Function Prediction for Hypothetical Proteins in Yeast *Saccharomyces cerevisiae* Using Multiple Sources of High-Throughput Data

Trupti Joshi  
Computer Science Department, University of Missouri-Columbia, Columbia, MO, USA.  
Yu Chen  
Computer Science Department, University of Missouri-Columbia, Columbia, MO, USA; UT-ORNL Graduate School of Genome Science and Technology, Oak Ridge, TN, USA.  
Jeffrey Becker  
UT-ORNL Graduate School of Genome Science and Technology, Oak Ridge, TN, USA; Department of Microbiology, University of Tennessee, Knoxville, TN, USA.  
Nickolai Alexandrov  
Ceres Inc., Malibu, CA, USA and  
Dong Xu  
Computer Science Department, University of Missouri-Columbia, Columbia, MO, USA; UT-ORNL Graduate School of Genome Science and Technology, Oak Ridge, TN, USA.

**ABSTRACT**

Characterizing gene function is one of the major challenging tasks in the post-genomic era. To address this challenge, we have developed a new integrated probabilistic method for cellular function prediction by combining information from protein-protein interactions, protein complexes, gene expression profiles and annotations of known proteins through an integrative statistical model. Our approach is based on a novel assessment for the relationship between (1) the interaction/correlation of two proteins’ high-throughput data and (2) their functional relationship in terms of their Gene Ontology (GO) hierarchy. We have developed a Web server (http://digbio.missouri.edu/genefas) for the predictions. We have applied our method to yeast *Saccharomyces cerevisiae* and predicted functions for 1548 out of 2472 unannotated proteins.

Keywords: function prediction, protein-protein interactions, microarray data, yeast, high-throughput data, *Saccharomyces cerevisiae*.

1. INTRODUCTION

Determination of protein function is one of the most challenging problems in the post-genomic era. The traditional wet laboratory experiments for this purpose are accurate, but the process is time-consuming and costly. Out of 6343 genes in yeast *Saccharomyces cerevisiae* (Baker’s yeast), only about 3866 genes have been annotated. Despite all the efforts, only 50-60 percent of genes have been annotated in most organisms. This leaves bioinformatics with the opportunity and challenge of predicting functions of unannotated proteins by developing efficient and automated methods.

Several approaches have been developed for predicting protein function. The classical way to infer function is based on sequence similarity using programs such as FASTA [1] and PSIBLAST [2]. Another method to predict function is based on sequence fusion information, i.e., the Rosetta-Stone approach [3]. Function can also be inferred based on the phylogenetic profiling of proteins in multiple genomes [4]. With ever-increasing flow of biological data generated by the high-throughput methods such as yeast two-hybrid systems [5], protein complexes identification by mass spectrometry [6][7], microarray gene expression profiles [8][9] and systematic synthetic lethal analysis [10][11], some computational approaches have been developed to use these data for gene function prediction. Cluster analysis of the gene-expression profiles is a common approach used to predict function based on the assumption that genes with similar functions are likely to be co-expressed [8][9][12]. Using protein-protein interaction data to assign function to novel proteins is another approach. Proteins often interact with one another in an interaction network to achieve a common objective. It is therefore possible to infer the functions of proteins based on the functions of their interaction partners. Schwikowski et al. [13] applied neighbor-counting method in predicting the function. They assigned function to an unknown protein based on the frequencies of its neighbors having certain functions. The method was improved by Hishigaki et al. [14], who used $\chi^2$ statistics. Both these approaches give equal significance to all the functions contributed by the neighbors of the protein. Other function prediction methods using high-throughput data include machine-learning and data-mining approaches [15] and Markov random fields [16][17]. Instead of searching for a simple consensus among the functions of the interacting partners, Deng et al. used the Bayesian approach to assign a probability for a hypothetical protein to have the annotated function. Another Bayesian approach for combining heterogeneous data in yeast for function assignment has been applied by Troyanskaya et al [18].

Although these methods have been developed for gene function prediction, we believe that the error in the high-throughput data has not been handled well and the rich information contained in high-throughput data has not been fully utilized given the complexity and the quality of high-throughput data [19]. Inherent in the high-throughput nature of the experimental techniques is heterogeneity in data quality. The data generated are noisy and incomplete, with many false positives and false negatives. For example, the yeast two-hybrid assays may not detect some protein-protein interactions due to post-translational modifications, while mass spectrometry may fail to identify some transient and weak interactions. In a microarray clustering analysis, the genes with similar functions may not be clustered together due to lack of similar expression profiles. Clearly, different types of high-throughput data indicate different aspects of the internal relationships between the same set of genes. Each type of high-throughput data has its strengths and weaknesses in revealing certain relationships. Therefore, different types of
high-throughput data complement each other and offer more information than a single source. The combination of high-throughput data from various sources also provides a basis for cross-validating the data. While most current methods use a single source of high-throughput data for function prediction, it is evident that integrating various types of high-throughput data will help handle the data quality issue and better retrieve the underlying information from the data for function prediction. Although a few attempts have been made along the line, better statistical models can be developed to retrieve more information from the data.

In this paper, we propose a statistical model for functional annotation of the hypothetical proteins in *Saccharomyces cerevisiae* using high-throughput biological data including yeast two-hybrid, protein complexes, genetic interactions and microarray gene expression profiles. In our approach, we develop a statistical model, which better quantifies the relationship between functional similarity and high-throughput data similarity than existing methods, and improve the function predictions. We use the yeast *Saccharomyces cerevisiae* for our study, as it is a well-studied model organism for the eukaryotic systems with rich high-throughput data available. Our ultimate aim is to extend the prediction method to assign function to proteins in other organisms.

2. MATERIALS

2.1 Sources of Data

The *Saccharomyces cerevisiae* data were acquired from multiple sources as listed in Table 1. The protein-protein interaction data are of three types, i.e., physical binary interactions, genetic binary interactions and protein complex interactions. In the protein complexes it is unclear which proteins are in physical contact, although the protein complexes data is a rich resource. For simplicity, we assigned binary interactions between any two proteins participating in a complex. Thus in general, if there are *n* proteins in a protein complex, we add *n*(*n*-1)/2 binary interactions. The protein complexes data we use consists of 232 complexes, involving 1440 distinct proteins. These data, when converted to binary interactions, yield 49,313 binary interactions [20].

Table 1. Sources of data for *Saccharomyces cerevisiae*.

<table>
<thead>
<tr>
<th>Data</th>
<th>Types</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-Throughput Data</td>
<td>Protein Interactions from Yeast Two-Hybrid Screening (6516 binary pairs)</td>
<td>MIPS [22], Uetz et al. [23], Ito et al. [24]</td>
</tr>
<tr>
<td>Supportive Data</td>
<td>Mutant Phenotype, Protein Classes, Motif, EC Number, Subcellular Localization (for 6034 proteins)</td>
<td>MIPS (<a href="http://mips.gsf.de">http://mips.gsf.de</a>)</td>
</tr>
<tr>
<td>High-Throughput Data</td>
<td>Genetic Protein Interactions (1019 binary pairs)</td>
<td>MIPS (<a href="http://mips.gsf.de">http://mips.gsf.de</a>) (includes data from synthetic lethal screens, suppression and over-expression experiments [10][11])</td>
</tr>
<tr>
<td>High-Throughput Data</td>
<td>Protein Complexes (49,313 binary pairs)</td>
<td>Gavin et al.[6] and Ho et al. [7]</td>
</tr>
<tr>
<td>Microarray Data</td>
<td>Predication of Subcellular Localization</td>
<td>Yeast Protein Localization Server. (<a href="http://bioinfo.mbb.yale.edu/genom">http://bioinfo.mbb.yale.edu/genom</a> e/localize/)</td>
</tr>
</tbody>
</table>

The microarray gene expression data [25], used as log ratio of the expression profile against the reference state, includes 56 experimental conditions. If there was a missing data point in the expression profile, we substituted it with the average value of all the genes under that specific experimental condition, to maintain the dimension of the observations. We calculated Pearson and Spearman correlation coefficient for each gene pair in the microarray data. Having compared the two, Pearson correlation coefficient had a better predictive capacity and so we decided to use it for microarray data analysis. The Pearson correlation coefficient is defined as,

\[
r(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n\right)\left(\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n\right)}}
\]

where, \(X = \{x_1, x_2, \ldots, x_n\}\) and \(Y = \{y_1, y_2, \ldots, y_n\}\) are expression profiles of gene \(X\) and \(Y\) respectively, of \(n\) genes in total.

For function assignment, type of the functional annotation is of utmost importance. A particular gene product can be characterized with respect to its molecular function at the biochemical level (e.g. cyclase or kinase, whose annotation is often more related to sequence similarity and protein structure) or the biological process which it contributes to (e.g. pyrimidine metabolism or signal transduction that is often revealed in the high-throughput data of protein interaction and gene expression profiles). In our study, function annotation of protein is defined by GO (Gene Ontology) biological process (The Gene Ontology Consortium, 2000). It has a hierarchical structure with 9 classes at the top level that are subdivided into more specific classes at subsequent levels. Another functional classification is MIPS, which has a coarse hierarchical functional classification scheme, compared to GO. Having looked at both the functional classification systems, GO functional annotation appears to be a more systematic, detailed and robust classification in comparison to MIPS. Therefore, we used GO biological process classification, as of November 26, 2002 (ftp://ftp.geneontology.org/pub/go/ontology-archive), to assign function to unannotated proteins in our study. After acquiring the biological process functional annotation for the known proteins along with their GO ID, we generated a numerical GO INDEX, which represents the hierarchical structure of the classification. The deepest level of hierarchy is 13 (excluding the first level, which always begin with 1, representing biological process, to distinguish them from the other molecular function and cellular component categories in the GO annotation). The following shows an example of GO hierarchy:

1-4 cell growth and/or maintenance GO:0008151
1-4-3 cell cycle GO:0007049
1-4-3-2 DNA replication and chromosome cycle GO:000067
1-4-3-2-4 DNA replication GO:0006260
1-4-3-2-4-2 DNA dependent DNA replication GO:0006261
1-4-3-2-4-2-2 DNA ligation GO:0006266

An ORF (Open Reading Frame) can (and usually does) belong to multiple indices at various index levels in the hierarchy, as the proteins may be involved in more than one function in a cell.

2.2 Creation of Yeast Database

We have created a YEAST Database for centralized storage, easy retrieval and processing of all the data. The YEAST Database is created in the XML (http://www.w3.org/XML) format. XML allow us to define tags for the various attributes of the ORF and for easy expansion of the database to accommodate new data in the future, without major changes to the basic architecture. The information for each ORF is stored in separate files. Some of data in the files, such as sub-cellular localization, mutant phenotype and motifs have not been used for our current function prediction. However, we plan to integrate them in our future predictions.
3. METHOD

Our function prediction method consists of two steps. In the first step we identify the relationship between interacting proteins and functional similarities. We achieve this by estimating the a-priori probabilities for two genes to share a similar function for each type of high-throughput data. In the second step we utilize these estimated a-priori probabilities to predict the functions of unannotated proteins. Figure 1 shows the architecture of the method, implemented in GeneFAS.

For the microarray gene expression profiles, we define a pair of “interacting” genes if their Pearson correlation coefficient is greater than a threshold. We calculated percentage of such pairs sharing the same function for each INDEX level, to quantify the gene-function relationship between the correlated gene expression pairs. Results show a higher probability of sharing the same function for broad functional categories or highly correlated genes (Figure 3A). Clearly, there exists dependence between correlated genes in microarray data and similarity in function, as indicated in Figure 3B and C. The normalized ratio of microarray correlation pairs against the random pairs for sharing the same function shows the presence of information in highly correlated pairs in comparison to random pairs, which can be used in function prediction. Such information is compensated by the cases of anti-correlated gene expression profiles, whose gene pairs tends to have different functions comparing to random pairs, as indicated in the region with correlation coefficient less than –0.4 in Figure 3B. Since the information from anti-correlation is weak, we did not use it in our function prediction. Based on Figures 3, we decided to consider pairs with correlation coefficient ≥ 0.8 in predictions. We use these a-priori probabilities estimated from the data analysis in our function predictions.

3.1 Estimation of a-priori Probabilities

The a-priori probability (\(P_a\)) is the observed frequency based on the information available from high-throughput data about the functions of already annotated proteins. We estimated a-priori probabilities by comparing the pairs in high-throughput data, where both the genes have annotated functions, and by simultaneously comparing the level of similarity in functions that the two genes share in terms of the GO INDEX. For example, consider a physical binary interaction pair between ORF1 and ORF2, both of which have annotated functions. Assume ORF1 has a function represented by GO INDEX 1-4-3-2 and ORF2 has a function represented by GO INDEX 1-4-3-2. When compared with each other for the level of matching GO INDEX, they match with each other through 1-4-3 i.e., INDEX level 1 (1-4) and INDEX level 2 (1-4-3).

The results of the analysis of protein-protein interaction data are shown in Figure 2. The plots for physical, genetic and complexes protein-protein interactions data, show a drop in the percentage of pairs sharing the same function with an increase in the INDEX level, as seen in Figure 2A. It can be seen that, more pairs share less specific, broader functional categories as represented by lower index levels and fewer pairs share very specific functions as represented by higher index levels.

For the microarray gene expression profiles, Figure 3B and 3C show Normalized ratio against the percentage of pairs sharing the same function for random pairs for indices 1-6 and indices 7-13 respectively.

3.2 Prediction Using a-priori Probabilities

Our predictions are based on the idea of “guilt by association”, i.e., if an interaction partner of the studied hypothetical protein X has a known function, X may share the same function, with a probability governed by the high-throughput data relationship between X and its partner. Knowledge of the functional class of more interacting proteins can lead to a more accurate prediction of function. Each protein can belong to one or more functional classes, depending upon its interaction partners and their

![Figure 1. The architecture of GeneFAS.](image1)

![Figure 2. Results of analysis of yeast protein-protein interaction data. Figure 2A shows the percentage of pairs with protein interaction sharing the same levels of GO indices. Figure 2B. shows the normalized ratio for yeast protein interaction data.](image2)

![Figure 3. A. Percentage of pairs sharing the same levels of GO indices against Pearson correlation coefficient of microarray gene expression profiles. Figure 3B and 3C show Normalized ratio against the percentage of pairs sharing the same function for random pairs for indices 1-6 and indices 7-13 respectively](image3)
functions. We assign functions to the unannotated proteins on
the basis of common functions identified among the annotated
interaction partners and the estimated a-priori probabilities.

In our approach we identify the possible interactors for the
hypothetical protein in every high-throughput data type. We
compare the function for the hypothetical protein and each
interactor in terms of the GO INDEX. For example, if the
interactor for a hypothetical protein has GO INDEX 1-3-4-2, the
possible GO function INDICES for the hypothetical protein are
1-3, 1-3-4 and 1-3-4-2. For multiple interactors with the same
function, a higher confidence is attributed to the predicted
function. For example, if among the interactors for a
hypothetical protein, interactor 1 has GO INDEX 1-3-4-2, and
interactor 2 has GO INDEX 1-3-4-3, then the potential GO
function INDICES for the hypothetical protein are 1-3 and 1-3-
4 with a higher confidence, while 1-3-4-2 and 1-3-4-3 with a
lower confidence.

A Reliability Score is assigned to each potential GO INDEX
based on the a-priori probabilities from the analysis of high-
throughput data for each INDEX level from 1 to 13. For each
GO INDEX let the a-priori probability for the predicted protein
to share a function annotated for one of its interacting partner
for different high-throughput data be,

\[ P_i = \text{a-priori probability from genetic interactions}, \]
\[ P_2 = \text{a-priori probability from physical interactions}, \]
\[ P_3 = \text{a-priori probability from complex interactions}, \]
\[ P_4 = \text{a-priori probability from microarray}, \]

We assume the above four factors are independent for function
prediction. When the predicted protein has one, and only one
interacting partner with a given function \( F \) (corresponding to a
particular GO INDEX) for each type of high-throughput data, the
Reliability Score for the predicted protein having function \( F \)
is estimated as,

\[ \text{Reliability Score} = 1 - \frac{(1-P_1)(1-P_2)(1-P_3)(1-P_4)}{\sum_{i=1}^{K} P_i} \]

where, \((1-P_i)\) gives the probability of a protein not to share the
same function as its genetic interaction partner, and respectively
for all the other types of data. If no interacting partner with
function \( F \) is found for a specific type of high-throughput data,
the corresponding \((1-P_i)\) \((i=1,2,3,4)\) value is set to 1. Since \((1-
P_i)\) can be close to 0, for the sake of computational precision we
computed a natural logarithm so that the Reliability Score is
calculated as follows,

\[ \text{Reliability Score} = 1 - \exp \left\{ \log(1-P_1) + \log(1-P_2) + \log(1-P_3) + \log(1-P_4) \right\} \]

Multiple interactors with function \( F \) for each type of data are
also treated similarly, as above. All the interactors with function
\( F \) for a particular type of high-throughput data can be combined
so that the score contributions in Equation (3) for each type of
data are,

\[ \log(1-P_1) = \log(1-P_{1-2}) + \log(1-P_{1-3}) + \log(1-P_{1-4}) + \ldots \]
\[ \log(1-P_2) = \log(1-P_{2-3}) + \log(1-P_{2-4}) + \log(1-P_{2-5}) + \ldots \]
\[ \log(1-P_3) = \log(1-P_{3-4}) + \log(1-P_{3-5}) + \log(1-P_{3-6}) + \ldots \]
\[ \log(1-P_4) = \log(1-P_{4-5}) + \log(1-P_{4-6}) + \log(1-P_{4-7}) + \ldots \]

The final predictions are sorted based on the Reliability Score
for each predicted GO INDEX. Reliability Score is an empirical
scoring function and does not necessarily indicate the accuracy
or confidence in the predictions. Our next step was to evaluate
the performance of the method in terms of the scoring function.
Validation results allowed us to estimate the confidence in the
final predictions, given the scoring function.

3.3 Validation

For validation, we divided the 3866 annotated proteins with
known GO INDEX into two sets. The training set comprised of
randomly selected 3766 proteins and the testing set contains
remaining 100. All a-priori probabilities were calculated for the
training set of 3766 and the corresponding values were used in
testing. 10 such validations were performed, with different
testing and training sets and each time the re-calculated a-priori
probabilities were used in the predictions.

4. RESULTS

We performed the prediction on each of the 10 testing sets,
using the a-priori probabilities calculated for the corresponding
training set. The results were evaluated using sensitivity and
specificity, which are two important measures to evaluate the
performance of a bioinformatics prediction method. We
estimate the sensitivity to determine the success rate of the
method and specificity to assess the confidence in the
predictions of the method. For a given set of proteins \( K \), let \( n_i \)
be the number of the known functions for protein \( P_i \). Let \( m_i \) be
the number of functions predicted for the protein \( P_i \) by the method.
Let \( k_i \) be the number of predicted functions that are correct (the
same as the known function). Thus sensitivity (SN) and
specificity (SP) are defined as,

\[ \text{SN} = \frac{\sum_{i=1}^{K} k_i}{\sum_{i=1}^{K} n_i} \]
\[ \text{SP} = \frac{\sum_{i=1}^{K} k_i}{\sum_{i=1}^{K} m_i} \]

As seen in Figure 4 the sensitivity and specificity match for the
testing and training sets, indicating there is no significant
memory effect. Since there are many predictions with the
Reliability Score above 0.9, we calculated specificity with finer
interval, i.e., we used an interval of 5 for \(-\log(1-\text{Reliability}
\text{score})\) to the base 10 as shown in Figure 5. Figure 6 shows the
sensitivity versus specificity of the method with Reliability score
cutoff from 0.0 to 0.9. Above 0.9 cutoff for Reliability Score, the
specificity reaches as high as 95%. The specificity is a
confidence measure of a prediction, and it represents the
estimated chance to be correct for a given prediction, where the
Reliability Score does not reflect the prediction confidence.

Using our method, we have been able to assign function to 1548
out of the 2472 unannotated proteins in yeast. The number of
hypothetical genes with function predictions with respect to the
specificity and Index levels can be found in Table 2. Table 3
shows the function predictions for 21 hypothetical genes for
INDEX 5 with prediction specificity more than 0.9. The
assigned GO molecular functions for some of these hypothetical
proteins, whose information was unused in our prediction,
support our function predictions. Table 4 illustrates some
examples, which emphasize how multiple sources of data help
in function prediction in comparison to using only one source of
data. The predicted functions for hypothetical proteins have
higher confidence, when multiple sources of high-throughput
data support the predictions.

5. WEB-INTERFACE

GeneFAS has a web-interface for functional annotation of yeast
genes using multiple sources of high-throughput data. All the
predictions and the Web server can be accessed at
http://digbio.missouri.edu/genefas. The predictions can be
searched either by complete or partial matching yeast ORF /
gene name or a protein sequence from any other organism in
raw or the FASTA format. The user can also select the type(s) of
high-throughput data to be used for the predictions. For a
given protein sequence the tool compares it against the database
of all yeast proteins using BLAST [2] and outputs a list of the
hits with significant sequence similarity from the yeast and
gives the user option to select a yeast protein to view its
prediction. This may give some idea about the function of the
query protein. The user must select the expectation value (E-
value), mutation matrix and the number of hits to be displayed.
Links are provided to the BLAST alignment for the query
sequence. Confidence estimates and links to GO hierarchy
Table 3. Few examples of yeast hypothetical genes with predicted function at INDEX level 5 with specificity > 0.9.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Predicted GO Index</th>
<th>1-RScore (\times 10^{-14})</th>
<th>Function Description</th>
<th>GO Molecular Function at SGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>YDR106C</td>
<td>1-4-6-2-4-2</td>
<td>0.1</td>
<td>ribosome biogenesis</td>
<td>Ribosomal large subunit biogenesis</td>
</tr>
<tr>
<td></td>
<td>1-4-12-24-11-3</td>
<td>0.3</td>
<td>RNA processing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1-4-12-24-12-2</td>
<td>1.5</td>
<td>transcription, DNA-dependent</td>
<td></td>
</tr>
<tr>
<td>YJR006W</td>
<td>1-4-6-2-4-2</td>
<td>0.2</td>
<td>ribosome biogenesis</td>
<td>RNA processing</td>
</tr>
<tr>
<td></td>
<td>1-4-12-24-12-2</td>
<td>0.6</td>
<td>transcription, DNA-dependent</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1-4-12-24-11-3</td>
<td>1.5</td>
<td>RNA processing</td>
<td></td>
</tr>
<tr>
<td>YNL175C</td>
<td>1-4-6-2-4-2</td>
<td>0.1</td>
<td>ribosome biogenesis</td>
<td>RNA binding</td>
</tr>
<tr>
<td></td>
<td>1-4-12-24-11-3</td>
<td>0.3</td>
<td>RNA processing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1-4-12-24-12-2</td>
<td>0.1</td>
<td>transcription, DNA-dependent</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1-4-12-6-22-2</td>
<td>1.9</td>
<td>protein biosynthesis</td>
<td></td>
</tr>
<tr>
<td>YOL028W</td>
<td>1-4-6-2-4-2</td>
<td>(10^{-16})</td>
<td>ribosome biogenesis</td>
<td>Ribosome assembly, ribosome nucleus</td>
</tr>
<tr>
<td></td>
<td>1-4-12-24-11-3</td>
<td>(10^{-16})</td>
<td>RNA processing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1-4-12-24-12-2</td>
<td>(10^{-16})</td>
<td>transcription, DNA-dependent</td>
<td></td>
</tr>
<tr>
<td>YPL122W</td>
<td>1-4-12-6-22-2</td>
<td>1.8</td>
<td>protein biosynthesis</td>
<td>ATP binding ABC transporter</td>
</tr>
</tbody>
</table>

Table 4. Examples of hypothetical genes with the high-throughput data used for their function prediction and highest prediction confidence (specificity) for Indices 1-3. The numbers in columns 2-5 indicate number of interacting partners with known functions for the hypothetical genes.

<table>
<thead>
<tr>
<th>Hypothetical Genes</th>
<th>Physical Interaction</th>
<th>Genetic Interaction</th>
<th>Complexes Interaction</th>
<th>Microarray</th>
<th>Prediction Confidence (specificity)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Index 1</td>
<td>Index 2</td>
<td>Index 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YGL245W</td>
<td>3</td>
<td>1</td>
<td>258</td>
<td>28</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YBL066C</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>-</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.583</td>
</tr>
<tr>
<td>YAL061W</td>
<td>-</td>
<td>-</td>
<td>1298</td>
<td>2</td>
<td>0.583</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.250</td>
</tr>
<tr>
<td>YFR006W</td>
<td>-</td>
<td>-</td>
<td>923</td>
<td>2</td>
<td>0.583</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.250</td>
</tr>
<tr>
<td>YAL036C</td>
<td>5</td>
<td>-</td>
<td>152</td>
<td>2</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.955</td>
</tr>
<tr>
<td>YKL123C</td>
<td>-</td>
<td>1</td>
<td>150</td>
<td>2</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.955</td>
</tr>
<tr>
<td>YAL049C</td>
<td>1</td>
<td>-</td>
<td>923</td>
<td>-</td>
<td>0.583</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.250</td>
</tr>
<tr>
<td>YCL058C</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.369</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.212</td>
</tr>
</tbody>
</table>

6. DISCUSSION

Systematic and automatic methods for predicting gene function using high-throughput data represent a major challenge in the post genomic era. To address this challenge, we developed a systematic method to assign function in an automated fashion using integrated computational analysis of yeast high-throughput data including yeast two-hybrid, genetic interaction, protein complexes and microarray data, together with the GO biological process functional annotation. In particular, this paper gives the first systematic study on the quantitatively relationship between the correlation of microarray gene expression profiles and the functional similarity. Such relationship provides a unique approach for function prediction. Our approach differs from pure computational methods (such as sequence comparison) to identify the relationship between a hypothetical protein and any protein with known function, since
our method is developed on the foundation of patterns and dependencies retrieved from the experimental data, thus giving higher confidence for the prediction. The integration of high-throughput data helps cross-validation and reduces the noise level for each type of data. Of course, considering the noisy nature of the high-throughput data, some predictions may not be correct and it is important to check the confidence levels for predictions. However, our predictions can provide biologists with hypotheses to study and design specific experiments to validate the predicted functions using tools such as mutagenesis. Such combination of computational methods and experiments may discover biological functions for hypothetical proteins much more efficiently than traditional methods. Our method can be applied to other species as well. We are currently applying this method to the Arabidopsis thaliana genome. We are also developing a more systematic Bayesian approach for assessing the probability of function prediction, predicting the functions of hypothetical proteins without direct interaction partners of known functions, and handling the dependence between the information for function prediction from different high-throughput data sources.

7. ACKNOWLEDGMENTS

This work has been supported by the Department of Energy’s “Genomes to Life” program under the project, “Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling” and a research contract with Ceres Inc., Malibu, CA. We would like to thank Drs. Victor Olman, Ying Xu, Guohui Lin, and Loren Hauser for helpful discussions.

8. REFERENCES