Protein-Protein Interaction Prediction: Feature Enrichment

Suyoun Kim and Lavanya Viswanathan: Madhavi Ganapathiraju's Group

1. Abstract

Protein-protein interactions (PPIs) play a significant role in identifying the function, the structural complex, and biological process of the protein. Since traditional experiments to determine PPIs are often costly in terms of resources, technical and scientific expertise and the time taken to characterize each interaction, **machine learning algorithms** have been applied to PPI prediction. **Computational algorithms treat this task as a binary classification problem.**

However, missing features occur in PPI data because very few proteins have been extensively studied. In the database of human interactome we consider, for the features approximately **15% of the data is missing**. In this work, we investigate the impact of **1) bioinformatic** and **2) computational methods** in dealing with missing features on the prediction of human PPIs. As part of **bioinformatic** enrichment, we present a thorough study on the impact of **transferring Gene Ontology information among orthologs** on prediction human PPIs. The results show that even less than 3% of features imputation can contribute to improve PPI prediction. In the **computational** method section, we examine the use of **linear regression to impute values** and show that it helps improve the PPI prediction.

Require the data to ideally have sufficient features of the protein pairs and *without missing values.*

2. Approach

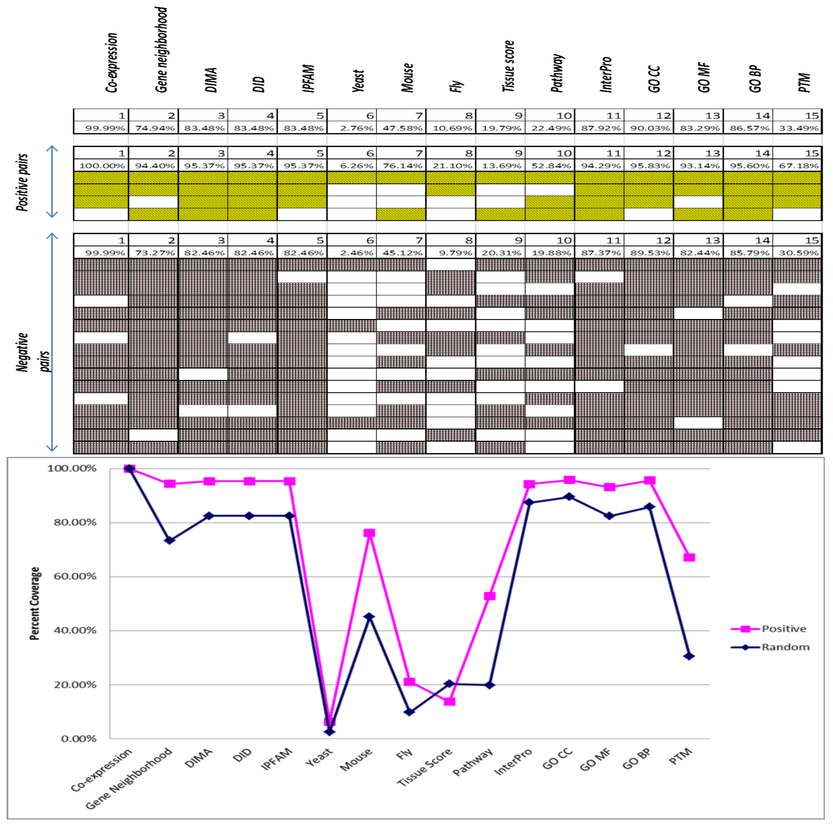
•***Transferring Gene Ontology knowledge among Orthologs.***

Orthologs are genes in different species that have evolved from a common ancestral gene by speciation. Generally, orthologs tend to **retain the same function** in the course of evolution. Therefore, since **orthologous genes have similar characteristic of biological processes, cellular components, and molecular functions,** the information of GO annotations are exchangeable among orthologous groups.

•***Imputation using Linear Regression:***

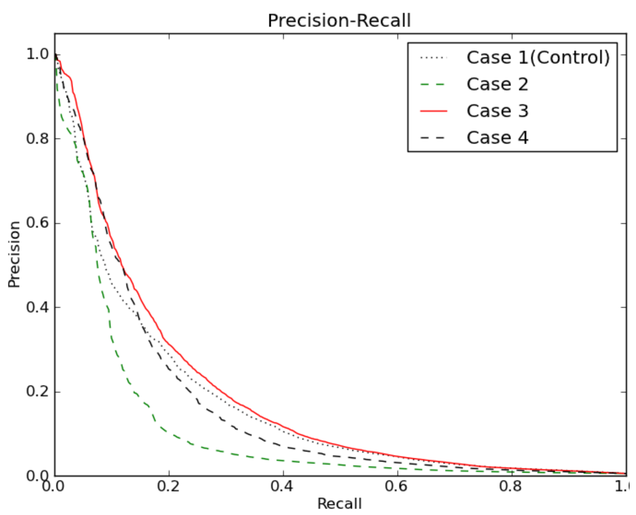
We applied the concept of **Linear Regression** to the data at hand. We estimate weights corresponding to **each column**, using the data in the training set. For this purpose, we **split the training set** itself into training and test sets. Imputation is then done on the test set, using these **estimated weights**. Thus this method uses the features whose values are present to estimate the values of the missing features.

3. Feature Coverage in Data

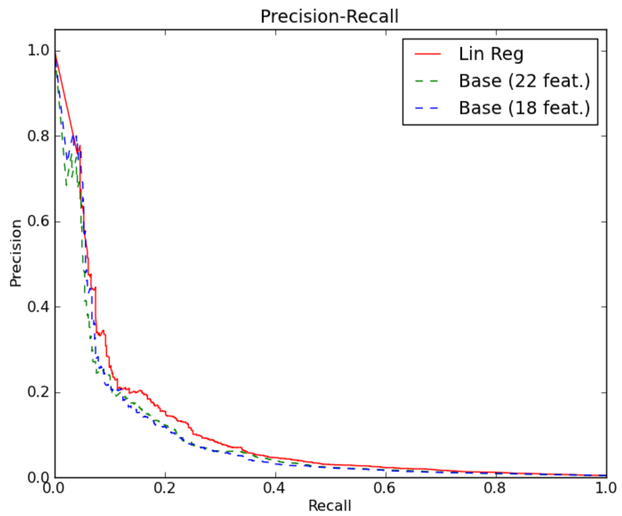


4. Impact of Our Feature Enrichment Algorithms on Human PPI Prediction

Bioinformatic Enrichment



Computational Enrichment



5. Results and Discussion

For our analysis, we use 260,458 protein pairs from the human interactome. From our results, it was clear that **incorporating information from other organisms improves the prediction of human PPIs.**

•**Using linear regression**, had an **AUC of around .109 as compared to the random model that had a value 0.093**. **This can be attributed to the fact that we use features whose values are known in the instances to impute the value of the unknown features which helps in enhancing accuracy.**

http://tonks.dbmi.pitt.edu/w

madhavi@cs.cmu.edu