Protein Function Prediction using Weak-label Learning

Guoxian Yu
School of Comp. Sci. & Tech.
South China Univ. of Tech.
Guangzhou,510006,China
guoxian.yu@mail.scut.edu.cn

Guoji Zhang
School of Sciences
South China Univ. of Tech.
Guangzhou,510640,China
magjzh@scut.edu.cn

Huzefa Rangwala
Dept. of Computer Science
George Mason University
Fairfax,22030,VA
rangwala@cs.gmu.edu

Carlotta Domeniconi
Dept. of Computer Science
George Mason University
Fairfax,22030,VA
carlotta@cs.gmu.edu

Zhiwen Yu
School of Comp. Sci. & Tech.
South China Univ. of Tech.
Guangzhou,510006,China
zhwyu@scut.edu.cn

ABSTRACT

Protein function prediction is one of the fundamental issues in the post-genomic era. Multi-label learning is widely used for predicting functions of proteins. Most multi-label learning methods assume that the proteins with annotation do not have any missing functions. However, in practice, we may have a subset of the ground-truth functions for a protein, and whether the protein has other functions is unknown. To complete the partial annotation of proteins, we propose a *Pro*tein Function Prediction method with *W*eak-label *L*earning (ProWL), and a variant of ProWL (ProWL-IF). Both ProWL and ProWL-IF replenish the functions of proteins under the assumption that proteins are partially annotated. In addition, ProWL-IF takes advantage of the knowledge that a protein cannot have certain functions (called *irrelevant functions*), which can further boost the performance of protein function prediction. Our experimental results on protein-protein interaction and gene microarray expression benchmarks validate the effectiveness of ProWL and ProWL-IF.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology- Classifier Design and Evaluation; J.3 [**Life and Medical Sciences**]: Biology and Genetics

General Terms

Algorithms, Performance, Experimentation

Keywords

Protein Function Prediction, Weak-label Learning, Multi-label Learning

ACM-BCB'12, October 7-10, 2012, Orlando, FL, USA

1. INTRODUCTION

High-throughput biological techniques provide information about the interaction of several thousands of proteins simultaneously. However, most of these proteins are not functionally annotated. As such, protein function annotation is one of the fundamental issues in the post-genomic era [17]. It is time-consuming and expensive to manually annotate proteins in biological experiments. For these reasons, it is necessary and promising to develop computational methods to automatically annotate proteins.

Various computational models (including classification and clustering methods) have been proposed for annotating proteins. Some approaches annotate proteins using the amino acid sequences associated with these proteins [7, 9]. Some methods take advantage of the protein protein interactions (PPI) in the cell to predict the functions of proteins [3, 17]. Some approaches annotate proteins by integrating various data sources (including amino acid sequences and PPI)[10, 20].

Proteins have multiple roles and functions; each function can be viewed as a label. Thus multi-label learning techniques are widely studied in protein function prediction [7, 19]. Some approaches, first train a classifier for each function and then combine these classifiers' predictions to annotate a protein [9]. In particular, some techniques organize the classifiers trained for each function according to the function catalogue hierarchical structure [1, 15] and then annotate proteins [2, 5, 21]. Another class of protein function prediction methods incorporate the correlations between the functions (labels) to improve the multi-label prediction accuracy [8, 23].

All these methods assume that the functions associated with proteins are complete and fixed, which is often not true for real-world PPI data. Often we know a subset of the functions of a protein, and whether an annotated protein has additional functions is unknown. This type of multi-label prediction problem is referred to as the 'weak label' or 'incomplete class assignment' problem [4, 18]. In this paper, unlike traditional multi-label learning methods [2, 8, 23], we develop a method, called *Pro*tein Function Prediction with *W*eak-label *L*earning (ProWL), which can annotate proteins with incomplete function assignment in the training set. The approach proposed in [4, 18] for multi-label learning with weak labels considers the specified labels of an instance as relevant labels, and all the unspecified labels of the instance as candidates for relevant labels. In practice, we may also know that a protein cannot have certain functions (hereinafter, we call these functions *irrelevant functions*). Both previous approaches [4, 18] ignore this prior knowledge, which can further boost the performance of protein function prediction. Here, we make use of these irrelevant functions and propose a variation of ProWL, called *Pro*tein Function Prediction with *W*eak-label *L*earning and Knowledge of *I*rrelevant *F*unction (ProWL-IF). ProWL-IF can not only leverage the functions associated with a protein, but also the irrelevant ones. We summarize our key contributions as follows:

1. We consider the incomplete annotation problem for protein function prediction.

2. We design the ProWL alogrithm to annotate proteins with incomplete annotations, and propose the ProWL-IF algorithm, which takes advantage of both relevant and irrelevant functions to replenish missing functions of proteins.

3. We compare the proposed methods against other related techniques using various metrics on public available protein datasets, and show their effectiveness.

2. RELATED WORK

Traditional multi-label learning approaches focus on predicting the multiple labels for each test instance simultaneously [19]. These methods utilize the label correlations among the different multi-labeled instances [23], and often assume that the given labels for the training instances are complete and accurate. However, in several real world applications, complete or full set of labels may be missing, noisy and not provided. A few weak label (or missing label) learning algorithms have been proposed in the literature within the single label or multiple label learning settings [4, 18, 22]. Prediction of the complete set of labels (i.e., predicting the missing labels), given partial or incomplete labels is defined as the "weak label learning problem".

Sun et. al. [18] developed a method called WEak Label Learning (WELL) for predicting missing labels for multi-labeled instances. WELL formulates a convex optimization, that first approximates similarities between labels by assuming a group of low-rank base similarities. WELL was validated on a set of text, image and bioinformatic applications. Buncak et. al. [4] studied the incomplete class assignments problem for annotating images, and developed an approach called MLR-GR. This method optimizes ranking errors and the group lasso loss. Qi et. al. [14] uses the Hierarchical Dirichlet Process to append missing labels for a set of images. In addition, Wang et. al. [22] developed an approach for annotating weakly labeled facial images. However, this approach is a single-label (or multi-class) method and focuses on refining the noisy labeled images.

Several computational approaches have been developed for protein function prediction, that differ in terms of methodology, input data, and even problem definition. We refer the reader to a comprehensive survey paper on this topic [12]. Relevant to our work, Chi et. al. [6] proposed an iterative protein function prediction method using partial annotations. At each iteration, using the most confident predicted functions, pairwise similarities between training proteins and testing proteins are updated. This updated similarity is used for predicting functions for test proteins at the next iteration.

In our paper, we develop a new weak labeled learning algorithm for predicting multiple functions (or labels) of proteins. We refer to our approach as ProWL, *Pro*tein function prediction with *W*eak-label *L*earning. We extend ProWL to incorporate irrelevant function (or labels) information of proteins and call this approach as ProWL-IF.

3. PROBLEM FORMULATION

In this paper, we study the weak-label problem in protein function prediction for two tasks as illustrated in Figure 1. In the first task, we have partially labeled proteins: Given a protein, some of its functions are specified, and some may be missing. The task we address is: How to use incomplete annotations to replenish the missing functions (cf. Figure 1(a) and 1(b))? We develop algorithms for two scenarios. In the first case, we develop ProWL to replenish missing functions by assuming that we have prior knowledge of *only the relevant* functions. As shown in Figure 1(a), we have known relevant functions denoted by 1 and missing functions denoted by "?", which are set to 0, and become candidates for being predicted as relevant i.e., ProWL may modify a 0 to 1.

We also consider the case (Fig 1(b)), where we have prior knowledge of both the *relevant* and *irrelevant* functions. We develop ProWL-IF to handle this case. In this case, the relevant functions are denoted by 1, irrelevant functions are denoted by -1 and the missing functions denoted by "?" are set to 0. ProWL-IF's objective is to identify if the missing functions (0s) are relevant or not i.e., 1 or -1, respectively.

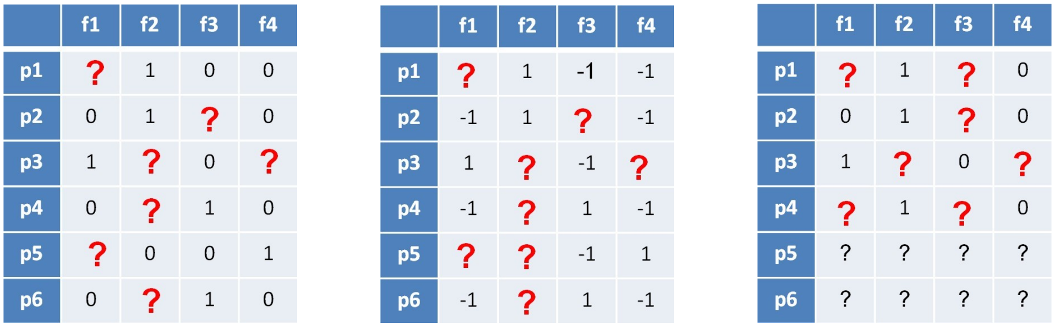
In the second task, we address the following issue: How to utilize the incomplete annotated proteins to annotate proteins which are completely unlabeled (cf. Figure 1(c))?. In this case the goal is to predict functions for completely unannotated proteins, e.g., proteins **p5** and **p6** in Figure 1(c). For this task, we assume that we have knowledge of only relevant functions for the set of incomplete annotated proteins (or training proteins).

3.1 Protein Function Prediction with Weak-label Learning

Given n proteins, let the number of distinct functions across all proteins be K. Let $Y = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n]$ be the original label set with $y_{ik} = 1$ if protein i has the k-th function, and $y_{ik} = 0$ otherwise. It is important to incorporate function correlation in protein function prediction [8, 13, 23]. Various methods are proposed to measure the function correlation. We define a function correlation matrix $C^{'} \in R^{K \times K}$ based on cosine similarity (also used in [8]) as follows:

$$C^{'}_{st} = \frac{\mathbf{Y}^T_{.s}\mathbf{Y}_{.t}}{\|\mathbf{Y}_{.s}\|\|\mathbf{Y}_{.t}\|} \quad (1)$$

where $C^{'}_{st}$ is the function correlation between functions s and t, and $\mathbf{Y}_{.s}$ represents the s-th column of Y. From Eq. (1), we can observe that, given functions s, t, and u, if functions s and t often co-exist in the same proteins, while functions s and u seldom co-exist, then $C^{'}_{st}$ will be larger than $C^{'}_{su}$.



	f1	f2	f3	f4
p1	?	1	0	0
p2	0	1	?	0
p3	1	?	0	?
p4	0	?	1	0
p5	?	0	0	1
p6	0	?	1	0

(a) Task1(ProWL)

	f1	f2	f3	f4
p1	?	1	-1	-1
p2	-1	1	?	-1
p3	1	?	-1	?
p4	-1	?	1	-1
p5	?	?	-1	1
p6	-1	?	1	-1

(b) Task1(ProWL-IF)

	f1	f2	f3	f4
p1	?	1	?	0
p2	0	1	?	0
p3	1	?	0	?
p4	?	1	?	0
p5	?	?	?	?
p6	?	?	?	?

(c) Task2

Figure 1: Task Summaries

We normalize $C^{'}$ as follows:

$$C_{st} = \frac{C^{'}_{st}}{\sum_{k=1}^{K} C^{'}_{sk}} \tag{2}$$

Thus, C_{st} can be viewed as the likelihood that a protein has function t given that it is annotated with function s.

We now consider the case with incomplete annotation, and define the weighted loss function as the first part of our objective function as follows:

$$\begin{aligned} \Phi_1(\mathbf{f}) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} M_{ik}(f_{ik} - \tilde{y}_{ik})^2 \\ &= \frac{1}{2}\|M \circ (F - \tilde{Y})^T (F - \tilde{Y})\|_2^2 \end{aligned} \tag{3}$$

where $\circ$ means element-wise multiplication (also called Hadamard product), $\tilde{Y} = [\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \ldots, \tilde{\mathbf{y}}_n]$ is the extended function set of n proteins, with $\tilde{Y} = YC$. f_{ik} is the predicted likelihood of protein i with respect to the k-th function. If functions s and t often co-exist in the same proteins, then, if a protein is annotated with function s, it is likely that it will also have function t. M_{ik} is the weight of protein i with respect to function k:

$$M_{ik} = \left\{ \begin{array}{l} 1, \quad y_{ik} = 1 \\ \mathbf{y}_i^T \mathbf{c}_{.k}, y_{ik} = 0 \end{array} \right. \tag{4}$$

where $\mathbf{c}_{.k}$ is the k-th column of C. As defined in Eq. (4), if the annotated functions of protein i have large correlation score with function k, the weight M_{ik} will be large. If the k-th function of protein i is missing, the minimization of Eq. (3) can help us to replenish this function.

Proteins with similar acid amino sequences tend to have similar functions, and the 'guilt by association' rule [16] assumes that interacting proteins are more likely to share similar functions. To make use of this kind of knowledge, as in semi-supervised learning [24], we incorporate a smoothness term within our objective function:

$$\begin{aligned} \Phi_2(\mathbf{f}) &= \frac{1}{2}\sum_{i,j=1}^{n} \|\frac{\mathbf{f}_i}{\sqrt{D_{ii}}} - \frac{\mathbf{f}_j}{\sqrt{D_{jj}}}\|_2^2 W_{ij} \\ &= tr(F^T(I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}})F) \\ &= tr(F^T L F) \end{aligned} \tag{5}$$

where $F = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n]$, D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} W_{ij}$. W_{ij} captures the similarity between proteins i and j. The matrix W can be set using the pairwise sequence similarities, or using the frequency of interactions found in multiple PPI studies, or as a kernel matrix derived from PPI studies. I is an $n \times n$ identity matrix, $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, and $tr(\cdot)$ is the matrix trace operation. The motiviation to minimize Eq. (5) is that if the similarity (or interaction weight) W_{ij} between protein i and j is high, these two proteins are more likely to share similar functions (i.e., $\|\mathbf{f}_i - \mathbf{f}_j\|_2^2$ should be small). Thus, by minimizing Eq. (5), we can ensure similar proteins to have similar functions, which is in accordance with the 'guilt by association' rule.

Our objective function to be minimized is:

$$\begin{aligned} \Phi(F) = \frac{1}{2}\|M \circ (F - \tilde{Y})^T (F - \tilde{Y})\|_2^2 \\ + \alpha tr(F^T L F) + \beta \|F^T F\|_2^2 \end{aligned} \tag{6}$$

The third term is added to control the sparsity of F, since each function is associated with a small number of proteins. α and β are parameters to balance the importance of the second and third terms, respectively.

Optimization:. Taking the derivative of Eq. (6) with respect to F, we have:

$$\frac{\partial \Phi(F)}{\partial F} = M \circ (F - \tilde{Y}) + \alpha L F + \beta I F \tag{7}$$

Eq. (7) can be divided into K problems and for the k-th problem it can be solved as:

$$(\tilde{M}_{.k} + \alpha L + \beta I)\mathbf{f}_{.k} = \mathbf{p}_k \tag{8}$$

where

$$\tilde{M}_{.k} = diag(M_{.k}), \ \mathbf{p}_k = \mathbf{M}_{.k} \circ \tilde{\mathbf{Y}}_{.k} \tag{9}$$

$diag(\cdot)$ is the vector diagonalization operation. Instead of computing the inverse of $(\tilde{M}_{.k} + \alpha L + \beta I)$, Eq. (8) can be solved with various existing fast iterative solvers [11]. We use the Conjugate Gradient (CG) solver, which is guaranteed to terminate in n steps. The most time-consuming step at each iteration of CG is a matrix vector product, whose time complexity is proportional to the number of non-zero elements in $\tilde{M}_{.k} + \alpha L + \beta I$. Since $\tilde{M}_{.k}$, L and I are sparse, positive definite, and with $O(n)$ non-zero elements, Eq. (8) can be efficiently solved. In our experiments, we find CG terminates in fewer than 30 iterations. The ProWL is described in **Algorithm 1**.

Algorithm 1 ProWL: *Pro*tein Function Prediction with *W*eak-label *L*earning

```
Input:
    Weight matrix $W$, incomplete annotations $Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$, $\alpha$, $\beta$
Output:
    Predicted likelihood score vectors $\{\mathbf{f}_i\}_{i=1}^n$
Compute $C$ using Eq. (3) and $L = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$.
Set $\tilde{Y} = YC$ and initialize $M$ using Eq. (4).
for $k = 1$ to $K$ do
    Set $\tilde{M}_{.k}$ and $\mathbf{p}_k$ using Eq. (9).
    Solve $\mathbf{f}_{.k}$ using Eq. (8)
end for
return $F = [\mathbf{f}_{.1}, \mathbf{f}_{.2}, \ldots, \mathbf{f}_{.K}]^T$.
```

3.2 Protein Function Prediction with Weak-label Learning and Knowledge of Irrelevant Functions

In practice, we may know that some functions are *not* associated with specific proteins. However, all the aforementioned multi-label learning methods with weak labels [4, 14, 18] consider the irrelevant functions as candidates for missing functions, thus ignoring this knowledge. We introduce ProWL-IF, a variation of ProWL, which takes advantage of both the annotated relevant and irrelevant functions, in addition to missing functions.

In this setting, we have a partially annotated function set $Z = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]$, with $z_{ik} = 1$ if protein i has the k-th function, $z_{ik} = -1$ if protein i does not have this function, and $z_{ik} = 0$ if it's unknown whether the protein has the function i.e., it is missing. At first, we transform Z into $\bar{Z} = [\bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2, \ldots, \bar{\mathbf{z}}_n]$ where $\bar{\mathbf{z}}_i = \frac{\mathbf{z}_i + |\mathbf{z}_i|}{2}$, and $|\mathbf{z}_i|$ is the absolute value of $\mathbf{z}_i$. Next, we define the correlation between functions s and t based on $\bar{Z}$ as follows:

$$\tilde{C}_{st} = \frac{\bar{\mathbf{Z}}_{.s}^T \bar{\mathbf{Z}}_{.t}}{\|\bar{\mathbf{Z}}_{.s}\| \|\bar{\mathbf{Z}}_{.t}\|} \tag{10}$$

where $\bar{\mathbf{Z}}_{.s}$ is the s-th column of $\bar{Z}$. We normalize $\tilde{C}$ as in Eq. (2).

Similarly to Eq. (3), the weighted loss function of ProWL-IF is defined as:

$$\Psi_1(\mathbf{f}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} M_{ik}^{'}(f_{ik} - \tilde{z}_{ik})^2 \tag{11}$$

where $\tilde{Z} = [\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \ldots, \tilde{\mathbf{z}}_n]$ is the extended label set of proteins. For the i-th protein with respect to the k-th function, $\tilde{z}_{ik}$ is specified as:

$$\tilde{z}_{ik} = \begin{cases} z_{ik}, & z_{ik} = 1 \ or \ z_{ik} = -1 \\ \bar{\mathbf{z}}_i^T \mathbf{c}_{.k}, & z_{ik} = 0 \end{cases} \tag{12}$$

where $\mathbf{c}_{.k}$ is the k-th column of the correlation matrix C, and $M_{ik}^{'}$ is the weight of protein i with respect to the k-th function:

$$M_{ik}^{'} = \begin{cases} 1, & z_{ik} = 1 \ or \ z_{ik} = -1 \\ \bar{\mathbf{z}}_i^T \mathbf{c}_{.k}, & z_{ik} = 0 \end{cases} \tag{13}$$

Eq. (11) looks similar to Eq. (3), but in Eq. (11) $\tilde{\mathbf{z}}_i \in [-1, 1]$ and in Eq. (3) $\tilde{\mathbf{y}}_i \in [0, 1]$. In addition, Eq. (11) does not consider the irrelevant functions as candidate missing functions, whereas Eq. (3) does. Therefore, ProWL-IF has the advantage of properly capturing the prior irrelevant and relevant function information.

Putting together Eq. (11) and Eq. (5), the objective of ProWL-IF is to minimize the following function:

$$\begin{aligned} \Psi(F) = &\frac{1}{2}\|M^{'} \circ (F - \tilde{Z})^T(F - \tilde{Z})\|_2^2 \\ &+ \alpha tr(F^T LF) + \beta\|(F + 1_{n\times K})^T(F + 1_{n\times K})\|_2^2 \end{aligned} \tag{14}$$

$\mathbf{1}_{n\times K}$ is an $n \times K$ matrix with all entries equal to 1. The third term controls the complexity and sparsity of F, since each protein has a large proportion of irrelevant functions (denoted by -1) and a small proportion of relevant functions (denoted by 1). α and β are scalar parameters to balance the importance of the smoothness and sparsity terms, respectively.

Taking the derivation of $\Psi(F)$ with respect to F, we have:

$$\frac{\partial\Psi(F)}{\partial F} = M^{'} \circ (F - \tilde{Z}) + \alpha LF + \beta I_{n\times n}(F + 1_{n\times K}) \tag{15}$$

where $I_{n\times n}$ is an $n \times n$ identity matrix. Similar to Eq. (7), Eq. (15) can be divided into K problems and solved as:

$$(\tilde{M}_{.k}^{'} + \alpha L + \beta I_{n\times n})\mathbf{f}_{.k} = \mathbf{q}_k \tag{16}$$

where

$$\tilde{M}_{.k}^{'} = diag(\mathbf{M}_{.k}^{'}), \ \mathbf{q}_k = \mathbf{M}_{.k}^{'} \circ \tilde{\mathbf{Z}}_{.k} - \beta I_{n\times n}\mathbf{1}_{n\times 1} \tag{17}$$

Eq. (16) can be efficiently solved in the same way as Eq. (8), and the learning procedure for ProWL-IF is similar to that of ProWL (**Algorithm 1**).

4. EXPERIMENTAL SETUP

4.1 Datasets

We evaluate the performance of the proposed methods on public available protein function prediction benchmarks, among which three are PPIs and one is a micro-array gene expression data. The first dataset (**DS1**) was extracted from BioGrid [1] with PubMed ID 17200106, and its largest connected component contains 1002 proteins annotated according to FunCat [15][2], across 33 functions. The functions in FunCat are organized in a tree structure. We use the most informative functions as defined in [8] and [23]. Informative functions are the ones that have at least 30 proteins as members, and within the tree structure these functions do no not have a particular descendant node with more than 30 proteins. The second dataset (**DS2**) was downloaded from BioGrid (2011-12-25). After the preprocessing and filtering, it contains 3041 proteins annotated with 86 informative functions. The weight matrix W of the second dataset are specified by the number of PubMed IDs, where 0 means no interaction between two proteins and $p > 0$ means this interaction is supported by p distinct publications. The third dataset (**DS3**) was extracted from heterogeneous data sources of humans [10][3]. We use its largest connected component, which includes 2950 proteins annotated according to the Gene Ontology [1]. Similar to [10], we use the functions that have at least 30 proteins annotated with them. The fourth dataset (**DS4**) was used in WELL [18][4] and includes

[1] http://thebiogrid.org/

[2] http://mips.helmholtz-muenchen.de/proj/funcatDB/

[3] http://morrislab.med.utoronto.ca/?sara/SW

[4] http://lamda.nju.edu.cn/files/WELL.rar

1500 proteins annotated with 14 functions. We specify the weight matrix W for proteins in the same way as it was done for WELL. The weight matrices W of DS1 and DS3 were specified by the providers, and the same matrices were used for ourselves. The statistics of the processed datasets are listed in Table 1.

Table 1: Statistics of datasets (Avg±Std means average number of functions for each protein and its standard deviation)

Dataset	#Proteins	#Functions	Avg±Std
DS1	1002	33	2.00 ± 1.37
DS2	3041	86	1.94 ± 1.60
DS3	2950	200	6.86 ± 3.77
DS4	1500	14	4.23 ± 1.58

We assume the datasets used in the experiments have complete functions annotations, and simulate the weak-label settings of ProWL and ProWL-IF on these functions. When evaluating ProWL, we simulate the incomplete annotation problem by masking the ground truth (or relevant) functions (1) to missing functions (?) based on a threshold called the *Incomplete Function* (IF) ratio. The IF ratio denotes the percentage of relevant functions (denoted by labels 1s), for a protein that are masked or set to missing or "?", see Figure 1(a) for more detail. When evaluating ProWL-IF, the IF ratio sets both the relevant (1s) and irrelevant (-1s) functions to missing i.e., "?". For consistency, the IF ratio is defined for the relevant functions and the same number of irrelevant functions are masked in this case. see Figure 1(b) for more detail.

There is no weak-label learning method proposed for protein function prediction domain. We compare our methods with WELL [18] and MLR-GL [4]. WELL and MLR-GL need an input kernel matrix, and we substitute the kernel with the PPI matrices, or specify it as in WELL[18]. The parameters of WELL are specified as the authors reported. For MLR-GL we use the default parameters in the package provided by the authors [5]. For ProWL and ProWL-IF, we set α and β to 0.01 and 0.001, respectively. We observe the performance with respect to various metrics does not change as we vary α and β around the fixed values. This setting is not optimal, and we will investigate how to adapt the parameter values in the future.

4.2 Evaluation Metrics

Various performance metrics have been developed for evaluating multi-label learning methods [19]. Here we introduce the metrics will be used in this paper (previously used in WELL and MLR-GL).

MacroF1 is the average $F1$ scores of different functions:

$$MacroF1 = \frac{1}{K}\sum_{k=1}^{K}\frac{2p_k r_k}{p_k + r_k}$$

where p_k and r_k are the precision and recall of the k-th function.

MicroF1 calculates the $F1$ measure on the predictions of different functions as a whole:

$$MicroF1 = \frac{1}{K}\frac{\sum_{k=1}^{K} 2p_k r_k}{\sum_{k=1}^{K} p_k + r_k}$$

Ranking loss evaluates the average fraction of function label pairs that are not correctly ordered.

$$\begin{aligned} RankingLoss = \frac{1}{n}\sum_{i=1}^{N}\frac{1}{|\mathbf{y}_i||\bar{\mathbf{y}}_i|}|\{(y_1, y_2) \in \\ \mathbf{y}_i \times \bar{\mathbf{y}}_i | F(i, y_1) \leq F(i, y_2)\}| \end{aligned}$$

where $\bar{\mathbf{y}}_i$ contains the labels that are not in $\mathbf{y}_i$ with $\bar{y}_{ic} = 1$ iff $y_{ic} = 0$, and $\bar{y}_{ic} = 0$ iff $y_{ic} = 1$. The performance is perfect when $RankingLoss = 0$.

The adapted *Area Under the Curve* (AUC) for multi-label learning was introduced in [4]. AUC first ranks all the functions for each test protein in descending order of their scores; it then varies the number of predicted functions from 1 to the total number of functions, and computes the receiver operator curve by calculating true positive rate and false positive rate for each number of predicted functions. It finally computes the area under the curve of all functions to evaluate the multi-label learning methods.

To maintain consistency with other evaluation metrics, we report *1-RankingLoss*. Thus, the higher the value of *1-RankingLoss*, the better the performance.

5. EXPERIMENTAL ANALYSIS

5.1 Performance on Replenishing Missing Functions

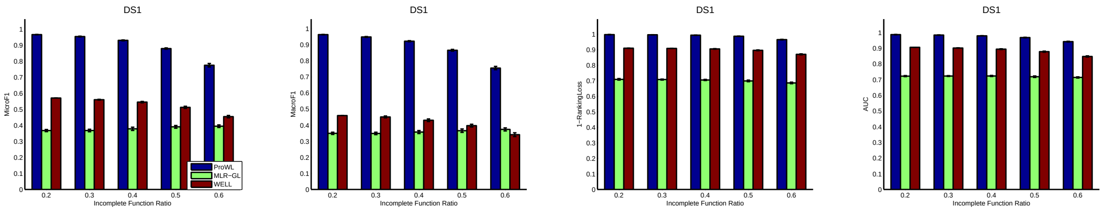
We performed experiments to investigate the performance of the proposed methods in replenishing the missing functions. In these experiments, we use all the proteins within the datasets and vary the IF ratio of each protein from 20% to 60%, with an interval of 10%, to study the performance of different methods. Some proteins in the PPI network do not have any true functions. To make use of the PPI network structure, we do not remove them, but we evaluate the performance of replenishing missing functions on only the annotated proteins. The experimental results (average of 20 independent runs and standard deviations) are shown in Figures 2 - 5. We were not able to run WELL to completion on DS3 (using 4GB RAM). *MicroF1* and *MacroF1* depend on a hard partitioning of $\mathbf{f}_i$ into relevant and irrelevant functions. Here we consider the functions corresponding to the largest s values of $\mathbf{f}_i$ as the relevant ones, and the remaining as irrelevant functions of protein i. s is determined by the number of ground-truth functions of the i-th protein.

From the figures, we can observe that ProWL outperforms WELL and MLR-GL in replenishing missing functions of proteins in almost all the metrics across the four datasets. For example, ProWL on average is 6.96% better than WELL, and 58.70% better than MLR-GL, when compared using *MicroF1* on DS4. These results confirm the effectiveness of ProWL in Task 1.

Another interesting observation from Figures 2 - 5 is that the multi-label *MicroF1* and *MacroF1* scores decrease as the IF ratio increases. However, for DS3 and DS4 the decrease in the F1 scores is not as evident as for DS1 and DS2. This can be explained by the fact that DS3 and DS4 have a larger number of functions per protein, and a higher IF ratio still allows proteins within the set to have a few relevant functions. ProWL uses these functions and replenishes the missing ones.

Since the setting of ProWL-IF is different from ProWL, WELL and MLR-GL, we conducted additional experiments

[5] http://www.cse.msu.edu/∼bucakser/



(a) MicroF1 (b) MacroF1 (c) 1-RankLoss (d) AUC

Figure 2: Replenishing missing functions of DS1.

on the four datasets to investigate the performance of ProWL-IF. We simulated the setting of ProWL-IF by masking some relevant functions (+1) and some irrelevant functions (-1) to missing functions (0). We set the incomplete function ratio with respect to the relevant functions (+1) similarly to ProWL and the number of masked irrelevant functions (-1) in ProWL-IF was the same as relevant functions. All the unmaksed functions were served as relevant functions or irrelevant functions. We just report the results with respect to MacroF1 and MicroF1 in Table 2 (other results will be provided on a supplementary webpage). The better performance in Table 2 are shown in **boldface** (statistical significance is examined via pairwise t-test at 95% significant level). We can observe that ProWL-IF generally outperforms ProWL. This observation shows the benefit in making use of irrelevant functions as prior knowledge.

5.2 Performance for Task 2 (Completely Unlabeled Test Proteins)

We wanted to assess the strengths of ProWL in leveraging the partially annotated proteins and making predictions for proteins that were completely unannotated. We performed another set of experiments to investigate the performance of ProWL in this scenario. We first partitioned our dataset into two parts: (i) training set with missing annotations and (ii) test set with no annotations (i.e., completely unannotated). For the training set we varied the IF ratio from 20% to 80% in increments of 20%, and used ProWL to report the prediction performance on the test set only.

To assess the advantage of the missing function assumption, we also include the results for another variation of ProWL called ProWL-Part. For ProWL-Part we assume that a missing label for a protein in the training set will be set as an irrelevant function for the protein (i.e., set M_{ik} to 0, if the k-th function is missing for protein i). We again report the performance of ProWL-Part for the test set i.e., proteins with no annotations. The experimental results (average of 20 independent runs) are reported in Tables 3- 4. The setting of missing functions for each protein is determined as in the first set of experiments, but s is specified as the average number of functions of all proteins. Due to space limit, we report only the results for *MicroF1*, *1-RankingLoss*, and AUC, and fix the training set to 80% of the dataset and the test set to 20% of DS3 and DS4 (other results will be provided on a supplementary webpage). The best performance and its comparable performance are shown in **boldface** (statistical significance is examined via pairwise t-test at 95% significant level).

From these tables, we can see that ProWL predicts the functions of proteins with higher accuracy than the other methods in most metrics. Considering MicroF1 on DS4, for example, ProWL on average is 9.93% better than MLR-GL and 2.72% better than WELL. With the same IF ratio in all the three metrics, ProWL outperforms ProWL-Part 18 times, ties with ProWL-Part 5 times, and loses to ProWL-Part only one time. This statistic corroborates the benefit in introducing the missing function assumption. The difference between ProWL and ProWL-Part diminishes as the IF ratio increases. This is because the estimated function correlation become inaccurate as the IF ratio increases.

6. CONCLUSION

In this paper, we studied the incomplete annotation problem for protein function prediction and propose ProWL to annotate proteins with incomplete annotation. To make use of irrelevant functions of proteins, we introduce a variant of ProWL, called ProWL-IF. Unlike traditional multi-label learning methods, which consider all the missing functions as candidates of relevant functions, ProWL-IF takes into account both relevant and irrelevant functions for prediction. Our experimental results demonstrate that the proposed methods have higher performance than other related methods.

We will investigate a function correlation scheme that can capture the correlation with a large ratio of missing functions.

7. ACKNOWLEDGEMENTS

This work is partially supported by grants from NSF IIS 0905117, Natural Science Foundation of China (Project Nos. 60973083, 61070090, 61003174 and 61170080), grants from the NSFC-Guangdong Joint Fund (Project No. U1035004 and U1135004) and China Scholarship Council (CSC).

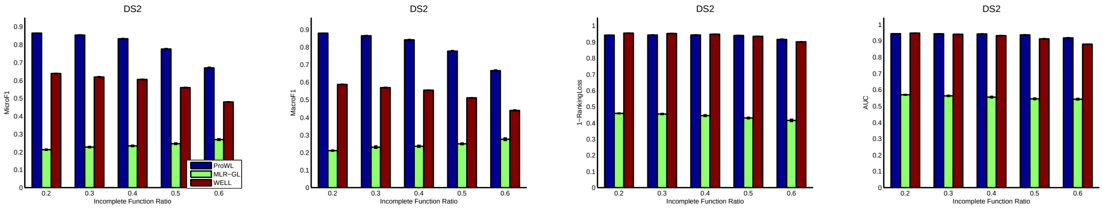
8. REFERENCES

[1] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25, 2000.

[2] Z. Barutcuoglu, R. Schapire, and O. Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

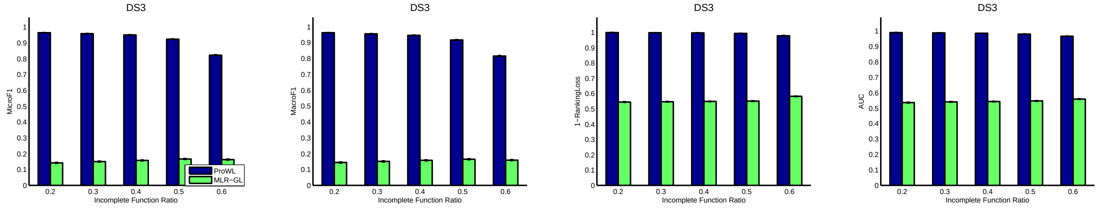
[3] E. Becker, B. Robisson, C. Chapple, A. Guénoche, and C. Brun. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, 28(1):84–90, 2012.

[4] S. Bucak, R. Jin, and A. Jain. Multi-label learning with incomplete class assignments. In *IEEE*



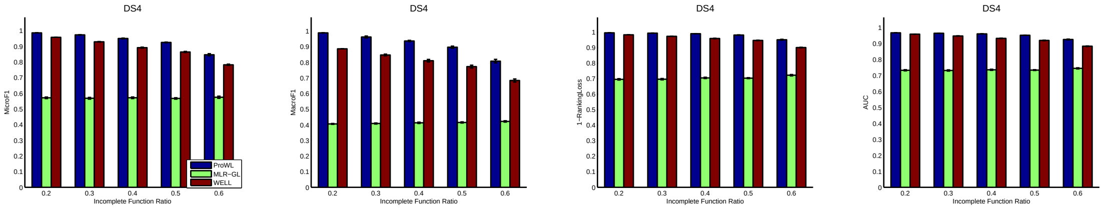
(a) MicroF1 (b) MacroF1 (c) 1-RankingLoss (d) AUC

Figure 3: Replenishing missing functions of DS2.



(a) MicroF1 (b) MacroF1 (c) 1-RankingLoss (d) AUC

Figure 4: Replenishing missing functions of DS3.



(a) MicroF1 (b) MacroF1 (c) 1-RankingLoss (d) AUC

Figure 5: Replenishing missing functions of DS4.

Table 2: Experimental results of ProWL-IF on four datasets.

Dataset	Method	MicroF1				MacroF1			
		20%	40%	60%	80%	20%	40%	60%	80%
DS1	ProWL	95.26	87.79	76.54	73.49	46.17	45.88	43.62	41.65
	ProWL-IF	**98.46**	**90.97**	**82.54**	**76.40**	**98.14**	**89.93**	**80.64**	**72.70**
DS2	ProWL	85.35	77.56	65.08	55.24	35.96	34.06	31.75	29.03
	ProWL-IF	**91.08**	**83.25**	**71.83**	**60.97**	**90.78**	**82.38**	**70.29**	**59.78**
DS3	ProWL	95.78	92.31	78.28	50.73	95.48	91.63	77.75	51.98
	ProWL-IF	**97.82**	**93.61**	**83.88**	**60.29**	**97.57**	**92.97**	**82.87**	**59.43**
DS4	ProWL	97.28	92.47	83.99	50.73	33.74	34.68	35.63	34.12
	ProWL-IF	**98.10**	92.69	83.86	**60.29**	**96.95**	**89.74**	**78.73**	**64.36**

Table 3: Experimental results (avg±std) on DS3.

Metric	IF Ratio	ProWL	ProWL-Part	MLR-GL
MicroF1	20%	**24.68±1.20**	23.70±1.07	14.86±0.94
	40%	**24.62±1.31**	23.08±1.29	14.77±0.92
	60%	**22.90±1.03**	**22.21±1.48**	13.35±1.03
	80%	**19.88±0.89**	**19.92±1.11**	10.37±1.06
1-RankingLoss	20%	**76.52±0.85**	**76.11±0.86**	70.72±0.96
	40%	**77.74±0.86**	**77.86±1.10**	69.27±1.16
	60%	77.31±1.15	**78.21±1.08**	67.00±1.16
	80%	**77.73±1.31**	**77.14±0.93**	64.58±1.05
AUC	20%	**78.70±0.78**	77.62±0.67	70.75±1.10
	40%	**78.78±0.71**	76.36±1.07	68.51±1.01
	60%	**77.07±1.10**	74.34±1.37	65.06±1.04
	80%	**73.70±1.25**	70.35±1.45	61.45±0.98

Table 4: Experimental results (avg±std) on DS4.

Metric	IF Ratio	ProWL	ProWL-Part	MLR-GL	WELL
MicroF1	20%	**63.04±1.39**	61.24±1.48	60.09±1.51	61.75±1.27
	40%	**63.41±1.72**	62.24±1.03	58.78±0.87	61.96±1.03
	60%	**63.88±1.52**	62.13±1.10	58.31±1.28	61.63±0.98
	80%	**62.16±1.17**	60.78±1.41	53.25±1.14	60.29±1.27
1-RankingLoss	20%	**81.15±1.09**	80.14±1.08	78.31±1.38	80.19±0.94
	40%	**81.18±1.23**	80.35±0.73	77.42±1.09	80.11±0.92
	60%	**81.64±1.21**	80.36±0.86	76.85±1.05	79.91±0.87
	80%	**80.67±1.00**	79.81±1.26	72.97±0.95	79.40±0.95
AUC	20%	**82.22±1.04**	81.41±0.88	80.33±1.06	81.09±0.84
	40%	**82.21±1.14**	81.51±0.60	79.47±0.79	81.20±0.87
	60%	**82.51±0.99**	81.55±0.84	78.88±0.98	80.82±0.81
	80%	**81.64±0.92**	80.77±0.96	75.55±0.77	80.48±0.69

Conference on Computer Vision and Pattern Recognition, pages 2801–2808. IEEE, 2011.

[5] N. Cesa-Bianchi, M. Re, and G. Valentini. Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference. *Machine Learning*, 88(2):209–241, 2012.

[6] X. Chi and J. Hou. An iterative approach of protein function prediction. *BMC Bioinformatics*, 12(1):437, 2011.

[7] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Proceedings of Advances in Neural Information Processing Systems*, pages 681–687. MIT Press, 2001.

[8] J. Jiang and L. McQuay. Predicting protein function by multi-label correlated semi-supervised learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1059–1069, 2012.

[9] C. Leslie, E. Eskin, A. Cohen, J. Weston, and W. Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, 2004.

[10] S. Mostafavi and Q. Morris. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics*, 26(14):1759–1765, 2010.

[11] J. Nocedal and S. Wright. *Numerical Optimization.* Springer Verlag, 1999.

[12] G. Pandey, V. Kumar, and M. Steinbach. Computational approaches for protein function prediction. Technical Report TR 06-028, Department of Computer Science and Engineering, University of Minnesota, Twin Cities, 2006.

[13] G. Pandey, C. Myers, and V. Kumar. Incorporating functional inter-relationships into protein function prediction algorithms. *BMC Bioinformatics*, 10(1):142, 2009.

[14] Z. Qi, M. Yang, Z. Zhang, and Z. Zhang. Mining partially annotated images. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1199–1207. ACM, 2011.

[15] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Güldener, G. Mannhaupt, M. Münsterkötter, et al. The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*, 32(18):5539–5545, 2004.

[16] B. Schwikowski, P. Uetz, S. Fields, et al. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18(12):1257–1261, 2000.

[17] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3(1), 2007.

[18] Y. Sun, Y. Zhang, and Z. Zhou. Multi-label learning with weak label. In *Proceedings of 24th AAAI Conference on Artificial Intelligence*, 2010.

[19] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. *Data Mining and Knowledge Discovery Handbook*, pages 667–685, 2010.

[20] K. Tsuda, H. Shin, and B. Schölkopf. Fast protein classification with multiple networks. *Bioinformatics*, 21(suppl 2):ii59, 2005.

[21] G. Valentini. True path rule hierarchical ensembles for genome-wide gene function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):832–847, 2011.

[22] D. Wang, S. Hoi, and Y. He. Mining weakly labeled web facial images for search-based face annotation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information*, pages 535–544. ACM, 2011.

[23] X. Zhang and D. Dai. A framework for incorporating functional inter-relationships into protein function prediction algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(3):740–753, 2012.

[24] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Proceedings of Advances in Neural Information Processing Systems*, pages 321–328. Vancouver, British Columbia, Canada, 2003.