Correlated Protein Function Prediction via Maximization of Data-Knowledge Consistency[*]

Hua Wang[1], Heng Huang[2,**], and Chris Ding[2]

[1] Department of Electrical Engineering and Computer Science
Colorado School of Mines, Golden, Colorado 80401, USA
[2] Department of Computer Science and Engineering
University of Texas at Arlington, Arlington, Texas 76019, USA
huawangcs@gmail.com, {heng,chqding}@uta.edu

Abstract. Protein function prediction in conventional computational approaches is usually conducted one function at a time, fundamentally. As a result, the functions are treated as separate target classes. However, biological processes are highly correlated, which makes functions assigned to proteins are not independent. Therefore, it would be beneficial to make use of function category correlations in predicting protein function. 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 elegantly incorporated through the knowledge similarity. Comprehensive experimental evaluations on *Saccharomyces cerevisiae* data demonstrate promising results that validate the performance of our methods.

Keywords: Protein Function Prediction, Mutli-Label Classification, 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 decade. Among numerous existing algorithms, graph based approaches and data integration based approaches have demonstrated effectiveness due to their clear connections to biological facts.

[*] This work was partially supported by US NSF IIS-1117965, IIS-1302675, IIS- 1344152.

[**] Corresponding author.

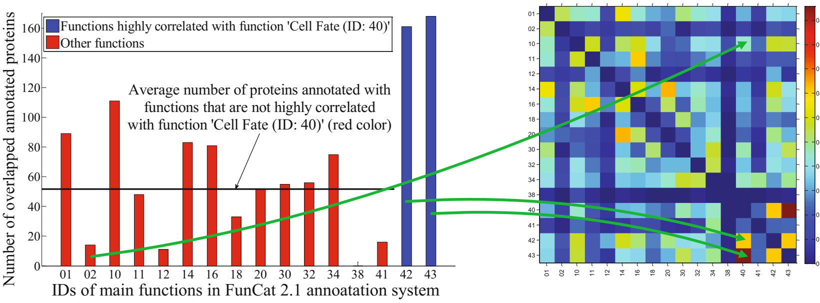
R. Sharan (Ed.): RECOMB 2014, LNBI 8394, pp. 311–325, 2014.

Since many biological experimental data can be readily represented as networks, graph-based approaches are the most natural way to predict protein function [1]. Neighborhood-based methods [2–5] 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 [6, 7] 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 [8, 9], and those derived from kernel methods [10]. More recently, the authors developed a graph-based protein function prediction method [11] using PPI graph to take advantage of the function-function correlations by considering protein function prediction as a multi-label classification problem, which takes the same perspective as this work. Experimental data from one single source often incomplete and sometimes even misleading [12], therefore predicting protein function using multiple biological data has attracted increased attention. [13] proposed a kernel-based data fusion approach to integrate multiple experimental data via a hybrid kernel and use support vector machine (SVM) for classification. [14] presented a locally constrained diffusion kernel approach to combine multiple types of biological networks. Artificial neural network is employed in [15] for the integration of different protein interaction data.

Most existing computational approaches usually consider protein function prediction as a standard classification problem [13, 16, 17]. Typically, these approaches make prediction one function at a time, fundamentally, *i.e.*, 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 [11, 18]. These function category correlations, albeit useful, are seldom utilized in predicting protein function. In this study, we explore this special characteristic of the protein functional categories and make use of the function-function correlations to improve the overall predictive accuracy of protein functions.

1.1 Multi-label 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 *multi-label classification* problem [19, 11, 20–22, 18]. Multi-label 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) [23] on yeast 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 yeast 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. As a result, we reasonably speculate that these two functions are statistically correlated in a stronger way. As a result, 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.



(a) Number of proteins annotated to both function "40" and one of the other functions

(b) Correlation matrix among the 17 main functions in Funcat 2.1 to yeast genome.

Fig. 1. Left: number of proteins annotated to both function 40 and one of the other functions. **Right:** visualization of the correlation values defined by Eq. (1) among the 17 main functions in FunCat 2.1 to yeast genome.

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, protein-protein interaction (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 are routinely used as supervision in the classification tasks, *i.e.*, 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 multi-label classification task (*i.e.*, protein function annotations in protein function prediction problems) convey important attribute information [21], 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 descriptions of two proteins abstracted from experimental data, and $\mathbf{f}_i$ and $\mathbf{f}_j$ be the labeling vectors that encode the annotated functions of the same 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 $\mathbf{x}_j$, *i.e.*, the data and the knowledge are consistent, we would expect that the two similarity measurements should be close given that they are normalized to the same scale, *i.e.*, $\mathcal{S}_D(\mathbf{x}_i, \mathbf{x}_j) \approx \mathcal{S}_K(\mathbf{f}_i, \mathbf{f}_j)$. With this assumption, we may determine the optimal function assignments to unannotated proteins by minimizing the difference between the two sets of similarities, *i.e.*, maximizing the consistency between experimental data and biological knowledge. In this paper, we formalize this assumption and propose 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, our goal is to predict functions for 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 constructed from certain biological experimental data, such as the amino acid histogram of a protein sequence. 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. W is usually seen as edge weight matrix of a graph where proteins correspond to vertices. In the simplest case of a PPI network, $W_{ij} = 1$ if protein $\mathbf{x}_i$ and protein $\mathbf{x}_j$ 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. $\{\mathbf{y}_i\}_{i=1}^{l}$ 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 $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T = \left[\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(K)}\right]$, where $\mathbf{y}^{(k)} \in \mathbb{R}^n$ is a class-wise function annotation vector. Besides the ground truth function assignment matrix Y, we also define $F = [\mathbf{f}_1, \ldots, \mathbf{f}_n]^T \in \mathbb{R}^{n \times K}$ 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, as they are one of our most important mechanism 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 is, the more closely the two functions are related. Therefore, functions assigned to a protein are no longer independent, but 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 following:

$$C_{kl} = \cos(\mathbf{y}^{(k)}, \mathbf{y}^{(l)}) = \frac{\langle \mathbf{y}^{(k)}, \mathbf{y}^{(l)} \rangle}{\|\mathbf{y}^{(k)}\| \, \|\mathbf{y}^{(l)}\|}, \tag{1}$$

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 yeast genome, function correlations defined in Eq. (1) 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 truths, which justifies the correctness of our formulation for function category correlations in Eq. (1) from biological perspective.

3 The 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 the functions annotated to proteins are consistent with their experimental data, we would expect the data similarity is close to the knowledge similarity:

$$\begin{aligned} \min \sum_{i,j} & \left[\mathcal{S}_D(\mathbf{x}_i, \mathbf{x}_j) - \mathcal{S}_K(\mathbf{f}_i, \mathbf{f}_j)\right]^2, \\ \text{s.t.} \quad & \mathbf{f}_i = \mathbf{y}_i, \ \forall\, 1 \le i \le l, \end{aligned} \tag{2}$$

where the constraint fixes the functions assigned to annotated proteins to be ground truth. The optimization objective in Eq. (2) 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 in 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 is to count the number of common annotated functions of two proteins, *i.e.* $\mathbf{f}_i^T \mathbf{f}_j$. 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 may 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 FunCat scheme used in this work, if protein $\mathbf{x}_i$ is annotated with one of the ancestor function of protein $\mathbf{x}_j$'s annotated function, the two proteins are closely related even they do not share any common functions. Therefore, in order to capture correlations among different functions, instead of the simple dot product, we compute the knowledge similarity as following:

$$\mathcal{S}_K(\mathbf{f}_i, \mathbf{f}_j) = \mathbf{f}_i^T C^{-1} \mathbf{f}_j = \mathbf{f}_i^T A \mathbf{f}_j, \tag{3}$$

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

Note that, compared to the dot product similarity defined by $\mathbf{f}_i^T \mathbf{f}_j$ based on the Euclidean distance, the knowledge similarity computed by Eq. (3) is based on the Mahalanobis distance, where C acts as the covariance matrix encoding 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 in a priori. In protein function prediction, the Euclidean distance based knowledge similarity is independent of the concerned biological species, whereas the Mahalanobis distance based knowledge similarity is specific to the biological species of interest thereby has increased statistical power. Most importantly, function-function correlations, the most important advantage of a multi-label 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 Eq. (3), we can formalize the data-knowledge consistency assumption in Eq. (2) by the following optimization problem:

$$\arg\min_{F} \sum_{i,j=1}^{n} (W_{ij} - \sum_{k,l=1}^{K} F_{ik} A_{kl} F_{jl})^2, \tag{4}$$

$$\text{s.t.} \quad F_{ik} = Y_{ik}, \ \forall \ 1 \leq i \leq l, \ 1 \leq k \leq K. \tag{5}$$

In standard classification problems in machine learning, $F_{ik}\,(1 \leq i \leq l)$ 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 its associated 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 have 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 Eq. (5), it is reasonable to relax the confidence variables $F_{ik}\,(1 \leq i \leq l)$ for annotated proteins to be dynamic variables, which approximate the ground truth function assignments. The constraint in Eq. (5) hence can be written to minimize the following penalty function:

$$\alpha \sum_{i=1}^{l} \sum_{k=1}^{K} (Y_{ik} - F_{ik})^2 \,, \tag{6}$$

where $\alpha > 0$ controls the relative importance of the penalty. Following the experiences in graph-based semi-supervised learning, 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:

$$\begin{aligned} J_{\text{MDKC}}(F) &= \|W - FAF^T\|_F^2 + 2\alpha \, \mathbf{tr}\left((Y - F)^T V (Y - F)\right), \\ s.t. \quad & F \geq 0, \end{aligned} \tag{7}$$

where $\|\cdot\|_F$ 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 Eq. (7), the constraint $F \geq 0$ is naturally enforced because W is nonnegative by definition. Most importantly, with this nonnegative constraint Eq. (7) will be enriched with clustering interpretation as detailed soon later, which makes the mathematical formulation of the proposed method more biologically meaningful.

We call Eq. (7) as our proposed Maximization of Data-Knowledge Consistency (MDKC) approach. Upon the solution of Eq. (7), we assign putative functions to unannotated proteins.

3.2 Computational Algorithm of MDKC Approach

Mathematically, Eq. (7) is a regularized NMF problem [24–26]. Although the optimization techniques for the NMF problem and its variants have been extensively studied in literature [27, 28, 24–26, 29, 30], solving Eq. (7) is challenging. Most, if not all, existing algorithms to solve NMF problems are only able to deal with rectangle input matrices (the number of rows of a matrix is different from that of columns) or asymmetric square matrices, but not symmetric input matrices such as the one used in our objective in Eq. (7). This is because the latter involves a fourth-order term due to the symmetric usage of the factor matrix F, which inevitably complicates the problem (More detailed analyses can be found in our earlier works [31, 32]). Traditional solutions to symmetric NMF typically rely on heuristics [27, 33], thus we introduce Algorithm 1 to solve Eq. (7) in a principled way. Due to space limit, the proofs of its correctness and convergence will be provided in the extended journal version.

Algorithm 1. Algorithm to solve Eq. (7)

```
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 [27].
repeat
    3. Compute $F_{ij} \leftarrow F_{ij}[\frac{(WFA+\alpha VY)_{ij}}{(FAF^TFA+\alpha VF)_{ij}}]^{\frac{1}{4}}$.
until Converges
```

4 Results and Discussion

We evaluate the proposed MDKC approach on *Saccharomyces cerevisiae* genome data. We apply the proposed method on protein sequence data, and an integration of protein sequence data and PPI network data, respectively.

We use MIPS Functional Catalogue (FunCat) system [23] to annotate proteins, 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 (version 2.1) consists of 27 main functional categories that cover general fields such as cellular transport, metabolism, cellular communication, *etc.* 17 main function categories in FunCat annotation scheme are involved to annotate yeast genome, which are listed in Table 1.

Table 1. Main functional categories in FunCat annotation scheme (version 2.1) and the corresponding number of annotated proteins to yeast species

Function ID	Function Description	Size
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

4.1 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 first evaluate the proposed MDKC approach using protein sequences. We compare the predictive accuracy of our approach against functional similarity weight (FS) approach [4] and kernel-based data fusion (KDF) approach [13]. We also report the performance of majority voting (MV) approach [2] as a baseline. We employ broadly used average precision and average F1 score [4] as performance metrics.

Adaptive Decision Boundary for Prediction. To predict specific putative functions for unannotated proteins we need a decision boundary (threshold) for learned ranking values, say $\mathbf{y}^{(k)}$, of each class. In many semi-supervised 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 as following [19, 11]:

$$b_k^{\mathrm{opt}} = \arg\min_{b_k} \left[\frac{e_+(b_k)}{|S_+|} + \frac{e_-(b_k)}{|S_-|} \right] . \tag{8}$$

And the decision rule to assign a function to protein $\mathbf{x}_i$ is given by:

$$\begin{cases} \mathbf{x}_i \text{ is annotated with the } k\text{-th function if } F^*_{ik} > b_k^{\text{opt}}; \\ \mathbf{x}_i \text{ is not annotated with the } k\text{-th function if } F^*_{ik} \leq b_k^{\text{opt}}; \end{cases} \tag{9}$$

Data Preparation. We obtain sequence data from GenBank [34], and describe a protein sequence through one kind of its elementary constituents, *i.e.*, 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, *e.g.*, "ART" is one unit, and "MEK" is another one. The trimer histogram of a sequence hence can be used to characterize a protein $\mathbf{x}_i$, which is denoted as P_i. Because histogram indeed is a probability distribution, we use Kullback-Leibler (KL) divergence [35], a standard way to assess the difference between two probability distributions, to measure the distance between two proteins, which is defined as:

$$D_{\text{KL}}\left(P_i \parallel P_j\right) = \sum_k P_i\left(k\right) \log \frac{P_i\left(k\right)}{P_j\left(k\right)}, \tag{10}$$

where k denotes the index of the k-th trimer. Because KL divergence is non-symmetric, *i.e.*, $D_{\text{KL}}\left(P_i \parallel P_j\right) \neq D_{\text{KL}}\left(P_j \parallel P_i\right)$, we use the symmetrized KL divergence as following:

$$D_{\text{S-KL}}\left(i, j\right) = \frac{D_{\text{KL}}\left(P_i \parallel P_j\right) + D_{\text{KL}}\left(P_j \parallel P_i\right)}{2}. \tag{11}$$

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

$$\begin{aligned} W_{ij} &= D_{\text{S-KL}}\left(i, i\right) + D_{\text{S-KL}}\left(j, j\right) - 2D_{\text{S-KL}}\left(i, j\right) \\ &= -\left[D_{\text{KL}}\left(P_i \parallel P_j\right) + D_{\text{KL}}\left(P_j \parallel P_i\right)\right]. \end{aligned} \tag{12}$$

Improved Predictive Capability. We perform standard 5-fold cross validation to evaluate the compared approaches and report the average performance of 5 trials in Table 2. For FS approach, because it does not supply a threshold, we use the one giving best F1 score to make prediction. We implement two versions of our method to evaluate the contributions of each of its components. First, we solve Eq. (7) by Algorithm 1, which is the proposed method. Second, we solve a degenerate version of the problem in Eq. (7) by not incorporating the correlations between functional categories. Specifically, we replace FAF^T in Eq. (7) by FF^T, which is denoted by MDKC-S.

The results in Table 2 show that the MDKC-S and MDKC approaches clearly outperform the other compared approaches, which concretely quantify the advantage of our approaches. The improvement on classification performance of MDKC approach over MDKC-S approach clearly justify the usefulness of function-function correlations in predict putative protein functions.

Table 2. 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%

4.2 Evaluation on Integrated Biological Data

As mentioned earlier, biological data from one single experimental source only convey information for a certain aspect, 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 [12]. 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.

Data Preparation. We download PPI data for *Saccharomyces cerevisiae* species from BioGRID (version 2.0.56) [36]. 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, *e.g.*, false positive rate in yeast two-hybrid experiments is estimated as high as 50% [37]. Therefore, we use the Topological Measurement (TM) method [38] to compute the data similarity matrix W_{PPI}, which takes into consideration paths with all possible lengths on a network and weights the influence of every path by its length. Specifically, $(W_{\text{PPI}})_{ij}$ between proteins $\mathbf{x}_i$ and $\mathbf{x}_j$ is computed as:

$$(W_{\text{PPI}})_{ij} = \sum_{k=2}^{|V|-2} \text{PR}^k(i,j),$$
$$\text{PR}^k(i,j) = \frac{\text{PS}^k(i,j)}{\text{MaxPS}^k(i,j)}, \tag{13}$$

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

$$\text{PS}^k(i,j) = \left(A^k\right)_{ij}, \tag{14}$$

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

$$\text{MaxPS}^k(i,j) = \begin{cases} \sqrt{d_i d_j}, & \text{if } k = 2, \\ d_i d_j, & \text{if } k = 3, \\ \sum_{k \in N(i), l \in N(j)} \text{MaxPS}^{k-2}(k,l), & \text{if } k > 3. \end{cases} \tag{15}$$

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 [38].

We compute the sequence data similarity following the same ways in Section 4.1, which is denoted as and W_{sequence} respectively. The integrated data similarity W is hence computed as following:

$$W = W_{\text{PPI}} + \gamma W_{\text{sequence}}, \tag{16}$$

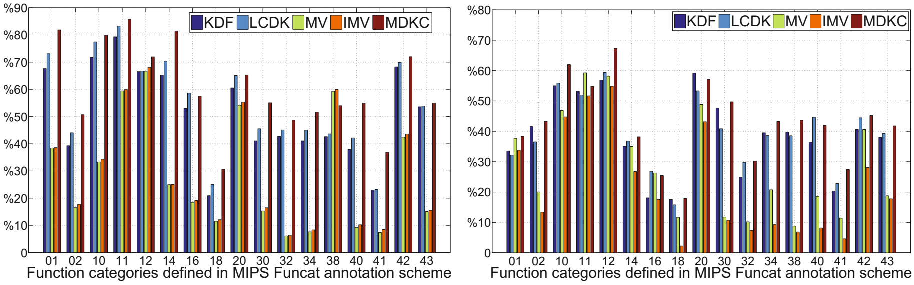
where γ is a parameter to balance the two data sources and we empirically select it as:

$$\gamma = \frac{\sum_{i,,i \neq j} W_{\text{PPI}}(i,j)}{\sum_{i,,i \neq j} W_{\text{sequence}}(i,j)}. \tag{17}$$

We compare the predictive performance of our MDKC approach to two data integration based protein function prediction approaches: kernel-based data fusion (KDF) approach [13] and locally constrained diffusion kernel (LCDK) approach [14], and two baseline approaches: majority voting (MV) approach [2] and iterative majority voting (IMV) approach [8]. The function-wise prediction performance measured by average precision and average F1 score in standard 5-fold cross validation are reported in Figure 2.

From the results in Figure 2(a) and Figure 2(b), we can see that the proposed MDKC approach consistently better than other compared approaches, sometimes very significantly, which again demonstrate the superiority of our approach.

A more careful examination on the results in Figure 2 shows that, although our approach outperforms the compared approaches in most functional categories, but not always, *e.g.*, the average precision for function “Transposable Elements,



(a) Precision.

(b) F1 score

Fig. 2. Performance of 5-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

Viral and Plasmid Proteins" (ID: 38). By scrutinizing the function category correlations, defined in Eq. (1) and illustrated in the right panel of Figure 1, we can see the average correlation of 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 testify the importance of function category correlations in predicting protein function.

5 Conclusions

In this paper, 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, we employed annotation knowledge in a novelly 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. Comprehensive empirical evaluations have been conducted on *Saccharomyces cerevisiae* genome, promising results in the experiments justified our analysis and validated the performance of our methods.

References

1. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Mol. System Biol. 3(1) (2007)
2. Schwikowski, B., Uetz, P., Fields, S.: A network of protein- protein interactions in yeast. Nat. Biotech. 18, 1257–1261 (2000)

3. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T.: Assessment of prediction accuracy of protein function from protein-protein interaction data. Yeast 18(6), 523–531 (2001)
4. Chua, H., Sung, W., Wong, L.: Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics 22(13), 1623–1630 (2006)
5. Chua, H., Sung, W., Wong, L.: Using indirect protein interactions for the prediction of Gene Ontology functions. BMC Bioinformatics 8(suppl. 4), S8 (2007)
6. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21, 302–310 (2005)
7. Weston, J., Elisseeff, A., Zhou, D., Leslie, C., Noble, W.: Protein ranking: from local to global structure in the protein similarity network. Proc. Natl. Acad. Sci. USA 101(17), 6559 (2004)
8. Vazquez, A., Flammini, A., Maritan, A., Vespignani, A.: Global protein function prediction from protein-protein interaction networks. Nat. Biotechnol. 21, 697–700 (2003)
9. Karaoz, U., Murali, T., Letovsky, S., Zheng, Y., Ding, C., Cantor, C., Kasif, S.: Whole-genome annotation by using evidence integration in functional-linkage networks. Proc. Natl Acad. Sci. USA 101(9), 2888–2893 (2004)
10. Liang, S., Shuiwang, J., Jieping, Y.: Adaptive diffusion kernel learning from biological networks for protein function prediction. BMC Bioinformatics 9, 162 (2008)
11. Wang, H., Huang, H., Ding, C.: Function-function correlated multi-label protein function prediction over interaction networks. In: Chor, B. (ed.) RECOMB 2012. LNCS, vol. 7262, pp. 302–313. Springer, Heidelberg (2012)
12. Whisstock, J., Lesk, A.: Prediction of protein function from protein sequence and structure. Q. Rev. Biophysics 36(3), 307–340 (2004)
13. Lanckriet, G., Deng, M., Cristianini, N., Jordan, M., Noble, W.: Kernel-based data fusion and its application to protein function prediction in yeast. In: Proc. of Pacific Symp. on Biocomputing, vol. 9, pp. 300–311 (2004)
14. Tsuda, K., Noble, W.: Learning kernels from biological networks by maximizing entropy. Bioinformatics 20, 326–333 (2004)
15. Shi, L., Cho, Y., Zhang, A.: ANN Based Protein Function Prediction Using Integrated Protein-Protein Interaction Data. In: Proc. of International Joint Conf. on Bioinformatics, Systems Biol. and Intelligent Comp., pp. 271–277 (2009)
16. Shin, H., Lisewski, A., Lichtarge, O.: Graph sharpening plus graph integration: a synergy that improves protein functional classification. Bioinformatics 23(23), 3217 (2007)
17. Sun, L., Ji, S., Ye, J.: Adaptive diffusion kernel learning from biological networks for protein function prediction. BMC Bioinformatics 9(1), 162 (2008)
18. Wang, H., Huang, H., Ding, C.: Protein function prediction via laplacian network partitioning incorporating function category correlations. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 2049–2055. AAAI Press (2013)
19. Wang, H., Huang, H., Ding, C.: Image Annotation Using Multi-label Correlated Green's Function. In: Proc. of IEEE ICCV 2009, pp. 2029–2034 (2009)
20. Wang, H., Ding, C., Huang, H.: Multi-label linear discriminant analysis. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 126–139. Springer, Heidelberg (2010)

21. Wang, H., Huang, H., Ding, C.: Multi-label feature transform for image classifications. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 793–806. Springer, Heidelberg (2010)
22. Wang, H., Huang, H., Ding, C.: Image annotation using bi-relational graph of images and semantic labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2011 (CVPR 2011), pp. 793–800 (2011)
23. Mewes, H., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., Frishman, D.: MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 27(1), 44 (1999)
24. Cai, D., He, X., Wu, X., Han, J.: Non-negative matrix factorization on manifold. In: Proc. of ICDM (2008)
25. Gu, Q., Zhou, J.: Co-clustering on manifolds. In: Proc. of SIGKDD (2009)
26. Cai, D., He, X., Han, J., Huang, T.S.: Graph regularized non-negative matrix factorization for data representation. IEEE Trans. Pattern Analysis Mach. Intell. 99 (2010)
27. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix t-factorizations for clustering. In: SIGKDD (2006)
28. Ding, C., Li, T., Jordan, M.: Convex and semi-nonnegative matrix factorizations. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(1), 45–55 (2010)
29. Wang, H., Nie, F., Huang, H., Makedon, F.: Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, vol. 2, pp. 1553–1558. AAAI Press (2011)
30. Wang, H., Nie, F., Huang, H., Ding, C.: Dyadic transfer learning for cross-domain image classification. In: Proc. of ICCV, pp. 551–556. IEEE (2011)
31. Wang, H., Nie, F., Huang, H., Ding, C.: Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation. In: Proceedings of ICDM (2011)
32. Wang, H., Huang, H., Ding, C., Nie, F.: Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. In: Chor, B. (ed.) RECOMB 2012. LNCS, vol. 7262, pp. 314–325. Springer, Heidelberg (2012)
33. Li, T., Ding, C., Jordan, M.: Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In: Proc. of ICDM (2007)
34. Benson, D., Karsch-Mizrachi, I., Lipman, D.: GenBank. Nucleic Acids Res. 34, D16–D20 (2006)
35. Kullback, S., Leibler, R.: On information and sufficiency. The Annals of Mathematical Statistics, 79–86 (1951)
36. Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34(database issue), D535 (2006)
37. Deane, C., Salwinski, L., Xenarios, I., Eisenberg, D.: Protein Interactions Two Methods for Assessment of the Reliability of High Throughput Observations. Mol. & Cellular Proteomics 1(5), 349–356 (2002)
38. Pei, P., Zhang, A.: A topological measurement for weighted protein interaction network. In: Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference, pp. 268–278 (2005)