JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 22, Number 6, 2015
© Mary Ann Liebert, Inc.
Pp. 546–562
DOI: 10.1089/cmb.2014.0172

Correlated Protein Function Prediction via Maximization of Data-Knowledge Consistency

HUA WANG[1] HENG HUANG[2] and CHRIS DING[2]

ABSTRACT

Conventional computational approaches for protein function prediction usually predict one function at a time, fundamentally. As a result, the protein functions are treated as separate target classes. However, biological processes are highly correlated in reality, which makes multiple functions assigned to a protein not independent. Therefore, it would be beneficial to make use of function category correlations when predicting protein functions. In this article, we propose a novel Maximization of Data-Knowledge Consistency (MDKC) approach to exploit function category correlations for protein function prediction. Our approach banks on the assumption that two proteins are likely to have large overlap in their annotated functions if they are highly similar according to certain experimental data. We first establish a new pairwise protein similarity using protein annotations from knowledge perspective. Then by maximizing the consistency between the established *knowledge similarity* upon annotations and the *data similarity* upon biological experiments, putative functions are assigned to unannotated proteins. Most importantly, function category correlations are gracefully incorporated into our learning objective through the knowledge similarity. Comprehensive experimental evaluations on the *Saccharomyces cerevisiae* species have demonstrated promising results that validate the performance of our methods.

Key words: protein function prediction, symmetric nonnegative matrix factorization.

1. INTRODUCTION

DUE TO ITS SIGNIFICANT IMPORTANCE IN post–genomic era, protein function prediction has been extensively studied and many computational approaches have been proposed in the past two decades. Among numerous existing algorithms, graph-based approaches and data integration–based approaches have demonstrated to be effective due to their clear connections to the biological facts.

Since many biological experimental data can be readily represented as networks, learning on graphs is one of the most natural ways to predict protein functions (Sharan et al., 2007). Neighborhood-based methods (Schwikowski et al., 2000; Hishigaki et al., 2001; Chua et al., 2006, 2007) assign functions to a protein based on the most frequent functions within a neighborhood of the protein, and they mainly differ in how the ‘‘neighborhood’’ of a protein is defined. Network diffusion-based methods (Nabieva et al., 2005;

[1]Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, Colorado.
[2]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, Texas.

Weston et al., 2004) view the interaction network as a flow network, on which protein functions are diffused from annotated proteins to their neighbors in various ways. Other function prediction approaches via biological networks include graph cut–based approaches (Vazquez et al., 2003; Karaoz et al., 2004), and those derived from kernel methods (Liang et al., 2008). Recently, two graph-based protein function prediction methods (Wang et al., 2012, 2013) using protein–protein interaction (PPI) graphs were developed to take advantage of the function–function correlations by considering protein function prediction as a multilabel classification problem, which took the same perspective as ours in the current work. Jiang et al. (2008) also proposed to utilize function–function similarity, though not explicitly, through the tree approximation of gene ontology.

Experimental data from one single source are usually incomplete, sometimes even misleading (Whisstock and Lesk, 2004); therefore, predicting protein function using multiple biological data has also attracted increased attention. For example, Lanckriet et al. (2004) proposed a kernel-based data fusion approach to integrate multiple experimental data via a hybrid kernel and use support vector machine (SVM) for classification. Tsuda and Noble (2004) presented a locally constrained diffusion kernel approach to combine multiple types of biological networks. Artificial neural network was employed in Shi et al. (2009) to integrate different protein interaction data.

All above conventional computational approaches usually consider protein function prediction as a standard classification problem (Lanckriet et al., 2004; Shin et al., 2007; Sun et al., 2008). That is, these approaches predict one function at a time, fundamentally. As a result, the classification for each functional category is conducted independently. However, in reality most biological functions are highly correlated, and protein functions can be inferred from one another through their interrelatedness. The function category correlations, albeit useful, are not fully utilized in predicting protein function in the earlier works. In this study, we explore this special characteristic of the protein functional categories and take advantage of the function–function correlations to improve the overall predictive accuracy of protein functions.

1.1. Multilabel correlated protein function prediction

Because a protein is usually observed to play several functional roles in different biological processes within the same organism, it is natural to annotate it with multiple functions. Therefore, protein function prediction is a *multilabel classification* (Wang et al., 2009, 2010a, 2011a) problem. The essential difference between single-label classification and multilabel classification lies in that (Wang et al., 2009, 2010a, 2011a) classes in the former are assumed to be mutually exclusive, while those in the latter are generally correlated to each other. Multilabel data, such as those used in protein function prediction, present a new opportunity to improve classification accuracy through label correlations, which are absent in single-label data. For example, when applying Functional Catalogue (FunCat) annotation scheme (version 2.1) (Mewes et al., 1999) on *Saccharomyces cerevisiae* genome, we observe that there is a big overlap between the proteins annotated to function ''Cell Fate'' (ID: 40) and those annotated to ''Cell Type Differentiation'' (ID: 43). As shown in the left panel of Figure 1, among 268 proteins annotated with function ''Cell Fate'' in

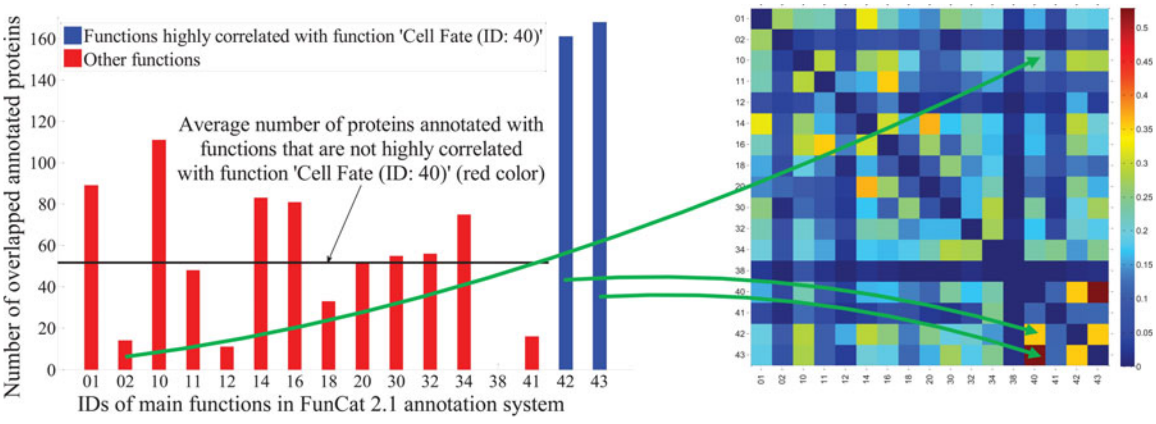

FIG. 1. (*Left*) Number of proteins annotated with both function ''Cell Fate'' (ID: 40) and one of the other functions. (*Right*) Visualization of the correlation values defined by Equation (2) among the 17 main functions defined in FunCat 2.1 to the *S. cerevisiae* genome.

S. cerevisiae genome, 168 proteins are also annotated with function ''Cell Type Differentiation,'' whereas the average number of proteins annotated with other functions is only about 51. From this observation, we can reasonably speculate that these two functions are statistically correlated in a stronger way. That is, if a protein is known to be annotated with function ''Cell Fate'' by either experimental or computational evidences, we have high confidence to annotate the same protein with function ''Cell Type Differentiation'' as well.

1.2. Data-knowledge consistency and our motivations

In protein function prediction, we need both experimental data and biological knowledge. Here we refer to *data* as original experimental measurements or results, such as protein sequences, PPI networks measured by yeast two-hybrid screening, gene expression profiles, etc. On the other hand, *knowledge* refers to human-curated research findings recorded in well structured databases or documented in biomedical literatures, such as human-encoded annotation databases, ontologies, etc.

In most existing approaches for protein function prediction, knowledge is routinely used as supervision in the classification tasks, that is, protein annotations are interpreted as labels assigned to data points. In this study, we employ knowledge information from a new perspective. Motivated by the observation that label indications in a multilabel classification task (i.e., protein function annotations in protein function prediction problems) convey important attribute information (Wang et al., 2010b), we use the function annotations of a protein as its description and assess pairwise protein similarities upon such descriptions. The key assumption of our work is that two proteins are likely to have large overlap in their annotated functions, if they are highly similar according to experimental data. More precisely, let $\mathbf{x}_i$ and $\mathbf{x}_j$ be the descriptions of two proteins abstracted from experimental data, and $\mathbf{f}_i$ and $\mathbf{f}_j$ be annotated functions of these two proteins respectively; we evaluate the similarity between the two proteins in the following two different ways. The first one is based upon experimental data and denoted as $\mathcal{S}_D(\mathbf{x}_i, \mathbf{x}_j)$, while the second one is based upon biological knowledge and denoted as $\mathcal{S}_K(\mathbf{f}_i, \mathbf{f}_j)$. If functions $\mathbf{f}_i$ and $\mathbf{f}_j$ are annotated appropriately to proteins $\mathbf{x}_i$ and $\mathbf{x}_j$—that is, the data and the knowledge are consistent—we expect that these two similarity measurements appear to be close:

$$\mathcal{S}_D(\mathbf{x}_i, \mathbf{x}_j) \approx \mathcal{S}_K(\mathbf{f}_i, \mathbf{f}_j). \tag{1}$$

With this assumption, we may determine the optimal function assignments to unannotated proteins by minimizing the difference between these two sets of similarities, that is, maximizing the consistency between experimental data and biological knowledge. In this article, we formalize this assumption by proposing our Maximization of Data-Knowledge Consistency (MDKC) approach. Through the knowledge similarity $\mathcal{S}_K(\mathbf{f}_i, \mathbf{f}_j)$, function category correlations are incorporated such that the predictive performance is expected to be enhanced.

1.3. Notations and problem formalization

In protein function prediction, we are given K biological functions and n proteins. Without losing generality, we assume the first l proteins are annotated, and our goal is to predict functions of the rest $n-l$ unannotated proteins.

Let $\mathbf{x}_i \in \mathbb{R}^p$ denote a protein, which is a vector description of the i-th protein abstracted from certain biological experimental data, such as the amino acid histogram of a protein sequence, etc. The pairwise similarities among the proteins are modeled as a symmetric matrix $W \in \mathbb{R}^{n\times n}$, where W_{ij} measures how similar proteins $\mathbf{x}_i$ and $\mathbf{x}_j$ are related. W is usually seen as edge weight matrix of a graph where proteins correspond to vertices. In the simplest case of a PPI network, $W_{ij}=1$ indicates that proteins $\mathbf{x}_i$ and $\mathbf{x}_j$ interact, and 0 otherwise. Every protein is assigned with a number of biological functions, which are described by a function annotation vector $\mathbf{y}_i \in \{0, 1\}^K$, such that $\mathbf{y}_i\ (k)=1$ if protein $\mathbf{x}_i$ is annotated with the k-th function, $\mathbf{y}_i\ (k)=0$ if it is not annotated with the k-th function or unannotated. The equation $\{\mathbf{y}_i\}_{i=1}^{l}$ for the first l annotated proteins are known, and our objective is to learn $\{\mathbf{y}_i\}_{i=l+1}^{n}$ for the $n-l$ unannotated proteins. We write $Y=[\mathbf{y}_1, \dots, \mathbf{y}_n]^T=[\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(K)}]$, where $\mathbf{y}^{(k)} \in \mathbb{R}^n$ is a classwise function annotation vector. Besides the ground truth function assignment matrix Y, we also define $F=[\mathbf{f}_1, \dots, \mathbf{f}_n]^T \in \mathbb{R}^{n\times K}$ as the predicted function assignment matrix, where $F_{ik}=\mathbf{f}_i\ (k)$ for $l+1 \le i \le n$ indicates our confidence to assign the k-th function to an unannotated protein $\mathbf{x}_i$.

2. FORMULATION OF FUNCTION CATEGORY CORRELATIONS

Before we proceed to the algorithm development of our new approach, we first explore and formalize the function category correlations, because they are one of our most important mechanisms to boost protein function prediction performance.

As shown in the left panel of Figure 1, proteins assigned to two different functions may overlap. Statistically, the bigger the overlap, the more closely two functions are related. Therefore, functions assigned to a protein are no longer independent. Instead, they can be inferred from one another. In the extreme case, such as in parent–child hierarchy of protein function annotation systems, once we know a protein is annotated to a child function, we can immediately annotate all the ancestor functions to the same protein.

Using cosine similarity, we define a function category correlation matrix, $C \in \mathbb{R}^{K \times K}$, where C_{kl} captures the correlation between the k-th and l-th functions as follows:

$$C_{kl} = \cos\left(\mathbf{y}^{(k)}, \mathbf{y}^{(l)}\right) = \frac{\langle \mathbf{y}^{(k)}, \mathbf{y}^{(l)} \rangle}{\|\mathbf{y}^{(k)}\| \|\mathbf{y}^{(l)}\|}, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors and $\|\cdot\|$ denotes the ℓ_2 norm of a vector.

Using FunCat annotation scheme on *S. cerevisiae* genome, function correlations defined in Equation (2) are illustrated in the right panel of Figure 1. The high correlation value between functions ‘‘Cell Fate’’ and ‘‘Cell Type Differentiation’’ shown in the figure implies that they are highly correlated, which agrees with the observations shown in the left panel. In addition, as can be seen in the right panel of Figure 1, some other function pairs are also highly correlated, such as ‘‘Transcription’’ and ‘‘Protein with Binding Function or Cofactor Requirement,’’ ‘‘Regulation of Metabolism and Protein Function,’’ and ‘‘Cellular Communication/Signal Transduction Mechanism,’’ etc. All these observations strictly comply with the biological truth, which firmly justifies the correctness of our formulation for function category correlations in Equation (2) from the biological perspective.

3. MAXIMIZATION OF DATA-KNOWLEDGE CONSISTENCY (MDKC) APPROACH

We assume that two proteins tend to have large overlap in their assigned functions if they are very similar in terms of some experimental data. In order to predict protein functions upon this assumption, we evaluate the similarity between two proteins in the following two ways, one by experimental data called as *data similarity*, and the other by biological knowledge called as *knowledge similarity*. We denote the former as $\mathcal{S}_D(\mathbf{x}_i, \mathbf{x}_j)$ and the latter as $\mathcal{S}_K(\mathbf{f}_i, \mathbf{f}_j)$. If functions annotated to proteins are consistent with their experimental data, we would expect the data similarity is close to the knowledge similarity:

$$\min \sum_{i,j} \left[\mathcal{S}_D(\mathbf{x}_i, \mathbf{x}_j) - \mathcal{S}_K(\mathbf{f}_i, \mathbf{f}_j)\right]^2, \tag{3}$$

$$s.t. \quad \mathbf{f}_i = \mathbf{y}_i, \forall 1 \le i \le l, \tag{4}$$

where the constraint in Equation (4) fixes the functions assigned to annotated proteins to be ground truth. The optimization objective in Equation (3) minimizes the overall difference between the two types of similarities, which thereby maximizes the data-knowledge consistency.

3.1. Optimization framework of the MDKC approach

In protein function prediction, the data similarity is already known *a priori*. Namely, $\mathcal{S}_D(\mathbf{x}_i, \mathbf{x}_j) = W$, and W depends on input experimental data. For example, when input data are a PPI network, W could be the adjacency matrix of the PPI graph in the simplest case or any derived topological similarity; when input data are protein sequences, W could be the inverse Euclidean distances of amino acid histogram vectors, etc. Because W is input dependent, we defer its detailed definitions to section 4 according to the experimental data used in the respective empirical evaluations.

Now we consider knowledge similarity. The simplest method to evaluate the knowledge similarity is to count the number of common annotated functions of two proteins: $\mathbf{f}_i^T \mathbf{f}_j$. However, the problem of this straightforward similarity measurement lies in that it considers all the biological functions to be independent and is unable to explore the correlations among them. In particular, it will give zero similarity whenever two proteins do not share any annotated functions, although they could be strongly related if their annotated functions are highly correlated. For example, given a pair of proteins, one annotated with function ''Cell Fate'' and the other annotated with function ''Cell Type Differentiation,'' although they may not share any common functions, they should still have certain similarities, either biologically or statistically, as illustrated in Figure 1. In the extreme case, in the parent–child annotation system, such as the FunCat scheme used in this work, if protein $\mathbf{x}_i$ is annotated with one of the ancestor functions of protein $\mathbf{x}_j$'s annotated function, the two proteins are closely related even though they do not share any common functions. Therefore, in order to capture correlations among different functions, instead of the dot product, we compute the knowledge similarity as following:

$$\mathcal{S}_K(\mathbf{f}_i, \mathbf{f}_j) = \mathbf{f}_i^T C^{-1} \mathbf{f}_j = \mathbf{f}_i^T A \mathbf{f}_j, \tag{5}$$

where, for notation brevity, we denote $A = C^{-1}$ in the sequel.

Note that compared to the inner product similarity defined by $\mathbf{f}_i^T \mathbf{f}_j$ based on the Euclidean distance,[1] the knowledge similarity computed by Equation (5) is based on the Mahalanobis distance,[2] where C acts as the covariance matrix that encodes the human-curated prior knowledge for the biological species of interest. Statistically speaking, because the Euclidean distance is independent of input data while the Mahalanobis distance captures the second-order statistics of the input data, the latter is able to better characterize the relationships between the data points of a given input data set when its distribution is known *a priori*. In protein function prediction, the Euclidean distance based knowledge similarity is independent of the concerned biological species. In contrast, the Mahalanobis distance based knowledge similarity is specific to the biological species of interest, which thereby has increased statistical power. Most importantly, function–function correlations, the most important advantage of a multilabel data set over the traditional single-label data set, are exploited for the later protein annotations tasks, which is an important contribution of the proposed method.

Utilizing the knowledge similarity defined in Equation (5), we can formalize the data-knowledge consistency assumption in Equation (3) by the following optimization problem:

$$\min_F \sum_{i,j=1}^{n} \left(W_{ij} - \sum_{k,l=1}^{K} F_{ik} A_{kl} F_{jl} \right)^2, \tag{6}$$

$$s.t. \quad F_{ik} = Y_{ik}, \forall 1 \le i \le l, \ 1 \le k \le K. \tag{7}$$

In standard classification problems in machine learning, F_{ik} $(1 \le i \le l)$are fixed for labeled data points. Specifically, a big F_{ik} indicates that data point $\mathbf{x}_i$ belongs to the k-th class, while a small F_{ik} indicates that $\mathbf{x}_i$ does not belong the k-th class. However, this assumption does not hold in the problem of protein function prediction. For an annotated protein, its associated functions refer to those who have certain experimental supports for the associations between this protein and the functions. On the other hand, the non-association between a protein and a function only means that we currently do not have any biological or computational evidence for the corresponding association. In reality, however, the protein could be annotated with the concerned function. And the exact goal of computational methods for protein function prediction is to identify putative protein functions, which could work as the candidates for further experimental screening. As a result, instead of using the hard constraints in Equation (7), it is reasonable to relax the confidence variables F_{ik} $(1 \le i \le l)$for annotated proteins to be dynamic variables, which approximate the ground truth function assignments. The constraint in Equation (7) hence can be written to minimize the following penalty function:

$$\alpha \sum_{i=1}^{l} \sum_{k=1}^{K} (Y_{ik} - F_{ik})^2, \tag{8}$$

[1] The Euclidean distance between two vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ is defined as $d(i,j) = \sqrt{(\mathbf{f}_i - \mathbf{f}_j)^T (\mathbf{f}_i - \mathbf{f}_j)} = \sqrt{\mathbf{f}_i^T \mathbf{f}_i + \mathbf{f}_j^T \mathbf{f}_j - 2\mathbf{f}_i^T \mathbf{f}_j}$. Following the standard way to convert a distance measurement to a similarity measurement, we can consider $\mathbf{f}_i^T \mathbf{f}_j$ to be a similarity measurement between the two vectors $\mathbf{f}_i$ and $\mathbf{f}_j$, which are based on the Euclidean distance.

[2] Given a covariance matrix C of a data distribution, the Mahalanobis distance between two vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ is defined as $d_C(i,j) = \sqrt{(\mathbf{f}_i - \mathbf{f}_j)^T C^{-1} (\mathbf{f}_i - \mathbf{f}_j)} = \sqrt{\mathbf{f}_i^T C^{-1} \mathbf{f}_i + \mathbf{f}_j^T C^{-1} \mathbf{f}_j - 2\mathbf{f}_i^T C^{-1} \mathbf{f}_j}$. Thus, $\mathbf{f}_i^T C^{-1} \mathbf{f}_j$ defines a similarity between the two vectors $\mathbf{f}_i$ and $\mathbf{f}_j$, which is based on the Mahalanobis distance specific to the input data.

where $\alpha > 0$ controls the relative importance of this penalty. Following the experiences in graph-based semi-supervised learning (Wang et al., 2009, 2012), we empirically set $\alpha = 0.1$ in all our experiments.

Finally, we write our objective in a more compact way using matrices to minimize the following:

$$J_{\text{MDKC}}(F) = \|W - FAF^T\|^2 + 2\alpha\mathbf{tr}\left((Y-F)^T V(Y-F)\right), \quad s.t. \quad F \geq 0, \tag{9}$$

where $\|\cdot\|$ denotes the Frobenius norm of a matrix and $\mathbf{tr}\,(\cdot)$ denotes the trace of a matrix. Here $V \in \mathbb{R}^{n\times n}$ is a diagonal indicator matrix, whose diagonal entry $V_{ii} = 1$ if the i-th protein is an annotated protein, while $V_{ii} = 0$ indicates that the i-th protein is unannotated. In Equation (9), the constraint $F \geq 0$ is naturally enforced because W is nonnegative by definition. Most importantly, with this nonnegative constraint, Equation (9) will be enriched with clustering interpretation as detailed later, which makes the mathematical formulation of the proposed method more meaningful from the machine learning perspective.

We call Equation (9) as our proposed Maximization of Data-Knowledge Consistency (MDKC) approach. Upon solving Equation (9), we can assign putative functions to an unannotated protein—say, the i-th protein—via $\mathbf{f}_i$.

3.1.1. A more in-depth look at MDKC approach. Equation (9) reveals the insight of our MDKC approach, that is, we attempt to approximate the data similarity using the knowledge similarity. The former is already known in biological experiments thereby fixed, while only part of the latter is known for annotated proteins and our task to identify the part for unannotated proteins. When non-negativity is enforced to F, min $\|W - FCF^T\|^2$ is a non-negativity matrix factorization (NMF) problem (Lee and Seung, 2001; Ding et al., 2010) as $W_{ij} > 0$ by definition.

The true power of our approach lies in that NMF is equivalent to spectral clustering when W is the edge weight matrix of a graph. This equivalence has been proved by Ding et al. (2006, 2010). Therefore, our approach can be seen to seek the shared structures of function assignments upon the topological modularity of an input graph from experimental data. Compared to traditional spectral clustering algorithms, which are unsupervised and their annotation information cannot be used, our approach is able to exploit the information conveyed by both labeled and unlabeled data. Most importantly, function category correlations are taken into account, thereby more information is used for protein function prediction.

3.2. Computational algorithm of the MDKC approach

Mathematically, Equation (9) is a regularized NMF problem (Cai et al., 2008; Gu and Zhou, 2009; Cai et al., 2010). Although the optimization techniques for the NMF problem and its variants have been extensively studied in literature (Ding et al., 2006, 2010; Cai et al., 2008; Gu and Zhou, 2009; Cai et al., 2010), solving Equation (9) is very challenging. Most, if not all, existing algorithms to solve NMF problems are only able to deal with rectangle input matrices (a rectangle matrix is a matrix whose number of the rows is different from that of its columns) or asymmetric square matrices, but not symmetric input matrices such as the one used in our objective in Equation (9). This is because the latter involves a fourth-order term due to the symmetric usage of the factor matrix F (Wang et al., 2011b,c), which inevitably complicates the problem. Traditional solutions to symmetric NMF typically rely on heuristics (Ding et al., 2006; Li et al., 2007), hence we propose a principled solution via proving a new generic matrix inequality as presented below. We introduce a new algorithm to solve Equation (9) in Algorithm 1. We prove its correctness and convergence as follows.

Algorithm 1: Algorithm to solve Eq. (9)

```
Data: 1. Data similarity matrix $W$.
2. Function–function correlations matrix $C$.
3. Indication matrix $Y$ derived from labels of annotated proteins.
Result: Factor matrices $F$.
1. Computer $A = C^{-1}$.
2. Initialize $F$ following (Ding et al., 2006).
repeat
    3. Compute $F_{ij} \leftarrow F_{ij}\left[\frac{(WFA + \alpha VY)_{ij}}{(FAF^T FA + \alpha VF)_{ij}}\right]^{\frac{1}{4}}$.
until Converges
```

3.2.1. Correctness of the algorithm. The following theorem guarantees the correctness of Algorithm 1.

Theorem 1 *If the update rules of F in Algorithm 1 converges, the final solution satisfies the Karush-Kuhn-Tucker (KKT) conditions.*

Proof. First, *J*MDKC can be written as follows:

$$J_{\mathrm{MDKC}}(F) = \mathbf{tr}\,(W^TW - 2FAF^TW + FAF^TFAF^T) \\ + 2\alpha\mathbf{tr}\,(Y^TVY - 2F^TVY + F^TVF), \tag{10}$$

where we exploit the properties of $W = W^T$, $A = A^T$, $V^T = V$, and for any matrix M we have $\mathbf{tr}\ (M) = \mathbf{tr}\ (M^T)$.

Then by removing the constant terms, we take the derivative of $J_{\mathrm{MDKC}}\ (F)$ with respect to F as follows:

$$\frac{dJ_{\mathrm{MDKC}}}{dF} = 4WFA + 4FAF^TFA + 4\alpha VY + 4\alpha VF. \tag{11}$$

Then the Karush-Kuhn-Tucker (KKT) conditions for nonnegativity of F is

$$(4WFA + 4FAF^TFA + 4\alpha VY + 4\alpha VF)_{ij}F_{ij} = 0, \tag{12}$$

which is the fixed point relationship that the solution must satisfy.

On the other hand, at the convergence of Algorithm 1, $F^{(\infty)} = F^{(t+1)} = F^{(t)}$, thus we can derive:

$$(4WFA + 4FAF^TFA + 4\alpha VY + 4\alpha VF)_{ij}F_{ij}^4 = 0, \tag{13}$$

which is identical to Equation (12) and proves Theorem 1. ■

3.2.2. Convergence of the algorithm. We use the auxiliary function approach (Lee and Seung, 2001) to prove the convergence of Algorithm 1. Here, we first introduce the definition of auxiliary function (Lee and Seung, 2001).

Lemma 1 *(Lee and Seung, 2001) Z (h, h′) is an auxiliary function of F (h) if the conditions Z (h, h′) ≥ F (h) and Z (h, h′) = F (h) are satisfied. (Lee and Seung, 2001) If Z is an auxiliary function for F, then F is nonincreasing under the update* $h^{(t+1)} = \mathit{argmin}_h\ Z\ (h, h')$.

The following inequality is also useful in our following proofs.

Lemma 2 *(Ding et al., 2006) For any matrices* $A \in \mathbb{R}_+^{n\times n}$, $B \in \mathbb{R}_+^{k\times k}$, $S \in \mathbb{R}_+^{n\times k}$, *and* $S' \in \mathbb{R}_+^{n\times k}$, *and A and B are symmetric, the following inequality holds:*

$$\sum_{ip}\frac{(AS'B)_{ip}S_{ip}^2}{S'_{ip}} \geq \boldsymbol{tr}\,(S^TASB). \tag{14}$$

Lemma 2 can only be used to deal with NMF with rectangle matrices or asymmetric square matrices, but not symmetric matrices, such as those used in our objective in Equation (9). As one of our theoretical contributions, we prove the following generic matrix inequality in Equation (15) to analyze objective functions involving fourth order matrix polynomials.

Lemma 3 *For any nonnegative symmetric matrices* $A \in \mathbb{R}_+^{k\times k}$ *and* $B \in \mathbb{R}_+^{k\times k}$, *for* $H \in \mathbb{R}_+^{n\times k}$ *the following inequality holds:*

$$\boldsymbol{tr}\,(HAH^THBH^T) \leq \sum_{ik}\left(\frac{H'AH'^TH'B + H'BH'^TH'A}{2}\right)_{ik}\frac{H_{ik}^4}{H'^3_{ik}}. \tag{15}$$

Proof. Let $H_{ik} = H'_{ik}u_{ik}$. The first term in RHS of Equation (15) is

$$\sum_{ik}(H'AH'^TH'B)_{ik}\frac{H_{ik}^4}{H'^3_{ik}} = \sum_{ijkrpq}H'_{jr}A_{rk}H'_{ik}H'_{ip}B_{pq}H'_{jq}u_{jq}^4 \tag{16}$$

Now, switching indexes: $i \Leftrightarrow j$, $p \Leftrightarrow q$, $r \Leftrightarrow k$, we obtain

$$\sum_{ik} (H'AH'^T H'B)_{ik} \frac{H_{ik}^4}{H_{ik}'^3} = \sum_{ijkrpq} H'_{ik}A_{kr}H'_{jr}H'_{jq}B_{qp}H'_{ip}u_{ip}^4 \tag{17}$$

The second term in RHS of Equation (15) is

$$\sum_{ik} (H'BH'^T H'A)_{ik} \frac{H_{ik}^4}{H_{ik}'^3} = \sum_{ijkrpq} H'_{ip}B_{pq}H'_{jq}H'_{jr}A_{rk}H'_{ik}u_{ik}^4. \tag{18}$$

Now, switching indexes: $i \Leftrightarrow j,p \Leftrightarrow q$, and $r \Leftrightarrow k$, we obtain

$$\sum_{ik} (H'BH'^T H'A)_{ik} \frac{H_{ik}^4}{H_{ik}'^3} = \sum_{ijkrpq} H'_{jq}B_{qp}H'_{ip}H'_{ik}A_{kr}H'_{jr}u_{jr}^4 \tag{19}$$

Careful examination of the RHS of Equations (16) (19) shows that they are identical except u^4 terms. Thus, the RHS of Equation (15) is

$$\sum_{ijkrpq} H'_{ip}B_{pq}H'_{jq}H'_{jr}A_{rk}H'_{ik}\frac{u_{ik}^4 + u_{jr}^4 + u_{jq}^4 + u_{ip}^4}{4}. \tag{20}$$

The LHS of Equation (15) is $\sum_{ijkrpq} H'_{ip}B_{pq}H'_{jq}H'_{jr}A_{rk}H'_{ik}u_{ik}u_{jr}u_{jq}u_{ip}$. For any a, b, c, $d > 0$, we have $a^4 + b^4 + c^4 + d^4 \geq 2(a^2b^2 + c^2d^2) \geq 4(ab)(cd)$, thus $u_{ik}u_{jr}u_{jq}u_{ip} \leq (u_{ik}^4 + u_{jr}^4 + u_{jq}^4 + u_{ip}^4)/4$, which proves Lemma 3. ■

Based on Lemmas 1–3, now we prove the convergence of Algorithm 1 by the following theorem.

Theorem 2 *Let*

$$J(F) = \mathbf{tr}(-2WFAF^T + FAF^TFAF^T - 4\alpha F^TVY + 2\alpha F^TVF),$$

then the following function

$$Z(F, F') = -2\sum_{ijkl} F'_{ji}A_{jk}F'_{kl}W_{li}\left(1 + \log\frac{F_{ji}F_{kl}}{F'_{ij}F'_{kl}}\right) + \sum_{ij}(F'AF'^TA')_{ij}\frac{F_{ij}^4}{F_{ij}'^3}$$
$$-4\alpha\sum_{ij}(VY)_{ij}F'_{ij}\left(1 + \log\frac{F_{ij}}{F'_{ij}}\right) + 2\alpha\sum_{ij}\frac{(VF')_{ij}F_{ij}^2}{F'_{ij}}$$

is an auxiliary function of J (G). Furthermore, it is a convex function in G and its global minimum is

$$F_{ij} = Fij\left[\frac{(WFA + \alpha VY)_{ij}}{(FAF^TFA + \alpha VF)_{ij}}\right]^{\frac{1}{4}} \tag{21}$$

Proof. First, by removing constant terms, the objective of the proposed method can be written as following:

$$J_{\mathrm{MDKC}}(F) = \mathbf{tr}\,(-2FAF^TW + FAF^TFAF^T - 4\alpha F^TVY + 2\alpha F^TVF) \tag{22}$$

By Lemma 3, we have

$$\mathbf{tr}\,(FAF^TFAF^T) \leq \sum_{ij}\left(F'AF'^TF'A\right)_{ij}\frac{F_{ij}^4}{F_{ij}'^3}. \tag{23}$$

Because of Lemma 2 and the inequality $2ab \leq a^2 + b^2$, we have

$$\mathbf{tr}(F^TVF) \leq \sum_{ij}\frac{(VF')_{ij}F_{ij}^2}{F'_{ij}}. \tag{24}$$

Because $z \leq 1 + \log z$, $\forall\, z > 0$, we have

$$\mathbf{tr}\,(FAF^T W) \geq \sum_{ijkl} F'_{ji} A_{jk} F'_{kl} W_{li} \left(1 + \log \frac{F_{ji} F_{kl}}{F'_{ij} F'_{kl}} \right) \tag{25}$$

$$\mathbf{tr}\,(F^T VY) \geq \sum_{ij} (VY)_{ij} F'_{ij} \left(1 + \log \frac{F_{ij}}{F'_{ij}} \right) \tag{26}$$

Summing over all these bounds, we get $Z\,(F, F')$, which clearly satisfies (1) $Z\,(F, F') \geq J\,(F)$ and (2) $Z\,(F, F) = J\,(F)$.

Following the same derivations as in Ding et al. (2010, 2006) and Gu and Zhou (2009), the Hessian matrix of $Z\,(F, F)$ is positive definite. Thus $Z\,(F, F')$ is a convex function of F. We obtain the global minimum of $Z\,(F, F')$ by setting $dZ\,(F, F')\,/\,dF_{ij} = 0$ and solving for F, from which we can get Equation (10) in the main text. This completes the proof of Theorem 2. ■

Because $J\,(F)$ in Eq. (9) is obviously lower bounded by 0, Lemma 1 and Theorem 2 guarantee the convergence of Algorithm 1.

Note that Lemma 3 and Equation (23) is the key step to prove the convergence of the proposed algorithm. Due to the symmetric usage of the side factor matrix F, the optimization objective Equation (9) involves a fourth-order term of F. Compared to the quadratic term involved in traditional NMTF for rectangle (asymmetric) input matrix, optimizing Equation (9) is definitely much harder. As a result, in existing works Ding et al. (2005, 2006;) and Li et al. (2007), heuristics are routinely used to tackle this difficulty. Only until our recent works (Wang et al., 2011b,c), by proving the generic matrix inequality in Lemma 3, we are able to rigorously derive an iterative algorithm as in Algorithm 1 and prove its convergence.

4. RESULTS AND DISCUSSION

We evaluate the proposed MDKC approach on *S. cerevisiae* genome data and apply it on PPI network data, protein sequence data, and an integration of these two types of experimental data, respectively.

We use MIPS Functional Catalogue (FunCat) system (Mewes et al., 1999) to annotate proteins, which is an annotation scheme for the functional description of proteins from prokaryotes, unicellular eukaryotes,

Table 1. Main Functional Categories in FunCat Annotation Scheme (Version 2.1) and the Corresponding Number of Annotated Proteins to Yeast Species

		Number of proteins
ID	*Function description*	*(yeast) annotated*
01	Metabolism	1397
02	Energy	336
10	Cell cycle and DNA processing	981
11	Transcription	1009
12	Protein synthesis	476
14	Protein fate	1125
16	Protein with binding function	1019
18	Regulation of metabolism and protein function	246
20	Transport facilitation and transport routes	995
30	Cellular communication and signal transduction	231
32	Cell rescue, defense, and virulence	515
34	Interaction with the environment	446
38	Transposable elements, viral, and plasmid proteins	59
40	Cell fate	268
41	Development	67
42	Biogenesis of cellular components	827
43	Cell type differentiation	437

plants, and animals. Taking into account the broad and highly diverse spectrum of known protein functions, FunCat (version 2.1) consists of 27 main functional categories that cover general fields such as cellular transport, metabolism, cellular communication, etc. The main branches exhibit a hierarchical and treelike structure with up to six levels of increasing specificity; 17 main function categories in FunCat annotation scheme are involved to annotate the yeast genome as listed in Table 1.

4.1. EVALUATION ON PPI NETWORK DATA

We first evaluate our MDKC approach on PPI network data, as they are the most popularly used experimental data for protein function prediction (Sharan et al., 2007). We compare our MDKC approach to four broadly used graph-based protein function prediction approaches: (1) Majority Voting (MV) approach (Schwikowski et al., 2000), (2) Iterative Majority Voting (IMV) approach (Vazquez et al., 2003), (3) χ^2 approach (Hishigaki et al., 2001), and (4) Functional Flow (FF) approach (Nabieva et al., 2005). We use *precision-recall* curves to assess prediction performance, which is the most widely used performance metric in existing literature.

4.1.1. Data preparation. We download PPI data for the *S. cerevisiae* species from BioGRID (version 2.0.56) (Stark et al., 2006). By removing the proteins connected by only one PPI, we end up with 4403 annotated proteins with 86167 PPIs. We represent the protein interaction network as a graph, with vertices corresponding to the proteins and edges corresponding to PPIs. The adjacency matrix of the graph is denoted as $B \in \{0, 1\}^{n \times n}$ where $n = 4403$, such that $B_{ij} = 1$ if proteins $\mathbf{x}_i$ and $\mathbf{x}_j$ interact, and 0 otherwise.

The adjacency matrix B itself measures the similarity among proteins in the sense that two proteins are related if they interact. However, two critical problems prevent us from directly using B as data similarity $\mathcal{S}_D(\mathbf{x}_i, \mathbf{x}_j)$ to predict protein function. First, B only measures the local connectivity of a graph and contains no information for connections via more than one edge. Therefore the important information contained in the global topology is simply ignored. Second, PPI data suffer from high noise due to the nature of high-throughput technologies, for example, false positive rate in yeast two-hybrid experiments is estimated as high as 50% (Deane et al., 2002). Therefore, we use the Topological Measurement (TM) method (Pei and Zhang, 2005) to compute the data similarity matrix W, which takes into consideration paths with all possible lengths on a network and weights the influence of every path by its length. Specifically, W_{ij} between proteins $\mathbf{x}_i$ and $\mathbf{x}_j$ is computed as Pei and Zhang (2005):

$$W_{ij} = \sum_{k=2}^{|V|-2} \mathrm{PR}^k(i, j), \text{ and } \mathrm{PR}^k(i, j) = \frac{\mathrm{PS}^k(i, j)}{\mathrm{MaxPS}^k(i, j)}, \tag{27}$$

where $|V|$ is the number of vertices in the PPI graph, $\mathrm{PR}^k(i, j)$ is the path ratio of the paths of length k between proteins $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\mathrm{PS}^k(i, j)$ and $\mathrm{MaxPS}^k(i, j)$ are defined as follows (Pei and Zhang, 2005):

$$\mathrm{PS}^k(i, j) = (A^k)_{ij}, \tag{28}$$

where $(\cdot)_{ij}$ denotes the ij-th entry of a matrix, and the following (Pei and Zhang, 2005)

$$\mathrm{MaxPS}^k(i, j) = \begin{cases} \sqrt{d_i d_j}, & \text{if } k = 2 \\ d_i d_j, & \text{if } k = 3, \\ \sum_{k \in N(i),\, l \in N(j)} \mathrm{MaxPS}^{k-2}(k, l), & \text{if } k > 3. \end{cases} \tag{29}$$

where $d_i = \sum_j B_{ij}$ is the degree of the i-th vertex, and $N(i)$ denotes its neighboring vertices. The detailed explanation of TM measurement can be referred to Pei and Zhang (2005).

4.1.2. Improved precision-recall performance. In order to generate the precision-recall curves for evaluation, we randomly select half of the proteins as annotated proteins and the rest as unannotated ones. The optimal result F^* of our MDKC produces a ranking list to indicate prediction confidence. Varying threshold, we obtain a precision-recall curve. Precision-recall curves for other compared approaches are produced following Chua et al. (2006). For MV approach, the k most frequent functions appearing in a protein's neighbors are assigned as the k most likely function, such that a precision-recall curve is obtained by varying k. For χ^2 approach, we take the k largest χ^2 statistics as the k most likely function to build the precision-recall curve. Because the solution to IVM approaches are not unique for every trial, we repeat the experiment several times to obtain the precision-recall pairs. FF produces a ranking list, which are used to generate the precision-recall curve.

The prediction results for function ''Metabolism'' are shown as in Figure 2(a), which has 1397 annotated proteins in the original data before random selection (the biggest number of annotated proteins among all 17 functional categories). The prediction results for two other functions ''Cellular Transport, Transport Facilitation and Transport Routes'' and ''Cell Type Differentiation'' are shown in Fig. 2b and c, which have 995 and 437 annotated proteins respectively. The precision-recall curves for other functions can not be presented due to space, from which the same observations can be seen. All these results show that the proposed MDKC approach significantly improves the prediction performance, which validates the effectiveness of our approach when the data similarity is derived from PPI network data.

In addition to reporting the protein function prediction performances over each individual biological function of the compared methods as above, we also report their average predictive capabilities over all the 17 biological functions. As shown in Figure 2, every compared method produces a precision-recall curve for a single biological function when we vary the thresholds, thus every curve comprises many different

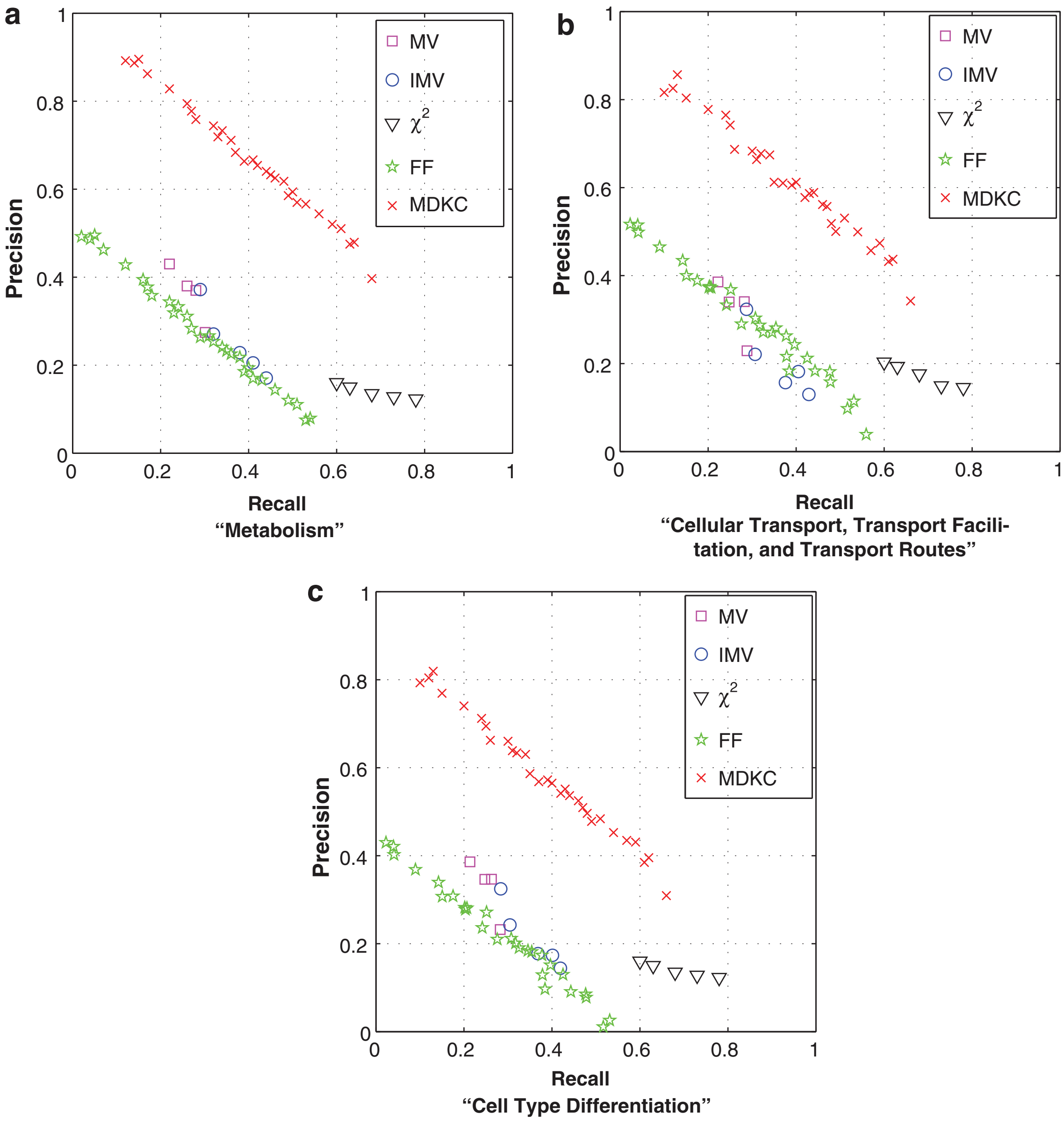

FIG. 2. Precision-recall curves by the five compared approaches for three functional categories in FunCat 2.1.

Table 2. Average F1 Scores by the Compared Approaches over All the 17 Main Functional Categories of FunCat Annotation Scheme when Using Protein–Protein Interaction Network Data Only

Approaches	*Average of the best F1 scores (%)*
MV	32.07
IMV	34.15
χ^2	33.03
FF	33.12
MDKC	40.03

FF, functional flow; IMV, iterative majority voting; MDKC, maximization of data-knowledge consistency; MV, majority voting.

pairs of precision and recall corresponding to setting a threshold value, where one pair corresponds to a point of the curve. In order to compare the overall predictive performance of the compared methods, from each precision-recall curve for a single biological function we pick up the precision-recall pair (point) that produces the best F1 score. Then we average the 17 best F1 scores from all the biological functions for a compared method and report it in Table 2. Obviously, the results in Table 2 demonstrate the better predictive capability of our new method.

4.2. Evaluation on protein sequence data

Because sequence is the most fundamental form to describe a protein, which contains important structural, characteristic, and genetic information, we evaluate the proposed MDKC approach using protein sequences. We compare the predictive accuracy of our approach against the (1) functional similarity weight (FS) approach (Chua et al., 2006) and (2) kernel-based data fusion (KDF) approach (Lanckriet et al., 2004). We also report the performance of majority voting (MV) approach (Schwikowski et al., 2000) as a baseline. We employ broadly used average precision and average F1 scores (Chua et al., 2006) as performance metrics.

4.2.1. Adaptive decision boundary for prediction. Although the evaluation through precision-recall curves in section 4.1 makes sense in scientific research, it does not make explicit predictions. In practice, however, specific putative functions for unannotated proteins are required for further post-genomic researches and applications, therefore a decision boundary (threshold) is necessary.

In many semisupervised learning algorithms, the threshold for classification is usually selected as 0, which, however, is not necessary to be the best choice. We use an adaptive decision boundary to achieve better predictive performance, which is adjusted such that the weighted training errors on annotated proteins are minimized.

Considering the binary classification problem for the k-th functional category, 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 as follows:

$$b_k^{\text{opt}} = \arg\min_{b_k} \left[\frac{e_+(b_k)}{|S_+|} + \frac{e_-(b_k)}{|S_-|} \right]. \tag{30}$$

And the decision rule to assign a function to protein x_i is given by:

$$\begin{cases} \mathbf{x}_i \text{ is annotated with the } k\text{-th function if } F^*_{ik} > b_k^{\text{opt}}; \\ \mathbf{x}_i \text{ is not annotated with the } k\text{-th function if } F^*_{ik} \leq b_k^{\text{opt}}; \end{cases} \tag{31}$$

4.2.2. Data preparation. We obtain sequence data from GenBank (Benson et al., 2006) and describe a protein sequence through one kind of its elementary constituents, that is, trimers of amino acids. Trimer, a type of k-mer (when $k=3$) broadly used in sequence analysis, considers the statistics of one amino acid and its vicinal amino acids, and regards any three consecutive amino acids as a unit to preserve order information, for example ''ART'' is one unit and ''MEK'' is another one. The trimer histogram of a sequence hence can be used to characterize a protein x_i, which is denoted as P_i. Because histogram indeed is a probability distribution, we use Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951), a standard way to assess the difference between two probability distributions, to measure the distance between two proteins, which is defined as:

$$D_{\mathrm{KL}}(P_i \| P_j) = \sum_k P_i(k) \log \frac{P_i(k)}{P_j(k)}, \tag{32}$$

where k denotes the index of the k-th trimer. Because KL divergence is nonsymmetric, that is, $D_{KL}(P_i \| P_j) \neq D_{KL}(P_j \| P_i)$, we use the symmetrized KL divergence as follows:

$$D_{\text{S-KL}}(i, j) = \frac{D_{\mathrm{KL}}(P_i \| P_j) + D_{\mathrm{KL}}(P_j \| P_i)}{2}. \tag{33}$$

Finally, the pairwise data similarity W is defined by converting the symmetrized KL divergences through the standard way:

$$\begin{aligned} W_{ij} &= D_{\text{S-KL}}(i, i) + D_{\text{S-KL}}(j, j) - 2D_{\text{S-KL}}(i, j) \\ &= -[D_{\mathrm{KL}}(P_i \| P_j) + D_{\mathrm{KL}}(P_j \| P_i)]. \end{aligned} \tag{34}$$

4.2.3. Improved predictive capability. We perform standard five-fold cross validation to evaluate the compared approaches and report the average performance of five trials in Table 3. For FS approach, because it does not supply a threshold, we use the one giving the best F1 score to make the prediction. We implement two versions of our method to evaluate the contributions of each of its components. First, we solve Equation (9) by Algorithm 1, which is the proposed method. Second, we solve a degenerate version of the problem in Equation (9) by not incorporating the correlations between functional categories. Specifically, we replace FAF^T in Equation (9) by FF^T, which is denoted by MDKC-S.

The results in Table 3 show that our MDKC-S and MDKC approach clearly outperform the other compared approaches, which concretely quantify the advantage of our approaches. The improvement on classification performance of the MDKC approach over the MDKC-S approach clearly justifies the usefulness of function–function correlations in predicting putative protein functions.

We further study the proposed method by using the knowledge similarity learned by using Gene Ontology (GO) (Consortium et al., 2008) terms. We randomly pick up 30 terms from every one of the three domains of GO and use the selected 90 terms to annotate the same set of proteins as above. We compute the function similarity using the following two different approaches. The first approach is the cosine similarity proposed in section 2, where we compute the function similarities using the training data (four folds out of the five folds) of each of the five experimental trails in the cross-validation. In addition, we also compute the function similarity using the semantic similarity between terms as proposed in Dotan-Cohen et al. (2009).We compare the proposed method against the same set of competing methods as before and report the results in Table 4. From the results, we can see that our new method once again performs better, especially when functional similarity is utilized. In addition, the classification performances of our method

Table 3. Average Precision and Average F1 Score by the Compared Approaches in 5-Fold Cross Validation on the Main Functional Categories of FunCat Annotation Scheme

Approaches	*Average precision (%)*	*Average F1 score (%)*
FS	33.65	22.78
KDF	53.45	38.10
MV	32.07	29.46
MDKC-S	56.51	39.04
MDKC	61.38	42.17

FS, functional similarity weight; KDF, kernel-based data fusion.

Table 4. Average Precision and Average F1 Score by the Compared Approaches in 5-Fold Cross Validation on 90 Randomly Selected Gene Ontology Terms

Approaches	*Average precision (%)*	*Average F1 score (%)*
FS	19.52	18.15
KDF	37.14	22.63
MV	20.11	17.55
MDKC-S	41.26	33.51
MDKC (cosine similarity)	48.15	39.62
MDKC [process linkage similarity (Dotan-Cohen et al., 2009)]	48.96	40.11

do not change much when using different function similarities, which demonstrates the robustness of our new method.

4.3. Evaluation on integrated biological data

As mentioned earlier, biological data from one single experimental source only convey information for certain aspects, which are usually incomplete and sometimes misleading. For example, similar sequences do not always have similar functions. In the extreme case, proteins with 100% sequence identity could perform different functional roles (Whisstock and Lesk, 2004). Therefore, integration of different biological data is necessary for more robust and complete protein function inferences. In general, results learned from a combination of different types of data are likely to lead to a more coherent model by consolidating information on various aspects of the same biological process. In this subsection, we evaluate the predictive performance using the integrated data from both PPI networks and protein sequences. We compute the data similarity for each individual data in the same ways as in section 4.1 and section 4.2, which are denoted as W_{PPI} and W_{sequence} respectively. The integrated data similarity W is hence computed as follows:

$$W = W_{\mathrm{PPI}} + \gamma W_{\mathrm{sequence,}} \tag{35}$$

where γ is a balance parameter and empirically selected as

$$\gamma = \frac{\sum_{i,\,,\,i \neq j} W_{\mathrm{PPI}}(i,j)}{\sum_{i,\,,\,i \neq j} W_{\mathrm{sequence}}(i,j)}. \tag{36}$$

We compare the predictive performance of our MDKC approach to two data integration based–protein function prediction approaches, (1) the kernel-based data fusion (KDF) approach (Lanckriet et al., 2004) and (2) the locally constrained diffusion kernel (LCDK) approach (Tsuda and Noble, 2004), and two baseline approaches, (3) the majority voting (MV) approach (Schwikowski et al., 2000) and (4) the iterative majority voting (IMV) approach (Vazquez et al., 2003). The function-wise prediction performance measured by average precision and average F1 score in standard five-fold cross validation is reported in Figure 3.

From the results in Figure 3a and Figure 3b, we can see that the proposed MDKC approach is consistently better than other compared approaches, sometimes very significantly, which again demonstrates the superiority of our approach.

A more careful examination on the results in Figure 3 shows that, although our approach outperforms the compared approaches in most functional categories, but not always, for example, the average precision for the function ‘‘Transposable Elements, Viral and Plasmid Proteins’’ (ID: 38). By scrutinizing the function category correlations, defined in Equation (2) and illustrated in the right panel of Figure 1, we can see the average correlation of the function ‘‘Transposable Elements, Viral and Plasmid Proteins’’ with other functional categories is among the lowest. As a result, the presence/absence of this function category can not benefit from other functional categories, because it only has weak correlations with them. In contrast, prediction for the function categories with high correlations to others generally can benefit from our approach. This observation firmly testifies the importance of function category correlations in predicting protein function.

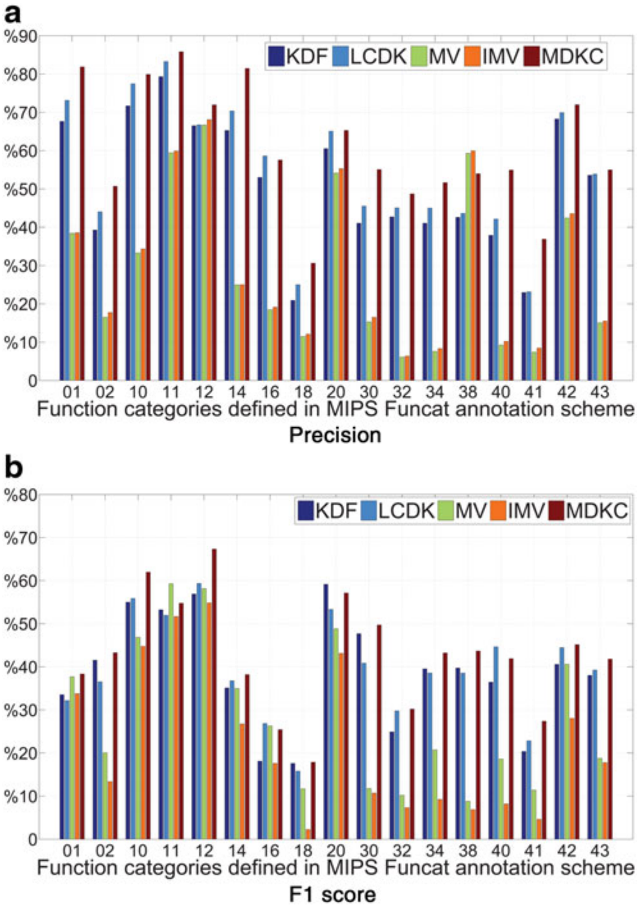

FIG. 3. Performance of five-fold cross validation for the 17 main functional categories in FunCat annotation scheme (version 2.1) by KDF, LCDK, MV, GMV, and the proposed MDKC approach.

5. CONCLUSIONS

In this article, we presented a novel Maximization of Data-Knowledge Consistency (MDKC) approach to predict protein function, which attempts to make use of function category correlations to improve the predictive accuracy. Different from traditional approaches in predicting protein function, which routinely use protein annotations as labels assigned to data points, we employed annotation knowledge in a completely different way to measure pairwise protein similarities. By maximizing consistency between the *knowledge similarity* computed from annotations and the *data similarity* computed from biological experimental data, optimal function assignments to unannotated proteins are obtained. Most importantly, function category correlations are incorporated in a natural way through the knowledge similarity. Using kernel mechanism, we further extend our approach to better fit more specific data for improved prediction performance. Comprehensive empirical evaluations have been conducted on *S. cerevisiae* genome using PPI network data, protein sequence data, and an integration of both of them respectively, promising results in the experiments justified our analysis and validated the performance of our methods.

ACKNOWLEDGMENTS

This work was partially supported by NSF-IIS 1117965, NSF-IIS 1302675, NSF-IIS 1344152, NSF-DBI 1356628, and NSF-IIS 1423591.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Benson, D., Karsch-Mizrachi, I., and Lipman, D. 2006. GenBank. *Nucleic Acids Res.* 34, D16–D20.

Cai, D., He, X., Han, J., and Huang, T.S. 2011. Graph regularized non-negative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 8, 1548–1560.

Cai, D., He, X., Wu, X., and Han, J. 2008. Non-negative matrix factorization on manifold. Proceedings of ICDM.

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.

Chua, H., Sung, W., and Wong, L. 2007. Using indirect protein interactions for the prediction of Gene Ontology functions. *BMC Bioinform.* 8, S8.

Deane, C., Salwinski, L., Xenarios, I., and Eisenberg, D. 2002. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* 1, 349–356.

Ding, C., He, X., and Simon, H. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. Proceedings of the SIAM Data Mining Conference, pp. 606–610.

Ding, C., Li, T., and Jordan, M. 2010. Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 45–55.

Ding, C., Li, T., Peng, W., and Park, H. 2006. Orthogonal nonnegative matrix t-factorizations for clustering. Proceedings of SIGKDD.

Dotan-Cohen, D., Letovsky, S., Melkman, A.A., and Kasif, S. 2009. Biological process linkage networks. *PloS One* 4, e5313.

Gene Ontology Consortium, et al. 2008. The gene ontology project in 2008. *Nucleic Acids Res*. 36, D440–D444.

Gu, Q., and Zhou, J. 2009. Co-clustering on manifolds. Proceedings of SIGKDD.

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.

Jiang, X., Nariai, N., Steffen, M., et al. 2008. Integration of relational and hierarchical network information for protein function prediction. *BMC Bioinform.* 9, 350.

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.

Kullback, S., and Leibler, R. 1951. On information and sufficiency. *Ann. Math. Stat.* 22, 79–86.

Lanckriet, G., Deng, M., Cristianini, N., et al. 2004. Kernel-based data fusion and its application to protein function prediction in yeast. Proceedings of the Pacific Symposium on Biocomputing, volume 9, pp. 300–311.

Lee, D., and Seung, H. 2001. Algorithms for non-negative matrix factorization. Proceedings of NIPS.

Li, T., Ding, C., and Jordan, M. 2007. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. Proceedings of ICDM.

Liang, S., Shuiwang, J., and Jieping, Y. 2008. Adaptive diffusion kernel learning from biological networks for protein function prediction. *BMC Bioinform.* 9, 162.

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. Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference, pp. 268–278.

Schwikowski, B., Uetz, P., and Fields, S. 2000. A network of protein-protein interactions in yeast. *Nat. Biotech.* 18, 1257–1261.

Sharan, R., Ulitsky, I., and Shamir, R. 2007. Network-based prediction of protein function. *Mol. Syst. Biol*. 3, 88.

Shi, L., Cho, Y., and Zhang, A. 2009. ANN based protein function prediction using integrated protein-protein interaction data. Proceedings of the International Joint Conference on Bioinformatics, Systems Biology, and Intelligent Computing, pp. 271–277.

Shin, H., Lisewski, A., and Lichtarge, O. 2007. Graph sharpening plus graph integration: a synergy that improves protein functional classification. *Bioinformatics* 23, 3217.

Stark, C., Breitkreutz, B., Reguly, T., et al. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535.

Sun, L., Ji, S., and Ye, J. 2008. Adaptive diffusion kernel learning from biological networks for protein function prediction. *BMC Bioinform.* 9, 162.

Tsuda, K., and Noble, W. 2004. Learning kernels from biological networks by maximizing entropy. *Bioinformatics* 20, 326–333.

Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. 2003. Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.* 21, 697–700.

Wang, H., Ding, C., and Huang, H. 2010a. Multi-label linear discriminant analysis. Proceedings of the 11th European Conference on Computer Vision (ECCV 2010), pp. 126–139.

Wang, H., Huang, H., and Ding, C. 2009. Image annotation using multi-label correlated Green's function. Proceedings of the 12th IEEE Conference on Computer Vision (IEEE ICCV 2009), pp. 2029–2034.

Wang, H., Huang, H., and Ding, C. 2010b. Multi-label feature transform for image classifications. Proceedings of the 11th European Conference on Computer Vision (ECCV 2010), pp. 793–806.

Wang, H., Huang, H., and Ding, C. 2011a. Image annotation using bi-relational graph of images and semantic labels. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2011 (CVPR 2011), pp. 793–800.

Wang, H., Huang, H., and Ding, C. 2011b. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization. Proceedings of the 20th ACM International Conference on Information and Knowledg Management (ACM CIKM 2011), pp. 279–284.

Wang, H., Huang, H., and Ding, C. 2012. Function-function correlated multi-label protein function prediction over interaction networks. Proceedings of the 16th International Conference on Research in Computational Molecular Biology (RECOMB 2012).

Wang, H., Huang, H., and Ding, C. 2013. Protein function prediction via laplacian network partition- ing incorporating function category correlations. Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013), pp. 2049–2055.

Wang, H., Nie, F., Huang, H., and Ding, C. 2011c. Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation. Proceedings of the 11th IEEE International Conference on Data Mining (IEEE ICDM 2011).

Weston, J., Elisseeff, A., Zhou, D., et al. 2004. Protein ranking: from local to global structure in the protein similarity network. *Proc. Natl. Acad. Sci. USA* 101, 6559.

Whisstock, J., and Lesk, A. 2004. Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* 36, 307–340.

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

E-mail: heng@uta.edu