# A Mixture of Feature Experts Approach for Protein-Protein Interaction Prediction

Yanjun Qi[1], Judith Klein-Seetharaman[1,2], Ziv Bar-Joseph[1]
[1]School of Computer Science, Carnegie Mellon University, Pittsburgh,PA 15213
[2]Department of Pharmacology, University of Pittsburgh School of Medicine, Pittsburgh,PA 15261

## Abstract

High-throughput methods can directly detect the set of interacting proteins in yeast but the results are often incomplete and exhibit high false positive and false negative rates. A number of researchers have recently presented methods for integrating direct and *indirect* data for predicting interactions. However, due to missing data and the high redundancy among the features used, different samples may benefit from different features based on the set of attributes available. In addition, in many cases it is hard to directly determine which of the datasets led to the prediction, which is an important issue for the biologists using these predications to design new experiments.

To address these challenges we use a Mixture-of-Experts method. We split the data into four (roughly) homogeneous sets. The individual experts use logistic regression and their scores are combined using another logistic regression. However, when combining the scores the weighting of each expert depends on the set of input attributes. Thus different experts will have different influence on the prediction depending on the available features.

We applied our method to predict the set of interacting proteins in yeast. Our method improved upon the best previous methods for this task. In addition, using the weighting of the experts the prediction can be easily evaluated by biologists based on the features that they feel are the most reliable.

## 1 Introduction

Correctly identifying the set of interacting proteins in an organism is useful for deciphering the molecular mechanisms underlying given biological functions and for assigning functions to unknown proteins based on their interacting partners. It is estimated that there are around 30,000 specific interactions in yeast, with the majority to be discovered [3].

A number of high-throughput experimental approaches have been applied to determine the set of interacting proteins on a proteome-wide scale in yeast. These include the two-hybrid (Y2H) screens and mass spectrometry methods. However, both methods suffer from high false positive and false negative rates [3]. Roughly 80,000 interactions have been predicted in yeast by various high-throughput methods, but only a small number ($\sim$2,400) are supported by more than one method.

In addition to direct interaction, there are many indirect sources that may contain information about protein interactions. For example, it has been shown that many interacting pairs are co-expressed [3]. These datasets provides partial information about the interacting pairs and suggest that direct data on protein interactions can be combined with indirect data to improve the success of protein interaction prediction when compared to direct data alone.

Researchers have recently suggested a number of methods to predict protein interactions by combining multiple data sources. Jansen et al. [4] combined direct and indirect data sources using a Bayes classifier. Lin et al.[6] compared Jansen's method with two other classifiers, Random Forest (RF) and Logistic Regression (LR) and found RF to be the best among them. Zhang et al.[10] constructed a decision tree to predict co-complexed protein pairs. Ben-Hur et al.[7] used kernels for protein interaction prediction. Yamanishi et al.[9] predicted pathway protein interactions using a variant of kernel canonical correlation analysis. All of the above methods were shown to improve the success of protein interaction prediction when compared to direct data alone.

While useful, the above methods do not address two important problems in this domain. First, these classification methods estimate a set of parameters that are used for all input pairs. However, the biological datasets used contain many missing values and highly correlated features. Thus, different samples may benefit from using different feature sets. The second problem is that biologists who want to use these methods to select experiments cannot easily determine which of the features contributed to the resulting prediction. Since different researchers may have different opinions regarding the reliability of the various features, it is useful if the method can indicate, for every pair, which feature contributed the most to the classification result.

In this paper we address the above challenges using a mixture of experts method. We divide the biological datasets into several groups. Each of the groups represents a specific data type and is used by an expert (classifier) to predict interactions. Results from all experts are combined such that the weight of each expert depends on the input sample and thus varies between input pairs. This weight can also be used to indicate the importance of the features used by this expert for predicting a pair. The importance can be used by biologists to determine their

Table 1: We used a total of 162 features from 17 different data sources. The first column lists which expert the feature source belongs to (Total four experts: P, E, F and S). The second column lists the name of the feature source. The third column lists the numbers of features in each source. The fourth column presents the percentage of pairs for which information is available using this feature.

| | Feature Source | Size | Coverage |
|---|---|---|---|
| P | HMS-PCI MS | 1 | 8.3 |
| P | TAP MS | 1 | 8.8 |
| P | Yeast-2-Hybrid | 1 | 3.9 |
| F | GO Function | 21 | 80.7 |
| F | GO Process | 33 | 76.1 |
| F | GO Component | 23 | 81.5 |
| F | Essentiality | 1 | 100 |
| F | MIPS protein class | 25 | 4.6 |
| F | MIPS mutant phenotype | 11 | 9.4 |
| S | Gene fusion/cooccurence | 1 | 100 |
| S | Sequence similarity | 1 | 100 |
| S | Homology derived PPI | 4 | 100 |
| S | Domain interaction | 1 | 100 |
| E | Gene Expression | 20 | 88.9 |
| E | Protein Expression | 1 | 42.8 |
| E | Trans Factor Binding | 16 | 98.0 |
| E | Synthetic Lethal | 1 | 7.6 |

confidence in the classification results.

We applied our method to predict protein interactions in yeast and the method improved upon previous methods. For a specific Yeast pathway, the pheromone pathway, we show that it is possible to extract information from the weight distribution, in addition to providing new predictions.

## 2 Feature Set

There are many biological sets related to protein-protein interaction. In this paper, we collected a total of 162 feature attributes from 17 different data sources (Table 1). Overall, these data sources can be divided into four feature experts: (1) P: Direct high throughput experimental PPI data, (2) E: indirect high throughput data (3) S: sequence based data sources and (4) F: functional properties of proteins.

In addition to the features, we also need a reference set to train/test the method. For the positive set (or interacting pairs) ∼3000 yeast protein pairs were extracted from the database of interacting proteins (DIP ). This set is composed of interacting protein pairs which have been experimentally validated and thus can serve as a reliable positive set. Unlike positive interactions, it is rare to find a confirmed report on non interacting yeast pairs. Here we follow [10] which have used a random set of protein pairs as their negative set instead. This selection is justified because of the small fraction of interacting pairs in the total set of potential protein pairs. It is estimated that roughly only 1 in 600 possible pairs actually interact.

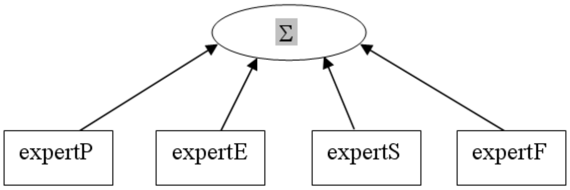

Figure 1: Mixture of Four Feature Experts

## 3 Method

**Mixture of Experts:** We apply the Mixture-of-experts model [2] to the protein-protein interaction prediction task. As Figure1 shows, our framework can be viewed as a single layer tree, with experts at the leaves. Each expert uses one of the dataset groups to predict interactions. A root gate is used to integrate predictions from multiple experts. The weights assigned to each of the experts by the root gate depends on the input set for a given pair.

For every protein pair we construct a $d$ dimensional input vector $X$ consisting of all the features presented in Table1. Given our model the conditional probability of the target variable $Y$ given the input data $X$ could be written as:

$$P(Y|X) = \sum_M P(M|X)P(Y|X, M) \qquad (1)$$

where $Y \in \{-1, 1\}$ is the label (does this pair interacts (1) or not (-1)). $M$ is a four dimensional indicator vector with each item in $\{0, 1\}$, representing the choice of the experts. The sum is over all configuration of $M$. In other words, target class label $Y$ is dependent on the input data $X$ and the choice of expert $M$. The choice of $M$ is also dependent on the input $X$. $P(M|X)$ is modeled using the root gate, while $P(Y|M, X)$ is modeled by each expert. Thus, the target $Y$ is dependent on the input $X$ and the multinomial random variable $M$.

For the $n$th training pair, the conditional probability $P(y^{(n)}|x^{(n)})$ is formulated using equation (1) as:

$$\sum_{i=1}^{4} P(m_i^{(n)} = 1|x^{(n)}, v) * P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \omega_i) \qquad (2)$$

The overall model parameters include the gate parameters $v$ and the experts parameters $\omega_i$.

Each expert uses binary logistic regression for predicting if a pair of proteins interact. For the $i$-th expert we set

$$P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \omega_i) = \frac{1}{1 + exp(-y^{(n)} * (w_i^T * x^{(n)}))} \qquad (3)$$

The root gate can take any functional form that is consistent with the estimation of probabilities. [2] used multinomial logit models for the gates. Here, we extend the binary logistic regression to model multinomial probability through voting. The binary logistic regression model is run once for each output branch of the gate. Next a modified probability is calculated by combining all the models. Assuming we have $i$ branches ($i = 1...4$ for our gate), set:

$$P(c_i^{(n)} = 1) = \frac{1}{1 + exp(-(v_i^T * x^{(n)}))} \tag{4}$$

then

$$P(m_i^{(n)} = 1|x^{(n)}, v) = \frac{P(c_i^{(n)} = 1)}{\sum_{j=1}^{C} P(c_j^{(n)} = 1)} \tag{5}$$

Note that $P(m_i^{(n)} = 1|x^{(n)}, v)$ depends on the input attributes ($x^{(n)}$) and represents the weight for expert $i$ when predicting the $n$th pair. In all of the above logistic regression steps, we use the ridge estimator to regularized parameters.

**Expectation-Maximization:** The overall set of parameters are trained using maximum likelihood estimation. The log-likelihood is,

$$\sum_{n=1}^{N} log(\sum_{i=1}^{4} P(m_i^{(n)} = 1|x^{(n)}, v) * P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \omega_i)) \tag{6}$$

[1] proposed an expectation-maximization (EM) algorithm for adjusting the parameters of the ME. The EM algorithm consists of two steps.

For the E-step, we compute $h_i^{(n)}$, the posterior weight for expert $i$ in predicting pair $n$:

$$h_i^{(n)} = P(m_i^{(n)} = 1|x^{(n)}, y^{(n)}, \Theta^t) = \frac{P(m_i^{(n)} = 1|x^{(n)}, v^t) * P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \omega_i^t)}{\sum_{j=1}^{4} P(m_j^{(n)} = 1|x^{(n)}, v^t) * P(y^{(n)}|x^{(n)}, m_j^{(n)} = 1, \omega_j^t)} \tag{7}$$

The M-step solves the following maximization problems:

$$v^{t+1} = argmax_v \sum_{n=1}^{N} \sum_{j=1}^{4} h_j^{(n)} * log(P(m_i^{(n)} = 1|x^{(n)}, v^t)) \tag{8}$$

and for each expert,

$$\omega_i^{t+1} = argmax_{\omega_i} \sum_{n=1}^{N} h_i^{(n)} * log(P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \omega_i^t)) \tag{9}$$

Therefore, the EM algorithm is summarized as

1. For each data pair $(x^{(n)}, y^{(n)})$, compute the posterior probability $h_i^{(n)}$ using the current values of the parameters.
2. For each expert $i$, solve a maximization problem in Eq.9 with observation $\{x^{(n)}, y^{(n)}\}_{n=1}^{N}$ and observation weights $\{h_i^{(n)}\}_{n=1}^{N}$.
3. For the root gate, solve the maximization problem in Eq.9 with observation $\{x^{(n)}, y^{(n)}\}_{n=1}^{N}$.
4. Iterate by using the updated parameter values.

**Handling missing values:** Biological datasets contain many missing values. The coverage (Table 1) ranges from 3.9% for Yeast Two-Hybrids to over 88.9% for gene expression data sets (in average) and 100% for sequence related features.

Using our input dependent method, missing values can be overcome by adding an additional feature column for each feature that has low feature coverage. This new feature uses 0 to represent missing values and 1 to represent present values. While this increases the size of our feature set, it is still very small (~200) compared to the total number of protein pairs (~18M). Since the weighting for the ME root gate depends on the input features, our classifiers can use the present / absent information when weighting the different features. We have found that this method works better than traditional methods that assign mean or median value to missing data (results not shown).

# 4 Results

**Performance Comparison:** We compared our ME method with five other classifiers that have been suggested in the past for this task: Decision Tree, Logistic Regression, Naiv̈e Bayes, Support Vector Machines and Random Forest. All comparisons were based on the following training and testing procedures. We randomly sampled a training set containing 100,000 yeast protein pairs to learn the decision model. Then we sampled a test set (another 30,000 pairs) from the remaining protein pairs, and used the trained model to evaluate the performance of the classifier. The above steps were repeated 10 times for each classifier and average values are reported.

We use AUC area and R50 partial area under the Receiver Operator Characteristic curve to evaluate performance. The area under the ROC curve (AUC) is commonly used as a summary measure of diagnostic accuracy. It can take values from 0.0 to 1.0. In our prediction task, we are predominantly concerned with the detection performance under conditions where the false positive rate is low. Here, we use 50 as a cut-off, i.e. R50 is a partial AUC score that measures the area under the ROC curve until reaching 50 negative predictions.

Table 2 lists the AUC score and R50 scores of the six methods. As can be seen the ME method achieves the best values for both criteria.

**Comparison of experts:** Next, we investigated the utility of the ME method in helping biologists analyze the interaction predictions with the goal of using them in the design of new experiments. For this purpose, we applied the ME method to a specific pathway, the yeast pheromone response and compared the contribution of different experts in the predictions made. We selected 25 proteins that are known to participate in this pathway and applied the ME algorithm (using a different training set) to classify the 300 (25*24/2) potentially interacting pairs. We determined a prediction threshold us-

Table 2: Average AUC and R50 scores. LR: Logistic regression; NB: Naive Bayes; RF: Random Forest; SVM: Support Vector Machine; ME: mixture of feature experts.

| **Method** | **AUC mean** | AUC std | **R50 mean** | R50 std |
|---|---|---|---|---|
| LR | 0.8823 | 0.033 | 0.2866 | 0.070 |
| NB | 0.9349 | 0.015 | 0.2486 | 0.047 |
| RF | 0.9321 | 0.014 | 0.2688 | 0.048 |
| SVM | 0.9159 | 0.024 | 0.2585 | 0.063 |
| ME | **0.9463** | 0.013 | **0.3080** | 0.078 |

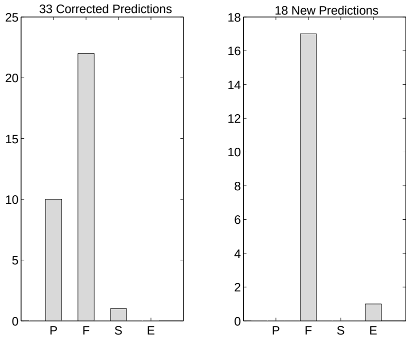

Figure 2: Frequency of each expert having maximum contribution. For definition of P,F,S,E experts, see Table1.

ing the training set. 51 of the pheromone response pairs were above the threshold and were thus predicted to be interacting. Among them, 33 interactions (64.7%) had been experimentally validated. The remaining 18 pairs are new predictions. Figure2(a) shows the frequency at which each of the four experts had maximum contribution among validated pairs. In line with biological intuition, the direct high-throughput evidence (expert P) and functional databases (expert F) are the predominant experts in the correct predictions. Figure2(b) shows that the majority of the 18 new predictions are based on recommendations by expert F. Based on the reliability of expert F in making correct predictions, this result indicates that the majority of the new predictions may turn out to be correct, once experimentally tested.

Interestingly, expert E (indirect experimental data category) is rarely used. This is seemingly in contradiction to previous estimations in which tree based feature ranking methods ranked gene expression features very highly [8]. Note that, when the feature sets are not grouped the wide availability of gene expression data and its high coverage may result in an increased use of this feature, even though it may lead to overfitting. As our results suggest, splitting the data into more homogeneous groups may help increase the prediction accuracy by decreasing its reliance on these high throughput data sources.

## 5 Conclusions and Future Work

In this paper we presented a mixture of experts method for predicting protein-protein interactions in Yeast. Diverse high-throughput biological datasets are split into four homogeneous experts. Each expert uses a subset of the data to predict protein interactions and experts predictions are combined such that the weight of each expert depends on the input data for the predicted protein pair. This method is useful for overcoming missing values which are a major issue when analyzing biological datasets. In addition, the weights can be used by biologists to determine confidence in the prediction for each pair. Using data from yeast we have shown that this algorithm improves upon previous methods suggested for this task.

We believe that as the prediction task becomes harder (for example, when analyzing human protein interactions [11]) the need for methods that can accommodate high levels of missing values and are directly interpretable by biologists increases. The next step will be to apply our method to the human protein interaction prediction task where missing values and the small number of positive examples are major obstacles in acquisition of an accurate protein interaction map.

## References

[1] Michael I. Jordan , Robert A. Jacobs, (1994) Hierarchical mixtures of experts and the EM algorithm, *Neural Computation*, v.6 n.2, 181-214

[2] Waterhouse, S.R. (1997) Classification and regression using mixtures of experts. Ph.D., Thesis, Department of Engineering, Cambridge University.

[3] von Mering C., Krause,R., Snel,B., Cornell,M., Oliver,S., Fields,S., & Bork,P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, **417**, 399-403.

[4] Jansen,R., Yu,H., Dreenbaum,D., Kluger,Y., *et. al*, (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science* **302**, 449-53.

[5] Lee,I., Date,SV., Adai,AT., Marcotte,EM. (2004) A probabilistic functional network of yeast genes. *Science*, **306** (5701):1555-8.

[6] Lin,N., Wu,B., Jansen,R., Gerstein,M., & Zhao,H., (2004) Information assessment on predicting protein-protein interactions, *BMC Bioinformatics* **5**, 154.

[7] Ben-Hur A., Noble W.S., (2005) Kernel methods for predicting protein-protein interactions, *Bioinformatics (Proceedings of the Intelligent Systems for Molecular Biology Conference)* **21** i38-i46

[8] Qi,Y., Klein-Seetharaman,J., & Bar-Joseph,Z., (2005) Random Forest Similarity for Protein-Protein Interaction Prediction from Multiple source, *Pacific Symposium on Biocomputing* **10**, 531-542.

[9] Yamanishi,Y., Vert,J.-P., Kanehisa,M., (2004) Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, **20**, 363-370.

[10] Zhang,L.,Wong,S.,King,OD.,Roth,FP.,(2004) Predicting co-complexed protein pairs using genomic and proteomic data integration, *BMC Bioinformatics*, **5**, 38.

[11] Rual,JF.,Venkatesan,K.,*et.al*,and Marc Vidal, (2005) Towards a proteome-scale map of the human proteinprotein interaction network, *Nature* **437**, 1173-1178