters we predict the functions of 2877 classified genes and 334 unclassified genes by assigning the function related with the smallest P-value.

S3) Finally, we use the MIPS November 2011 classification to check if classification is now available for the genes, unclassified in MIPS October 2009 classification but now functional predictions are made using our method.

We divided our predictions in three sets, based on P-value cutoff, which one can modify according to the desired level of statistical significance of the predictions. While, the first set contains functional predictions of 39 unclassified and 492 classified genes, predicted with 0.91 PPV using a *P*-value cutoff 10^{-16}, the second set contains functional predictions of 99 unclassified and 1088 classified genes, predicted with 0.85 PPV using a *P*-value cutoff 10^{-13}. The third set contains functional predictions of 334 unclassified and 2877 classified genes, predicted with 0.75 PPV using a *P*-value cutoff 10^{-5}. These results can also be regenerated by using our MATLAB code, available at http://www.isical.ac.in/∼shubhra/WPBS/WPBS.html.

Table II summarizes the predicted functions for 39 unclassified genes from 12 clusters in the first set. For each of the predicted functions, the related *p*-values, number of function-related genes in the cluster and the genome, is also shown in the table. The related genes within each cluster and the PPV values (between target gene and other cluster members) arising from various data sources are available in tabular form (tab delimited file) at http://www.isical.ac.in/∼shubhra/WPBS/unclassified1.xls. Tables with similar format, containing the predicted functions of unclassified yeast genes in set 2 and 3, are available at http://www.isical.ac.in/∼shubhra/WPBS/unclassified2.xls and http://www.isical.ac.in/∼shubhra/WPBS/unclassified3.xls, respectively. The predicted functions for classified genes in set 1, 2, and 3 are available at http://www.isical.ac.in/∼shubhra/WPBS/classified1.xls, http://www.isical.ac.in/∼shubhra/WPBS/classified2.xls, and http://www.isical.ac.in/∼shubhra/WPBS/classified3.xls, respectively.

From our prediction results for unclassified genes in Table II, we find that the gene YIL041W shows functional enrichment in MIPS category "CELLULAR TRANSPORT." Our examination revealed that the protein, coded by this gene, is a BAR domain-containing protein that localizes to both, early and late golgi vesicles, and required for vacuole biogenesis and fluid-phase endocytosis [6]. Now, endocytosis is one of the processes, required for cellular transport, by which cells absorb molecules by engulfing them. Therefore, involvement of YIL041W in cellular transport is a likely one.

The gene YHR219W has 28 cluster members and 14 out of 28 genes show functional enrichment in MIPS category "DNA topology." Our analysis predicts that YHR219W's involvement in "DNA topology" is likely due to its relation to helicase proteins [6] which, it encodes within the telomeric Y' element. These proteins play important roles in various cellular processes including DNA replication, DNA repair, RNA processing, chromosomal segregation, and maintenance of chromosome stability. It has been well known that the amino acid sequences of

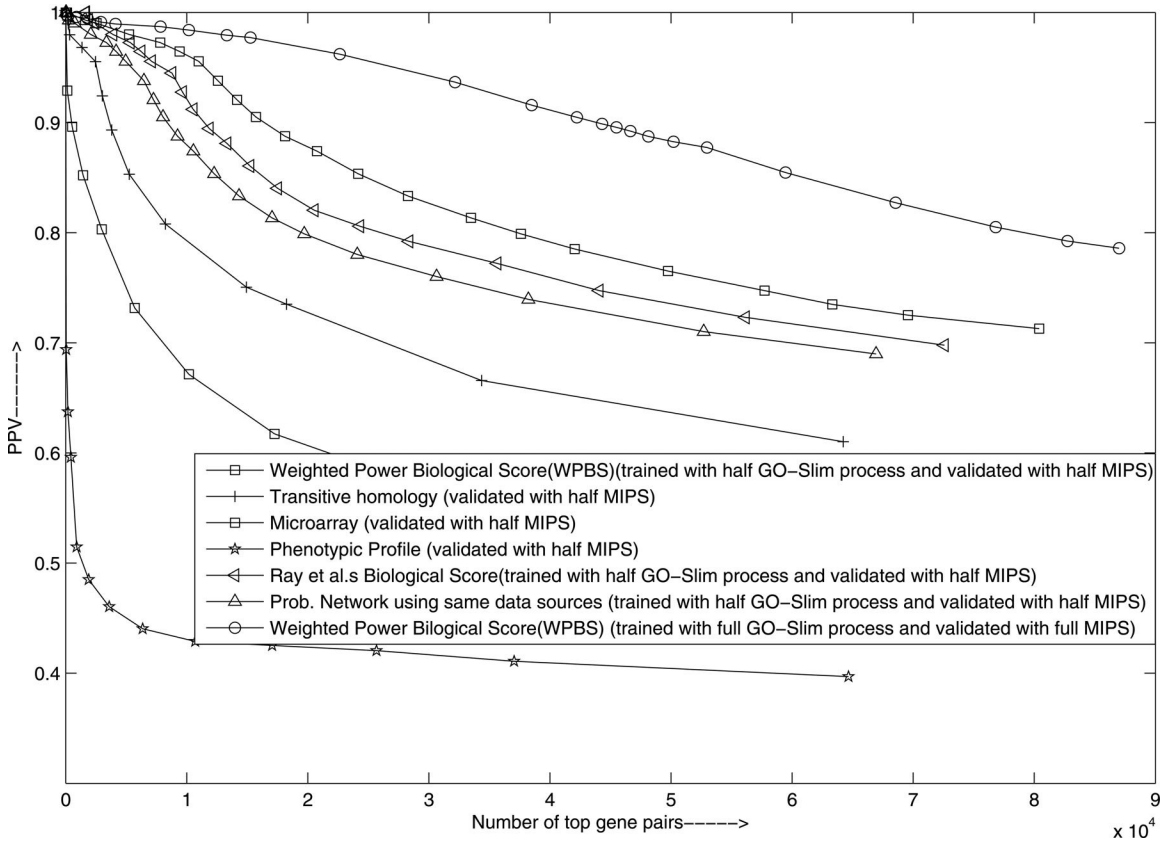

Fig. 4. Comparison between the WPBS and related methods in terms of PPV versus the number of top gene pairs using cross-validation procedure. To compare the performance of cross-validation results with the proposed score (WPBS), the top curve is provided. The top curves is also presented in Fig. 3, and is provided here for convenience. For the 2nd, 3rd, and 4th curve in the present figure, half of the genes with yeast GO-Slim process annotations are used to determine the power and weight coefficients, and the other half of the genes with MIPS annotations are used to evaluate the gene pairs. The curves 5th, 6th, and 7th, show the performance of some data sources when half of the genes with MIPS annotations are used to evaluate them.

these proteins contain several conserved motifs, and that the open reading frames (ORFs) that encode helicase-related proteins make up several gene families [22].

The gene YPR127W, along with its cluster members shows functional enrichment in MIPS category "C-compound and carbohydrate metabolism." According to SGD, this gene is differentially expressed during alcoholic fermentation, a biological process in which sugars (C-compounds), such as glucose, fructose, and sucrose, are converted into cellular energy and thereby, produce alcohol and carbon dioxide as metabolic waste products. Hence, our prediction for category "C-compound and carbohydrate metabolism" is a highly possible one.

Out of 39 unclassified genes, YCR070W is now renamed as YCR069W, and classified within category "protein fate" in MIPS. Our functional prediction, "modification by phosphorylation, dephosphorylation, autophosphorylation", is a subcategory of the function "protein fate" in MIPS. In SGD, YCR070W has been deleted and and merged with YCR069W.

C. Evaluation Based on Independent Training and Test Sets

For evaluating the results further, we randomly separated the set of 6072 genes into two disjoint training and test subsets of 3036 genes each. While, the training subset of 3036 genes is chosen randomly from yeast GO slim process annotations, the the test subset of 3036 genes is chosen from MIPS annotations. The training set is then used to determine the weight and power coefficients and the independent test set is used to evaluate the gene pairs. The values of the coefficients are observed to be close to the values mentioned in Section II-D, which, are obtained using the full set (6072) of genes. The process is repeated 10 times and the cross-validation results are reported. Fig. 4 shows curves for one of the cross-validation procedure and the curves demonstrate that WPBS performs better than Lee *et al.*'s probabilistic network, our previous integration method and individual data sources.

IV. Conclusion

In this investigation, we proposed a weighted power method, where, the weight and power coefficients for different data sources are determined in a systematic and adaptive manner using functional annotations of classified genes, available from yeast GO-Slim process annotations. Functional categories of 39, 99, and 334 unclassified yeast *Saccharomyces cerevisiae* genes are also predicted with 0.91, 0.85, and 0.75 PPV, respectively.

We also want to mention that, function prediction of some dubious ORFs and pseudogenes is a limitation of our approach and all other computational approaches, which involve

methods, based on similarity searching. In this regard, we computationally filtered out all pseudogenes and dubious ORFs from our prediction results by preparing a list of them from the file orf_coding_all.fasta, available at ftp://ftp.yeastgenome.org/sequence/S288C_reference/orf_dna/orf_coding_all.fasta.gz. Hence, our computational method will never be able to pick a pseudogene or dubious ORF as a candidate for function prediction. It is also evident from the experimental results that WPBS can be used for efficient gene function prediction.

ACKNOWLEDGMENT

The authors would like to aknowledge the support of the Department of Science and Technology, Government of India to the Center for Soft Computing Research. Authors would also like to thank the anonymous reviewers for their suggestions in improving the quality of research. The part of this work was completed by S. K. Pal as a J. C. Bose fellow of the Government 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," *Mol. Syst. Biol.*, 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] W. C. Barker *et al.*, "The Protein Information Resource (PIR)," *Nucl. Acids Res.*, vol. 28, no. 1, pp. 41–44, 2000.

[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," *Nucl. Acids Res.*, vol. 34, pp. D354–D357, 2006.

[5] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: A general repository for interaction datasets," *Nucl. Acids Res.*, vol. 34, pp. D535–D539, 2006.

[6] 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)," *Nucl. Acids Res.*, vol. 30, no. 1, pp. 69–72, 2002.

[7] Munich information for protein sequences. (2012). [Online]. Available: http://www.mips.com

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proc. Natl Acad. Sci. USA*, vol. 100, no. 21, pp. 12123–12128, 2003.

[14] 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.

[15] S. Bandyopadhyay S. S. Ray and S. K. Pal, "Combining multi-source information through functional annotation based weighting: Gene function prediction in yeast," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 2, pp. 229–236, Feb. 2009.

[16] Website, (2012). [Online]. Available: http://rana.lbl.gov/EisenData.htm

[17] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucl. Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[18] 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 Bioinformat.*, vol. 6, pp. 1–13, 2005.

[19] T. Pawson and P. Nash, "Assembly of cell regulatory systems through protein interaction domains," *Science*, vol. 300, pp. 445–452, 2003.

[20] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, pp. C47–C52, 1999.

[21] I. Lee, R. Narayanaswamy, and E. M. Marcotte, "Bioinformatic prediction of yeast gene function," in *Yeast Gene Analysis*, Amsterdam: Elsevier Press, 2006.

[22] 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.

Authors' photographs and biographies not available at the time of publication.