BIOINFORMATICS

Vol. 26 ECCB 2010, pages i645–i652
doi:10.1093/bioinformatics/btq394

Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins

Yanjun Qi[1,*], Oznur Tastan[2], Jaime G. Carbonell[2], Judith Klein-Seetharaman[2] and Jason Weston[3]

[1]NEC Labs America, 4 Independence Way, Princeton, NJ 08540, [2]School of Computer Science, Carnegie Mellon University, PA 15213 and [3]Google Research NY, 75 Ninth Avenue, New York, NY 10011, USA

ABSTRACT

Motivation: Protein–protein interactions (PPIs) are critical for virtually every biological function. Recently, researchers suggested to use supervised learning for the task of classifying pairs of proteins as interacting or not. However, its performance is largely restricted by the availability of truly interacting proteins (*labeled*). Meanwhile, there exists a considerable amount of protein pairs where an association appears between two partners, but not enough experimental evidence to support it as a direct interaction (*partially labeled*).

Results: We propose a semi-supervised multi-task framework for predicting PPIs from not only *labeled*, but also *partially labeled* reference sets. The basic idea is to perform multi-task learning on a supervised classification task and a semi-supervised auxiliary task. The supervised classifier trains a multi-layer perceptron network for PPI predictions from *labeled* examples. The semi-supervised auxiliary task shares network layers of the supervised classifier and trains with *partially labeled* examples. Semi-supervision could be utilized in multiple ways. We tried three approaches in this article, (i) classification (to distinguish partial positives with negatives); (ii) ranking (to rate partial positive more likely than negatives); (iii) embedding (to make data clusters get similar labels). We applied this framework to improve the identification of interacting pairs between HIV-1 and human proteins. Our method improved upon the state-of-the-art method for this task indicating the benefits of semi-supervised multi-task learning using auxiliary information.

Availability: http://www.cs.cmu.edu/∼qyj/HIVsemi

Contact: qyj@cs.cmu.edu

1 INTRODUCTION

Identifying protein–protein interactions (PPIs) in a comprehensive manner is essential for understanding the molecular basis underlying biological functions. Because of their importance in development and disease, PPIs have been the subject of intense research in recent years, both *computationally* and *experimentally*.

Experimental techniques for detecting PPIs have been reviewed in Shoemaker and Panchenko (2007a). Traditionally, PPIs have been studied individually through the use of genetic, biochemical and biophysical experimental techniques (also termed *small-scale* methods). Experiments in this paradigm are typically expensive and time-consuming (months for detecting just one PPI). In recent years, *large-scale* biological PPI experiments have been introduced to directly detect hundreds or thousands of protein interactions at a time. The two-hybrid (Y2H) screens (Ito *et al.*, 2001; Rual *et al.*, 2005; Stelzl *et al.*, 2005; Uetz *et al.*, 2000) and complex purification detection techniques using mass spectrometry (Gavin *et al.*, 2002, 2006; Ho *et al.*, 2002) are the two most popular approaches thus far applied successfully on a large scale . However, their resulting data sets are often incomplete and exhibit high false positive and false negative rates (von Mering *et al.*, 2002; Yu *et al.*, 2008).

Computational methods have been successfully applied to predict protein interactions (reviewed in Shoemaker and Panchenko, 2007b). Taking into account that indirect sources may contain partial evidence about protein interations, several approaches derive their predictions on particular types of information, such as overrepresented domain pairs in interacting proteins (Wang *et al.*, 2007). An alternative attractive approach is to integrate various indirect or direct sources of evidence in a statistical learning framework. A classifier is trained to distinguish between positive examples of truly interacting protein pairs and negative examples of non-interacting pairs. *Various methods* have been explored in this framework, including naive Bayes classifier by Jansen *et al.* (2003), decision tree from Zhang *et al.* (2004), kernel-based methods from Ben-Hur and Noble (2005) and Yamanishi *et al.* (2004), random forest-based method (Qi *et al.*, 2005), logistic regression (Lin *et al.*, 2004), and the strategies of summing likelihood ratio scores (Lee *et al.*, 2004; Rhodes *et al.*, 2005; Scott and Barton, 2007). Most of these studies have been carried out in yeast or human. They aimed to predict PPIs within a single organism (termed 'intra-species PPI prediction'). Recent work extends to predicting PPIs between organisms ('inter-species PPI prediction'), especially between host and pathogens. Tastan *et al.* (2009) extended the supervised learning framework to predict PPIs between HIV-1 and human proteins. A random forest-based classifier was used to integrate multiple biological information sources and defined the state-of-art performance for this task. Additionally, Davis *et al.* (2007) studied 10 host–pathogen PPIs using structural information. Later, Evans *et al.* (2009) searched for host protein motifs along virus protein sequences to obtain a list of host proteins highly enriched with those targeted by HIV-1 proteins.

While the supervised framework was shown to enrich current PPI data with additional inferred PPIs, its applicability is still limited. Supervised PPI detection requires a large number of labeled training examples (truely interacting proteins) in order for the statistical classifier to predict with proper accuracy. Except several well-studied organisms, such as yeast or human, most inter or intra-species PPI prediction tasks do not have a large number of reliable PPIs available as training data. For instance, no reliable global set of interacting pairs exist between HIV-1 and human proteins (see Section 2.2). This limitation largely restricts the prediction ability of current computational PPI algorithms. To conquer this

*To whom correspondence should be addressed.

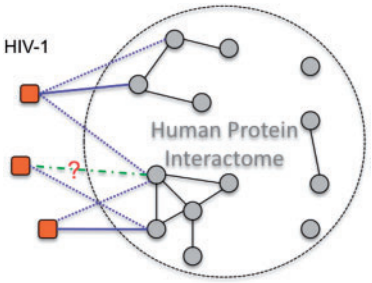

Fig. 1. Target problem: predicting protein interactions between HIV-1 (organge squre) and human (gray circle). There exist weakly labeled interaction pairs from NIAID (dashed blue edges) and labeled interaction pairs from experts' annotation (solid blue edges). We aim to predict whether a given unknown human to HIV-1 protein pair (dashed green) interacts or not.

limitation, recently semi-supervised approaches were proposed for computational PPI predictions. Yip and Gerstein (2009) proposed to improve supervised predictions by adding pseudo examples from previous runs of predictions. However, the predictions are prone to noise using this strategy.

It is sometimes possible to infer a relatively larger number of potentially interacting proteins, which may not have enough evidence to be confirmed as true positive labels. For instance, in the task of predicting PPIs between HIV-1 and human proteins, NIAID (Fu *et al.*, 2008) database retrieved protein pairs between HIV-1 protein and human protein from the scientific literature (details in Section 2). The extracted pairs are not experimentally confirmed PPIs, but are very likely to have interaction relationships. From a 'machine learning' perspective, these pairs are weakly labeled positive examples (see Fig. 1). In this case, an interesting question to ask is how to detect and add weakly labeled pairs to improve computational PPI predictions.

In this article, we present a multi-task learning framework to make use of weakly labeled examples together with conventionally labeled PPI pairs. A semi-supervised task is introduced in a network consisting of multiple layer perceptron as an *auxiliary task*. We train supervised PPI classification and the semi-supervised auxiliary task under the same network *simultaneously*. We apply our method to predict the set of interacting proteins between HIV-1 and human proteins by information integration of multiple biological sources. Our method improves upon the previous approach applied for this task. The results indicate that with the proposed semi-supervised multi-task approach, auxiliary information (weak labels) is able to improve the accuracy of the predictive models for PPIs between HIV-1 and human proteins.

The rest of the article is structured as follows. Section 2 describes the task of predicting PPIs between HIV-1 and human proteins and the available interacting data set in more details. Section 3 describes the semi-supervised multi-task learning framework. Section 4 presents the experimental results, and Section 5 concludes.

2 TARGET PROBLEM

HIV-1 causes the disease of acquired immune deficiency syndrome (AIDS), which remains a serious and growing threat to public health (Trkola, 2004). Both HIV-1 transmission and infection are complex processes, where much remains to be elucidated. The HIV-1 RNA encodes only a handful of proteins; however, it subverts the cellular machinery for its benefit. Virus–host PPIs are key in deciphering virus strategies, and such understanding may lead to designing novel ways to impede viral protein functions and thus reduce or eliminate HIV-1's potency as a deadly pathogen.

2.1 Information integration with multiple data sources

Recently, we made an attempt to predict the global set of interactions between HIV-1 and human host cellular proteins in Tastan *et al.* (2009). The task was to predict whether a given human to HIV-1 protein pair interacts or not. Thus it was formulated as a binary classification problem, where each protein pair belongs to either the 'interaction' or 'non-interaction' class. A random forest classifier was trained on a rich set of features including:

- co-occurrence counts of binding motifs to matched interacting domains;
- gene expression profile reflecting human gene expression patterns across HIV-1 samples: infected versus uninfected;
- similarity in terms of cellular location, molecular function and biological process;
- similarity of HIV-1 protein to human protein's known binding human partners (in terms of localization/function/process);
- pairwise sequence similarity between HIV-1 and human protein or its known human binding partners;
- if the HIV-1 protein shares any post-translational modification with human binding partners of the human protein;
- similarity of tissue distributions;
- topological properties of the human protein in human protein interaction network, such as node degree;
- HIV-1 protein type.

All data sources and how they were converted into features representing protein pair between HIV-1 and human have been described previously in Tastan *et al.* (2009) and are available for download in our supplement web site.

2.2 Partial positive labels from NIAID

The gold standard positive set we used in Tastan *et al.* (2009) were collected from NIAID (Fu *et al.*, 2008) database where interactions between HIV-1 and human proteins reported in the scientific literature were manually curated. It includes 2620 protein pairs involving 1406 human proteins and 17 HIV-1 proteins [15 HIV-1 proteins plus precursors of the envelope (env gp160) and gag (gag pr55)]. Each interaction in the database is associated with keywords extracted from scientific literature reporting the interaction. Some of these keywords are strong such as 'interacts with' and 'binds' (we named this set as 'GroupI' containing 955 protein pairs). While some other keywords are rather weak indicators of direct interactions such as 'upregulates' or 'inhibits' (this set of pairs was named as 'GroupII' and included 1665 protein pairs). Our previous work (Tastan *et al.*, 2009) used those 'GroupI' interactions (associated with strong keywords[1]) as training positive examples for binary PPI predictions.

[1]Small difference exists in keyword splits here, to (Tastan *et al.*, 2009)

Table 1. Basic statistics of feature and 'gold standard' set

Features	Positive PPIs (experts)	Partial positive	Remaining pairs	HIV-1 protein	Human protein
18	**158**	**2119**	352 338*	17	20 873

*This also excludes 226 pair experts labeled as 'unsure'. Bold values means related to PPI.

Recent studies from Cusick *et al.* (2008) pointed out that the literature-curated protein interaction experiments can be error-prone and possibly of lower quality than commonly assumed. The 'gold standard' reference set used in our previous work (Tastan *et al.*, 2009) was *ad hocly* built from 'GroupI' of NIAID. Clearly there exists not enough evidence supporting the reliabilities of these interaction labels from NIAID database.

2.3 Positive labels from experts' annotations

To increase the data quality, we consulted 16 HIV-1 experts about the validity of the interactions reported in the NIAID database. 15 of the experts are professors well known in the HIV-1 field and the last expert is a PHD student, who had extensively worked on HIV1 for 5 years. More details of experts' annotation process is provided in our supplementary web. HIV-1 experts were sent lists of interacting pairs along with the interaction keywords and the links to the articles reporting the interactions in NIAID. Experts are asked to annotate each pair with the 'interact' label if they believe the reported pair is a true direct interaction. If, on the other hand, either they do not believe two proteins interact, annotating it with the label 'not interacting', or they think the interaction might be indirect or they are unsure about the label, annotating it as 'unsure or indirect'. For each HIV-1 protein, the rules to select potential interaction partners sent to experts are *different*. If for a certain HIV-1 protein, the total number of interactions reported in NIAID is not many, we sent all of the interactions reported in the database. In other cases, only the subset of interactions associated with keywords 'binds' or 'interacts with' was sent (the longer the list is, the slower and more reluctant the experts' responses were). In this way, 361 interacting pairs were annotated. Most of the interactions (256/361) were annotated by a single expert and the rest received labels from multiple experts. In cases where there was a disagreement between experts' opinions on the labels, the 'majority vote' strategy was used to decide which label should be assigned. Finally, this resulted in **158** protein pairs that HIV-1 experts annotated as direct interactions between HIV-1 and human proteins.

Thus, this set serves as our positive 'gold standard' set. The rest of the NIAID dataset are treated as 'partial positives' examples since not enough evidence is yet accumulated for them to be considered as direct interactions but they are likely candidates.

In summary, this binary classification task contains **158** 'experts-annotated' positive example and **2119** partial positive (with **552** from 'groupI' and **1567** from 'groupII') PPI pairs (after removing those pairs labeled as 'not interact' and 'interact' from the experts). Each HIV-1 human protein pair is represented with 18 features. Related statistics of data sets used for this task are listed in Table 1.

The feature set used in our previous work (Tastan *et al.*, 2009) contains totally 35 attributes for each potential HIV-1 to Human protein pair. Among them, 17 items represent which one (assuming *i*) of the 17 HIV-1 proteins this pair involves with (with the *i* dimension set to 1 and all the other 16 dimensions set to zero). As mentioned above, since the creation process of *experts annotated (positive)* labels is correlated non-randomly with the type of HIV-1 proteins, we have to remove these 17 features, and use the remaining 18 features to describe each HIV-1 human protein pair. All labeled & partially labeled examples are shared in our supplementary web.

3 METHOD

A d-dimensional ($d=18$) feature vector x was constructed for every protein pair (between a HIV-1 protein and a human protein). Each entry in the feature vector summarizes one biological evidence (asking, for example, 'Does this HIV-1 protein include a certain motif that is highly likely to interact with one of the domains in the human protein?' (see Section 2.1). The target variable $y \in \{\pm 1\}$ represents whether this pair interacts (1) or not (-1). Thus, the problem of predicting protein interactions is handled as a binary classification task.

Considering the small number of positive labels (**158**) and a larger set of partial labels (**2119**), we propose to design semi-supervised multi-task learning (SML) strategies for making use of both sets, to achieve better prediction performance.

Given a set of labeled examples $(x_1,\ldots,x_L)$ and corresponding labels $(y_1,\ldots,y_L)$, our goal is to learn a supervised classifier (e.g. choose a discriminant function) $f(x)$, such that '$f(x_i)>0$ if $y_i=1$' or '$f(x_i)<0$ if $y_i=-1$'.

3.1 A multi-layer perceptron network for supervised PPI prediction

The supervised classifier we chose is a multi-layer perceptron (MLP) network with M layers of hidden units that gives a 1-dimensional output:

$$f(x)=\sum_j w_j^O \; h_j^M(x)+b^O, \tag{1}$$

where w^O is the weight vector for the output layer. The m-th hidden layer is defined as

$$h_i^m(x) \;=\; S\left(\sum_j w_j^{m,i} \; h_j^{m-1}(x)+b^{m,i}\right), m=2,...,M \tag{2}$$

and S is a non-linear squashing function like 'tanh'. To train this supervised classifier, we employ the Hinge loss (on labeled examples):

$$\sum_{i=1}^{L} \ell(f(x_i),y_i)=\sum_{i=1}^{L} \max(0,1-y_i f(x_i)). \tag{3}$$

3.2 Multi-task learning with semi-supervised auxiliary task

According to the available labels, we could formalize our objective as two tasks: (i) supervised classification with positive (from experts) and negative labels; (ii) the usage of partial positive labels in order to improve the supervised classification. One natural way to combine two objectives is through multi-task learning.

Multitask learning is the procedure of learning several tasks at the same time with the aim of mutual benefits. A good overview of multi-task learning, especially focusing on neural networks, can be found in Caruana (1997). The idea of sharing information learnt across sub-tasks seems a more economical use of data, where presumably all tasks are learnt *jointly*. A typical example is a MLP network where first layers will be shared to all tasks and typically learn levels of feature processing that are useful to all tasks.

In our problem, the second task aims to make use of weak positive labels and is auxiliary to the main classification. We call them 'semi-supervised auxiliary task' in this article since the task, uses just weak labels with

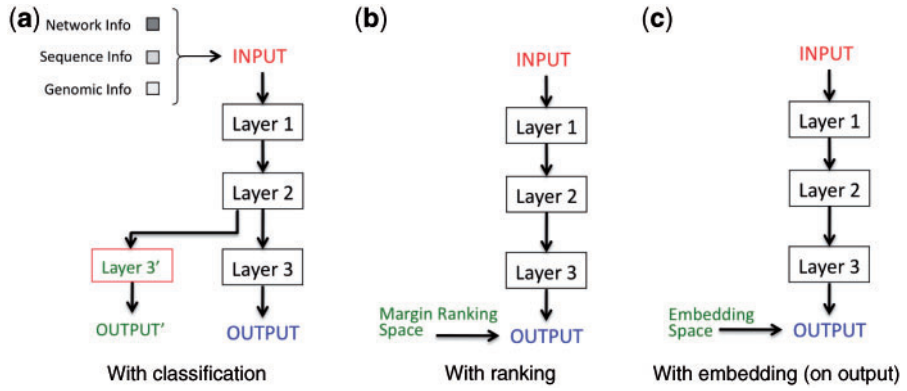

Fig. 2. To perform multi-task learning with the supervised PPI classification, three semi-supervised tasks have been proposed to extend the network structure of multi-layer perceptron: (**a**) training another classifier to distinguish partial positive and negative examples; (**b**) training a ranker to sort partial positive and negative data; (**c**) training an embedding on the output of the supervised classifier.

diverse levels of confidence (e.g. various keywords associated in NIAID database). Typical semi-supervised learning refers to the use of both labeled and unlabeled data during training. For our task, though not the typical semi-supervised setting, we view it as a similar setup and the proposed auxiliary tasks could be naturally extended to unlabeled data with side information, e.g. functional association between proteins.

Formally speaking, multi-task learning of supervised classification and semi-supervised auxiliary task equals to learning two tasks jointly with the optimization of the following loss function:

$$\sum_{i=1}^{L} \ell(f(x_i), y_i) + \lambda \quad \text{Loss (Auxiliary Task)} \tag{4}$$

There exist many ways to build this *auxiliary task* using MLP networks (e.g. different network structure and/or distinct loss function). In the following, we propose three possibilities (in Fig. 2).

3.3 Auxiliary Task I: classification

Figure 3a illustrates the first strategy to use partial labels. This is the classical way of multi-tasking in the MLP framework. Our auxiliary task shares the first m layers of the original MLP, but have a new output layer:

$$g(x) = \sum_j w_j^{AUX} h_j^M(x) + b^{AUX} \tag{5}$$

This network is trained to *distinguish* partial positive examples from negative examples (e.g. classification), simultaneously as we train the original network on *labeled* data. Assuming a set of partially labeled examples $(x_{L+1}, \ldots, x_{L+U})$. In this auxiliary task, they are assigned with corresponding pseudo labels $(y'_{1+U}, \ldots, y'_{L+U})$. We train this pseudo classficiation with hinge loss as well, which means,

$$\text{Loss (Auxiliary Task)} = \sum_{j=L+1}^{L+U} \max(0, 1 - y'_j g(x_j)) \tag{6}$$

3.4 Auxiliary Task II: ranking

Illustrated in Figure 3b, this time we use the same network architecture for both two tasks. The auxiliary information we know for the second task is 'partial labeled PPI pairs are more likely to be true than negative pairs'. This could be formalized as a 'ranking' task using MLP: to rank 'weak positive examples' highly than 'negative examples' if ordering them by the output $f(x)$ from the MLP. Naturally, the above assumption comes to minimize a ranking-type margin cost:

$$\text{Loss(Aux.)} = \sum_{p \in P} \sum_{n \in N} \max\left(0, 1 - f(x_p) + f(x_n)\right), \tag{7}$$

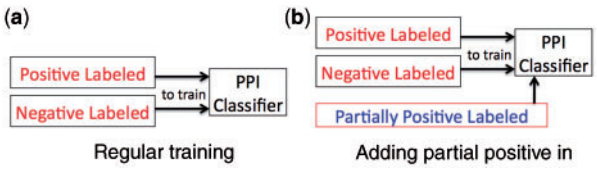

Fig. 3. Two ways to train baseline classifiers for performance comparison. (**a**) train with positive + negative; (**b**) train with positive + partial positive (treat as positive) + negative.

where P means the index of partial positives and N represents the set of negative examples. The training is handled with stochastic gradient descent which samples the cost *online* w.r.t. (p, n).

3.5 Auxiliary Task III: embedding on output

One key assumption used by many semi-supervised algorithms is the structure assumption, which assumes that points within the same structure (such as a cluster or a manifold) are likely to have the same label (Chapelle *et al.*, 2006). In Figure 2c, we explore the partial labeled examples as a guidance to explore the hidden structure assumption in our data.

This could be pursued through an embedding technique proposed in Weston *et al.* (2008). The proposed model contains a network with two identical copies of the same function, with the same weights, and with outputs fed into a 'distance'-measuring layer. Given two examples x_i and x_j, we can feed each of them into these two identical networks, and use the last 'distance' layer to compute whether the two examples are similar or not (i.e. in terms of their network outputs). If we know in advance whether they are similar or not, this pairwise 'labeling' could function as 'hidden structure' guidance and can be used for learning of parameters in the network (Fig. 2c). A margin-based loss following (Weston *et al.*, 2008) is chosen for training:

$$L(f_i, f_j, W_{ij}) = \begin{cases} ||f_i - f_j||_2 & \text{if } W_{ij} = 1, \\ \max(0, m - ||f_i - f_j||_2) & \text{if } W_{ij} = 0 \end{cases} \tag{8}$$

This loss function encourages similar examples (where $W_{ij} = 1$) to be close in output space, and dissimilar ones to have a distance of at least m from each other's output. W_{ij} specifies the similarity or dissimilarity between examples x_i and x_j, which serves as the 'pairwise labeling' guidance for the embedding loss function.

With *partially labeled* examples in our data, we could derive a set of W_{ij} labels to embedding training. Three strategies are considered in our

experiments to derive W_{ij}:

- $W_{ij}=1$, if both examples from the *partially labeled* set;
- $W_{ij}=1$, if one *partially labeled* example and the other from the positively *labeled* set;
- $W_{ij}=0$, if one *partially labeled* example and the other a negatively *labeled* example;

The main motivation is that even though examples of partial positive PPI sets have not enough evidence to be considered as direct interactions, they are highly likely candidates. Thus in the embedding of output space, these examples should be similar to each other, and dissimilar to negative examples. The embedding model is trained by the pairs of examples with W_{ij} labels. The training is also handled with stochastic gradient descent which samples the cost online w.r.t (i,j). Training steps either 'push' similar examples together or 'pull' dissimilar examples apart from each other. This hidden structure is exactly what we want to preserve in the space of output $f(\cdot)$ in our data.

It is very natural to multi-task embedding task with our main supervised classification task. Since embedding model makes use of a neural network with two identical copies and an extra 'distance measuring' layer, we can just use our supervised classifier MLP as the base network for embedding. This equals to add the embedding as a regularizer on our main classifier MLP. In Figure 2c, a semi-supervised regularizer is added on the supervised loss measured on the entire network's output (1):

$$\sum_{i=1}^{L}\ell(f(x_i),y_i)+\lambda\sum_{i,j=1}^{L+U}L(f(x_i),f(x_j),W_{ij}) \tag{9}$$

Here, labeled training examples are denoted as x_i, $i=1,\ldots,L$ and partially labeled examples are denoted as x_i, $i=L+1,\ldots,L+U$. Essentially, multi-tasking tries to classify labeled examples, whilst simultaneously the embedding tries to push the classification score of partial positive examples close to the scores of positive examples, and apart from those scores of negative labeled examples.

3.6 Semi-supervised multi-task learning

The overall goal of the auxiliary task is to improve accuracy on the supervised task by uncovering hidden structures in the original data. All tasks, including classification, ranking and embedding, are trained by stochastic gradient descent. The training cooperation between the main task and the auxiliary task could be summaried as looping over two tasks:

(1) Select the next task.

(2) Select a random training example for this task.

(3) Update the MLP network parameters for this task by taking a gradient step with respect to this example.

(4) Go to 1.

To give a concrete example, the pseudocode of multitasking with 'embedding output' case is given in Algorithm 1.

Algorithm 1 Multi-tasking with embedding on layer

Input: labeled data (x_i,y_i), $i=1,\ldots,L$, partially labeled data x_i, $i=L+1,\ldots,L+U$, set of functions $f(\cdot)$, see Eq. (1):
repeat
  Pick a random *labeled* example (x_i,y_i).
  Make a gradient step to optimize $\ell(f(x_i),y_i)$, see Eq. (3).
  Pick a random *partially labeled* example x_p.
  Pick a random example x_q, where $W_{pq}=1$.
  Make a gradient step for $\lambda L(f(x_p),f(x_q),1)$, see Eq. (9).
  Pick a random *partially labeled* example x_m.
  Pick a random example x_n, where $W_{mn}=0$.
  Make a gradient step for $\lambda L(f(x_m),f(x_n),0)$, see Eq. (9).
until stopping criteria is met.

4 RESULTS

4.1 Experimental setting

When training the classification model, negative (non-interacting) examples are required. However, it is almost impossible to show two proteins do not interact, a large set of non-interacting protein pairs does not exist. A commonly applied strategy is to randomly select protein pairs from all possible protein pairs as the negative set, excluding those known interacting ones. Here, we exclude all those pairs that are in NIAID database. For interacting pairs between HIV-1 and human proteins, it is estimated that roughly only 1 in about 100 possible pairs actually interacts (Tastan *et al.*, 2009). This is an extremely unbalanced ratio between positive and negative sets. We use this ratio to build the negative set which includes $\sim$16 000 random negative pairs.

The positive pairs in our setting include only those PPIs pairs confirmed by the HIV-1 experts as 'interacting' (**158** pairs). The partial positive pairs (**2119** left pairs of NIAID) function as auxillary information in the training phase only.

The experimental evaluation is based on five folds cross-validation (CV) with 20 randomly repeated CV runs to obtain average performance scores. The reason we repeat cross-validation runs is that randomness exists when sampling the negative training set. To conquer this random effect, we pursued multiple CV runs on multiple independently sampled negative sets. Averaged performance scores are used for comparisons.

To measure the predictive power of SML for identifying protein interactions between HIV-1 and Human, we compared three variants of SML with two baseline classifiers. The SML models are named as: (i) SMLC: SML with auxiliary classification task; (ii) SMLR: SML with auxiliary ranking task; (iii) SMLE: SML with embedding on output space. Two baselines include: (i) RF: Random Forest; (ii) MLP: Multi-Layer Perceptron Neural Net. Three SML methods and the MLP model are implemented using Torch 5 package (Weston *et al.*, 2008). Random Forest was from the Berkeley RF package (Breiman, 2001).

4.2 Baselines to compare

Our task formulation is closely related to the framework of supervised classification of protein pairs through information integration. Tastan *et al.* (2009) showed that RF give the state-of-art performance for the HIV-1 to human PPI prediction task (though partial labels 'GroupI' used as *positive* for training in that case). Our SML models are built on MLP networks. Thus, it is worth to compare and investigate how much improvement we could achieve beyond the baselines: MLP network classifier and the state-of-the-art RF classifier.

Moreover, we also evaluate the performance of both two base classifiers when adding those partially labeled positive pairs into the training positive (from experts). Ideally, these partial labels should be weighted differently in the training compared to those experts' labels. But since partial positive pairs are associated with different keywords in NIAID, it is tricky to select the weights. We finally used a simple strategy in evaluation: just adding them as

training positive examples into the current positive set. Figure 3 summarized two ways we utilized in training baseline classifiers. For the case in Figure 3b, two baselines are named as: (i) RF-P: Random Forest adding partial positives in training; (ii) MLP-P: Multi-Layer Perceptron Neural Net adding partial positives in training.

For each classifier, parameter optimization was carried out independently in identical cross-valiation fashion. Each method has distinct sets of parameters to tune. For SMLE, we need to learn the underline MLP network structure (hidden layer, hidden units, etc.), the learning rate, choices of embedding pairs, ratio between embedding and classification during the joint training. For SMLC, we need to learn the underline MLP network structure, ratio between the main classification and the pseudo classification, and the learning rate. For SMLR, we need to learn the underline MLP network structure, ratio between the main classification and the pseudo ranking, the learning rate and the choices of pairwise ranking pairs. To avoid overfitting, we did not try very deep MLP architecture. Thus either linear (if possible) or adding one hidden layer was tried for MLP architecture. The best parameters found for the classification auxiliar model is with one hidden layer, 15 hidden units and learning rate 0.005. For the ranking model, the best setup is with linear layer with learning rate 0.01. For the SML 'embedding output' model, the choice is one hidden layer, five hidden units, learning rate 0.005 and we train embedding with only the pushing apart step.

4.3 Evaluation metrics

When evaluating the performance of a classifier on an imbalanced test set such as is the case here, computing accuracy is not useful because a high true negative (TN) rate can easily be obtained by chance. Therefore, we evaluated the quality of our predictive model using four metrics which ignore the success on the TN rate and summarize prediction performance over a range of output thresholds. (i) Mean average precision (MAP) score is used to summarize the precision recall curve and is the mean of the average precision scores across recall levels. Precision refers to the fraction of interacting pairs predicted by the classifier that are truly interacting ('true positives'). Recall measures how many of the known pairs of interacting proteins have been identified by the learning model. (ii) Precision recall breakpoint (PRB) score is the value of when precision is equal to recall across different cutoffs on the predicted score. (iii) Receiver operator characteristic curves plot the true positive rate against the false positive rate for different cut-off values of the predicted score. The area under the ROC curve (AUC) is commonly used as a summary measure of diagnostic accuracy. (iv) R50 is a partial AUC score that measures the area under the ROC curve until reaching 50 negative predictions, e.g. low false positive rate. For our prediction task where classes are extremely unbalanced, we are predominantly concerned with the condition where false positive rate is low.

All these score range between 0 and 1, where values close to 1 indicates more successful predictions.

4.4 Performance

Table 2 compares three proposed SML models and two baseline classifiers (each have two cases of training) using AUC R50, MAP, PRB and AUC scores. The scores are averaged from 20 randomly repeated 5-folds CV runs.

Table 2. Performance comparison (with multiple metric scores)

Method	R50	MAP	PRB	AUC
SMLC	0.277	0.263	0.312	0.905
SMLR	**0.310**	0.268	0.311	**0.919**
SMLE	0.309	**0.277**	**0.326**	0.908
RF	0.199	0.135	0.180	0.893
RF-P	0.230	0.213	0.281	0.896
MLP	0.204	0.197	0.257	0.859
MLP-P	0.229	0.210	0.282	0.893

SMLC, SML with classification task; SMLR, SML with ranking task; SMLE, SML with embedding on output; RF, Random Forest; MLP, Multi-Layer Perceptron Net. RF-P, RF adding partial positive; MLP-P, MLP adding partial positive. Bold values gives the best performance in the column.

For two baselines, the second type of training (adding partial labels in) achieves better performance than the regular training, which is not surprising. MLP model (MAP-P 0.21) makes comparable performance to the state-of-the-art RF (MAP 0.213) model.

All SML models perform better than baseline strategies (the best MAP achieves 0.277, e.g. about 0.06 better than RF-P; the best R50 gets 0.310, about 0.08 better than RF-P). This is expected since our partial positive examples are associated with keywords describing PPIs. SML auxiliary tasks tried to capture the intrinsic patterns underlying these weak labels, from either labels themselves, or their pairwise relationships with other examples. Multi-tasking with MLP improve the performance compared to MLP alone. We conclude that SML achieves the state-of-art performance on the task of predicting interactions between HIV-1 and human protein.

Among three SML models, the SMLR—'ranking' and SMLE—'embedding on output' task seem to capture the patterns of partial labels better compared SMLC—'classification'. We think this observation makes sense since essentially the only reliable assumption we could derive from weak positive labels is 'partial positives are more likely to be interacting than negative random pairs'. The ranking auxiliar task—SMLR performed training on this assumption exactly, which achieved the best R50 (0.310) and the best AUC (0.919) scores. Under the best parameter setup (learned by CV), the 'embedding output'—SMLE task is similar to SMLR where it tried to push the network output value of partial positives apart from the output of random negatives. This achieved the best MAP (0.277, about 0.067 increase to RF-P) and the best PRB (0.326, about 0.045 increase to RF-P) scores. SMLC model could not capture this assumption directly, consequently resulting in less improvement from multi-tasking.

Furthermore, we tried to compare SML models directly with previous results in Tastan *et al.* (2009). Our current 'gold standard' positive set uses the 158 experts annotated interactions between HIV-1 and human proteins. Differently, (Tastan *et al.*, 2009) used 955 'GroupI' pairs as training positive. We tried to apply SMLR model on the same supervised PPI prediction runs in Tastan *et al.* (2009) and multi-task with the ranking task using 'GroupII' 1665 pairs as partially labeled examples. This model gets an averaged 0.253 MAP score and RF achieved 0.230 MAP score in Tastan *et al.* (2009). The improvement is less impressive in this setting and we guess this is because 'GroupII' set is not much larger than'GroupI' set.

Table 3. Statistics of overlaps between top predicted human partners to those found in (i) (Brass *et al.*, 2008) siRNA screen list, (ii) (Ott, 2008) virion screen list, (iii) combined four siRNA screens (Brass *et al.*, 2008; König *et al.*, 2008; Yeung *et al.*, 2009; Zhou *et al.*, 2008)

Score cutoff	Num Predicted interactions	Confirmed by NIAID	Novel Interactions	No. human protein in predInteractions	Overlap siRNA	overlap virion	overlap CombineFourSiRNA
−1.8	3428	259	3123	1027	24	72	96
−1.5	2434	223	2172	721	21	61	72

4.5 Validation

A final model was trained with all available expert labeled interactions using the best parameter setting we found for SMLE. Since randomness exists when sampling the negative training set, we utilized multiple independently sampled negative sets to overcome this random effect and to reduce the potential bias inherent in using a single training set. Through bagging models trained with five randomly sampled negative sets, our final score is obtained through value averaging. We then ranked all HIV-1 to human protein pairs according to their classification score. The derived ranked order list were thresholded and the top ranked 2500 pairs build our list of predicted PPIs. This list is downloadable from our supplementary web.

Following Tastan *et al.* (2009), we carry two validations by checking whether the human proteins reported in the functional siRNA screen are ranked high in our predicted list. The siRNA screen identified 282 human genes to have an effect on HIV-1 infection (Brass *et al.*, 2008). Also we check the human proteins in our top ranked PPI list, whether they have been detected in virion (Ott, 2008) or not. This functional assay found that 316 human proteins are highjacked by HIV-1 in its virion. The predicted pairs that involve with the virion related human genes would be of great interest to HIV-1 virologists. Table 3 gives the statistics of overlapps between our predicted human partners to proteins found in the two reported functional screens. Clearly, there is a good portion of predictions confirmed by these functional screens. Recently, three other functional screened human gene lists (König *et al.*, 2008; Yeung *et al.*, 2009; Zhou *et al.*, 2008) related to HIV-1 become available online. We combine these three human gene lists with the one in Brass *et al.* (2008) to form a combined list called as 'CombineFourSiRNA' in Table 3. We then check the overlap of this combined list to our top predicted human protein partners. The last column in Table 3 describes that nearly 10% predicted partners are validated by four siRNA screens, which gives strong indications of how good our predictions are.

5 CONCLUSIONS

Supervised learning methods have been used for the task of classifying pairs of proteins as interacting or not. However their performance is restricted by the availability of labeled training examples, i.e. known PPIs. In many cases, there exist considerable amount of protein pairs, where an association is proposed in the literature but not enough experimental evidence is available to determine the existence of a direct interaction. Such is the case for the task of predicting human to HIV-1 inter-species interactome.

In this article, we designed a semi-supervised multi-task learning framework to integrate a larger set of potentially interacting protein pairs retrieved from literature (*weak labels*) and a smaller set of interactions annotated by experts. The proposed SML combine a semi-supervised auxiliary task with a supervised PPI classifier. A multi-layer perceptron network is trained for PPI classification on *labeled* examples. Simultaneously, we multi-task this network with an auxiliary task which aims to use weak positive labels to improve the supervised classification. Three auxiliary strategies are evaluated on the task of predicting interactions between HIV-1 and human proteins. Through CV, our method was shown to improve upon the best previous method for this task indicating the benefits of multi-tasking with auxiliary information.

In addition to improved performance on inferring human HIV-1 PPIs, the proposed SML structure provides a flexible framework for general computational PPI prediction tasks. SML models could be easily extended to other species or pairs of species, or to incorporate other auxiliary information, such as other kinds of weak labels or supporting information between unlabeled protein pairs. For instance, the noisy interaction pairs from high throughput experiments in human could be used to build neighbor pairs for training SML model (e.g. embedding on output) very naturally and thus the method has significant potential for intra-species PPI predictions such as in human.

Funding: This work is supported in part by NEC Labs America Core Research Funding, National Institutes of Health grants P50-GM082251 and LM07994-01.

Conflict of Interest: none declared.

REFERENCES

Ben-Hur,A. and Noble,W. (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21**, i38–i46.

Brass,A.L. *et al*. (2008) Identification of host proteins required for hiv infection through a functional genomic screen. *Science*, **319**, 921–926.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32. article.

Caruana,R. (1997) Multitask learning. *Mach. Learn.*, **28**, 41–75.

Chapelle,O. *et al*. eds (2006) *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. MIT Press, MA, USA.

Cusick,M.E. *et al*. (2008) Literature-curated protein interaction datasets. *Nat. Methods*, **6**, 39–46.

Davis,F.P. *et al*. (2007) Host pathogen protein interactions predicted by comparative modeling. *Protein Sci.*, **16**, 2585–2596.

Evans,P. *et al*. (2009) Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med. Genomics*, **2**, 27.

Fu,W. *et al*. (2008) Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res.*, **37**, D417–D422.

Gavin,A.-C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Gavin,A. *et al*. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.

Ho,Y. *et al*. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

Ito,T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast proteininteractome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jansen,R. *et al*. (2003) A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.

König,R. *et al*. (2008) Global analysis of host-pathogen interactions that regulate early-stage hiv-1 replication. *Cell*, **135**, 49–60.

Lee,I. *et al*. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.

Lin,N. *et al*. (2004) Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, **5**, 154.

Ott,D.E. (2008) Cellular proteins detected in hiv-1. *Rev. Med. Virol.*, **18**, 159–175.

Qi,Y. *et al*. (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. *Proc. Pac. Symp. Biocomput.*, **10**, 531–542.

Rhodes,D.R. *et al*. (2005) Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.*, **8**, 951–959.

Rual,J.F. *et al*. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.

Scott,M.S. and Barton,G.J. (2007) Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinformatics*, **8**, 239.

Shoemaker,B.A. and Panchenko,A.R. (2007a) Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLoS Comput. Biol.*, **3**, e42.

Shoemaker,B.A. and Panchenko,A.R. (2007b) Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, **3**, e43.

Stelzl,U. *et al*. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 830–832.

Tastan,O. *et al*. (2009) Prediction of interactions between HIV-1 and human proteins by information integration. *Pac. Symp. Biocomput.*, **14**, 516–527.

Trkola,A. (2004) HIV-host interactions: vital to the virus and key to its inhibition. *Curr. Opin. Microbiol.*, **7**, 555–559.

Uetz,P. *et al*. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

von Mering,C. *et al*. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

Wang,H. *et al*. (2007) InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biol.*, **8**, R192.1–R192.18.

Weston,J. *et al*. (2008) Deep learning via semi-supervised embedding. In *ICML '08: Proceedings of the 25th International Conference on Machine Learning*. ACM, New York, NY, USA, pp. 1168–1175.

Yamanishi,Y. *et al*. (2004) Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, **20**, 363–370.

Yeung,M.L. *et al*. (2009) A genome-wide short hairpin rna screening of jurkat T-cells for human proteins contributing to productive HIV-1 replication. *J. Biol. Chem.*, **284**, 19463–19473.

Yip,K.Y. and Gerstein,M. (2009) Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. *Bioinformatics*, **25**, 243–250.

Yu,H. *et al*. (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.

Zhang,L. *et al*. (2004) Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, **5**, 38.

Zhou,H. *et al*. (2008) Genome-scale rnai screen for host factors required for hiv replication. *Cell Host Microbe*, **4**, 495–504.