

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

Trupti Joshi<sup>1</sup>, Yu Chen<sup>1,2</sup>, Jeffrey M. Becker<sup>2,3</sup>, Nikolai Alexandrov<sup>4</sup>, Dong Xu<sup>1,2\*</sup>

<sup>1</sup> Digital Biology Laboratory, Computer Science Department, University of Missouri-Columbia, Columbia, MO, USA.

<sup>2</sup> UT-ORNL Graduate School of Genome Science and Technology, Oak Ridge, TN, USA.

<sup>3</sup> Departments of Microbiology, Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN, USA.

<sup>4</sup> Ceres Inc., Malibu, CA, USA.

\* To whom correspondence should be addressed

Address: Digital Biology Laboratory, Computer Science Department,  
201 Engineering Building West,  
University of Missouri, Columbia, MO, 65211

E-mail: xudong@missouri.edu

Fax: 573-882-8318

Phone: 573-882-7064

Running Title : Function Prediction Using High-Throughput Data

## **ABSTRACT**

**Motivation:** Characterizing gene function is one of the major challenging tasks in the post genomic era. To address this challenge, we have developed GeneFAS (Gene Function Annotation System) using an integrated probabilistic method for function prediction. Our method combines the information from protein-protein interactions, protein complexes, microarray gene-expression profiles and functional annotations for known proteins.

**Results:** We followed the GO hierarchical biological process functional annotation. Our approach differs from the others in a variety of ways. Firstly, we developed a more systematic statistical model for estimating the probability for a protein to have a certain function. Secondly, our approach allows confidence assessment, based on the statistical model. Finally, we have also developed a Web server and made the predictions freely available. We have applied our method to yeast *Saccharomyces cerevisiae* and predicted functions for 1548 out of 2472 unannotated proteins.

**Availability:** The program and documents are available at <http://digbio.missouri.edu/genefas>.

**Contact:** [xudong@missouri.edu](mailto:xudong@missouri.edu)

**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. As of November 2003, 126 bacterial, 16 archaeal and 9 eukaryotic genome sequences are complete, while the sequencing of 121 bacterial, 2 archaeal and 35 eukaryotic genomes is in progress ([http://www.ncbi.nlm.nih.gov/sutils/genom\\_table.cgi](http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi)). The traditional wet laboratory experiments can assign functions for the genes in these genomes accurately. However, 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 using the information derived from sequence similarity, phylogenetic profiles, protein-protein interactions, protein complexes and gene expression profiles. The classical way to infer function is based on sequence similarity using programs such as FASTA (Pearson *et al.* 1998) and PSI-BLAST (Altschul *et al.* 1997). Another method to predict function is based on sequence fusion information, i.e., the Rosetta-Stone approach (Marcotte *et al.* 1999). Function can also be inferred based on the phylogenetic profiling of proteins in multiple genomes (Pellegrini *et al.* 1999). With ever-increasing flow of biological data generated by the high-throughput methods such as yeast two-hybrid systems (Chien *et al.* 1991), protein complexes identification by mass spectrometry (Gavin *et al.* 2002; Ho *et al.* 2002), microarray gene expression profiles (Eisen *et al.* 1998; Brown *et al.* 2000) and systematic synthetic lethal analysis (Tong *et al.* 2001; Goehring *et al.* 2003), 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 (Eisen *et al.* 1998; Brown *et al.* 2000; Pavlidis *et al.* 2001). 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. (2000) 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. (2001), 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 (Clare *et al.* 2003) and Markov random fields (Deng *et al.* 2002; 2003). 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 (2003).

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 (Chen and Xu, 2003). 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 (yeast two-hybrid data), 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 (Chen *et al.* 2003). The microarray gene expression

data (Roberts *et al.* 2000), 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,

$$r(X, Y) = \frac{\sum x_i y_i - (\sum x_i \sum y_i) / n}{\sqrt{[(\sum x_i^2 - (\sum x_i)^2 / n)(\sum y_i^2 - (\sum y_i)^2 / n)]}} \quad (1)$$

where,  $X = \{x_1, x_2, \dots, x_n\}$  and  $Y = \{y_1, y_2, \dots, y_n\}$  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:0000067
                  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. The architecture of the function prediction method, which is implemented in GeneFAS, is represented in Figure 1.

### 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 1-4-3-3-4 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. Comparison of our results with similar analysis on random pairs as seen in Figure 2B, shows a normalized ratio of protein-protein interaction pairs against the random pairs for sharing the same index level. Since the value is highly biased above 1, there clearly exists a relationship between the protein-protein interaction data and similarity in function, which can be

utilized to make function predictions based on these data. Such relationship is stronger for more specific functions with higher index levels.

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  $\geq 0.8$  in predictions. We use these *a-priori* probabilities estimated from the data analysis in our function predictions.

### **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 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 (physical interactions, genetic interactions, protein complexes and microarray gene expression with correlation coefficient  $\geq 0.8$ ). 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,

$P_1 = a\text{-priori}$  probability from genetic interactions,

$P_2 = a\text{-priori}$  probability from physical interactions,

$P_3 = a\text{-priori}$  probability from complex interactions,

$P_4 = a\text{-priori}$  probability from microarray gene expression,

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,

$$\text{Reliability Score} = 1 - (1 - P_1)(1 - P_2)(1 - P_3)(1 - P_4) \quad (2)$$

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,

$$\mathbf{Reliability\ Score = 1 - exp [ ( Log(1-P_1) + Log(1-P_2) + Log(1-P_3) + Log(1-P_4) ) ] \quad (3)}$$

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,

$$\text{Log}(1-P_1) = \text{Log}(1-P_{11}) + \text{Log}(1-P_{12}) + \text{Log}(1-P_{13}) + \dots \quad (4)$$

$$\text{Log}(1-P_2) = \text{Log}(1-P_{21}) + \text{Log}(1-P_{22}) + \text{Log}(1-P_{23}) + \dots \quad (5)$$

$$\text{Log}(1-P_3) = \text{Log}(1-P_{31}) + \text{Log}(1-P_{32}) + \text{Log}(1-P_{33}) + \dots \quad (6)$$

$$\text{Log}(1-P_4) = \text{Log}(1-P_{41}) + \text{Log}(1-P_{42}) + \text{Log}(1-P_{43}) + \dots \quad (7)$$

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,

$$SN = \frac{\sum_1^K k_i}{\sum_1^K n_i} \quad (8) \quad SP = \frac{\sum_1^K k_i}{\sum_1^K m_i} \quad (9)$$

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  $-\text{Log}(1-\text{Reliability score})$  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 as well as a stand-alone command line tool 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 (Altschul *et al.* 1997) 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 (<http://www.godatabase.org/cgi-bin/go.cgi>), Saccharomyces Genome Database (Dwight *et al.* 2002) records and evidences used are provided for each prediction.

## 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, genetics interaction, protein complexes and gene expression 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 prediction 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.

#### **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. We also like to thank Rajkumar Bondugula for a critical reading of this paper.

## FIGURE LEGENDS

Figure 1. The architecture of GeneFAS.

Figure 2. Results of analysis of yeast protein-protein interaction data. The physical, genetic and complexes interaction data, were compared for similarity in function in terms of GO index levels they share. Figure 2A shows the percentage of pairs with protein interaction sharing the same levels of GO indices. To estimate the significance of the information in the protein-protein interaction data, we compared it with random pairs and computed ratio of the two. Figure 2B. shows the normalized ratio for yeast protein interaction data.

Figure 3. A. Percentage of pairs sharing the same levels of GO indices against Pearson correlation coefficient of microarray gene expression profiles. B and C. Normalized ratio for the percentage of gene pairs sharing the same levels of GO indices, against the percentage of pairs sharing the same function for random pairs versus Pearson correlation coefficient of microarray gene expression profiles. Figure 3B shows ratio for indices 1-6. Figure 3C shows for indices 7-13 for Pearson correlation coefficient between 0-1, where the values are close to 1 in the range  $-1$  to  $0$ .

Figure 4. *Sensitivity* and *specificity* for the predictions using the training and testing data sets versus the Reliability Score, with an interval of 0.1.

Figure 5. *Specificity* for predictions with Reliability Score above 0.9.

Figure 6. *Sensitivity* versus *specificity* of the method.

FIGURES

Fig 1

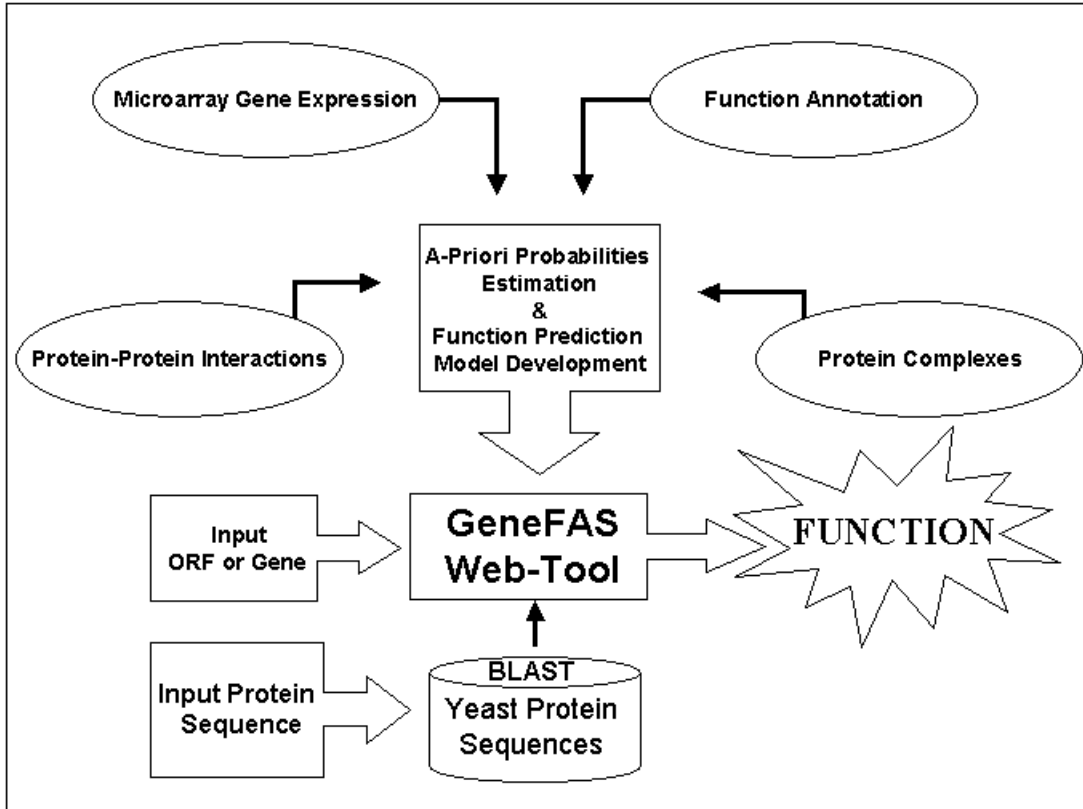


Fig 2.

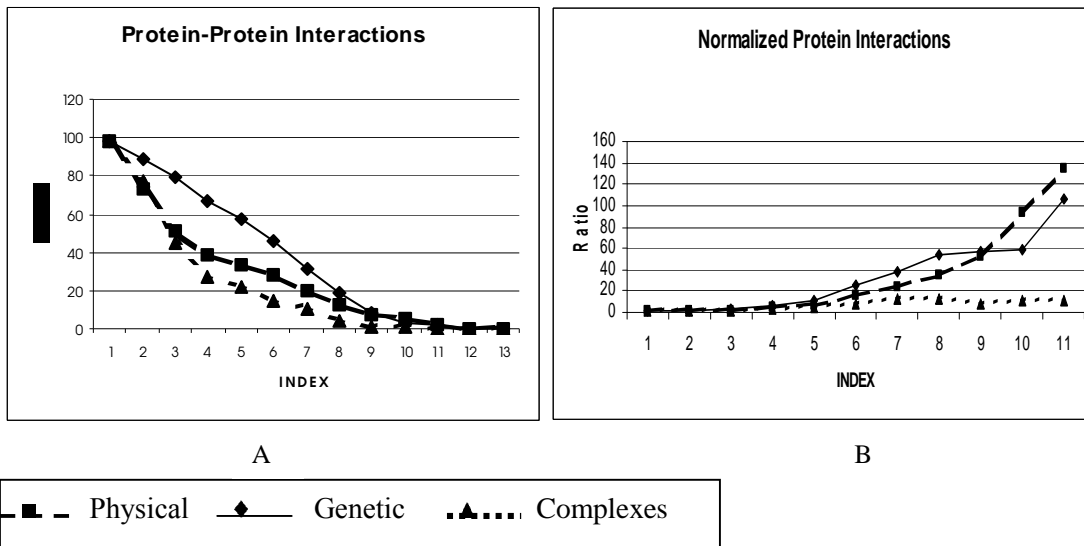
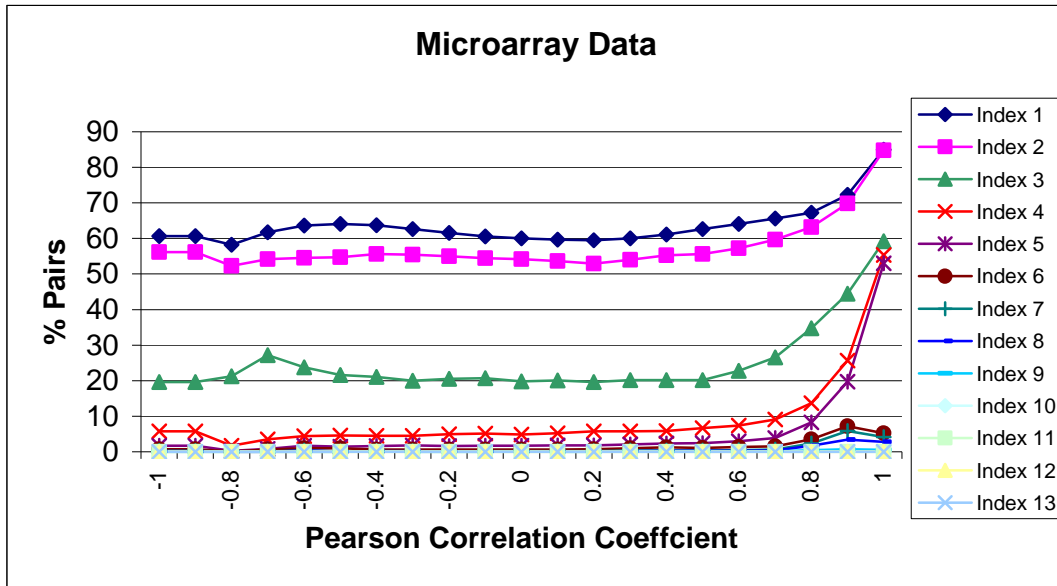
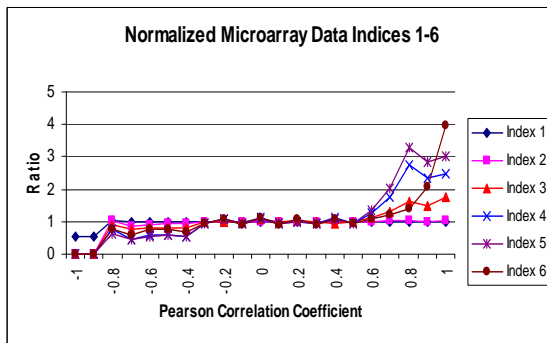


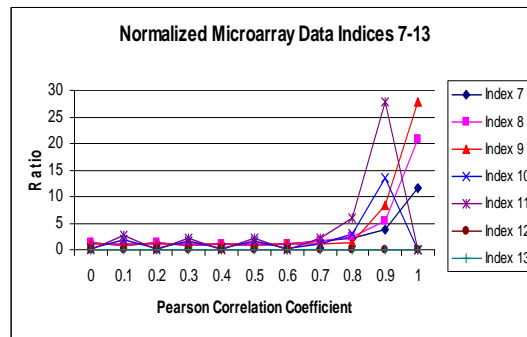
Fig 3



A



B



C

Fig 4.

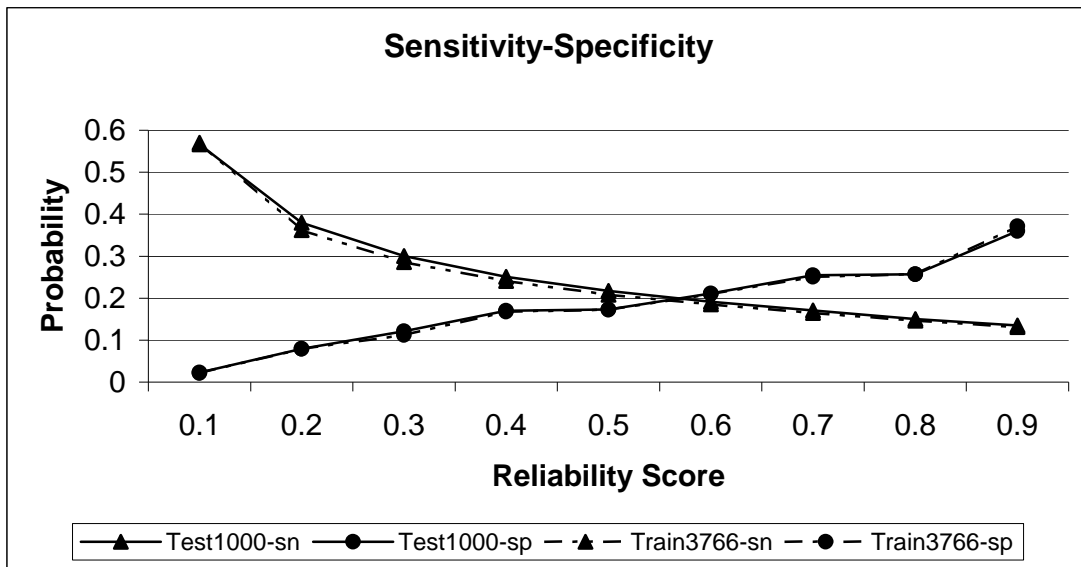


Fig 5.

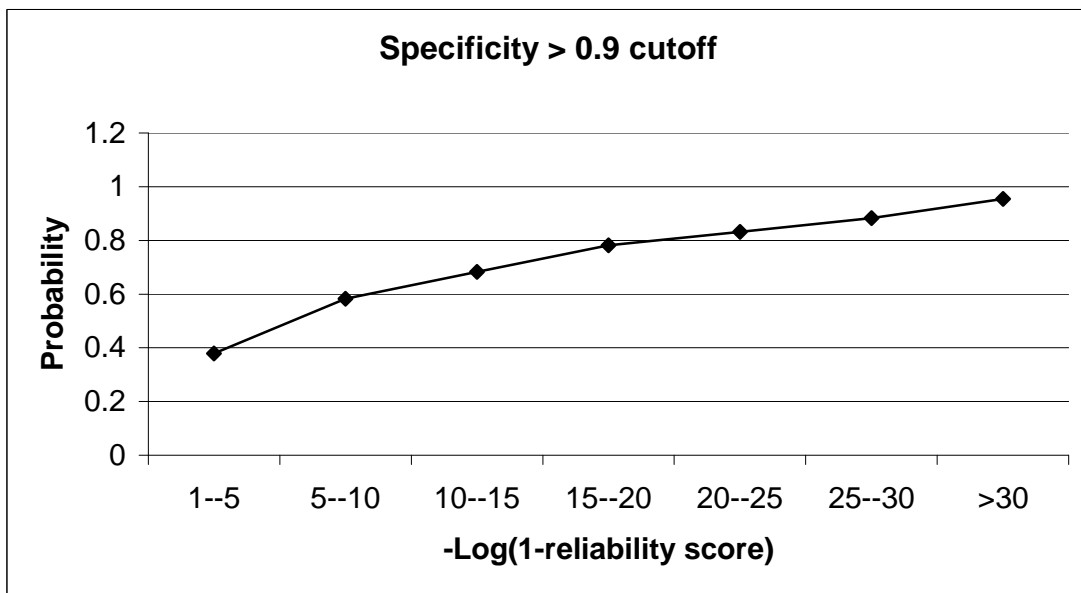
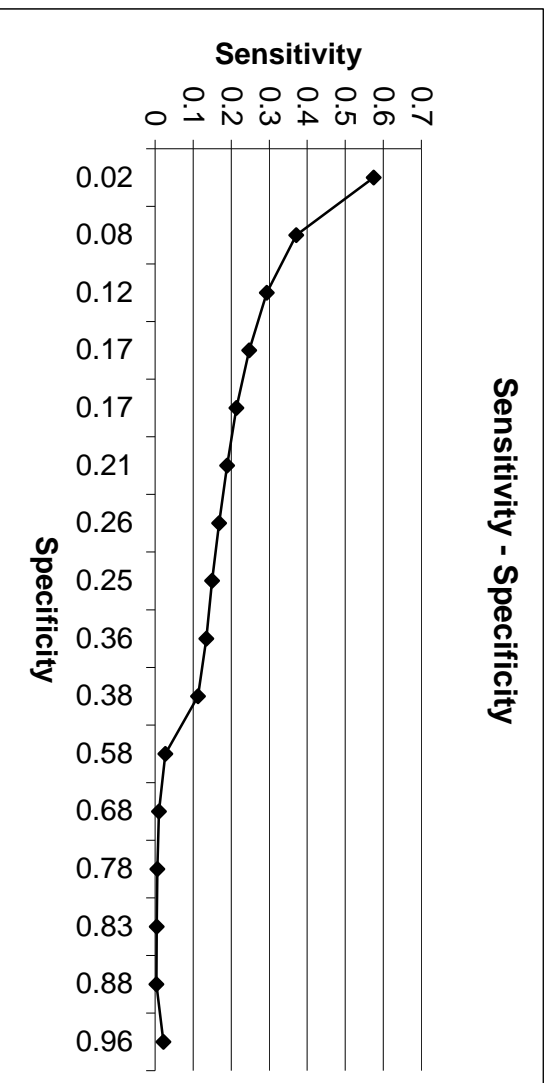


Fig 6.



## TABLES

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

Data	Types	Source
General Information	Gene names, ORF names, Functional annotation	The Gene Ontology Consortium (2000).
High-Throughput Data	Protein Interactions from Yeast Two-Hybrid Screening (6516 binary pairs)	MIPS (Mewes et al. 2002), Uetz et al. 2000, Ito et al. 2001
	Genetic Protein Interactions (1019 binary pairs)	MIPS ( <a href="http://mips.gsf.de">http://mips.gsf.de</a> ) (includes data from synthetic lethal screens, suppression and over-expression experiments. Tong et al. 2001 and Goehring et al. 2003)
	Protein Complexes (49,313 binary pairs)	Gavin <i>et al.</i> 2002 and Ho <i>et al.</i> 2002
	Microarray Data	Roberts et al. 2000, including 56 experiments conditions.
Supportive Data	Mutant Phenotype, Protein Classes, Motif, EC Number, Subcellular Localization (for 6034 proteins)	MIPS (CYGD) ( <a href="http://mips.gsf.de/genre/proj/yeast/index.jsp">http://mips.gsf.de/genre/proj/yeast/index.jsp</a> )
	Prediction of Subcellular Localization	Yeast Protein Localization Server. ( <a href="http://bioinfo.mbb.yale.edu/genome/localize/">http://bioinfo.mbb.yale.edu/genome/localize/</a> )

Table 2. Number of hypothetical genes with function predictions with respect to prediction confidence probabilities and index levels.

INDEX	Probability									
	□0.9	□0.8	□0.7	□0.6	□0.5	□0.4	□0.3	□0.2	□0.1	□0.0
<b>1</b>	385	468	549	669	947	947	1500	1548	1548	1548
<b>2</b>	211	272	322	397	548	548	892	1548	1548	1548
<b>3</b>	80	98	124	178	280	280	549	1477	1542	1542
<b>4</b>	33	55	67	83	144	144	328	622	1511	1532
<b>5</b>	21	35	54	72	110	110	243	504	1439	1486
<b>6</b>	0	0	0	6	28	28	110	253	1278	1439
<b>7</b>	0	0	0	0	10	10	61	133	326	1295
<b>8</b>	0	0	0	0	0	0	4	32	154	975
<b>9</b>	0	0	0	0	0	0	0	0	1	631
<b>10</b>	0	0	0	0	0	0	0	0	0	414
<b>11</b>	0	0	0	0	0	0	0	0	0	196
<b>12</b>	0	0	0	0	0	0	0	0	0	60

Table 3. List of yeast hypothetical genes for which GeneFAS can predict function at INDEX level 5 with specificity > 0.9.

Gene	Predicted GO Index	1-RScore	Function Description of Predicted GO Index	GO id	GO Molecular Function at SGD
YDR101C	1-4-6-2-4-2	$0.1 * 10^{-14}$	ribosome biogenesis	GO:0007046	Ribosomal large subunit biogenesis
	1-4-12-24-11-3	$0.3 * 10^{-14}$	RNA processing	GO:0006396, GO:0006394	
	1-4-12-24-12-2	$1.5 * 10^{-14}$	transcription, DNA-dependent	GO:0006351	
YDR417C	1-4-12-6-22-2	$10^{-30}$	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
YDR496C	1-4-6-2-4-2	$10^{-30}$	ribosome biogenesis	GO:0007046	
	1-4-12-24-11-3	$10^{-30}$	RNA processing	GO:0006396, GO:0006394	
	1-4-12-24-12-2	$10^{-30}$	transcription, DNA-dependent	GO:0006351	
YER006W	1-4-6-2-4-2	$0.2 * 10^{-14}$	ribosome biogenesis	GO:0007046	rRNAprocessing
	1-4-12-24-12-2	$0.6 * 10^{-14}$	transcription, DNA-dependent	GO:0006351	
	1-4-12-24-11-3	$1.5 * 10^{-14}$	RNA processing	GO:0006396, GO:0006394	
YGL102C	1-4-12-6-22-2	$10^{-30}$	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
YGR145W	1-4-6-2-4-2	$10^{-30}$	ribosome biogenesis	GO:0007046	
	1-4-12-24-12-2	$0.1 * 10^{-14}$	transcription, DNA-dependent	GO:0006351	
	1-4-12-24-11-3	$0.1 * 10^{-14}$	RNA processing	GO:0006396, GO:0006394	
YKL056C	1-4-12-6-22-2	$10^{-30}$	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
YLL044W	1-4-12-6-22-2	$0.2 * 10^{-14}$	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
YLR196W	1-4-6-2-4-2	$10^{-30}$	ribosome biogenesis	GO:0007046	
	1-4-12-24-11-3	$10^{-30}$	RNA processing	GO:0006396, GO:0006394	
	1-4-12-24-12-2	$10^{-30}$	transcription, DNA-dependent	GO:0006351	
YLR198C	1-4-12-6-22-2	$0.4 * 10^{-14}$	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
YMR290C	1-4-12-6-22-2	$10^{-30}$	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
	1-4-12-24-11-3	$10^{-30}$	RNA processing	GO:0006396, GO:0006394	
	1-4-12-24-12-2	$10^{-30}$	transcription, DNA-dependent	GO:0006351	
	1-4-6-2-4-2	$10^{-30}$	ribosome biogenesis	GO:0007046	
YNL119W	1-4-12-6-22-2	$7.5 * 10^{-14}$	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
YNL132W	1-4-6-2-4-2	$10^{-30}$	ribosome biogenesis	GO:0007046	
	1-4-12-24-11-3	$10^{-30}$	RNA processing	GO:0006396, GO:0006394	
	1-4-12-24-12-2	$10^{-30}$	transcription, DNA-dependent	GO:0006351	
YNL175C	1-4-6-2-4-2	$0.1 * 10^{-14}$	ribosome biogenesis	GO:0007046	RNA binding

	1-4-12-24-11-3	0.3 *10 <sup>-14</sup>	RNA processing	GO:0006396, GO:0006394	
	1-4-12-24-12-2	0.1 *10 <sup>-14</sup>	transcription, DNA- dependent	GO:0006351	
	1-4-12-6-22-2	1.9 *10 <sup>-14</sup>	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
YOR206W	1-4-6-2-4-2	10 <sup>-30</sup>	ribosome biogenesis	GO:0007046	Ribosome assembly, ribosome nucleus export
	1-4-12-24-11-3	10 <sup>-30</sup>	RNA processing	GO:0006396, GO:0006394	
	1-4-12-24-12-2	10 <sup>-30</sup>	transcription, DNA- dependent	GO:0006351	
YOR277C	1-4-12-6-22-2	10 <sup>-30</sup>	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
YOR309C	1-4-12-6-22-2	10 <sup>-30</sup>	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
YPL142C	1-4-12-6-22-2	10 <sup>-30</sup>	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
YPL226W	1-4-12-6-22-2	1.8 *10 <sup>-14</sup>	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	ATP binding ABC transporter
YPL238C	1-4-12-6-22-2	1.1 *10 <sup>-14</sup>	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	
YPR044C	1-4-12-6-22-2	10 <sup>-30</sup>	protein biosynthesis	GO:0006412, GO:0006416, GO:0006453	

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.

Hypothetical Genes	Physical Interactions	Genetic Interactions	Complexes (Binary Interactions)	Microarray	Prediction Confidence (specificity)		
					Index 1	Index 2	Index 3
YGL245W	3	1	258	28	0.955	0.955	0.955
YBL066C	2	1	9	-	0.955	0.583	0.360
YAL061W	-	-	1	9	0.583	0.369	0.250
YFR006W	-	-	33	-	0.955	0.955	0.583
YAL036C	5	-	16	152	0.955	0.955	0.955
YKL128C	-	1	-	-	0.369	0.360	-
YAL049C	1	-	4	-	0.583	0.250	0.212
YCL058C	1	-	-	-	0.369	0.250	0.212

## REFERENCES

- Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **25**: 3389 -- 3402.
- Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M., and Haussler, D. 2000. Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. *Proc. Natl. Acad. Sci. USA.* **97**: 262 -- 267.
- Chen, Y., Joshi, T., Xu, Y., and Xu, D. 2003. Towards Automated Derivation of Biological Pathways Using High-throughput Biological Data. In Proceedings of the IEEE Conference on Bioinformatics and Biotechnology, pp. 18--25. IEEE/CS Press.
- Chen, Y., and Xu, D. 2003. Computation Analysis of High-throughput Protein-protein Interaction Data. *Current Peptide and Protein Science.* **4**: 159 -- 181.
- Chien, C., Bartel, P., Sternglanz, R., and Fields, S. 1991. The Two-hybrid System: A Method to Identify and Clone Genes for Proteins that Interact with a Protein of Interest. *Proc. Natl. Acad. Sci. USA.* **88**: 9578 -- 9582.
- Clare, A., and King, R. D. 2003. Predicting Gene Function in *Saccharomyces cerevisiae*. ECCB 2003 (also published as a journal supplement in Bioinformatics. 19: ii42-ii49).
- Deng, M., Zhang, K., Mehta, S., Chen, T., and Sun, F. 2002. Prediction of Protein Function Using Protein-protein Interaction Data. In Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB2002), pp. 197 - 206. IEEE Computer Society, Los Alamitos, California.
- Deng, M., Chen, T., and Sun, F. 2003. Integrated Probabilistic Model for Functional Prediction of Proteins, RECOMB2003.
- Dwight, S., Harris, M., Dolinski, K., Ball, C., Binkley, G., Christie, K., Fisk, D., Issel-Tarver, L., Schroeder, M., Sherlock, G., et al. 2002. *Saccharomyces* Genome Database (SGD)

provides Secondary Gene Annotation Using the Gene Ontology (GO). *Nucleic Acids Res.* **30**: 69 - 72.

Eisen, M., Spellman, P., Brown, P., and Bostein, D. 1998. Cluster Analysis and Display of Genome-wide Expression Patterns. *Proc. Natl. Acad. Sci. USA.* **95**: 14863 -- 14868.

Gavin, A., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J., Michon, A., Cruciat, C. 2002. Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes. *Nature.* **415**: 141 -- 147.

Goehring, A., Mitchell, D., Tong, A., Keniry, M., Boone, C., and Sprague, G. 2003. Synthetic Lethal Analysis implicates Ste20p, a p21-activated Protein Kinase, in Polarisome Activation. *Mol. Bio. Cell.* **4**: 1501 – 1516.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. 2002. Systematic Identification of Protein Complexes in *Saccharomyces cerevisiae* by Mass Spectrometry. *Nature.* **415**: 180 -- 183.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T. 2001. Assessment of Prediction Accuracy of Protein Function from Protein-protein Interaction Data. *Yeast.* **18**: 523 -- 531.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. 2001. Toward a Protein-protein Interaction Map of the Budding Yeast: A Comprehensive System to Examine Two-Hybrid Interactions in All Possible Combinations Between the Yeast Proteins. *Proc. Natl. Acad. Sci. USA.* **98**: 4569 -- 4574.

Marcotte, E., Pellegrini, M., Ng, H., Rice, D., Yeates, T., and Eisenberg, D. 1999. Detecting Protein Function and Protein-protein Interactions from Genome Sequences. *Science.* **285**: 751 -- 753.

Marcotte, E., Pellegrini, M., Thompson, M., Yeates, T., and Eisenberg, D. 1999. A Combined Algorithm for Genome-wide Prediction of Protein Function. *Nature.* **402**: 83 -- 86.

- Mewes, H., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. 2002. MIPS: A Database for Genomes and Protein Sequences. *Nucleic Acids Res.* **30**: 31 -- 34.
- Pavlidis, P., and Weston, J. 2001. Gene Functional Classification from Heterogeneous Data. In *Proceedings of the Fifth International Conference on Computational Molecular Biology (RECOMB2001)*, pp. 249 - 255.
- Pearson, W., and Lipman, D. 1998. Improved Tools for Biological Sequence Comparison. *Proc. Natl Acad. Sci. USA.* **85**: 2444 -- 2448.
- Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D., Yeates, T. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, **96**: 4285-- 4288.
- Roberts, C., Nelson, B., Marton, M., Stoughton, R., Meyer, M., Bennett, H., He, Y. D. D., Dai, H., Walker, W., Hughes, T., et al. 2000. Signaling and Circuitry of Multiple MAPK Pathways Revealed by a Matrix of Global Gene Expression Profiles. *Science.* **287**: 873 -- 880.
- Schwikowski, B., Uetz, P., and Fields, S. 2000. A Network of Protein-protein Interactions in Yeast. *Nature Biotechnology.* **18**: 1257 -- 1261.
- The Gene Ontology Consortium. 2000. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics.* **25**: 25 -- 29.
- Tong, A., Evangelista, M., Parsons, A., Xu, H., Bader, G., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C., Bussey, H., Andrews, B., Tyers, M., and Boone, C. 2001 Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. *Science.* **294**: 2364 -- 2368.
- Troyanskaya, O., Dolinski, K., Owen, A., Altman, R., and Botstein, D. 2003. A Bayesian Framework for Combining Heterogeneous Data Sources for Gene Function Prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci.* **100**: 8348 – 8353.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. 2000. A Comprehensive Analysis of Protein-protein Interactions in *Saccharomyces cerevisiae*. *Nature*. **403**: 623 -- 627.

## **WEB SITE REFERENCES**

[http://www.ncbi.nlm.nih.gov/sutils/genom\\_table.cgi](http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi), National Center for Biotechnology Information (NCBI).

<http://www.w3.org/XML>, XML Extensible Markup Language (XML).

<http://bioinfo.mbb.yale.edu/genome/localize>, Yeast Protein Localization Server.

<http://www.yeastgenome.org>, Saccharomyces Genome Database (SGD).

<http://mips.gsf.de>, Munich Information Center for Protein Sequences (MIPS).

<http://digbio.missouri.edu/genefas>, GeneFAS (Gene Function Annotation System).