# Protein Function Prediction via Laplacian Network Partitioning Incorporating Function Category Correlations

**Hua Wang[†], Heng Huang[‡*], Chris Ding[‡]**
[†]Department of Electrical Engineering and Computer Science
Colorado School of Mines, Golden, Colorado 80401, USA
[‡]Department of Computer Science and Engineering
University of Texas at Arlington, Arlington, Texas 76019, USA
huawangcs@gmail.com, heng@uta.edu, chqding@uta.edu

## Abstract

Understanding the molecular mechanisms of life requires decoding the functions of the proteins in an organism. Various high-throughput experimental techniques have been developed to characterize biological systems at the genome scale. A fundamental challenge of the post-genomic era is to assign biological functions to all the proteins encoded by the genome using high-throughput biological data. To address this challenge, we propose a novel Laplacian Network Partitioning incorporating function category Correlations (LNPC) method to predict protein function on protein-protein interaction (PPI) networks by optimizing a Laplacian based quotient objective function that seeks the optimal network configuration to maximize consistent function assignments over edges on the whole graph. Unlike the existing approaches that have no unique optimization solutions, our optimization problem has unique global solution by eigen-decomposition methods. The correlations among protein function categories are quantified and incorporated into a correlated protein affinity graph which is integrated into the PPI graph to significantly improve the protein function prediction accuracy. We apply our new method to the BioGRID dataset for the $Saccharomyces\ Cerevisiae$ species using the MIPS annotation scheme. Our new method outperforms other related state-of-the-art approaches more than 63% by the average precision of function prediction and 53% by the average F1 score.

## 1 Introduction

Discovering biological functions of an organism is a central goal of functional genomics. Although function assignment for every protein using traditional experimental techniques could take decades, the current accumulated data from different biological sources make it possible to automatically predict protein functions to guide laboratory experiments and speed up the annotation process. Existing methods typically make prediction one function at a time, fundamentally, although in reality most biological functions are intertwined to be carried out together. For example, "Metabolism" and "Protein Fate" [Mewes *et al.*, 1999] are closely related and often annotated to the same protein. Therefore, the underlying relationships among biological functions convey valuable information which could be utilized to improve the overall protein function prediction accuracy [Wang *et al.*, 2012; 2013]. However, to the best of our knowledge, very limited computational research has been done to make use of the functional correlations. In this study, we propose a Laplacian Network Partitioning Incorporating Function Category Correlations (LNPC) approach to incorporate the functional correlations into a network based method for protein function prediction. We first propose a novel Laplacian Network Partitioning (LNP) method, as part of LNPC, to formulate protein function prediction as an optimization problem to maximize the function assignment consistency over edges on the whole protein interaction network. After that we introduce a statistical model to quantify the function category correlations, based on which we construct a correlated protein affinity graph and integrate it into biological protein interaction networks. The experimental results show that our LNPC method outperforms other state-of-the-art methods.

### 1.1 Network Based Methods for Protein Function Prediction

High-throughput technologies for protein-protein interaction (PPI) screening have created large-scale data across human and many model species [Ito *et al.*, 2001; Uetz *et al.*, 2000; Ho *et al.*, 2002; Tong *et al.*, 2001; Edgar *et al.*, 2002; Pellegrini *et al.*, 1999; Enright *et al.*, 1999; Harbison *et al.*, 2004], which are routinely represented as networks, with nodes representing proteins and edges representing the detected PPIs. The availability of protein interaction networks has spurred on the development of network based computational methods to elucidate protein functions [Sharan *et al.*, 2007].

The most natural and straightforward method to predict protein function on PPI networks is neighbor counting, because it is observed that 70–80% of proteins share at least one function with its interacting partner [Titz *et al.*, 2004]. Schwikowski *et al*. [Schwikowski *et al.*, 2000] suggested a

*Corresponding Author. This work was partially supported by NSF IIS-1117965.

majority voting (MV) approach that labels a protein with the functions occurring most frequently in its interacting partners. Hishigaki *et al*. [Hishigaki *et al.*, 2001] used $\chi^2$ statistics to identify the functions that are over-represented in the interacting partners of a protein. However, only using the immediate interaction partners limits predictions to proteins with at least one interaction partner with known annotation. Moreover, the possible annotations for an unknown protein are limited by the annotations of its interacting partners. Chua *et al*. [Chua *et al.*, 2006] tackled these two problems by further considering the functions annotated to the indirect interacting partners of a protein in addition to the direct interacting partners. In general, such neighbor counting approaches only take advantage of the local structures of a PPI network.

In contrast, several methods have been proposed toward global optimization by taking into account the full topology of the network. Vazquez *et al*. [Vazquez *et al.*, 2003] aimed at assigning a function to each unannotated proteins so as to maximize the number of edges that connect proteins assigned with the same function, and solve the problem by simulated annealing. Karaoz *et al*. [Karaoz *et al.*, 2004] developed the same idea with an iterative local search method. Nabieva *et al*. [Nabieva *et al.*, 2005] treated the optimization problem as a generalization of multi-way $k$-cut problem, and used integer linear programming method to solve the problem. Although these researches have shown effective experimental results, the presented methods are not able to produce a unique solution for the optimization problem, which makes them of less practical use in real applications. In this study, we not only adopt the ideas in these previous work to seek the optimal network configuration to maximize consistent function assignments over edges on the whole graph, but also reformulate the optimization objective as a quotient discriminant function. Most importantly, the proposed optimization objective can be efficiently and uniquely resolved by the standard generalized eigen-decomposition method.

### 1.2 Protein Function Prediction Using Function Category Correlations

A protein is usually observed to play several functional roles in different biological processes within an organism, thus it is natural to annotate a protein with multiple functions. In protein function prediction, the biological functions are usually correlated to each other, because most biological processes accomplish together with other processes but seldom happen individually. If a protein is known to be annotated to one category, it is very likely also annotated to those categories highly correlated to the annotated one. Apparently, we can leverage the underlying category relationships to improve the protein function prediction. For example, when applying Functional Catalogue (FunCat) 2.1 annotation scheme [Mewes *et al.*, 1999] to the yeast genome, we observe that there is a big overlap between the proteins annotated to function "Cell Fate" (ID: 40) and "Cell Type Differentiation" (ID: 43). As shown in Fig. 1, among 268 proteins annotated with function "Cell Fate" in the yeast genome, 168 proteins are also annotated with function"Cell Type Differentiation", but the average number of proteins annotated with other functions is only about 51. Thus, we speculate these two functions are highly correlated from the statistics of the annotated data. If a protein is assigned with function "Cell Fate" by experimental or computational algorithms, we can infer that this protein also has a high probability to be annotated with function "Cell Type Differentiation". With this recognition, we propose a correlated protein affinity graph to incorporate the function category correlations and integrate it into biological protein interaction networks to significantly enhance the overall prediction performance.

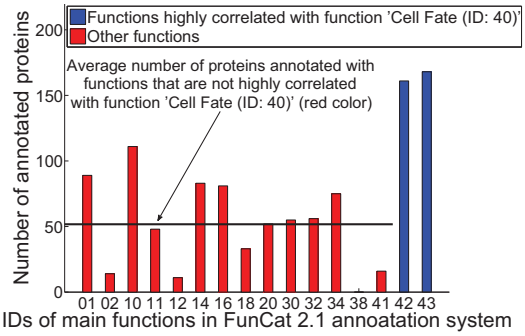


Figure 1: Distribution of numbers of proteins known to be annotated with function "Cell Fate" (ID: 40) and also annotated with other main functional categories in Funcat 2.1 annotation scheme.

## 2 Materials

**Physical interaction network.** We construct the PPI network using the protein interaction dataset compiled by BioGRID [Stark *et al.*, 2006] for the *Saccharomyces Cerevisiae* (yeast) species. The resulting network is an undirected graph, where each vertex represent a protein and an edge represent an observed physical link. For all the reported results, we consider only the proteins making up the largest connected component of the physical interaction map from the BioGRID 2.0.56 dataset, which end upwith 4403 proteins and 86167 physical interaction links.

**Function annotation dataset.** In this study, we use the FunCat annotation scheme (version 2.1) by Munich Information Center for Protein Sequences (MIPS) [Mewes *et al.*, 1999] due to its clear tree-like hierarchical structure. 27 main functional categories are defined in FunCat 2.1 annotation scheme, among which 17 functions are annotated to the yeast genome. The function identifiers (ID) and the descriptions of these 17 main functional categories are listed in Table 1.

## 3 Protein Function Prediction via Laplacian Network Partitioning Incorporating Function Category Correlations

### 3.1 Problem Formalization

We formalize the protein function prediction problem. We have $n$ proteins, of which $l$ proteins have known annotations. Our task is to annotate the rest $n-l$ proteins using the network data. In this problem, there are $K$ biological functions. Thus

Table 1: Main functional categories in Funcat 2.1 annotation scheme.

| Function ID | Function Description | Number of proteins (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 (Folding, Modification, Destination) | 1125 |
| 16 | Protein with Binding Function or Cofactor Requirement (Structural or Catalytic) | 1019 |
| 18 | Regulation of Metabolism and Protein Function | 246 |
| 20 | Cellular Transport, Transport Facilitation and Transport Routes | 995 |
| 30 | Cellular Communication/Signal Transduction Mechanism | 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 (Systemic) | 67 |
| 42 | Biogenesis of Cellular Components | 827 |
| 43 | Cell Type Differentiation | 437 |

the task is to assign one or more biological functions to each of the unannotated proteins.

Mathematically, we represent the annotated proteins as $(x_1, \cdots, x_l)$, and unannotated proteins as $(x_{l+1}, \cdots, x_n)$. The protein network data is given by the an $n \times n$ affinity matrix $W$ with $W_{ij}$ indicating the affinity between $x_i$ and $x_j$. Using graph theory terminology, we say the protein network is a graph $G = (V, E)$, where the nodes $V$ corresponds to proteins $\{x_1, \ldots, x_n\}$, and the edges $E$ are edge weights $W$.

As in most previous approaches [Schwikowski *et al.*, 2000; Hishigaki *et al.*, 2001; Chua *et al.*, 2006; Vazquez *et al.*, 2003; Karaoz *et al.*, 2004; Nabieva *et al.*, 2005; Sharan *et al.*, 2007], the function prediction is carried out for one function at a time, repeating $K$ time for $K$ functions. In each prediction task, we use indicator $y_i = \pm 1$ for protein $x_i$, where $y_i = +1$ indicates that protein $x_i$ has the function in question, $y_i = -1$ indicates that protein $x_i$ does not have the function in question.

Our work begins with an observation that in essence, under the one-function-at-a-time prediction framework, the function prediction becomes a graph node partitioning problem under the constraints that annotated proteins are fixed to the known function assignments. It is known that [Pothen *et al.*, 1990; Chung, 1997; Shi and Malik, 2000; Ding *et al.*, 2007] "spectral graph partitioning" has been shown to be a state-of-the-art partitioning approach.

Motivated by this observation, (1) we further recognized that the original prediction model [Vazquez *et al.*, 2003; Karaoz *et al.*, 2004; Nabieva *et al.*, 2005] can be transformed into a Laplacian network partitioning (LNP) model closely related to spectral graph partitioning model. (2) In the LNP model, we relax the indicators $\{y_i\}$. This requires a proper re-formulation to enforce the constraints (enforce annotated proteins to have their known functions). We provide an effective quadratic formulations to resolve this problem. (3) In this LNP approach, we can easily incorporate the protein function category correlations as detailed later in Sect. 3.3. These 3 steps are the main contributions of this work.

### 3.2 Computational Algorithms

Compared with local neighbor counting approaches [Schwikowski *et al.*, 2000; Hishigaki *et al.*, 2001; Chua *et al.*, 2006], global optimization approaches usually demonstrate better performance in predicting protein functions [Vazquez *et al.*, 2003; Karaoz *et al.*, 2004; Nabieva *et al.*, 2005]. Vazquez *et al*. [Vazquez *et al.*, 2003] proposed to exploit the global topology of the whole PPI network by annotating proteins to minimize the number of times different annotations are associated with neighboring proteins. Karaoz *et al*. [Karaoz *et al.*, 2004] developed a similar approach by augmenting the physical protein interaction networks using gene-expression data. Their optimization objective can be formulated as following:

$$\begin{aligned} \max_{\vec{y}} \quad & \left( \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} W_{ij} y_i y_j \right) \;=\; \max_{\vec{y}} \left( \vec{y}^T W \vec{y} \right), \\ s.t. \quad & y_i = 1 \quad \forall i \in S_+, \\ & y_i = -1 \quad \forall i \in S_-, \end{aligned} \tag{1}$$

where in the constraints (the "boundary condition"), $S_+$ is the set of proteins that are annotated to have the function of interest (positive samples), and $S_-$ is the set of those annotated proteins which do not have the function (negative samples). Here we denote $\vec{y} = [y_1, \ldots, y_n]^T$ and assume $W_{ii} = 0$.

In this study, we reformulate the optimization objective in a better way and provide the unique optimal solution this problem. We first note that the constraints in Eq. (1) can be satis-

fied by following penalty function:

$$p(\vec{y}) = \left( \frac{\sum_{x_i \in S_+} y_i}{|S_+|} - \frac{\sum_{x_i \in S_-} y_i}{|S_-|} \right)^2 = \vec{y}^T E \vec{y},$$

$$E_{ij} = \begin{cases} \frac{1}{|S_+|^2} & \text{if } x_i \in S_+ \text{ and } x_j \in S_+ \\ \frac{1}{|S_-|^2} & \text{if } x_i \in S_- \text{ and } x_j \in S_- \\ \frac{-2}{|S_+||S_-|} & \text{if } \begin{cases} x_i \in S_+ \text{ and } x_j \in S_- \\ x_i \in S_- \text{ and } x_j \in S_+ \end{cases} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

This quadratic function $p(\vec{y})$ reaches the maximum value of 4 when the constraints are completely satisfied. Our solution let those annotated proteins be also dynamic variables. Thus the constraints becomes:

$$\max_{\vec{y}} p(\vec{y}) \quad (3)$$

Second, we note that the maximization of $\vec{y}^T W \vec{y}$ can be equivalently written as:

$$\min_{\vec{y}} \ \vec{y}^T (D - W) \vec{y}, \quad (4)$$

where $D = \text{diag}\,(d_1, d_2, \cdots, d_n)$ and $d_i = \sum_j W_{ij}$ is the degree of vertex (protein) $x_i$. This is because $\vec{y}^T D \vec{y} - \vec{y}^T W \vec{y} = \sum_{i=1}^n D_{ii} y_i^2 - \vec{y}^T W \vec{y} = \sum_{i=1}^n d_i - \vec{y}^T W \vec{y} = \text{const} - \vec{y}^T W \vec{y}$, noting that $y_i \pm 1$ implies $y_i^2 = 1$ and $\sum_{i=1}^n d_i$ is a constant.

Generally speaking, we need to combine both Eqs. (3–4) into a single objective to minimize them. For example, we can optimize:

$$\min_{\vec{y}} \left[ (1 - \beta)\, \vec{y}^T (D - W) \vec{y} - \beta \vec{y}^T E \vec{y} \right] \quad (5)$$

where $0 < \beta < 1$ is a tradeoff parameter. The problem with the this formulation lies in that the tradeoff parameter $\beta$ is hard to select in practice. Different choice of $\beta$ leads to different solution. Here we propose a way to bypass this difficulty by defining the following optimization objective [Wang *et al.*, 2010d; 2010c]:

$$\max_{\vec{y}} \frac{\vec{y}^T E \vec{y}}{\vec{y}^T (D - W) \vec{y}} \quad (6)$$

This formulation achieves the two objectives Eqs. (3–4) without additional parameters. The solution to the problem in Eq. (6) is well established. It is given by the generalized eigenvalue problem, $\lambda E \vec{v} = (D - W) \vec{v}$, the eigenvectors $\vec{v}_k$ corresponds to the eigenvalues $\lambda_k$ where $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_n$. Because $\vec{v}_1$ is a constant vector [Chung, 1997], we use the second smallest eigenvector $\vec{v}_2$ as the desired solution. We call this approach for protein function prediction as Laplacian Network Partitioning (LNP) because it partitions the nodes of the network into two parts. Note that the solution to problem Eq. (6) is unique, while in many previous works the global unique solution to the optimization problem is not guaranteed.

Once $\vec{v}_2$ from Eq. (6) is computed, we can obtain the final the function assignment for an unannotated protein $x_i$ by the simple decision

$$y_i = \begin{cases} +1, & \text{if } \vec{v}_2(i) > 0 \\ -1, & \text{if } \vec{v}_2(i) < 0 \end{cases} \quad (7)$$

This corresponds to use 0 as the decision value and is not necessarily optimal, especially in the case in protein function prediction, where only a small fraction of proteins are annotated to a given biological function. Taking into account the unbalanced distribution of training data, we adjust the decision boundary such that the weighted training errors are minimized. Let $e_+(b)$ and $e_-(b)$ be the numbers of misclassified positive and negative training samples for a given decision boundary $b$. We choose the optimal decision boundary as following [Wang *et al.*, 2009; 2012; 2013]:

$$b^{\text{opt}} = \arg\min_b \left( \frac{e_+(b)}{|S_+|} + \frac{e_-(b)}{|S_-|} \right). \quad (8)$$

The function assignment for an unannotated protein $x_i$ is then determined by:

$$y_i = \begin{cases} +1, & \text{if } \vec{v}_2(i) > b^{\text{opt}} \\ -1, & \text{if } \vec{v}_2(i) < b^{\text{opt}} \end{cases} \quad (9)$$

### 3.3 Construction of Correlated Protein Affinity Matrix ($W$)

Existing approaches construct the protein affinity matrix, $W$ in Eq. (1), only from the biological experimental data, such as those from the high-throughput technologies, while the correlations among protein functions are usually overlooked. Since the function correlations convey valuable information to infer protein function assignment as discussed earlier in Sect. 1.2, it is expected to improve the the overall predictive accuracy by making use of them. We thus propose the following scheme to construct a Correlated Protein Affinity Matrix (CPAM), $W$, to incorporate the function correlations as following:

$$W = W_0 + \gamma W_{\text{L}} \quad (10)$$

where $W_0$ is the affinity matrix built upon the biological experimental data same as that in previous approaches, $W_{\text{L}}$ is the pairwise function annotation similarity matrix to incorporate the correlations among protein functions, and $\gamma$ is a parameter used to balance the influence of the two affinity matrices and empirically selected as $\gamma = \frac{\sum_{i,j,i \neq j} W_0(i,j)}{\sum_{i,j,i \neq j} W_{\text{L}}(i,j)}$.

In protein function prediction, one protein can be assigned to multiple functions simultaneously, therefore the proteins assigned to two different functions may overlap. Intuitively, the bigger the overlap is, the more closely the two functions are related to each other. Considering this function correlations, the function assignments to a protein are no longer independent, but can be inferred one another. In the extreme case, such as parent-child hierarchy in the protein function annotation systems, once we know a protein is annotated to a child protein function, we can immediately annotate the parent function to the same protein. In this subsection, we concentrate on modeling the correlations among protein functions and incorporating them into protein affinity graphs.

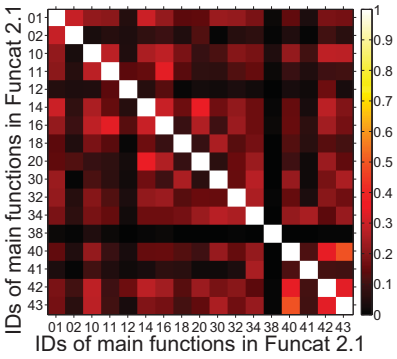


Figure 2: Correlation matrix among the 17 functions in FunCat 2.1 annotated to yeast genome.

The overlap of protein annotations measures the correlation between function categories. We use the cosine similarity to measure correlations between these two protein functions. Let $K$-vector $\vec{z}_i$ represents the function assignment indication vector for protein $x_i$, such that $\vec{z}_i(k) = +1$ if protein $x_i$ is known to be annotated with the $k$th function, and $\vec{z}_i(k) = -1$ if protein $x_i$ is known to be not annotated with the $k$th function. We write $Z = [\vec{z}_1, \ldots, \vec{z}_n]$. Let $\vec{r}_k$ and $\vec{r}_l$ represent the $k$th and $l$th rows of $Z$ respectively. We define the function category correlation matrix, $C \in \mathbb{R}^{K \times K}$, to capture the function correlation between two categories as following:

$$C_{kl} = \cos(\vec{r}_k, \vec{r}_l) = \frac{\langle \vec{r}_k, \vec{r}_l \rangle}{\|\vec{r}_k\| \|\vec{r}_l\|} \tag{11}$$

Using the FunCat 2.1 annotation dataset for yeast genome, the function correlations defined in Eq. (11) are illustrated in Fig. 2. The high correlation value between functions "Cell Fate" and "Cell Type Differentiation" depicted in Fig. 2 shows that they are highly correlated, which agrees with the observations in Sect. 1.2. In addition, as shown in Fig. 2 some other function pairs are 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 comply with the biological nature, which justifies of the utility of the function correlations from biological perspective.

A simple form to measure the overlap of the annotated functions to two proteins is $\vec{z}_i^T \vec{z}_j$. The bigger the overlap is, the more similar the proteins are. The problem with this straightforward similarity measurement is that it treats all the protein functions independently and therefore is unable to explore the correlations among them. In particular, it will give zero similarity whenever two proteins do not share annotated functions. However, two proteins with no common annotated functions can still be strongly related if their annotated functions are highly correlated. Therefore, instead of computing the function annotation similarity by the dot product, we compute it by $\vec{z}_i^T \, C \, \vec{z}_j$. By normalization, the pairwise function annotation similarity, $W_{\text{L}}$, is defined as following:

$$W_{\text{L}}(i,j) = \frac{\vec{z}_i^T \, C \, \vec{z}_j}{\| \vec{z}_i \| \, \| \vec{z}_j \|} \tag{12}$$

Unannotated proteins are first initialized using MV approach.

Applying $W_{\text{L}}$ into Eq. (10), the correlated protein affinity matrix is constructed, and the function correlations are naturally incorporated. We call this approach Laplacian Network Partitioning Incorporating Function Category Correlation (LNPC) (in contrast to LNP where function correlation is not used).

We note that, by taking into account the function correlations, protein function prediction is an ideal case of multi-label classification [Wang *et al.*, 2012; 2013], which was recently formalized in machine learning. Due to its wide applicability, multi-label learning [Wang *et al.*, 2009; 2010d; 2010a; 2010b; 2011] has attracted a lot of attention in scientific research in recent years.

## 4 Experimental Results and Discussion

We applied the proposed approach to the yeast interaction data with 4403 proteins and 86167 interactions. We use 17 protein function categories which are annotated to yeast genome (See Section 2). We evaluate our method using the standard 5-fold cross validation, as in many previous studies. We also implemented four methods proposed in previous studies, including Majority Voting (MV) approach [Schwikowski *et al.*, 2000], Iterative Majority Voting (IMV) approach [Vazquez *et al.*, 2003], $\chi^2$ approach [Hishigaki *et al.*, 2001], FunctionalFlow (FF) approach [Nabieva *et al.*, 2005].

**Cross validation**. We use standard 5-fold cross validation method. The proteins are divided into 5 equal-size groups randomly. One group is assumed to be unannotated and the rest 4 groups are annotated. We run all 5 prediction methods to predict the functions for the kept-out group of proteins. The predicted results are compared to the true functions of these proteins. This is repeated 5 times to keep each group as unannotated in turn, and final results are averaged.

**Performance metrics**. As in many previous studies, we choose *precision* and *F1 score* to evaluate the prediction performance. Let TP (true positive) be the number of proteins which we correctly predict to have a given function, FP (false positive) be the number of proteins which we incorrectly predict to have the function, and FN (false negative) be the number of proteins which we incorrectly predict to not have the function. The "precision" is defined as $\text{TP}/(\text{TP} + \text{FP})$, and the "recall" (also known as "sensitivity") is defined as $\text{TP}/(\text{TP} + \text{FN})$. In addition, we also use the "F1 score" to evaluate precision and recall together, which is the harmonic mean of precision and recall: defined as $(2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$. F1 score is extensively used in the related works and other domains such as information retrieval. Typically, improving the precision of an algorithm decreases its recall and vice versa, therefore F1 score is a balanced performance metric. To measure the overall prediction performance, we use average precision and average F1 score over all 17 main functional categories to evaluate our algorithm.

### 4.1 Function Prediction

We compare the performances of Majority Voting (MV) approach [Schwikowski *et al.*, 2000], Iterative Majority Voting

Table 2: Average precision and average F1 score by the five approaches in comparison over the main functional categories by FunCat 2.1.

| Approaches | Average Precision | Average F1 score |
|---|---|---|
| MV | 30.12% | 28.56% |
| IMV | 30.92% | 21.69% |
| $\chi^2$ | 13.76% | 7.32% |
| FunctionalFlow | 17.99% | 18.21% |
| LNPC | **49.20%** | **43.70%** |

(IMV) approach [Vazquez *et al.*, 2003], $\chi^2$ approach [Hishigaki *et al.*, 2001], FunctionalFlow (FF) approach [Nabieva *et al.*, 2005], and proposed Laplacian Network Partitioning Incorporating Function Category Correlations (LNPC) approach, on the PPI graph built from BioGRID data of version 2.0.54 with annotation by MIPS Funcat scheme of version 2.1, using 5-fold cross validation.

The overall prediction performance measured by average precision and average F1 score are listed in Table 2. The LNPC results are improved significantly over other approaches. This quantifies the advantages of the proposed LNPC approach, and demonstrates that the reformulation on the optimization objective with the additional constraint on the annotated proteins does improve the prediction of protein functions. The relative improvement on average precision of the proposed approach over the best of the other approaches is about $(49.20\% - 30.12\%)/30.12\% = 63.05\%$, and that on average F1 score is more than 53%.

When calculating the overall performance of four previous approaches as shown in Table 2, we use their respective optimal parameters: in MV approach we select the 3 most frequently occurring functions in a protein's neighbors; in $\chi^2$ approach radius $= 1$ gives the best performance; in FF approach we assign functions according to the proportions of positive and negative training samples as suggested by [Nabieva *et al.*, 2005].

### 4.2 Effectiveness of utilizing Correlations among Protein Function categories

We further analyze the effectiveness of using the function correlations in protein function prediction. We compare the prediction results of LNPC (utilizing the correlations among function categories) against the baseline LNP (not using the correlations).

Fig. 3 illustrates the LNPC results vs LNP results across all 17 function categories. LNPC approach consistently outperform the LNP approach in every function category. We also notice that the improvements for "Metabolism", "Biogenesis of Cellular Components" and "Cell Type Differentiation" are among the highest. By examining the correlation matrix $C$ defined in Eq. (11), the correlations among these function categories are relatively high, indicating the function correlation is the direct cause of the improvements observed in Fig. 3.

Table 3 presents the overall prediction performance comparisons of LNP and LNPC by 5-fold cross validation. The results show that the LNPC approach clearly outperform the

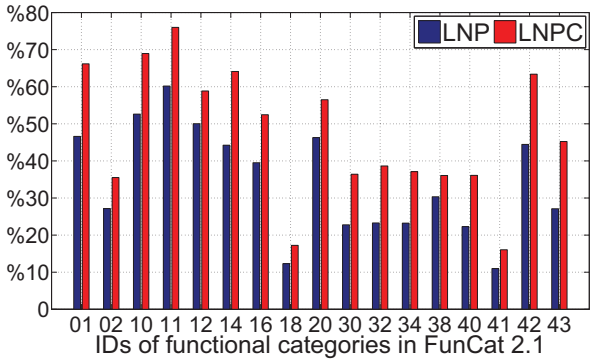


Figure 3: Prediction performance utilizing correlations among function categories (LNPC) vs the baseline (LNP) not utilizing the correlation. LNPC results are improved across all 17 function categories.

Table 3: Prediction performance comparison between LNP and LNPC.

| | Average Precision | Average F1 Score |
|---|---|---|
| LNP: $W = W_0$ | 33.20% | 36.24% |
| LNPC: $W = W_0 + \gamma W_{\mathrm{L}}$ | 48.75% | 42.50% |

LNP approach. The result of Table 3 and Fig. 3 conclusively demonstrate that utilizing the correlations among function categories improve function prediction significantly.

## 5 Conclusions

We proposed a new Laplacian network partitioning approach incorporating function category correlations for protein function prediction, and showed its promising performance compared to other related state-of-the-art approaches. Our proposed approach aims at global optimization to utilize the full topology of the whole protein interaction networks. Unlike the existing global optimization approaches, we formulate the optimization objectives as a Laplacian based model, which places the protein function prediction under the spectral clustering framework from graph theory perspective and provides state-of-the-art partitioning capabilities such that the prediction performance is enhanced. The optimization problem by the proposed approach is parameter free and can be efficiently and uniquely solved by eigen-decomposition methods. However, most existing related work only used heuristic or simulating methods to solve the problem and cannot give out a unique solution. Moreover, we proposed a statistical scheme to model measure the correlations among protein function categories, by which we introduced a protein affinity graph to naturally incorporate the function category correlations. After such correlated protein affinity graph is integrated into proposed Laplacian network partitioning method, both overall and function-wise prediction performance are significantly improved.

## References


[Chua *et al.*, 2006] H.N. Chua, W.K. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006.

[Chung, 1997] F.R.K. Chung. *Spectral graph theory*. Amer Mathematical Society, 1997.

[Ding *et al.*, 2007] Chris Ding, Horst D Simon, Rong Jin, and Tao Li. A learning framework using green's function and kernel regularization with application to recommender system. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 260–269. ACM, 2007.

[Edgar *et al.*, 2002] R. Edgar, M. Domrachev, and A.E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207, 2002.

[Enright *et al.*, 1999] A.J. Enright, I. Iliopoulos, N.C. Kyrpides, and C.A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, 1999.

[Harbison *et al.*, 2004] C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J.B. Tagne, D.B. Reynolds, J. Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.

[Hishigaki *et al.*, 2001] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 18(6):523–531, 2001.

[Ho *et al.*, 2002] Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, 2002.

[Ito *et al.*, 2001] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

[Karaoz *et al.*, 2004] U. Karaoz, TM Murali, S. Letovsky, Y. Zheng, C. Ding, C.R. Cantor, and S. Kasif. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl Acad. Sci. USA*, 101(9):2888–2893, 2004.

[Mewes *et al.*, 1999] HW Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, and D. Frishman. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, 27(1):44, 1999.

[Nabieva *et al.*, 2005] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps Bioinformatics, 21. *Bioinformatics*, 21:302–310, 2005.

[Pellegrini *et al.*, 1999] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, R. Grothe, and T.O. Yeates. Assigning protein functions by comparative genome analysis protein phylogenetic profiles, 1999.

[Pothen *et al.*, 1990] A. Pothen, H. D. Simon, and K. P. Liou. Partitioning sparse matrices with egenvectors of graph. *SIAM Journal of Matrix Anal. Appl.*, 11:430–452, 1990.

[Schwikowski *et al.*, 2000] B. Schwikowski, P. Uetz, and S. Fields. A network of protein- protein interactions in yeast. *Nat. Biotechnol.*, 18:1257–1261, 2000.

[Sharan *et al.*, 2007] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Molecular systems biology*, 3(1), 2007.

[Shi and Malik, 2000] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE. Trans. on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[Stark *et al.*, 2006] C. Stark, B.J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34(Database Issue):D535, 2006.

[Titz *et al.*, 2004] B. Titz, M. Schlesner, and P. Uetz. What do we learn from high-throughput protein interaction data? *Expert Review of Proteomics*, 1(1):111–121, 2004.

[Tong *et al.*, 2001] A.H.Y. Tong, M. Evangelista, A.B. Parsons, H. Xu, G.D. Bader, N. Page, M. Robinson, S. Raghibizadeh, C.W.V. Hogue, H. Bussey, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants, 2001.

[Uetz *et al.*, 2000] P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, et al. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.

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

[Wang *et al.*, 2009] Hua Wang, Heng Huang, and Chris Ding. Image annotation using multi-label correlated green's function. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2029–2034. IEEE, 2009.

[Wang *et al.*, 2010a] Hua Wang, Chris Ding, and Heng Huang. Multi-label classification: Inconsistency and class balanced k-nearest neighbor. In *Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010)*, 2010.

[Wang *et al.*, 2010b] Hua Wang, Chris Ding, and Heng Huang. Multi-label linear discriminant analysis. In *ECCV 2010*, pages 126–139. Springer, 2010.

[Wang *et al.*, 2010c] Hua Wang, Heng Huang, and Chris Ding. Discriminant laplacian embedding. In *Proc. AAAI Conf. Artificial Intelligence (AAAI 2010)*, pages 618–623, 2010.

[Wang *et al.*, 2010d] Hua Wang, Heng Huang, and Chris Ding. Multi-label feature transform for image classifications. In *ECCV 2010*, pages 793–806. Springer, 2010.

[Wang *et al.*, 2011] Hua Wang, Heng Huang, and Chris Ding. Image annotation using bi-relational graph of images and semantic labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2011)*, pages 793–800. IEEE, 2011.

[Wang *et al.*, 2012] Hua Wang, Heng Huang, and Chris Ding. Function-function correlated multi-label protein function prediction over interaction networks. In *Research in Computational Molecular Biology (RECOMB 2012)*, pages 302–313. Springer, 2012.

[Wang *et al.*, 2013] Hua Wang, Heng Huang, and Chris Ding. Function-function correlated multi-label protein function prediction over interaction networks. *Journal of Computational Biology*, 20(4):322–333, 2013.