

# GENOME-SCALE PROTEIN FUNCTION PREDICTION IN YEAST *SACCHAROMYCES CEREVISIAE* THROUGH INTEGRATING MULTIPLE SOURCES OF HIGH-THROUGHPUT DATA

YU CHEN<sup>1,2†</sup> AND DONG XU<sup>1,2\*</sup>

<sup>1</sup> *UT-ORNL Graduate School of Genome Science and Technology, Oak Ridge, TN, USA.*

<sup>2</sup> *Digital Biology Laboratory, Computer Science Department, University of Missouri-Columbia, Columbia, MO, USA.*

As we are moving into the post genome-sequencing era, various high-throughput experimental techniques have been developed to characterize biological systems at the genome scale. Discovering new biological knowledge from high-throughput biological data is a major challenge for bioinformatics today. To address this challenge, we developed a Bayesian statistical method together with Boltzmann machine and simulated annealing for protein function prediction in the yeast *Saccharomyces cerevisiae* through integrating various high-throughput biological data, including protein binary interactions, protein complexes and microarray gene expression profiles. In our approach, we quantified the relationship between functional similarity and high-throughput data. Based on our method, 1802 out of 2280 unannotated proteins in the yeast were assigned functions systematically. The related computer package is available upon request.

## 1. Introduction

An immediate challenge of the post-genomic era is to assign biological functions to all the proteins encoded by the genome. For example, only one-third of all 6200 predicted genes in yeast *Saccharomyces cerevisiae* (Baker's yeast) were functionally characterized when the complete sequence of yeast genome became available<sup>1</sup>. At present, 4044 yeast genes have been annotated out of 6324 genes. Despite all the efforts, only 50-60% of genes have been annotated in most organisms with complete genomes. This leaves bioinformatics with the opportunity and challenge of predicting functions for unannotated proteins by developing effective and automated methods.

With ever-increasing flow of biological data generated by high-throughput methods, such as yeast two-hybrid systems<sup>2</sup>, protein complexes identification by mass spectrometry<sup>3,4</sup> and microarray gene expression profiles<sup>5</sup>, 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 for predicting functions based on the assumption that genes with similar functions are likely to be co-expressed<sup>6</sup>. Using protein-protein

---

<sup>†</sup>Current address: Oncology Business Unit, Novartis Pharmaceuticals Corp. One Health Plaza, East Hanover, NJ 07936, USA.

\* Corresponding author (xudong@missouri.edu).

interaction data to assign function for 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, also known as “guilt by association”<sup>7</sup>. Schwikowski et al.<sup>7</sup> applied a 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.<sup>8</sup> who used  $\chi^2$  statistics. Both these approaches give equal significance to all the functions contributed by the protein neighbors in the interaction network. Other function prediction methods using high-throughput data include machine-learning approach<sup>9</sup> and Markov random fields<sup>10,11</sup>. MAGIC (Multisource Association of Genes by Integration of Clusters) approach to combine heterogeneous data for function assignment has been applied in yeast by Troyanskaya et al.<sup>12</sup>.

One major challenge for protein function prediction is that the errors (noises) in the high-throughput data have 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<sup>13</sup>. A possible solution for this problem is Bayesian probabilistic model<sup>14</sup>, which could lead to a coherent function prediction and reduce the effect of noise by combining information from diverse data sources within a common probabilistic framework and naturally weighs each information source according to the conditional probability relationship among information sources. Another major limitation of current function prediction methods based on “majority rule” assignment<sup>7</sup> is that the global properties of interaction network are underutilized since current methods often do not take into account the links among proteins of unknown functions. Vazquez et al.<sup>15</sup> recently proposed a global method to assign protein functions based on protein interaction network by minimizing the number of protein interactions among different functional categories.

To further overcome these limitations, we developed a computational framework for systematic protein function annotation at the genome scale. Our current study focuses on yeast *Saccharomyces cerevisiae*, where rich high throughput data are available. Compared with current methods, our method is distinctive in the following aspects: (1) Unannotated proteins can be assigned to various function categories of GO biological processes with probabilities. This is in contrast to many other prediction methods where proteins were predicted as yes or no without confidence assessment to a limited number of function categories (e.g., MIPS<sup>16</sup>, which is less detailed than GO). (2) We quantitatively measured functional dependencies underlying each type of high-throughput data, including protein binary interactions, protein complexes, and microarray gene expression profiles) and coded them into “functional linkage graphs” (interaction network), where each node represents one protein and Bayesian probabilities were calculated to represent the function similarity using the weight of each edge between two proteins. (3) We developed a novel global

function prediction method based on Boltzmann machine for function prediction with integration of functional linkage evidences from different types of high-throughput data. We may predict the function of an unannotated gene, even if none of its neighbors in the network has known function. Our method is robust for combining and propagating information systematically across the entire network based on the global optimization of the network configuration.

## 2. Data sources

All the high-throughput data were coded into an interaction network, which can be viewed as a weighted non-directed graph  $G_p(D) = (V_p, E_p)$  with the vertex set  $V_p = \{d_i \mid d_i \in D\}$  and the edge set  $E_p = \{(d_i, d_j) \mid \text{for } d_i, d_j \in D \text{ and } i \neq j\}$ . Each vertex represents one protein and each edge represents one measured connection between the two linked proteins from any high-throughput data.

### Protein-protein binary interaction data

The protein-protein interaction data from high-throughput yeast two-hybrid interaction experiments were from Uetz et al.<sup>17</sup> and Ito et al.<sup>18</sup>, together with 5075 unique interactions among 3567 proteins. We combined the yeast two-hybrid data with the protein-protein interaction data in the MIPS database (<http://mips.gsf.de/proj/yeast/CYGD/db/>). In total, 6516 unique binary interactions among 3989 proteins were used in this study.

### Protein complexes

The protein complex data were obtained from Gavin et al.<sup>3</sup> and Ho et al.<sup>4</sup>. In the protein complexes, although it is unclear which proteins are in physical contact, the protein complex data contain rich information about functional relationship among involved proteins. 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. This yields 49,313 edges to the interaction network.

### Microarray gene expression data

The gene-expression profiles of microarray data were from Gasch et al.<sup>19</sup>, which included 174 experimental conditions for all the genes in yeast. A Pearson correlation coefficient was calculated for each possible ORF pairs to quantify the correlation between the gene pairs.

## 3. Methods

### *3.1 Measurement of protein function similarity*

A particular gene product can be characterized with different types of function, including molecular function at the biochemical level (e.g. cyclase or kinase,

whose annotation is often more related to sequence similarity and protein structure) and the biological process at the cellular level (e.g. pyrimidine metabolism or signal transduction, which 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<sup>20</sup>. The GO biological process ontology is available at <http://www.geneontology.org>. It has a hierarchical structure with multiple inheritance. After acquiring the biological process functional annotation for the known proteins along with their GO Identification (ID), we generated a numerical GO INDEX, which represents the hierarchical structure of the classification. The more detailed level of the GO INDEX, the more specific function a protein belongs to. The maximum level of GO INDEX is 12. The following shows an example of GO INDEX hierarchy, with the numbers on the left giving GO INDICES and the numbers in the brackets indicating GO IDs:

```

2          cellular process (GO:0009987)
2-1       cell communication (GO:0007154)
2-1-8    signal transduction (GO:0007165)
2-1-8-1  cell surface receptor linked signal transduction (GO:0007166)
2-1-8-1-4 G-protein coupled receptor protein signaling pathway (GO:0030454)
2-1-8-4-4-12 signal transduction during conjugation with cellular fusion (GO:0000750)

```

In the SGD data (<http://www.yeastgenome.org/>), 4044 yeast proteins have been assigned one or more GO biological process IDs. We calculated protein function similarity by comparing the level of similarity that the two proteins share in terms of their GO INDICES. For example, if both ORF1 and ORF2 have annotated functions, assuming that ORF1 has a function represented by GO INDEX 2-1-8-1 and ORF2 has a function represented by GO INDEX 2-1-8. When compared with each other for the level of matching GO INDEX, they match with each other through 2-1-8, i.e., INDEX level 1 (2), INDEX level 2 (2-1) and INDEX level 3 (2-1-8).

### 3.2 Calculation of Bayesian probabilities

We calculated probabilities for two genes to share the same function based on different types of high-throughput data, i.e., microarray data, protein binary interaction data and protein complex data. With the assumption that  $H = \{M, B, C\}$  denotes the interaction events in different types of high-throughput data, where  $M$  represents two genes correlated in gene expression profiles with Pearson correlation coefficient  $r$  in microarray data,  $B$  represents a protein binary interaction and  $C$  represents a protein complex interaction, the *posterior* probability that two proteins have the same function,  $p(S/H)$ , is computed using the Bayes' formulas:

$$p(S/H) = \frac{P(H|S)P(S)}{P(H)} \quad (1)$$

where  $S$  represents the event that two genes/proteins have the same function at a given level of GO INDEX. The probability  $p(S)$  is the relative frequency of

genes/proteins whose functions are the same at the given level of GO INDEX by chance.  $p(H|S)$  is the conditional (*a priori*) probability that two genes/proteins to have the event  $H$  given that they have the same function at a given level of GO INDEX. The probability  $p(H)$  is the frequency of  $H$  in the entire data set, e.g., the frequency of gene expression correlated with coefficient  $r$  over all gene pairs in yeast, which is calculated from the genome-wide gene expression profiles ( $H = M$ ) or the relative frequency of two proteins having a known binary interaction over all possible pairs in yeast, which is estimated from the known protein interaction data set ( $H = B$ ). The probabilities  $p(H/S)$ , and  $p(S)$  are computed based on a set of proteins whose functions have been annotated in GO biological process.

To quantify the gene function relationship between the correlated gene expression pairs, we calculated the probabilities of such gene expression correlated pairs sharing the same function at each GO INDEX level, based on our early study<sup>21</sup>. Results show a higher probability of sharing the same function for broad functional categories (the high-order GO INDEX levels) or highly correlated genes in expression profiles (Figure 1A). Figure 1B shows the presence of information in highly correlated gene-expression pairs for their gene functional relationship in comparison to random pairs. In Figure 1, we only show the curves of GO INDEX 1, 2, 3 and 4. The other higher GO INDEX levels (from 5 to 10) have the same trend. Based on Figures 1, we decided to consider pairs with gene expression profile correlation coefficient  $\geq 0.7$  for function predictions, as other pairs have little information for function prediction. The estimated probabilities of sharing the same function corresponding to gene pairs with  $r \geq 0.7$  were smoothed by using a monotone regression function (the pool-adjacent-violators algorithm<sup>22</sup>) for function prediction of unannotated proteins.

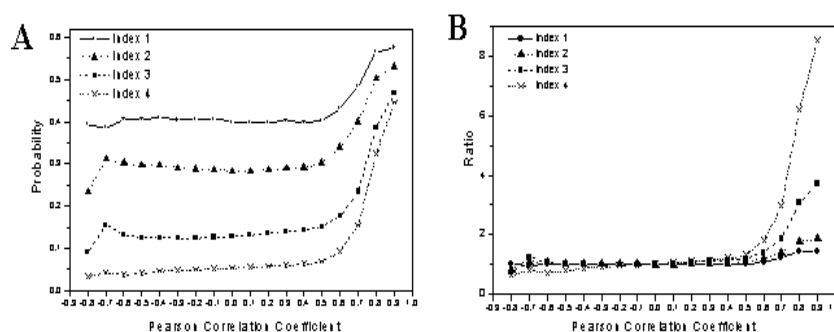


Figure 1. A: Probabilities of pairs sharing the same level of GO indices against Pearson correlation coefficient of microarray gene expression profiles. B: Normalized ratio for the percentage of gene pairs sharing the same level of GO indices ( $p(S|M)$ ), against the percentage of random pairs sharing the same function ( $p(S)$ ) versus Pearson correlation coefficient of microarray gene expression profiles.

The analysis result of the protein-protein interaction data is shown in Figure 2. The plots for protein binary interaction and complex interaction data, show a drop of probabilities of sharing the same function with an increase in the GO INDEX level, as seen in Figure 2A. A higher probability to share less specific, broader functional categories as represented by lower GO INDEX levels is observed. Comparison of our results with similar analysis on random pairs, shows a normalized ratio of protein-protein interaction pairs against the random pairs for sharing the same GO INDEX level (as seen in Figure 2B). Since the value is highly above 1, particularly for more specific function categories, there clearly exists a relationship between the protein-protein interaction data and function similarity. Such relationship can be utilized to make function predictions.

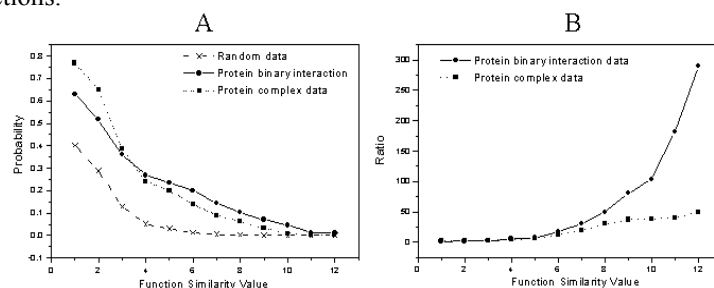


Figure 2. Functional relationship in yeast protein-protein interaction data. The horizontal axis in both plots shows the GO INDEX levels that two proteins share. (A) The probability of interacting proteins sharing the same levels of GO INDEX. (B) The normalized ratio of (A) compared with random pairs.

### 3.3 Protein function prediction

#### 3.3.1 Local prediction

In the local prediction of an unannotated protein using its immediate neighbors in the network graph, we follow the idea of “guilt by association”<sup>7</sup>, i.e., if an interaction partner of an unannotated protein  $x$  has a known function,  $x$  may share the same function, with a probability underlying the high-throughput data between  $x$  and its partner. We identify the possible interactors for  $x$  in each high-throughput data type (protein binary interaction, protein complex interaction and microarray gene expression with correlation coefficient  $r \geq 0.7$ ). We assign functions to the unannotated proteins on the basis of common functions identified among the annotated interaction partners using the probabilities described in section 3.2. Furthermore, we assume that the information contents for protein function prediction from different sources of high-throughput data or different interaction partners are independent based on the early suggestion that the information from different high-throughput data are conditionally uncorrelated<sup>23,24</sup>. A protein can belong to one or more functional classes, depending upon its interaction partners and their functions. For example, protein  $x$  is an unannotated protein with several interaction partners having known functions. With the assumption that  $F_i$ ,  $i = 1, 2, \dots, n$ , represents

a collection of all the functions that the interaction partners have, a likelihood score function for protein  $x$  to have function  $F_i$ ,  $G(F_i, x)$ , is calculated as:

$$G(F_i, x) = 1 - (1 - P(S_l|M)) * (1 - P(S_l|B)) * (1 - P(S_l|C)) \quad (2)$$

where  $S_l$  represents the event that two proteins have the same function,  $F_i$ , whose GO INDEX has  $l$  levels,  $l = 1, 2, \dots, 12$ .  $P(S_l|M)$ ,  $P(S_l|B)$  and  $P(S_l|C)$  are the probabilities of interaction pairs to have the same function for gene expression correlation coefficient  $\geq 0.7$  ( $M$ ), protein binary interaction ( $B$ ) and protein complex interaction ( $C$ ), respectively. In each type of high-throughput data, one unannotated protein might have multiple interaction partners with function  $F_i$ . Suppose that there are  $n_M$ ,  $n_B$ , and  $n_C$  interaction partners with function  $F_i$  in the three types of high throughput data, respectively. The combined probabilities  $P(S_l|M)$ ,  $P(S_l|B)$ , and  $P(S_l|C)$  in equation (2) are calculated as:

$$P(S_l|M) = 1 - \prod [1 - P_j(S_l|M)], \quad j = 1, 2, \dots, n_M. \quad (3)$$

$$P(S_l|B) = 1 - \prod [1 - P_j(S_l|B)], \quad j = 1, 2, \dots, n_B. \quad (4)$$

$$P(S_l|C) = 1 - \prod [1 - P_j(S_l|C)], \quad j = 1, 2, \dots, n_C. \quad (5)$$

$P_j(S_l|M)$ ,  $P_j(S_l|B)$ , and  $P_j(S_l|C)$  were estimated probabilities retrieved from the probability curves calculated in section 3.2 for a single pair of genes/proteins. We also defined the likelihood score  $G(F_i, x)$  as *Reliability Score* for each function,  $F_i$ . The final predictions are sorted based on the *Reliability Score* for each predicted GO INDEX. The *Reliability Score* represents the probability for the unannotated protein to have a function  $F_i$ , assuming that all the evidences from the high-throughput data are independent and only applicable to immediate neighbors in the network.

### 3.3.2 Global Prediction

The major limitation of the local prediction method is that it only uses the information of immediate neighbors in a graph to predict a protein's function. In some cases, the uncharacterized proteins may not have any interacting partner with known function annotation and its function cannot be predicted based on the local prediction method. In addition, the global properties of the graph are underutilized since this analysis does not include the links among proteins of unknown function. In Figure 3 proteins 1, 2, 3 and 4 are annotated and proteins 5, 6, 7 and 8 have unknown functions. If we only use the local prediction method, the functions of proteins 3 and 4 can be predicted but the functions of proteins while 1 and 2 cannot since all the neighbors of proteins 1 and 2 are unannotated proteins. Moreover, the contributions of function assignment for protein 4 is not only from the neighbor proteins 7 and 8 whose functions are already known, but also from protein 1 when its functions is predicted through the following information propagation: proteins 5 and 6  $\rightarrow$  protein 3  $\rightarrow$  protein 2  $\rightarrow$  protein 1. Hence, the functional annotation of uncharacterized proteins should not only be decided by their direct neighbors but also controlled by the global configuration of the interaction network. Based on such global optimization strategies, we developed a new approach for predicting protein

function. We used the Boltzmann machine to characterize the global stochastic behaviors of the network. A protein can be assigned to multiple functional classes, each with a certain probability.

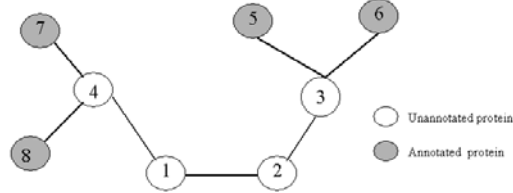


Figure 3: Illustration of protein function prediction from interaction network. Protein 1, 2, 3 and 4 are unannotated proteins. Protein 5, 6, 7 and 8 are annotated proteins with known functions.

In Boltzmann machine (BM), we consider a physical system with a set of states,  $\alpha$ , each of which has an energy,  $H_\alpha$ . In thermal equilibrium, given a temperature  $T$ , each of the possible states  $\alpha$  occurs with probability:

$$P_\alpha = \frac{1}{R} e^{-H_\alpha / K_B T} \quad (6)$$

where the normalizing factor  $R = \sum_\alpha e^{-H_\alpha / K_B T}$  and  $K_B$  is the Boltzmann's

constant. This is called the Boltzmann-Gibbs distribution. It is derived based on the general assumptions about microscopic dynamics, and it can be applied to a stochastic network. In an undirected graphical model with binary-valued nodes, each node (protein)  $i$  in the network has only one state value  $Z$  (1 or 0). In our case,  $Z = 1$  means that corresponding node (protein/gene) either has a known function, or it is ready for a function prediction. Now we consider the system going through a dynamic process from non-equilibrium to equilibrium, which corresponds the optimization process for the function prediction. For the state at time  $t$  (optimization integration step  $t$ ), node  $i$  has the probability for  $Z_{t,i}$  to be 1,  $P(Z_{t,i} = 1 | Z_{t-1, j \neq i})$  and the probability is given as a sigmoid-function of the inputs from all the other nodes at time  $t-1$ :

$$P(Z_{t,i} = 1 | Z_{t-1, j \neq i}) = \frac{1}{1 + e^{-\beta \sum_{j \neq i} W_{ij} Z_{t-1, j \neq i}}} \quad (7)$$

where  $\beta$  is a parameter reversely proportional to the annealing temperature and  $W_{ij}$  is the weight of the edge connecting proteins  $i$  and  $j$  in the interaction graph.  $W_{ij}$  is calculated by combining the evidence from gene expression correlation coefficient  $\geq 0.7$  ( $M$ ), protein binary interaction ( $B$ ) and protein complex interaction ( $C$ ):

$$W_{ij} = \delta_j \sum_{k=1}^{12} G(F_k, i | j) = \delta_j \sum_{k=1}^{12} (1 - (1 - P(S_k | M))(1 - P(S_k | B))(1 - P(S_k | C))) \quad (8)$$

where  $S_k$  represents the event that two proteins  $i$  and  $j$  have the same function ( $F_k$ ) at the GO INDEX level  $k$ ,  $k = 1, 2, \dots, 12$ .  $G(F_k, i | j)$  is the reliability score

for proteins  $i$  and  $j$  sharing the same function  $F_k$ .  $P(S_k|M)$ ,  $P(S_k|B)$  and  $P(S_kC)$  were estimated probabilities retrieved from the probability curves calculated in section 3.2.  $\delta_i$  is the modifying weight:

$$\delta_j = \begin{cases} 1 & \text{if } j \in \text{annotated proteins} \\ P(Z_{t-1,j} = 1) & \text{otherwise} \end{cases} \quad (9)$$

To achieve the global optimization, we conducted simulated annealing technique as the following process. First we set the initial state of all unannotated proteins (nodes) randomly to be 0 or 1. The state of any annotated protein is always 1. If an unannotated protein is assigned with the state 1, its function will be predicted based on its immediate neighbors with known functions using the local prediction method. Next starting with a high temperature, pick a node  $i$  and compute its value  $P_i$  according to equation (6), then update its state, till all the nodes in the network reach equilibrium. With gradually cooling down, the system might settle in a global optimization of network configuration if the sum of weights associated to all the unannotated proteins reaches the maximum value.

#### 4. Results

We have implemented local and global methods to predict functions for unannotated proteins. We used *sensitivity* and *specificity* to measure the performance of our methods using ten-fold cross validation. We labeled all 4044 annotated proteins with known GO INDICES into folds 1 to 10. Each time, we pick one fold as the test dataset and the other nine folds as training data to calculate prior probabilities. We estimate the *sensitivity* to determine the success rate of the method and *specificity* to assess the confidence in the predictions. 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:

$$SN = \frac{\sum_1^K k_i}{\sum_1^K n_i} \quad (10) \quad SP = \frac{\sum_1^K k_i}{\sum_1^K m_i} \quad (11)$$

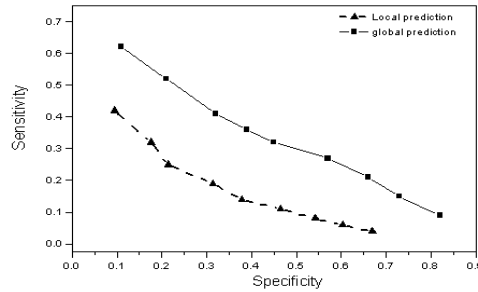


Figure 4: *Sensitivity-specificity* plot on the test set for the three prediction methods.

Figure 4 shows the *sensitivity* versus *specificity* of the methods with *Reliability score* cutoff from 0.1 to 0.9. It showed that the global prediction has a significantly better sensitivity-specificity plot than the local prediction. In our method, the highest specificity can only reach 70%. Some false positives generated in our method might be caused by the independence assumption of different sources of high-throughput data. Such assumption could be oversimplified due to biases inherent in data. For example, protein binary interactions are likely to be correlated in gene expression profiles. On the other hand, the predicted functions from our method could be true but they have not been determined by experiments yet, thus, they are not included in GO annotation.

INDEX	Reliability Score								
	$\geq 0.9$	$\geq 0.8$	$\geq 0.7$	$\geq 0.6$	$\geq 0.5$	$\geq 0.4$	$\geq 0.3$	$\geq 0.2$	$\geq 0.1$
1	897	964	1045	1116	1185	1264	1331	1530	1707
2	847	922	978	1052	1130	1217	1315	1519	1707
3	710	801	883	955	1018	1102	1236	1491	1693
4	627	714	789	870	949	1052	1151	1433	1673
5	605	691	761	836	918	1016	1120	1405	1659
6	271	378	472	447	622	707	849	1128	1495
7	104	173	248	316	395	483	595	722	1159
8	14	31	48	68	103	147	194	299	680
9	0	1	2	3	4	4	11	20	105
10	0	0	0	0	0	0	0	0	6

Table 1. Number of unannotated genes with function predictions with respect to prediction confidence probabilities and index levels.

Using all the 4044 annotated proteins with known GO INDICES as the training set, we are able to assign functions to 1802 out of the 2280 unannotated proteins in yeast at different level of functions (different levels of GO INDICES). The detail prediction results can be queried at <http://digbio.missouri.edu/~ychen/ProFunPred>. The number of unannotated genes with function predictions with respect to the specificity and GO INDEX levels can be found in Table 1. Using our method, we not only assign general functional categories to unannotated protein but also assign very specific functions to some unannotated proteins. 104 unannotated proteins were assigned functions with reliability score  $\geq 0.9$  and GO INDEX level  $\geq 7$ . Moreover, using our global prediction method we can assign functions for the proteins whose interacting partners are all unannotated proteins with unknown function. For example, all interacting partners of *YBR100W* have unknown functions. Using global function prediction, *YBR100W* was assigned function as “nucleic acid metabolism” (reliability score 0.8) and “response to DNA damage” (reliability score 0.8).

## 5. Discussion

Systematic and automated prediction of protein function using high-throughput data represents 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 binary interaction, protein complexes and gene expression microarray data, together with the GO biological process functional annotation. We applied Boltzmann machine for the global protein function annotation by combining and propagating information across the entire network. Our method is robust to obtain global optimization using simulated annealing. With six different sets of randomly selected starting points, we obtained exactly the same result as shown in Table 1.

Future work includes exploring better optimization methods and statistical models. To solve the optimization problem in Boltzmann machine, in contrast to the simulated annealing technique, a Bayesian learning of posterior distributions over parameters<sup>25</sup> provides a more elaborate and systematic estimation of maximum likelihood. In addition, supervised learning methods such as Conditional Random Fields<sup>26</sup> can also be alternative schemes to model this stochastic learning process. Furthermore, we will develop more elaborate model-based integrations to address the dependencies among different high-throughput data for protein function prediction.

## Acknowledgments

We would like to thank Drs. Jeffery Becker, Ying Xu, Loren Hausser, and Qiang Zhao for helpful discussions. This research is supported in part by the US Department of Energy's Genomes to Life program (<http://www.doegenomestolife.org>) under project, "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling" ([www.genomes-to-life.org](http://www.genomes-to-life.org)). This work is also partially funded by Nation Science Foundation (EIA-0325386).

## References

- 
- <sup>1</sup> Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, et al *Science*. **546**:346-352, (1996).
  - <sup>2</sup> Chien, C., Bartel, P., Sternglanz, R., and Fields, S. *Proc. Natl. Acad. Sci. USA*. **88**:9578-9582, (1991).
  - <sup>3</sup> Gavin, A., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J., Michon, A., Cruciat, C. *Nature*. **415**:141-147, (2002).
  - <sup>4</sup> Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. *Nature*. **415**: 180-183, (2002).

- 
- <sup>5</sup> Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M., and Haussler, D. *Proc. Natl. Acad. Sci. USA.* **97**:262-267, (2000)
- <sup>6</sup> Pavlidis, P., and Weston, J. In Proceedings of the Fifth International Conference on Computational Molecular Biology (RECOMB2001), pp. 249 – 255, (2001)
- <sup>7</sup> Schwikowski, B., Uetz, P., and Fields, S. *Nature Biotechnology.* **18**:1257-1261, (2000)
- <sup>8</sup> Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T. *Yeast.* **18**: 523-531, (2001).
- <sup>9</sup> Clare, A. and King, R. D. ECCB 2003 (also published as a journal supplement in Bioinformatics. 19: ii42-ii49), (2003).
- <sup>10</sup> Deng MH, Zhang K, Mehta S, Chen T, Sun FZ, The first IEEE Computer Society bioinformatics conference, CSB2002, 117-126, (2002).
- <sup>11</sup> Letovsky, S. and Kasif, S. *Bioinformatics*, **19** Suppl 1:1197-1204, (2003).
- <sup>12</sup> Troyanskaya, O., Dolinski, K., Owen, A., Altman, R., and Botstein, D. *Proc. Natl. Acad. Sci.* **100**:8348–8353, (2003)
- <sup>13</sup> Chen, Y., and Xu, D. *Current Peptide and Protein Science.* **4**:159-181, (2003).
- <sup>14</sup> Winkler, R. L. An introduction to Bayesian Inference and Decision, Holt, Rinehart and Winston Inc, (1972).
- <sup>15</sup> Vazquez, A., Flammini, A., Maritan, A. and Vespignani, A. *Nat Biotechnol.* **21**:697-700, (2003).
- <sup>16</sup> Mewes, H., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M. et al. *Nucleic Acids Res.* **30**:31-34, (2002).
- <sup>17</sup> 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. *Nature.* **403**:623-627, (2000).
- <sup>18</sup> Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. *Proc. Natl. Acad. Sci. USA.* **98**:4569-4574, (2001).
- <sup>19</sup> Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. and Brown, P. O, *Mol Cell Biol*, **11**:4241-4257, (2000).
- <sup>20</sup> The Gene Ontology Consortium. *Nature Genetics.* **25**:25-29, (2000).
- <sup>21</sup> Joshi, T., Chen, Y., Becker, J. M., Alexandrov, N., Xu, D. *OMICS*, in press.
- <sup>22</sup> Haerdle, W. Applied Nonparametric Regression, Cambridge, UK. (1995).
- <sup>23</sup> Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., et al. *Science.* **302**:449-453, (2003).
- <sup>24</sup> Asthana S., King O. D., Gibbons F. D. and Roth F. P., *Genome Res.* **14**:1170-1175, (2004).
- <sup>25</sup> Ackley D. H., Hinton G. E. and Sejnowski T. J., *Cognitive Science*, **9**:147-169, (1985).
- <sup>26</sup> Lafferty J., McCallum A., and Pereira F., *International Conference on Machine Learning (ICML)* (2001).