

Function–Function Correlated Multi-label Protein Function Prediction over Interaction Networks

HUA WANG, HENG HUANG, and CHRIS DING

ABSTRACT

Many previous works in protein function prediction make predictions one function at a time, fundamentally, which assumes the functional categories to be isolated. However, biological processes are highly correlated and usually intertwined together to happen at the same time; therefore, it would be beneficial to consider protein function prediction as one indivisible task and treat all the functional categories as an integral and correlated prediction target. By leveraging the function–function correlations, it is expected to achieve improved overall predictive accuracy. To this end, we develop a network-based protein function prediction approach, under the framework of multi-label classification in machine learning, to utilize the function–function correlations. Besides formulating the function–function correlations in the optimization objective explicitly, we also exploit them as part of the pairwise protein–protein similarities implicitly. The algorithm is built upon the Green’s function over a graph, which not only employs the global topology of a network but also captures its local structures. In addition, we propose an adaptive decision boundary method to deal with the unbalanced distribution of protein annotation data. Finally, we quantify the statistical confidence of predicted functions to facilitate post-processing of proteomic analysis. We evaluate the proposed approach on *Saccharomyces cerevisiae* data, and the experimental results demonstrate very encouraging results.

Key words: algorithms, biochemical networks, gene clusters.

1. INTRODUCTION

MANY EXISTING METHODS IN PREDICTING protein function from protein interaction network data typically make predictions one function at a time, fundamentally. This turns the problem into a convenient form for using existing machine-learning algorithms, which, however, abstract the function correlations, although most biological functions are interdependent from one another. For example, “Transcription” and “Protein Synthesis” (Mewes et al., 1999) usually appear together, one after another i.e., they tend to appear in the biological processes involving the same protein. As a result, if a protein is known to be annotated with the function “Transcription,” it is highly probable to annotate the same protein with function “Protein Synthesis” as well. In other words, the *function–function correlations* convey valuable information toward understanding the biological processes, which provides a potential opportunity to improve the protein

function prediction accuracy. To this end, how to effectively exploit function–function correlations presents a challenging, yet important, problem in proteomic analysis for protein function prediction. In this study, we tackle this new problem by placing protein function prediction under the framework of multi-label classification, an emerging topic in machine learning, to develop a new graph-based protein function prediction method to take advantage of function–function correlations.

1.1. Network-based protein function prediction

Recent availability of protein interaction networks for many species has spurred on the development of network-based computational methods in protein function prediction. Typically, an interaction network is first modeled as a graph, with the vertices representing proteins and the edges representing the detected protein–protein interactions (PPI), followed by a graph-based statistical learning method to infer putative protein functions.

Review of related works The most straightforward method using network data to predict protein determines the functions of a protein from the known functions of its neighboring proteins on a PPI network (Schwikowski et al., 2000; Hishigaki et al., 2001; Chua et al., 2006), which leverages only local information of a network. Later researchers used global optimization approaches to improve the protein function predictions by taking into account the full topology of networks (Vazquez et al., 2003; Karaoz et al., 2004; Nabieva et al., 2005). All these approaches can be summarized as the following common schemes: (1) compute a set of ranking lists, and (2) make predictions using certain thresholds on the ranking lists. In step 1, which is the most critical part of the algorithms, they all compute the ranking lists one function at a time and ignore the relationships among the functions. A broad variety of network-based approaches using other models for protein function prediction are surveyed in Sharan et al. (2007).

We use an example to illustrate the deficiencies of the aforementioned methods. A small part of the PPI graph constructed from the BioGRID data (Stark et al., 2006) and annotated by MIPS Funcat scheme (Mewes et al., 1999) is shown in Figure 1. The clear oval vertices are unannotated proteins while the elliptical ones are proteins annotated with function “Metabolism” and the rectangular ones are proteins not annotated with the same function. The task is to determine whether the unannotated proteins have the functionality of “Metabolism.” When neighbor-counting approaches are applied, only the annotated proteins contribute to the annotation of an unannotated protein. For example, the functions of “YIL152W” is determined solely by those of “HSP82” and “BUD4,” but the rest of the annotated proteins and their unannotated neighbor “YER071C” are not used. In global optimization approaches, the annotated proteins are always treated the same, no matter how far they are from and how many links they are connected to the unannotated proteins. For example, when the global optimization approaches applied to annotate “YER071C,” “SSB2,” and “BUD4” are treated the same, although the former is closer to “YER071C”; “HSP82” and “CAP1” are also treated the same, although there are two connections from the former to “YER071C” while there is only one from the latter. Function-flow approach (Nabieva et al., 2005) takes care of the distance and link patterns, but it restricts the propagation to a fixed number of steps.

Motivation to use the Green’s function approach From the example in Figure 1, we can see that the above existing approaches bank on two assumptions: local consistency and global consistency, which are the exact foundations of the label propagation approaches for classification in machine learning. This

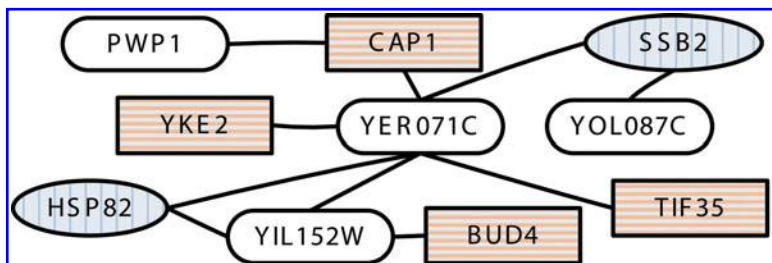


FIG. 1. A part of (PPI) graph constructed by BioGRID data, which illustrates the deficiencies of some existing approaches. The oval vertices without background color are the unannotated proteins, while the elliptical vertices with blue background are the annotated proteins associated with function “metabolism,” and the rectangular vertices are the annotated proteins not associated with function “metabolism”.

motivates us to formulate protein function prediction over a PPI network as a label propagation problem on a graph. Among existing label propagation methods, we choose to develop our new method from the Green's function approach (Ding et al., 2007; Wang et al., 2009) due to its demonstrated effectiveness in other applications and clear intuitions. Most importantly, the weaknesses in previous methods can be perfectly solved by the Green's function approach as detailed in Section 2.4.

1.2. Multi-label correlated protein function prediction

Because a protein is usually observed to play several functional roles in different biological processes within an organism, it is natural to annotate it with multiple functions. Thus, protein function prediction is an ideal example of *multi-label classification* (Wang et al., 2009, 2010b–d, 2011) in machine learning. Multi-label classification, in which each object may belong to more than one class, is an emerging topic driven by the advances of modern technologies in recent years. Placing protein function prediction under the framework of multi-label classification, we use the Green's function approach to integrate the function–function correlations from the theory of reproducing kernel Hilbert space (RKHS) (in Section 2.4). Besides incorporating the function–function correlations as a regularizer in the optimization objective explicitly, we also take advantage of them as part of the pairwise protein similarities implicitly (in Section 2.5). In addition, we propose an adaptive decision boundary method to deal with the unbalanced distribution of protein annotation data (in Section 2.6), and quantify the statistical confidence of predicted putative functions for post-processing of proteomic analysis (in Section 2.7).

2. METHODS

In this section, we propose a function–function correlated multi-label (FCML) approach using the Green's function on a graph to predict protein functions, which incorporates the function–function correlations in two levels: one from the function perspective to formulate the functionwise similarities explicitly in the optimization objective (in Section 2.4), and the other from the protein similarity perspective using function assignments to model the function correlations implicitly (in Section 2.5). Besides being used in the proposed approach, the latter also provides a means for all other previous related works to exploit the function correlations.

2.1. Notations and problem formalization

In protein function prediction, given K biological functions and n proteins, each protein x_i is associated with a set of labels represented by a function assignment indication vector $\mathbf{y}_i \in \{-1, 0, 1\}^K$ such that $\mathbf{y}_i(k) = 1$ if protein x_i has the k -th function, $\mathbf{y}_i(k) = -1$ if it does not have the k -th function, and $\mathbf{y}_i(k) = 0$ if its function assignment is not yet known *a priori*. Given l annotated proteins $\{(x_1, \mathbf{y}_1), \dots, (x_l, \mathbf{y}_l)\}$ where $l < n$, the goal is to predict functions $\{\mathbf{y}_i\}_{i=l+1}^n$ for the unannotated proteins $\{x_i\}_{i=l+1}^n$. We write $Y = [\mathbf{y}_1, \dots, \mathbf{y}_n]^T$, and $Y_l = [\mathbf{y}_1, \dots, \mathbf{y}_l]^T = [\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(K)}]$, where $\mathbf{y}^{(k)} \in \mathbb{R}^l$ is a classwise function assignment vector. We also define $F = [\mathbf{f}_1, \dots, \mathbf{f}_n]^T \in \mathbb{R}^{n \times K}$ as the decision matrix for prediction, and $\{\mathbf{f}_i\}_{i=l+1}^n$ includes the decision values for prediction.

We formalize a protein interaction network as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The vertices \mathcal{V} corresponds to the proteins $\{x_1, \dots, x_n\}$, and the edges \mathcal{E} are weighted by an $n \times n$ similarity matrix W with W_{ij} indicating the similarity between x_i and x_j . In the simplest case, W is the adjacency matrix of the PPI graph where $W_{ij} = 1$ if proteins x_i and x_j interact, and 0 otherwise. In this work, W is computed in Equation (14) to incorporate more useful information.

In summary, for a protein function prediction task, we are given W and Y_l as input, and the outputs of our method are decision values for the predicted putative functions assigned to the unannotated proteins, that is, $\{\mathbf{f}_i\}_{i=l+1}^n$.

2.2. Protein function prediction using the Green's function over a graph

In this section, we first briefly review the Green's function approach for label propagation over a graph, from which we will develop the proposed FCML method in Section 2.4.

The Green's function is of significant importance in solving partial differential equations, because it transforms them into integral equations. In physics, the Green's function $G(\mathbf{r}, \mathbf{r}')$ represents the field

response (i.e., influence) at location \mathbf{r} to the presence of a charge at local \mathbf{r}' . In machine learning, $G(\mathbf{r}, \mathbf{r}')$ quantifies the influence of a labeled data point at \mathbf{r}' to another unlabeled data point at \mathbf{r} .

To be more specific, given a graph with edge weights W , its combinatorial Laplacian is defined as $L = D - W$ (Chung, 1997), where $D = \text{diag}(We)$ and $\mathbf{e} = [1 \dots 1]^T$. The Green's function over the graph is defined as the inverse of L with zero-mode discarded, which is computed as following (Ding et al., 2007; Wang et al., 2009, 2010a):

$$G = L_+^{-1} = \frac{1}{(D - W)_+} = \sum_{i=2}^n \frac{\mathbf{v}_i \mathbf{v}_i^T}{\lambda_i}, \quad (1)$$

where L_+^{-1} and $(D - W)_+$ indicate that zero eigen-mode is discarded, and $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ are the eigenvalues of L , and \mathbf{v}_i are the corresponding eigenvectors. Because the Green's function G in Equation (1) is a kernel (Ding et al., 2007), from the theory of RKHS the optimization objective of the Green's function approach is to minimize (Ding et al., 2007):

$$J_0(F) = (1 - \mu) \|F - Y\|^2 + \mu \text{tr}(F^T \mathcal{K}^{-1} F), \quad (2)$$

where $\mathcal{K} = G$ is the kernel, $\mu \in (0, 1)$ is a constant to control the smoothness regularizer $\text{tr}(F^T \mathcal{K}^{-1} F)$, and $\text{tr}(\cdot)$ denotes the trace of a matrix. If we only consider one biological function, protein function prediction amounts to a two-class classification problem, where the function assignment vector is then reduced as a scalar, i.e., $y_i \in \{1, 0, -1\}$. Given labeled data $\{(x_i, y_i)\}_{i=1}^l$ and unlabeled data $\{x_i\}_{i=l+1}^n$, the labels of unlabeled data are computed by influence propagation from labeled data to those unlabeled (Ding et al., 2007):

$$y_j = \text{sign} \left(\sum_{i=1}^l G_{ji} y_i \right), \quad l < j \leq n. \quad (3)$$

Now we consider all the K functions. Extending Equation (3), we may assign functions to unannotated proteins as follows. Given K biological functions, we may assign functions to unannotated proteins as Ding et al. (2007):

$$\mathbf{y}_j = \text{sign}(\mathbf{f}_j), \quad l < j \leq n, \quad \text{where } F = GY. \quad (4)$$

We name Equation (4) simply as multi-label Green's function (MLGF) approach, beyond which we will propose a novel function–function correlated multi-label Green's function approach.

2.3. Utilizing both local and global structure of a PPI network by the Green's function

Before proceeding to propose our new approach, we first point out that the Green's function is closely related to a well-established distance metric on a generic weighted graph, where the edge weight measures the similarity between two end vertices. Based on the derivation of the distance metric, it is easy to see that the Green's function approach not only takes advantage of global topology of a network but also leverages its local structures.

We view a generic weighted graph as a network of electric resistors, where the edge connecting vertices x_i and x_j is a resistor with resistance r_{ij} . The graph edge weight (the pairwise similarity) between vertices x_i and x_j is $w_{ij} = 1/r_{ij}$. Two vertices not connected by a resistor are viewed as equivalently connected by a resistor with $r_{ij} = \infty$ or $w_{ij} = 0$. The most common task on a resistor network is to calculate the effective resistance between different vertices. The effective resistance R_{ij} between vertices x_i and x_j is equal to $1/(\text{total current between } x_i \text{ and } x_j)$ when x_i is connected to voltage 1 and x_j is connected to voltage 0. Let $G = (D - W)_+^{-1}$ be the Green's function on the graph, a remarkable result established in 1970s (Klein and Randić, 1993):

$$R_{ij} = (\mathbf{e}_i - \mathbf{e}_j)^T G (\mathbf{e}_i - \mathbf{e}_j) = G_{ii} + G_{jj} - 2G_{ij}, \quad (5)$$

where \mathbf{e}_i is a vector of all 0's except a "1" at i th entry. Recall the Mahalanobis distance in a metric space is $d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T M (\mathbf{x}_i - \mathbf{x}_j)$, we can view R_{ij} as a distance on a graph. The same conclusion can also be drawn from the random walk perspective of view (Ding et al., 2007). In statistics, given pairwise similarity $S = (s_{ij})$, a standard way to convert to distance is $d_{ij} = s_{ii} + s_{jj} - s_{ij}$. From Equation (5), we have

$s_{ij} = G_{ij} + \text{const}$. By ignoring the additive constant, G is the similarity metric underlying the effective resistor distance. Therefore, by simulating label propagation over a graph as current flowing on an electric network, the voltage of each vertices (i.e., the function assigned to a protein) is determined by both global topology and local structures of the electric network. In other words, for protein function prediction, the Green's function approach not only targets on global optimization but also rewards the local linkage patterns and distance impacts as illustrated in Figure 1.

From the analysis above, we can list some properties of the Green's function. First, G is clearly a semipositive definite function. Second, any function $\mathbf{f} \in \mathbb{R}^n$ can be expanded in the basis of G , that is $(\mathbf{v}_2, \dots, \mathbf{v}_n)$ plus a constant $\mathbf{e}/\sqrt{n} = \mathbf{v}_1$. Third, for a kernel function \mathcal{K} , \mathcal{K}_{ij} measures the similarity between two objects i and j . Therefore, Green's function is a bona fide kernel, which will be used to derive the proposed approach from RKHS theory (Section 2.4).

2.4. Function–function correlated RKHS approach for multi-label classification

Although the function–function correlations are useful to infer putative functions of unannotated proteins, MLGF approach defined in Equation (4) neglects them because it treats the biological functions as isolated. In multi-label scenarios, however, we concentrate on making use of the function–function correlations, which could be defined as $C \in \mathbb{R}^{K \times K}$ using cosine similarity as follows (Wang et al., 2009, 2010b–d, 2011):

$$C_{kl} = \cos(\mathbf{y}^{(k)}, \mathbf{y}^{(l)}) = \frac{\langle \mathbf{y}^{(k)}, \mathbf{y}^{(l)} \rangle}{\|\mathbf{y}^{(k)}\| \|\mathbf{y}^{(l)}\|}. \quad (6)$$

Following Wang et al., (2009), we expect to maximize $\mathbf{tr}(FCF^T)$. In order to make connection with the theory of RKHS, instead of directly using F , we use kernel-assisted decision matrix $\mathcal{K}^{-\frac{1}{2}}F$, which leads to the following objective to maximize (Wang et al., 2009):

$$J_C(F) = \mathbf{tr}(\mathcal{K}^{-\frac{1}{2}}FCF^T\mathcal{K}^{-\frac{1}{2}}). \quad (7)$$

Combining Equation (7) with the original RKHS objective in Equation (2), we minimize the following objective:

$$J(F) = \beta \|F - Y\|^2 + \mathbf{tr}(F^T\mathcal{K}^{-1}F) - \alpha \mathbf{tr}(\mathcal{K}^{-\frac{1}{2}}FCF^T\mathcal{K}^{-\frac{1}{2}}), \quad (8)$$

where $\alpha \in (0, 1)$ balances the two objectives, and $\beta = \frac{1-\mu}{\mu}$.

Differentiating J with respect to F , we have:

$$\frac{\partial J}{\partial F} = 2\beta(F - Y) + 2\mathcal{K}^{-1}F - 2\alpha\mathcal{K}^{-1}FC = 0 \implies F = \frac{1}{\beta I + \mathcal{K}^{-1}}\beta Y + \alpha \frac{1}{\beta I + \mathcal{K}^{-1}}\mathcal{K}^{-1}FC. \quad (9)$$

Because β is usually very small in typical empirical settings, we have:

$$\frac{F}{\beta} = \mathcal{K}Y + \alpha \frac{F}{\beta}C \implies \tilde{F} = \mathcal{K}Y + \alpha \tilde{F}C = GY + \alpha \tilde{F}C, \quad (10)$$

where $\tilde{F} = \frac{F}{\beta}$. Thus, we have

$$\tilde{F} = GY(I - \alpha C)^{-1}. \quad (11)$$

We name Equation (11) as our proposed function–function correlated multi-label (FCML) approach for protein function prediction. By Equation (11) we can compute F in a closed form without iterations, which is more mathematically elegant than related approaches. Moreover, $(I - \alpha C)^{-1}$ can be seen as another graph, which propagates label influence through the label correlations over the whole network.

A more in-depth look at FCML method. In practice, we select $\alpha < \frac{1}{\max(\zeta_k)}$, where $\zeta_k (0 < k < K)$ are the eigenvalues of C . Under this condition, Equation (11) can be written as follows:

$$\tilde{F} = GY(I + \alpha C + \alpha^2 C^2 + \dots), \quad (12)$$

which can be further seen as the following iterative process:

$$\begin{cases} \tilde{F}^{(0)} = GY, \\ \tilde{F}^{(t+1)} = GY + \alpha \tilde{F}^{(t)} C. \end{cases} \quad (13)$$

Equation (13) reveals the insight of the proposed FCML method. At the initialization step, the decision matrix $\tilde{F}^{(0)}$ is first initialized via label propagation using the Green's function method, which is exactly the same as the MLGF method defined in Equation (4). Then at each iteration step, besides retaining the initial information (the first term), the influence by function–function correlations are also taken into account (the second term). For example, we consider the case when protein x_i is annotated with the k_{s1} -th function but not the k_2 -th function, that is, $Y_{ik_1} = 1$ and $Y_{ik_2} = 0$. If through label propagation on the interaction network, protein x_i still can not acquire the k_2 -th function upon the network topology, after the initialization step we have $\tilde{F}_{ik_1}^{(0)} > 0$ and $\tilde{F}_{ik_2}^{(0)} = 0$. On the other hand, if these two functions are correlated, $C_{k_1 k_2} > 0$, then through the iteration step, $\tilde{F}_{ik_2}^{(1)} = \tilde{F}_{ik_2}^{(0)} + (\alpha \tilde{F}^{(0)} C)_{ik_2} > 0$, protein x_i is likely to be annotated with the k_2 -th function due to the function–function correlations.

2.5. Correlation augmented interaction network

Traditional network-based protein function prediction approaches only use biological interaction networks obtained from experimental data such as those from high-throughput technologies. When viewing protein function prediction as a multi-label classification problem, we can also build a computational interaction network $W_L \in \mathbb{R}^{n \times n}$ from label assignment perspective. As one of our important contributions, we make use of this new computational interaction network and propose a *correlation augmented interaction network* as follows:

$$W = W_{\text{Bio}} + \gamma W_L, \quad (14)$$

where W_{Bio} is the biological interaction network, which is same as in existing approaches. γ controls the relative importance of W_L , and empirically selected as $\gamma = \frac{\sum_{i,j,i \neq j} W_{\text{Bio}}(i,j)}{\sum_{i,j,i \neq j} W_L(i,j)}$.

The true power of the correlation augmented interaction network construction scheme defined in Equation (14) lies in that, the original biological similarities among proteins are augmented by the function assignment similarities, thereby label propagation pathways over a graph are reinforced. Moreover, with this interaction network construction scheme, the correlations among the functional categories are encoded into the graph weights, such that the resulted hybrid graph can be directly used in previous works to enhance their prediction performance. In this work, we use W defined in Equation (14) to compute the Green's function in Equation (1).

Protein–protein similarity from function assignments (W_L) Because multiple functions could be assigned to one single protein, the overlap between the function assignments of two proteins can be used to evaluate their similarity. The more functions shared by two proteins, the more similar they are. As a result, besides the class membership indications, the label assignment vector \mathbf{y}_i is enriched with characteristic meaning and can be used as an attribute vector to characterize protein x_i . Using cosine similarity, the function assignment similarity between two proteins is computed as:

$$W_L(i,j) = \cos(\mathbf{y}_i, \mathbf{y}_j) = \frac{\langle \mathbf{y}_i, \mathbf{y}_j \rangle}{\|\mathbf{y}_i\| \|\mathbf{y}_j\|}. \quad (15)$$

Our task in protein function prediction is to assign functions to unannotated proteins upon annotated ones. In order to compute W_L , however, we need function assignments of all the proteins including those annotated and unannotated. Therefore, we first initialize unannotated proteins through a majority voting (Schwikowski et al., 2000) approach, which makes predictions using the top three frequent functions of the protein's interacting partners. Note that the class similarity defined in Equation (6) is different from the protein similarity defined here in Eq. (15). The former is a function-wise similarity matrix of size $K \times K$, whereas the latter is a proteinwise similarity matrix of size $n \times n$, although essentially they both convey label correlations.

Biological protein–protein similarity (W_{Bio}) and multisource integration W_{Bio} in Equation (14) computes the protein–protein similarity from biological experimental data, which is the same as existing

works and can integrate multiple experimental sources. Let $W^{(1)}$ be the graph built from BioGRID PPI data (Ho et al., 2002; Giot et al., 2003), $W^{(2)}$ be that from synthetic lethal data (Tong et al., 2004), $W^{(3)}$ be that from gene coexpression data (Edgar et al., 2002), $W^{(4)}$ be that from gene regulation data (Harbison et al., 2004), etc., W_{Bio} is computed as follows (Pei and Zhang, 2005):

$$W_{\text{Bio}}(i, j) = 1 - \prod_k [1 - r^{(k)} W^{(k)}(i, j)], \quad (16)$$

where $r^{(k)}$ is estimated reliabilities of the corresponding network by expression profile reliability (EPR) index (Deane et al., 2002). Equation (16) reflects the fact that interactions detected in multiple experiments are generally more reliable than those detected by a single experiment (Von Mering et al., 2002).

Because in reality the overlap among different biological networks typically is very small, and the BioGRID PPI network data are fairly comprehensive, in this work we set $W_{\text{Bio}} = W^{(1)}$, where $W^{(1)}(i, j) = 1$ if protein x_i and x_j interact, and 0 otherwise.

By using the graph constructed from Equation (14), in addition to explicitly modeling the function–function correlations as in Equation (11), the correlations are also implicitly incorporated into the network linkages, so that the predictive accuracy can be further enhanced.

2.6. Adaptive decision boundary for function assignment

The MLGF approach defined in Equation (4) and FCML approach defined in Equation (11) produce ranked lists for function/label assignment, therefore decision boundaries are required to make predictions. Most existing research works using ranking lists to predict protein functions normally do not supply a threshold explicitly. Instead, they use a set of ROC curves (or the variant “precision”–“recall” curves) to evaluate the prediction performance. In some of these approaches, a heuristic cutoff point is given at the function assignment step e.g., in the majority voting (MV) approach, Schwikowski et al. (2000) assigned the three most frequently occurring functions among its neighbors to an unannotated protein. However, such threshold might not be the optimal one.

In many semi-supervised learning algorithms, the threshold for classification is usually selected as 0, which again is not necessarily the best choice. We propose an adaptive decision boundary to achieve better performance, which is adjusted such that the weighted training errors of all positive and negative samples are minimized.

Considering the binary classification problem for the k -th class, we denote b_k as the decision boundary, S_+ and S_- as the sets of positive and negative samples for the k -th class, and $e_+(b_k)$ and $e_-(b_k)$ as the numbers of misclassified positive and negative training samples. The adaptive (optimal) decision boundary is given by the Bayes’ rule $b_k^{\text{opt}} = \arg \min_{b_k} \left[\frac{e_+(b_k)}{|S_+|} + \frac{e_-(b_k)}{|S_-|} \right]$. And the decision rule is given by:

$$x_i \text{ acquires label } k = \begin{cases} +1, & \text{if } \tilde{F}_{ik} > b_k^{\text{opt}}; \\ -1, & \text{if } \tilde{F}_{ik} < b_k^{\text{opt}}. \end{cases} \quad (17)$$

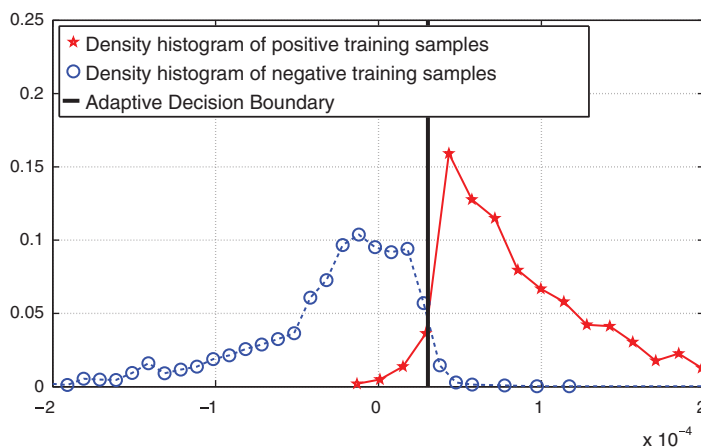


FIG. 2. Optimal decision boundary to minimize misclassification for function “11” (Transcription) (the black vertical line) is different from 0.

Figure 2 shows the adaptive decision boundary for function “11” (transcription) defined in MIPS Funcat annotation scheme (version 2.1) using BioGRID PPI data (version 2.0.45). In the figure, the areas (probability likelihood) of misclassifications are minimized, and the adaptive decision boundary is different from 0.

2.7. Statistical confidence of putative protein function

Many existing protein function prediction approaches only assign “yes” or “no” to a protein when deciding its membership to a function. However, due to the high noise in the biological experiments to generate PPI data, it would be better to estimate the probability of a given prediction rather than simply saying “yes” or “no.” Namely, the statistical confidence to a given prediction is necessary and often of great use in post-processing of proteomic analysis. For example, in order to minimize the experimental time, biologists would decide the order of biological experiments according to the confidence values of the putative protein functions.

Quantitatively evaluating the confidence of a prediction is usually not easy, because the underlying probability model and the actual training and testing data distribution are constantly changing for different biological functions. In this study, we adopt the posterior probability as a metric of the confidence for a prediction due to its clear statistical meaning and explicit computational formula.

Let Y_{ik} be the ground truth membership of protein x_i for the k -th biological function, and $Z_{ik} \in \pm 1$ be the predicted membership, we denoted the confidence for the prediction as $c(Z_{ik})$. Given the prior probabilities, $P(Y_{ik} = \pm 1)$, either computed from the training data or set equally to be 0.5, and the class-conditional densities $p(Z_{ik}|Y_{ik} = \pm 1)$, the posterior probability $P(Y_{ik} = +1|Z_{ik})$ (i.e., the confidence $c(Z_{ik})$) is given by the Bayes’ rule as follows:

$$c(Z_{ik}) = P(Y_{ik} = +1|Z_{ik}) = \frac{p(Z_{ik}|Y_{ik} = +1)P(Y_{ik} = +1)}{\sum_{l=\pm 1} p(Z_{ik}|Y_{ik} = l)P(Y_{ik} = l)}. \quad (18)$$

Hastie et al. (Hastie and Tibshirani, 1998) propose to fit Gaussians to the class-conditional densities $p(Z_{ik}|Y_{ik} = \pm 1)$. The posterior probability is thus a sigmoid, whose slope is determined by the tied variance. Despite its clear intuition and explicit formulation, this approach is seldom useful in real applications because the assumption of Gaussian class-conditional densities is often violated. Figure 3 shows a plot of class-conditional densities $p(Z_{ik}|Y_{ik} = \pm 1)$ for the training data of function “02” (Energy) in MIPS Funcat annotation system. The plot shows histograms of the densities (with bin 0.1 wide), derived from 10-fold cross-validation. Obviously, these densities are far away from Gaussian. In order to tackle this problem, inspired by empirical data, Platt et al. (1999) proposed to fit the conditional-class probabilities implicitly and fit the posterior probability to a parametric form of a sigmoid:

$$c(Z_{ik}) = P(Y_{ik} = +1|Z_{ik}) = \frac{1}{1 + \exp(AZ_{ik} + B)}. \quad (19)$$

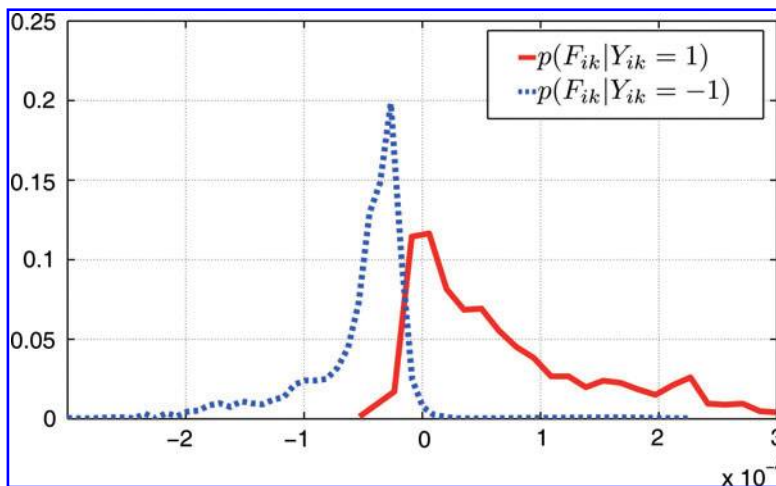
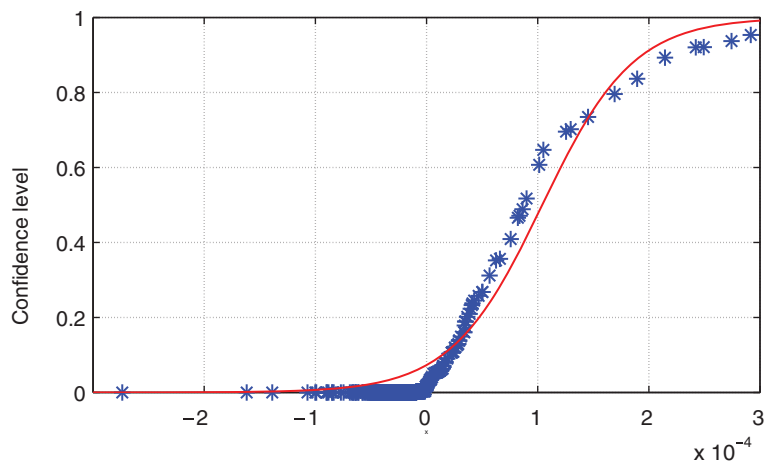


FIG. 3. Class-conditional densities for the training data of function “02” (Energy). The solid red line is $p(Z_{sik}|Y_{sik} = +1)$ while the dashed blue line is $p(Z_{sik}|Y_{sik} = -1)$. Obviously, these two histograms are not Gaussian.

FIG. 4. The fit of a sigmoid to the training data of function “02” (Energy) as shown in Figure 3. Blue markers are the average posterior probabilities computed for the examples falling into a bin of width 0.1. The solid red line is the best-fit sigmoid to the posterior probabilities where $A = -2.124e + 004$ and $B = 2.916$.



The posterior probability defined in Equation (19) takes the form of logistic function, which represents the cumulative distribution function of a big family of exponential distributions. Because of this statistical enrichment, we adopt Equation (19) as our quantitative metric to measure the prediction confidence. The fitted sigmoid curve for the data shown in Figure 3 is plotted in Figure 4.

3. MATERIALS AND DATA SETS

Two types of data are involved in the experimental evaluations for protein function prediction: function annotation data and PPI data. In this section, we describe the data used in this work.

The functional catalogue (FunCat) (Mewes et al., 1999) is a project under the Munich Information Center for Protein Sequences (MIPS), which is an annotation scheme for the functional description of proteins from prokaryotes, unicellular eukaryotes, plants, and animals. Taking into account the broad and highly diverse spectrum of known protein functions, FunCat of version 2.1 consists of 27 main functional categories. Seventeen of them are involved in annotating *Saccharomyces cerevisiae*, which covers general fields such as cellular transport, metabolism, and cellular communication/signal transduction. The main branches exhibit a hierarchical, treelike structure with up to six levels of increasing specificity. Although

TABLE 1. FUNCTION IDS AND NAMES BY FUNCAT SCHEME VERSION 2.1.

<i>ID</i>	<i>Name</i>
'01'	Metabolism
'02'	Energy
'10'	Cell cycle and dna processing
'11'	Transcription
'12'	Protein synthesis
'14'	Protein fate (folding, modification, destination)
'16'	Protein with binding function or cofactor requirement (structural or catalytic)
'18'	Regulation of metabolism and protein function
'20'	Cellular transport, transport facilitation and transport routes
'30'	Cellular communication/signal transduction mechanism
'32'	Cell rescue, defense and virulence
'34'	Interaction with the environment
'38'	Transposable elements, viral and plasmid proteins
'40'	Cell fate
'41'	Development (systemic)
'42'	Biogenesis of cellular components
'43'	Cell type differentiation

there are still other protein annotation systems such, as the Gene Ontology (Ashburner et al., 2000), we use the FunCat annotation system due to its clear treelike hierarchical structure.

The protein–protein interaction data can be downloaded from the BioGRID database (Stark et al., 2006), and we focus on the *S. cerevisiae*. By removing the proteins connected by only one PPI, there are 4299 proteins with 72624 PPIs in the BioGRDI database of version 2.0.45 annotated by FunCat annotation scheme, together with 1997 unannotated proteins. All related 17 level-1 biological functions are listed in Table 1.

4. RESULTS AND DISCUSSIONS

In this article, we proposed a function–function correlated multi-label (FCML) approach for protein function prediction to utilize the correlations among the biological functions to improve the overall prediction performance. Using the PPI data from BioGRID database (Stark et al., 2006) and FunCat annotation scheme (Mewes et al., 1999) on *S. cerevisiae* data, we evaluate our proposed approach and make predictions for unknown proteins.

For statistical metrics, we use the standard *precision* and *F1 score* that have been widely used in previous protein function prediction research work. Let TP (true positive) be the number of proteins that we correctly predict to have a given function, FP (false positive) be the number of proteins that we incorrectly predict to have the function, and FN (false negative) be the number of proteins which we incorrectly predict to not have the function. The “precision” is defined as $\frac{TP}{TP+FP}$, and the “recall” (also known as “sensitivity”) is defined as $\frac{TP}{TP+FN}$. We do not report the specificity of the procedures, because even a trivial algorithm that assigns all proteins to membership of -1 will achieve high specificity due to the unbalanced distribution of positive and negative samples in protein annotation data. In addition, we also use the “F1 score” to evaluate precision and recall together, which is the harmonic mean of precision and recall and defined as following:

$$F1 \text{ score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 score is extensively used in previous related work and other domains such as information retrieval. Typically, improving the precision of an algorithm decreases its recall and vice versa, therefore F1 score is a balanced performance metric.

4.1. Evaluation of the function–function correlations

Because function–function correlations are one of the most important mechanisms to improve the prediction performance in our proposed approach, we first evaluate its correctness. Using the FunCat 2.1 annotation data set for *S. cerevisiae* genome, the function–function correlations defined in Equation (6) are illustrated in the right panel of Figure 5. The high correlation value between functions “40” (Cell Fate) and “43” (Cell Type Differentiation) depicted in this figure shows that they are highly correlated. In addition, as shown in this figure, some other function pairs are also highly correlated, such as functions “11” (Transcription) and “16” (Protein with Binding Function or Cofactor Requirement), “18” (Regulation of Metabolism and Protein Function), and “30” (Cellular Communication/Signal Transduction Mechanism), *etc.* All these observations comply with the biological nature, which justifies the utility of the function correlations from a biological perspective. Figure 5 will also be used to demonstrate the power of function–function correlations in Section 4.4.

4.2. Robustness of adaptive decision boundary

Adaptive decision boundary is another important contribution of this article. Instead of heuristically selecting the thresholds by experience like many existing approaches, adaptive decision boundary method principally computes the thresholds for function assignment from the training data to deal with the unbalanced distribution problem between positive and negative training data. We need to evaluate whether the adaptive decision boundary is robust to the amount of training data to compute it.

In order to evaluate the robustness of the adaptive decision boundaries, we compute them using different amounts of training data and report the corresponding prediction performance in five-fold cross-validation. As a demonstration, we randomly select function “11” (Transcription) and conduct the experiments on its

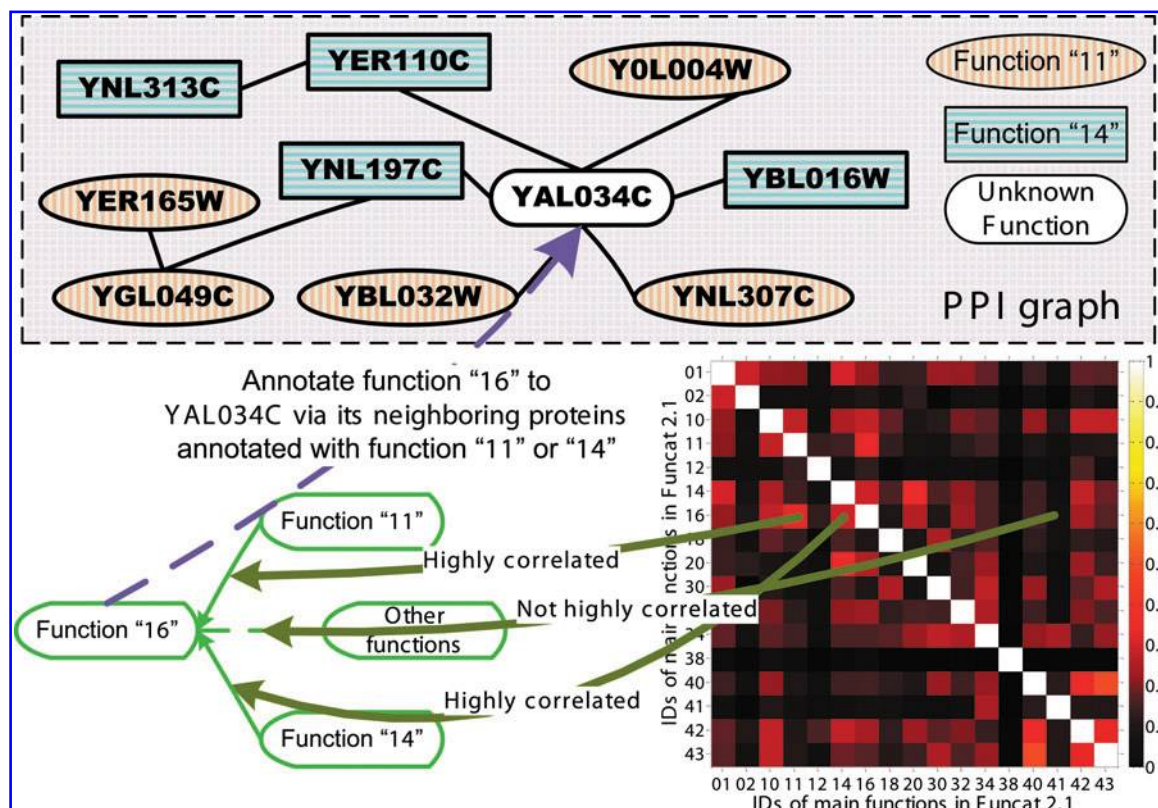


FIG. 5. Right panel: Illustration of the correlation matrix defined in Equation (6) among the 17 main functional categories in FunCat 2.1 annotated to *S. cerevisiae* genome. Left panel: protein, “FUN19” is the testing protein and other neighboring proteins are training data that have functions “11” or “14.” Middle panel: Our FCML method correctly annotates protein “FUN19” with function “16” utilizing function–function correlations.

data. We randomly select 5%, 10%, 20%, 40%, or 80% of the training data to compute the decision boundaries. The prediction performance measured by F1 score is plotted in Figure 6. From the experimental results, we can see that the prediction performance does not degrade much with the decrease of the amount of training data to compute the adaptive decision boundaries. In other words, adaptive decision boundary is a robust thresholding method as long as the amount of data used to compute it is not very small. For example, 5% of the training data is enough to calculate a valid adaptive decision boundary in the demonstration experiments.

4.3. Improved function prediction in cross-validation

We compare the performances of our proposed multi-label Green’s function (MLGF) approach and function–function correlated multi-label (FCML) approach to related commonly used methods, such as majority voting (MV) approach (Schwikowski et al., 2000), global majority voting (GMV) approach (Vazquez et al., 2003), χ^2 approach (Hishigaki et al., 2001), functional flow (FF) approach (Nabieva et al., 2005), and kernel logistic regression (KLR) method (Lee et al., 2006). The PPI graph is built from BioGRID data of version 2.0.45 with annotation by MIPS FunCat scheme of version 2.1. The ten-fold cross-validation is used. For four other approaches, we use their respective optimal parameters. In MV approach, we select the three most frequently occurring functions in a protein’s neighbors. In χ^2 approach, radius = 1 gives the best performance. In FF approach, we assign functions according to the proportions of positive and negative training samples as suggested by Nabieva et al. (2005).

We report the overall prediction performance over all functions using the microaverages of two performance values to address multi-label scenario. The microaverage is computed from the sum of per class contingency table, which can be seen as a weighted average that emphasizes more on the accuracy of classes/functions with more positive samples. The microaveraged precision and F1 score by the compared approaches over all 17 level-1 biological functions are listed in Table 2, which quantify the advantages of

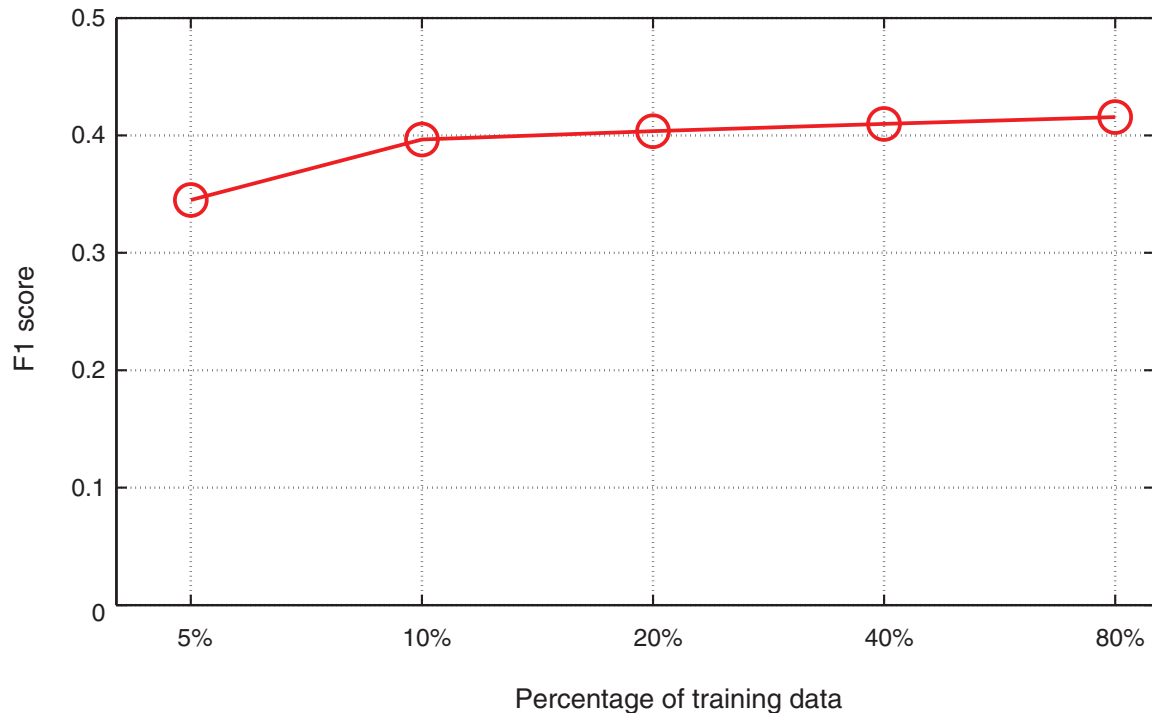


FIG. 6. Evaluation on the effectiveness and robustness of adaptive decision boundary. The performance measured by F1 score vs. the percentage of training data used to compute the adaptive decision boundary for function “11” (transcription).

our FCML approach over the others with more concrete evidence. Moreover, the performances by FCML approach are consistently better than those of MLGF, which demonstrates that incorporating the inherent correlations among biological functions can improve the prediction performance significantly.

4.4. Demonstration of the effectiveness of function–function correlations

In the above experiments, our FCML method outperforms all other methods. We carefully check those testing proteins, which are incorrectly annotated by other methods but correctly annotated by the FCML method. We find function–function correlations absolutely help the function prediction results. Now we will show one example to demonstrate the effectiveness of function–function correlations. For example, in the left panel of Figure 5, protein “FUN19” is the testing protein and other neighboring proteins are training data that have functions “11” (Transcription) or “14” (Protein Fate [Folding, Modification, Destination]).

TABLE 2. MICROAVERAGE OF PRECISION AND F1 SCORE BY SIX APPROACHES IN COMPARISON OVER ALL MAIN FUNCTIONAL CATEGORIES BY FUNCAT SCHEME (MEAN \pm STD)

Approaches	Average precision	Average F1 score
MV	30.69% \pm 1.12%	29.04% \pm 1.05%
GMV	31.13% \pm 2.14%	22.41% \pm 1.75%
χ^2	14.8% \pm 1.21%	7.60% \pm 0.67%
FF	28.01% \pm 1.69%	27.05% \pm 1.54%
KLR	36.81% \pm 2.31%	37.54% \pm 2.69%
MLGF	32.45% \pm 2.45%	36.36% \pm 2.61%
FCML	54.83% \pm 2.78%	43.74% \pm 2.61%

TABLE 3. STATISTICAL CONFIDENCE OF PREDICTED PUTATIVE FUNCTIONS FOR UNANNOTATED PROTEINS

Proteins	Function categories defined in MIPS Funcat annotation scheme																
	"1"	"2"	"10"	"11"	"12"	"14"	"16"	"18"	"20"	"30"	"32"	"34"	"38"	"40"	"41"	"42"	"43"
YAL027W	0.83	0.86	0.55	0.19	0.32			0.12		0.39	0.44	0.49					0.63
YAL034C							0.36	0.12									
YAL053W			0.62				0.87	0.61	0.99	0.77	0.88						
YAR027W	0.49	0.38	0.31	0.19			0.32	0.18	0.55		0.32	0.18					0.44
YAR028W	0.54	0.27				0.72											
YBL046W			0.51	0.67		0.81	0.27	0.24			0.17	0.32		0.17			
YBL049W		0.58	0.64		0.22	0.81	0.51	0.11				0.30	0.03	0.15		0.37	0.28
YBL060W						0.29	0.72	0.29	0.85	0.19		0.72				0.92	
YBL104C					0.56	0.70	0.56		0.56								0.65
YBR025C		0.11	0.59		0.38	0.60	0.21	0.11				0.21					
YBR062C	0.85										0.75	0.23					
YBR094W			0.77		0.44		0.78	0.42			0.18						
YBR096W						0.68	0.48		0.88	0.23							
YBR108W	0.60					0.26			0.39	0.26	0.19					0.48	0.23
YBR137W						0.72	0.47		0.44	0.11							
YBR162C			0.62		0.57					0.16		0.54		0.23			0.41
YBR187W				0.91	0.29		0.48	0.12					0.17		0.39		
YBR194W			0.42	0.45			0.42										
YBR225W	0.66		0.84				0.28				0.15					0.40	0.43
YBR246W		0.50			0.55		0.37				0.39	0.28					
YBR255W	0.40					0.39	0.22		0.34			0.20					
YBR270C	0.44	0.17			0.45				0.58			0.28					
YBR273C	0.34			0.26			0.38							0.10			
YBR280C	0.44					0.40											
YBR287W				0.47			0.26						0.12		0.29		
YCL028W		0.07															
YCL045C	0.71					0.40		0.08	0.66	0.17	0.39	0.22					0.43
YCL056C	0.48		0.79	0.28		0.39	0.21				0.44		0.10				0.20
YCR007C		0.38			0.43	0.33		0.35	0.96	0.18		0.87					0.18
YCR016W	0.79	0.09		0.63		0.35	0.86			0.82	0.23		0.03				
YCR030C								0.13	0.62		0.83				0.50		
YCR043C	0.39		0.72					0.52	0.37	0.23			0.02	0.51	0.64	0.64	
YCR061W						0.42			0.61					0.36	0.86	0.45	
YCR076C		0.13				0.93	0.36		0.49							0.33	
YCR082W			0.62	0.55		0.38	0.33	0.30		0.12	0.12	0.21					
YCR095C	0.50					0.25		0.12		0.10	0.18						
YDL012C	0.92	0.88			0.30		0.71		0.82		0.89			0.13	0.22	0.83	0.27
YDL063C					0.18	0.29	0.46		0.61								
YDL072C						0.55			0.69								
YDL089W		0.07			0.24		0.56				0.25						
YDL091C	0.65	0.17					0.34		0.47				0.01			0.33	
YDL099W									0.28								
YDL121C	0.39	0.08				0.76	0.65		0.34	0.12						0.31	
YDL123W	0.77	0.29		0.31	0.70	0.81		0.17			0.26		0.01	0.75		0.88	
YDL139C						0.67		0.10			0.52		0.14			0.56	
YDL156W	0.48	0.14	0.75	0.39				0.66	0.61				0.26				
YDL167C			0.59	0.55	0.40		0.26										
YDL189W	0.44			0.54	0.25		0.60			0.49	0.38						0.34
YDL204W	0.70		0.61	0.61		0.35	0.27	0.14		0.22	0.56	0.57		0.46		0.68	0.44
YDR049W			0.33	0.97			0.68		0.92	0.52			0.02				0.34
YDR051C			0.35	0.32		0.73	0.52	0.51			0.80	0.26		0.19			
YDR056C	0.93	0.58															
YDR063W	0.45	0.10				0.51							0.13		0.29		

(continued)

TABLE 3. (CONTINUED)

Proteins	Function categories defined in MIPS Funcat annotation scheme																
	"1"	"2"	"10"	"11"	"12"	"14"	"16"	"18"	"20"	"30"	"32"	"34"	"38"	"40"	"41"	"42"	"43"
YDR067C			0.57	0.42										0.01			
YDR068W		0.63		0.67			0.96	0.64	0.34					0.01			0.39
YDR078C			0.70				0.41					0.26					0.48
YDR084C			0.56				0.34	0.40	0.64								
YDR100W	0.74		0.48			0.63	0.68	0.08	0.71	0.54	0.14			0.26		0.64	0.14
YDR105C	0.67	0.42				0.58	0.81		0.80				0.22		0.18		
YDR106W												0.23		0.01			
YDR126W	0.99	0.49	0.99	0.98			0.94	0.44		0.27	0.97	0.28					0.97
YDR128W		0.18			0.21					0.35	0.40	0.58					0.65
YDR132C			0.74	0.61			0.53	0.12					0.01				0.29
YDR134C	0.46		0.36	0.41			0.27										
YDR152W	0.38			0.54			0.26										
YDR161W			0.42		0.49		0.44	0.75		0.53							0.37
YDR186C	0.89	0.22	0.61			0.62		0.74	0.58	0.89	0.32			0.40		0.40	0.69
YDR198C		0.21	0.48	0.31						0.16	0.16	0.39					0.37
YDR222W	0.40	0.15				0.29			0.74			0.42					0.37
YDR233C	0.42	0.07				0.31	0.22		0.68								
YDR239C	0.65					0.76		0.15	0.93	0.68	0.29	0.82					0.14
YDR266C	0.41		0.62	0.49	0.30	0.38	0.33	0.23	0.46		0.36						0.23
YDR326C		0.40			0.34			0.08		0.11		0.28					
YDR339C				0.82		0.62	0.40		0.98		0.29			0.13	0.23		0.21
YDR346C	0.41		0.64	0.53			0.48										
YDR348C	0.37		0.63	0.25		0.26	0.57	0.14									
YDR357C	0.42			0.72		0.55	0.75	0.12				0.13					
YDR361C	0.91	0.42	0.42				0.52			0.26							0.72 0.39
YDR367W					0.16							0.34					
YDR374C	0.50	0.14		0.46	0.17	0.63	0.68	0.08	0.63	0.43	0.50	0.52		0.47			0.68
YDR383C	0.43	0.11	0.51	0.39		0.46	0.26	0.31		0.21	0.29	0.32		0.23		0.66	0.50
YDR411C	0.36							0.11									
YDR458C			0.91	0.32		0.80		0.26		0.26	0.51	0.20	0.06				0.28
YDR475C	0.58			0.41		0.45	0.22	0.09		0.18	0.19	0.22					
YDR476C			0.42				0.60				0.46		0.04				0.48
YDR482C	0.62			0.68			0.41	0.20		0.43			0.04				
YDR486C	0.37	0.07				0.27			0.45								
YDR505C			0.39	0.49		0.32	0.31										
YDR520C	0.50	0.06		0.60								0.29					
YDR532C					0.28	0.65											0.32
YEL001C	0.58	0.07		0.43		0.76		0.16	0.42		0.15						0.40
YEL018W						0.41											0.36
YEL043W			0.57			0.26	0.34	0.08	0.43								0.41 0.26
YEL044W			0.60	0.66				0.08					0.27		0.29		0.43
YEL048C		0.10	0.37		0.16			0.08									
YER004W					0.56	0.97	0.78	0.17	0.98		0.41			0.45	0.60		
YER030W			0.53	0.55			0.30										0.53
YER033C				0.51			0.28			0.18							
YER048W-A	0.73	0.59				0.49		0.12	0.56	0.10				0.11			0.28
YER049W		0.08					0.55	0.13				0.33					
YER067W	0.45	0.17		0.26		0.48	0.53	0.13									
YER071C	0.53	0.13	0.28		0.20	0.57	0.60				0.95	0.15			0.27	0.69	
YER092W			0.40	0.26	0.22	0.58	0.53				0.24						0.43
YER113C		0.09				0.33	0.90		0.32			0.15					0.36
YER128W	0.39	0.08				0.40			0.69								
YER139C	0.44		0.36	0.75		0.54		0.28	0.29	0.76	0.11	0.74					0.17

(continued)

TABLE 3. (CONTINUED)

Proteins	Function categories defined in MIPS Funcat annotation scheme																		
	"1"	"2"	"10"	"11"	"12"	"14"	"16"	"18"	"20"	"30"	"32"	"34"	"38"	"40"	"41"	"42"	"43"		
YER182W					0.79					0.12		0.20		0.15					
YFL034W	0.50	0.08							0.69					0.17		0.31			
YFL062W		0.09			0.18	0.82			0.93		0.38	0.80		0.19	0.25		0.50		
YFR016C			0.41			0.81			0.54	0.16	0.24	0.31					0.64	0.41	
YFR017C	0.48	0.15	0.37																
YFR042W	0.53		0.65			0.57	0.38	0.20		0.17	0.60			0.10				0.28	
YFR043C	0.56		0.59	0.65			0.53	0.50		0.38	0.55								
YFR048W		0.07	0.62			0.33	0.31	0.13			0.15								
YGL010W	0.49		0.61			0.30	0.38						0.25					0.18	
YGL036W			0.92		0.62		0.54		0.18						0.15			0.49	
YGL060W			0.85		0.84								0.63		0.27			0.47	
YGL081W		0.24		0.48			0.76								0.20		0.70		
YGL083W		0.17		0.29		0.62	0.30				0.50				0.22				
YGL108C				0.27			0.76	0.08	0.57					0.01				0.15	
YGL131C				0.32	0.42								0.18						
YGL168W		0.07	0.32			0.32	0.21												
YGL220W		0.13				0.70	0.45	0.34	0.85	0.13	0.12	0.48		0.30				0.78	
YGL231C						0.25	0.43		0.42		0.20							0.27	
YGL242C	0.69	0.21		0.36										0.04					
YGR017W	0.40		0.48	0.65		0.65	0.41	0.21			0.64		0.02	0.13					
YGR058W	0.53	0.10		0.64									0.01					0.17	
YGR068C		0.17	0.28	0.47		0.34	0.72	0.23			0.47								
YGR071C	0.37		0.60			0.32	0.67			0.27	0.24							0.37	0.15
YGR093W	0.53	0.44		0.52	0.93			0.08							0.73				
YGR106C	0.41	0.17							0.69		0.22	0.32							
YGR122W	0.47	0.13		0.32				0.10							0.15				
YGR126W				0.55	0.28		0.49												
YGR130C						0.41	0.78											0.45	
YGR149W						0.82												0.42	
YGR163W	0.80	0.10					0.81	0.18			0.60	0.60		0.58					
YGR187C					0.45	0.80	0.45	0.08			0.19		0.02						
YGR189C	0.83	0.15			0.40	0.43				0.18				0.58				0.67	
YGR196C						0.58		0.30	0.70	0.56	0.36	0.82		0.23		0.45	0.66		
YGR206W	0.44					0.70		0.09	0.32	0.11		0.26							
YGR237C	0.64		0.63	0.82		0.40	0.45	0.11				0.12	0.44						
YGR263C		0.33		0.35		0.25	0.27								0.11				
YGR266W		0.34		0.29									0.20						
YGR271C-A	0.83	0.32		0.92			0.51	0.60										0.40	
YGR283C	0.34	0.08		0.68	0.75	0.35	0.44	0.88											
YGR295C						0.67							0.19						
YHL006C					0.32	0.24		0.35											
YHL014C					0.29			0.19				0.32		0.01					
YHL021C	0.36		0.53	0.64	0.32	0.70	0.54		0.48	0.17	0.36	0.16						0.22	
YHL029C	0.43					0.36	0.20					0.13							
YHL039W	0.65			0.49		0.24	0.47												
YHL042W				0.34							0.18					0.50			
YHR009C						0.39			0.58					0.19		0.44	0.14		
YHR029C	0.94	0.55				0.90	0.99		0.97			0.57		0.98		0.41	0.89		
YHR045W	0.74	0.11					0.29	0.14	0.40		0.34		0.06						
YHR059W				0.91	0.21		0.39	0.22											
YHR087W			0.32	0.33	0.30	0.52	0.45	0.09	0.50										
YHR097C							0.71	0.19	0.69					0.12		0.38			
YHR105W					0.30	0.75			0.48	0.42								0.26	

(continued)

TABLE 3. (CONTINUED)

Function categories defined in MIPS Funcat annotation scheme

<i>Proteins</i>	<i>"1"</i>	<i>"2"</i>	<i>"10"</i>	<i>"11"</i>	<i>"12"</i>	<i>"14"</i>	<i>"16"</i>	<i>"18"</i>	<i>"20"</i>	<i>"30"</i>	<i>"32"</i>	<i>"34"</i>	<i>"38"</i>	<i>"40"</i>	<i>"41"</i>	<i>"42"</i>	<i>"43"</i>
YHR131C			0.31				0.55		0.11								
YHR140W			0.28				0.67		0.80	0.17	0.43					0.80	0.69
YHR151C			0.48	0.64			0.50	0.12									
YHR199C		0.10					0.53	0.08	0.82								0.32
YHR207C			0.40	0.73			0.44							0.01			
YIL023C		0.58	0.98	0.97			0.87	0.32	0.56		0.17		0.15	0.26	0.50	0.93	0.95
YIL027C	0.72	0.41			0.16	0.34							0.02	0.10			0.19
YIL039W	0.52					0.27			0.38								
YIL096C		0.28					0.85	0.75	0.43								
YIL108W						0.40	0.26	0.14	0.82		0.18						0.14
YIL127C	0.92	0.55															
YIL151C			0.56	0.72				0.32						0.21			0.18
YIL152W			0.67			0.41	0.77	0.14		0.23	0.38	0.18		0.55		0.90	0.64
YIL157C	0.74	0.61				0.58	0.86		0.36	0.40							0.50
YIL161W				0.51			0.60	0.11	0.40								
YIR003W	0.46	0.07				0.68	0.29	0.15		0.13	0.16	0.57					0.42
YIR007W		0.33			0.52		0.43	0.21		0.25		0.23					
YJL048C	0.61		0.39	0.66	0.34	0.24			0.37					0.24		0.48	0.22
YJL051W	0.75	0.07	0.70	0.69	0.43	0.82	0.68			0.29						0.51	0.20
YJL057C		0.26		0.70		0.83	0.83			0.19				0.46			0.44
YJL058C		0.12					0.42										
YJL066C						0.79	0.62		0.93	0.40		0.52		0.37		0.30	0.39
YJL082W	0.47	0.14					0.23	0.39			0.19						
YJL097W	0.68	0.18				0.39	0.27		0.41		0.46						0.64
YJL105W	0.50			0.89			0.30	0.18			0.11						
YJL107C	0.46		0.67				0.24			0.10	0.29						0.43
YJL122W		0.21			0.71									0.17			
YJL123C	0.66		0.32		0.23	0.34		0.31	0.47								
YJL149W			0.31			0.29	0.44			0.09				0.12			0.30
YJL151C	0.41	0.09				0.50			0.51		0.23						
YJL162C	0.46	0.27		0.93	0.60		0.80	0.13			0.50	0.41					
YJL171C		0.22					0.75				0.27	0.54	0.11	0.10			
YJL181W			0.40	0.83			0.66	0.08					0.01				0.26
YJL185C		0.09	0.53				0.41	0.20		0.25	0.12	0.38		0.22		0.71	0.31
YJL207C	0.35	0.60						0.11	0.88	0.11	0.36	0.70	0.01	0.24		0.93	0.23
YJR011C	0.38			0.88	0.26	0.41	0.61	0.32		0.09							
YJR061W	0.77					0.58		0.09		0.12	0.15		0.01	0.17			0.17
YJR067C		0.14			0.75		0.26		0.53								
YJR082C			0.63		0.71	0.46	0.70			0.17							0.39
YJR088C	0.50					0.38		0.09									
YJR118C			0.51		0.20		0.40	0.28	0.30	0.21	0.27	0.18	0.01				
YJR134C						0.50			0.36	0.20		0.30					
YKL023W				0.42	0.25		0.22										
YKL037W		0.23				0.61	0.22	0.10	0.33		0.41	0.16					0.49
YKL050C	0.81	0.26	0.47								0.54						
YKL061W	0.42					0.44					0.39	0.27	0.02	0.14			0.48
YKL063C			0.57				0.77	0.48	0.62								
YKL065C							0.26										
YKL069W		0.10				0.36	0.40										
YKL075C			0.60	0.74							0.13						
YKL094W	0.86	0.23							0.71		0.66	0.22					
YKL098W	0.64	0.60	0.44	0.70	0.32						0.29	0.01	0.58			0.33	0.15
YKL151C	0.74			0.51		0.32	0.25		0.47	0.24		0.30					

(continued)

TABLE 3. (CONTINUED)

Proteins	Function categories defined in MIPS Funccat annotation scheme																	
	"1"	"2"	"10"	"11"	"12"	"14"	"16"	"18"	"20"	"30"	"32"	"34"	"38"	"40"	"41"	"42"	"43"	
YKL183W	0.81	0.55		0.68	0.83			0.88	0.49	0.70			0.01	0.11			0.42	
YKL206C			0.39		0.54	0.75			0.48									
YKR071C	0.53	0.20					0.81			0.10	0.47	0.33						
YKR077W	0.41		0.49	0.73	0.26	0.40	0.77					0.17	0.02				0.31	0.16
YKR088C						0.29			0.50									
YKR100C	0.64						0.31				0.11	0.15						
YLL014W				0.52														
YLL023C		0.63	0.83	0.79		0.75	0.55	0.73	0.73	0.30	0.88		0.03		0.63		0.42	
YLL032C		0.73			0.23										0.36	0.56		
YLR021W	0.72	0.07	0.60	0.75				0.33			0.48	0.35		0.26			0.34	
YLR030W			0.54			0.37				0.12	0.13							
YLR031W		0.51	0.52			0.24	0.27	0.07	0.38									0.14
YLR036C	0.76		0.92	0.97		0.41	0.81						0.02		0.49			
YLR050C		0.09				0.52		0.97										
YLR064W		0.16				0.77	0.27	0.68		0.50	0.46		0.16		0.62	0.15		
YLR065C						0.70	0.16	0.81										
YLR072W	0.54				0.19	0.56		0.08			0.15							
YLR108C		0.08			0.66								0.02					
YLR114C	0.64			0.69			0.26				0.13							
YLR173W				0.66	0.70					0.32								
YLR177W				0.71			0.88											
YLR187W	0.68	0.09																
YLR190W			0.33	0.31	0.22	0.60												0.14
YLR196W		0.23	0.61	0.51		0.66	0.34		0.78		0.61	0.58	0.02	0.18		0.77	0.61	
YLR199C		0.14				0.79				0.18	0.31			0.12				0.68
YLR218C	0.50		0.30			0.30		0.33		0.16							0.30	
YLR241W						0.57		0.40	0.09									0.14
YLR253W	0.60	0.26	0.62			0.37				0.20	0.47							0.44
YLR254C				0.45		0.65					0.28	0.75	0.02					
YLR257W	0.80	0.31				0.37	0.24		0.34									0.21
YLR267W			0.74			0.80	0.80	0.51		0.10	0.31	0.18		0.66		0.42	0.43	
YLR287C	0.52			0.37	0.58					0.39								
YLR315W				0.85			0.51	0.16			0.19							
YLR326W		0.11			0.29				0.93	0.49	0.77	0.94		0.17				0.33
YLR352W			0.33			0.27	0.48			0.12				0.13		0.32	0.14	
YLR376C					0.61	0.58					0.25							
YLR392C			0.31	0.32			0.48	0.12		0.22								
YLR407W	0.64	0.11	0.57	0.68			0.33			0.10	0.16	0.21						
YLR408C			0.53	0.34		0.34		0.08				0.20						0.57
YLR413W		0.11				0.46	0.52	0.14	0.73									
YLR426W	0.44	0.07				0.52	0.30			0.21	0.24	0.18		0.13		0.29		
YLR437C	0.44	0.16	0.42			0.38					0.22							0.40
YLR446W		0.29		0.97	0.22						0.46	0.41						
YLR455W	0.64		0.33	0.72		0.31	0.57	0.16						0.15				
YML011C			0.92	0.57						0.19		0.83	0.02	0.51		0.67	0.85	
YML018C		0.08				0.50	0.38	0.27	0.45	0.17		0.20						0.14
YML030W						0.46	0.37	0.08			0.17							
YML036W		0.93		0.93	0.23		0.92	0.59	0.98	0.48	0.81		0.07				0.97	
YML072C		0.10					0.39			0.10			0.02	0.18		0.26		
YML101C	0.48		0.42			0.63				0.12		0.24						
YML119W			0.65		0.36	0.67		0.25	0.64	0.35								0.28
YMR003W		0.13			0.21		0.82		0.39	0.45			0.04					
YMR010W						0.39	0.32				0.17							

(continued)

TABLE 3. (CONTINUED)

Proteins	Function categories defined in MIPS Funcat annotation scheme																
	"1"	"2"	"10"	"11"	"12"	"14"	"16"	"18"	"20"	"30"	"32"	"34"	"38"	"40"	"41"	"42"	"43"
YMR031C	0.58		0.71	0.41		0.55	0.24				0.29	0.20		0.11		0.55	
YMR067C			0.59			0.33	0.25									0.32	0.17
YMR071C		0.07				0.30	0.36	0.12	0.73			0.35					
YMR074C		0.08	0.99	0.90	0.43	0.91	0.98		0.98	0.40	0.97	0.36	0.57	0.79		0.99	0.95
YMR075W	0.42		0.76	0.58		0.47	0.27					0.39					0.34
YMR086W	0.84	0.26	0.51			0.36	0.63	0.12		0.24		0.32		0.27		0.36	0.30
YMR099C	0.67	0.10	0.57			0.36				0.35							
YMR102C	0.45		0.61				0.30	0.09				0.13					
YMR110C			0.48			0.25	0.34					0.13					
YMR111C	0.49	0.09	0.67	0.25			0.51	0.09									
YMR122W-A	0.42	0.09	0.52					0.39						0.02		0.42	
YMR124W	0.45		0.48	0.50		0.33	0.33	0.14			0.45	0.25	0.01	0.17			0.19
YMR144W		0.46	0.65		0.16	0.81	0.49		0.68		0.44	0.37					0.79
YMR163C						0.67			0.82	0.13							
YMR191W	0.53		0.49			0.66	0.42	0.27			0.39			0.14			0.24
YMR221C		0.06				0.90			0.94		0.45	0.43				0.38	
YMR233W			0.33			0.66	0.57	0.22			0.19						
YMR253C	0.66	0.10				0.63		0.27		0.14	0.18	0.15				0.32	
YMR258C						0.25		0.08									
YMR259C	0.38	0.06	0.67	0.72		0.25					0.36	0.38				0.42	
YMR310C	0.80		0.82	0.76			0.54			0.53	0.69	0.52		0.47		0.40	0.17
YNL022C				0.75										0.44			
YNL024C		0.09	0.48	0.46	0.16	0.55	0.24	0.31				0.32		0.14		0.65	
YNL035C	0.66	0.24				0.38	0.44	0.09			0.27	0.20	0.04	0.20		0.65	
YNL046W			0.68		0.45	0.29	0.70	0.09	0.48	0.36		0.78	0.04	0.77		0.32	0.77
YNL056W	0.37					0.40											
YNL087W		0.26		0.37			0.58			0.19	0.19						0.59
YNL092W		0.25	0.74				0.93		0.59				0.05	0.88		0.94	0.81
YNL095C					0.26	0.53		0.19	0.40					0.14			0.31
YNL122C									0.40	0.14							
YNL146W					0.45				0.40					0.10		0.92	
YNL149C	0.63	0.08					0.28					0.16					
YNL155W	0.63	0.22		0.50			0.50	0.20			0.14	0.75		0.79			
YNL157W		0.69	0.41				0.70							0.39		0.92	0.54
YNL181W		0.08				0.38	0.26	0.14	0.60								0.58
YNL212W						0.54	0.35			0.10		0.16					
YNL215W	0.61		0.57		0.79	0.91				0.42		0.36		0.15			0.32
YNL224C				0.77		0.77		0.25	0.90	0.89	0.15	0.61				0.95	0.88
YNL260C	0.65		0.70	0.63	0.47	0.65	0.26	0.32			0.52	0.25					
YNL279W	0.44	0.07				0.25	0.50	0.37						0.13			
YNL300W	0.48			0.39		0.37	0.25	0.07									
YNL310C	0.52	0.07				0.57	0.40	0.23		0.12							0.19
YNL321W			0.35						0.63					0.01			
YNR004W		0.07	0.44	0.65	0.32		0.71	0.29			0.15			0.12		0.36	
YNR009W	0.60		0.65			0.45					0.44	0.30		0.18		0.46	
YNR014W	0.45	0.51	0.61	0.68				0.13			0.35					0.29	
YNR020C			0.49														
YNR021W		0.13							0.57								
YNR024W						0.83					0.12						
YNR065C	0.75			0.64					0.41	0.50	0.51	0.71		0.15		0.42	0.25
YOL070C	0.58		0.89				0.52	0.26		0.11	0.27	0.30	0.01			0.49	
YOL087C		0.83		0.92					0.42			0.80		0.23		0.88	
YOL098C			0.43			0.60	0.51	0.09			0.49	0.20				0.26	0.37

(continued)

TABLE 3. (CONTINUED)

Proteins	Function categories defined in MIPS Funccat annotation scheme																
	"1"	"2"	"10"	"11"	"12"	"14"	"16"	"18"	"20"	"30"	"32"	"34"	"38"	"40"	"41"	"42"	"43"
YOL107W	0.66	0.07				0.55			0.45		0.16	0.22					
YOL131W			0.56	0.68		0.25	0.32	0.16			0.20						0.28
YOL137W						0.38											
YOR006C	0.91	0.96			0.67	0.93		0.16			0.92	0.63					0.91
YOR007C	0.36	0.57				0.95	0.22		0.29					0.35			
YOR042W							0.77			0.38							
YOR044W	0.54	0.16				0.80						0.25					
YOR051C						0.79	0.36		0.36	0.13	0.25			0.17			
YOR059C				0.29			0.32		0.29		0.36	0.53					0.55
YOR066W	0.36		0.61	0.68			0.68										0.27 0.23
YOR086C					0.20		0.32				0.26						0.48 0.15
YOR091W			0.37		0.43		0.58				0.19			0.11			0.26
YOR111W				0.35		0.34	0.61	0.37	0.63	0.16		0.45	0.01	0.27			0.18
YOR112W						0.63	0.56		0.52								
YOR141C		0.07	0.43							0.18	0.26	0.21					0.32
YOR164C	0.48					0.52	0.34										
YOR173W	0.46		0.45	0.46													
YOR175C	0.60	0.08				0.32	0.23		0.80		0.35	0.22	0.01				0.29
YOR189W	0.79		0.46	0.74									0.02				0.45
YOR220W			0.69	0.29		0.39	0.38										0.57 0.14
YOR227W							0.33	0.17			0.16	0.19					
YOR252W		0.08			0.37	0.74	0.54	0.09	0.42								
YOR264W			0.85					0.17			0.35						
YOR289W		0.11			0.71		0.20										
YOR311C						0.55			0.41				0.01				
YOR342C			0.92		0.62	0.41	0.31	0.14			0.12						
YOR352W	0.94	0.83				0.27	0.99	0.08	0.95	0.40	0.35	0.72					
YPL005W				0.30		0.38					0.12						
YPL009C			0.81	0.38			0.67	0.31		0.09	0.34		0.02	0.11			
YPL030W				0.41		0.62	0.71	0.61	0.18		0.82	0.05				0.83	0.20
YPL064C	0.50			0.60			0.29										
YPL066W				0.58			0.21										0.42
YPL077C			0.93	0.45	0.28	0.24	0.59		0.46			0.68					0.83
YPL105C					0.83						0.61						
YPL109C				0.74	0.75		0.91							0.11			0.29
YPL137C	0.78	0.36	0.79		0.29	0.90						0.32					0.77
YPL144W		0.13	0.41	0.88				0.20			0.44	0.18		0.27			
YPL162C						0.96			0.98		0.57						0.56
YPL165C			0.63	0.72				0.09									
YPL166W						0.87	0.54		0.32	0.13				0.13		0.52	0.29
YPL183C			0.85	0.93			0.77										
YPL189C-A	0.44	0.75				0.38	0.85	0.11	0.93	0.33							0.79
YPL199C			0.74	0.61	0.68	0.76	0.97			0.55	0.80	0.73		0.29	0.30		
YPL206C	0.62			0.44		0.34	0.23			0.20		0.19		0.42		0.47	0.14
YPL207W	0.60	0.07			0.19		0.34	0.08			0.20		0.01				
YPL222W				0.67			0.52	0.35									
YPL247C						0.76	0.31	0.12	0.69		0.17						0.20
YPL263C				0.32		0.45	0.34				0.29						
YPL267W			0.68			0.80	0.29	0.52		0.12	0.30						0.66
YPR045C				0.56			0.65	0.15			0.13						
YPR063C	0.67	0.23				0.35			0.63		0.39						0.15
YPR071W	0.85					0.78		0.17	0.93	0.16	0.21						
YPR114W			0.75						0.29				0.01	0.14			

(continued)

TABLE 3. (CONTINUED)

Proteins	Function categories defined in MIPS Funcat annotation scheme																	
	"1"	"2"	"10"	"11"	"12"	"14"	"16"	"18"	"20"	"30"	"32"	"34"	"38"	"40"	"41"	"42"	"43"	
YPR116W	0.22			0.48	0.48								0.02	0.13			0.29	
YPR148C	0.10	0.43			0.29	0.34		0.71	0.19		0.19		0.33				0.57	
YPR152C				0.67	0.16		0.28	0.08		0.13			0.12					
YPR153W	0.98	0.57	0.91	0.85		0.98	0.96	0.62	0.98		0.98						0.89	0.72
YPR174C		0.49	0.36	0.23	0.41	0.51	0.41				0.17							

In experimental results, protein "FUN19" is annotated with functions "11" and "14" by all six methods. But it is annotated with function "16" (Protein with Binding Function or Cofactor Requirement [Structural or Catalytic]) only by our FCML approach and not by the other approaches. We observe that no proteins directly interacting with "FUN19" are annotated with function "16," and only a small fraction (90 out of 355) of proteins indirectly interacting with "FUN19" via an intermediate protein are annotated with function "16." Thus, all five other methods fail to annotate protein "FUN19" with function "16."

However, a majority of proteins directly interacting with "FUN19" are annotated with either function "11" or function "14." By scrutinizing the function–function correlation matrix computed from Equation (6) as shown in the right panel of Figure 5, we can see that function "16" has the highest statistical correlations with functions "11" and "14." Utilizing such function–function correlations, our FCML method correctly annotates protein "FUN19" with function "16" as shown in the middle panel of Figure 5. In other words, the functionwise correlations play a significant role to improve overall predictive accuracy in protein function annotations.

4.5. Prediction and putative functions of unannotated proteins

We apply the proposed FCML approach on the BioGRID data annotated by the MIPS Funcat scheme and predict functions for the unannotated proteins. A list of all putative function predictions for level-1 functions in MIPS Funcat scheme by our algorithm is provided in Table 3, which is supplied in the Appendix of this article. In addition to predicted functions, we also report the corresponding statistical confidence values. For example, we annotate function "11" (Transcription) with statistical confidence of 0.83 and function "32" (Cell Rescue, Defense and Virulence) with statistical confidence of 0.12 to protein "YNR024W." Namely, our experimental results suggest that protein "YNR024W" is more likely to be annotated with function "11" than function "32."

5. CONCLUSIONS

We proposed a novel function–function correlated multi-label (FCML) protein function prediction approach and showed its promising performance, which outperforms other related approaches. Different from most existing approaches that divide protein function prediction into multiple separate tasks and make predictions fundamentally one function at a time, the proposed FCML approach considers all the biological functions as a single correlated prediction target and predict protein functions via an integral procedure. In the proposed approach, correlations among the functional categories are leveraged. By formulating protein function prediction as a multi-label classification problem, we use the Green's function over a graph to efficiently resolve the problem. The Green's function approach takes advantage of both the full topology of the interaction network toward global optimization and the local structures, such that the deficiencies lying in the existing approaches can be overcome. In addition, we propose an adaptive decision boundary method to deal with the unbalanced distribution of protein annotation data and quantify the statistical confidence of predicted functions for post-processing of proteomic analysis.

6. APPENDIX

6.1. Predicted putative functions for unannotated proteins and corresponding statistical confidence

We apply the proposed FCML approach on the BioGRID data and predict functions for the unannotated proteins. We use MIPS Funcat annotation scheme. A list of all putative function predictions for level-1 functions in MIPS Funcat scheme by our algorithm is provided in Table 3. The nonempty cells indicate the predicted putative functions of the corresponding protein. For example, protein “YAL034C” is predicted to have functions “16” (Protein with Binding Function) and “18” (Regulation of Metabolism and Protein Function).

In addition to predicted putative functions, we also report the corresponding statistical confidence values. For example, we annotate function “11” (Transcription) with statistical confidence of 0.83 and function “32” (Cell Rescue, Defense and Virulence) with statistical confidence of 0.12 to protein “YNR024W”. Namely, our experimental results suggest that protein “YNR024W” is more likely to be annotated with function “11” than function “32”.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

- Ashburner, M., Ball, C., Blake, J., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25.
- Chua, H., Sung, W., and Wong, L. 2006. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* 22, 1623–1630.
- Chung, F. 1997. *Spectral Graph Theory*. American Mathematical Society, Providence, RI.
- Deane, C., Salwinski, L., Xenarios, I., and Eisenberg, D. 2002. Protein interactions two methods for assessment of the reliability of high throughput observations*. *Molecular & Cellular Proteomics* 1, 349–356.
- Ding, C., Simon, H., Jin, R., et al. 2007. A learning framework using Green’s function and kernel regularization with application to recommender system. In *Proc. of ACM SIGKDD 2007*, 260–269.
- Edgar, R., Domrachev, M., and Lash, A. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207.
- Giot, L., Bader, J., Brouwer, C., et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* 302, 1727–1736.
- Harbison, C., Gordon, D., Lee, T., et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104.
- Hastie, T., and Tibshirani, R. 1998. Classification by pairwise coupling. *Annals of Statistics* 451–471.
- Hishigaki, H., Nakai, K., Ono, T., et al. 2001. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast* 18, 523–531.
- Ho, Y., Gruhler, A., Heilbut, A., et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183.
- Karaoz, U., Murali, T., Letovsky, S., et al. 2004. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl. Acad. Sci. USA* 101, 2888–2893.
- Klein, D., and Randić, M. 1993. Resistance distance. *J. Math. Chem.* 12, 81–95.
- Lee, H., Tu, Z., Deng, M., et al. 2006. Diffusion kernel-based logistic regression models for protein function prediction. *Omics* 10, 40–55.
- Mewes, H., Heumann, K., Kaps, A., et al. 1999. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 27, 44.
- Nabieva, E., Jim, K., Agarwal, A., et al. 2005. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 21, 302–310.
- Pei, P., and Zhang, A. 2005. A topological measurement for weighted protein interaction network. In *2005 IEEE Computational Systems Bioinformatics Conference*, 268–278.
- Platt, J. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*.

- Schwikowski, B., Uetz, P., and Fields, S. 2000. A network of protein-protein interactions in yeast. *Nat. Biotechnol.* 18, 1257–1261.
- Sharan, R., Ulitsky, I., and Shamir, R. 2007. Network-based prediction of protein function. *Mol. System Biol.* 3.
- Stark, C., Breitkreutz, B., Reguly, T., et al. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535.
- Tong, A., Lesage, G., Bader, G., et al. 2004. Global mapping of the yeast genetic interaction network. *Science* 303, 808–813.
- Vazquez, A., Flammini, A., Maritan, A., et al. 2003. Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.* 21, 697–700.
- Von Mering, C., Krause, R., Snel, B., et al. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399–403.
- Wang, H., Ding, C., and Huang, H. 2010a. Directed graph learning via high-order co-linkage analysis. In *Proc. of ECML/PKDD 2010*, 451–466.
- Wang, H., Ding, C., and Huang, H. 2010b. Multi-label classification: Inconsistency and class balanced k-nearest neighbor. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.
- Wang, H., Ding, C., and Huang, H. 2010c. Multi-label linear discriminant analysis. In *Proc. of ECCV 2010*, 126–139.
- Wang, H., Huang, H., and Ding, C. 2009. Image annotation using multi-label correlated greens function. In *Proc. of IEEE ICCV 2009*, 2029–2034.
- Wang, H., Huang, H., and Ding, C. 2010d. Multi-label feature transform for image classifications. In *Proc. of ECCV 2010*, 793–806.
- Wang, H., Huang, H., and Ding, C. 2011. Image annotation using bi-relational graph of images and semantic labels. In *Proc. of IEEE CVPR 2011*, 793–800.

Address correspondence to:
Dr. Heng Huang
University of Texas at Arlington
Computer Science and Engineering
Box 19015
416 Yates St.
Arlington, TX 76019

E-mail: heng@uta.edu