Supplemental Methods

1. Overview

The earlier version of interaction prediction methods to calculate PAIR V3.3 (Predicted Arabidopsis Interactome Resource) has been reported in (##Lin et al. (2009) Lin et al. (2011) cite##). The user accessibility and data quality have been discussed in main body of the article before. The detailed information of the PAIR including protocols, accuracy evaluation results, and the Internet accessibility can be found in the PAIR website (http://www.cls.zju.edu.cn/pair/).

The last main version of PAIR (PAIR V3.0) was calculated based on 6 types of indirect evidence, which has been reduced to 5 in V4.0 because of the newly applied qualification method (##cite## ROC). Part of the algorithm of feature extraction from indirect evidence has been changed (##table.1##). Meanwhile, the base scale of the golden data set supporting direct evidence is significantly increased form 4500 to 5785 with higher reliability (23000 by the last version filter). With several thoughtful filters added to the pipeline, the final direct evidence data set contains 5800 highly trustable physical protein interactions reported by at least 2 bibliographies with different low-throughout experimental methods.

With well-designed pipeline, newly updated data sources (updated in the end of 2014), and a more rigid quality control protocol, xxxx interactions involving xxxx proteins were predicted by PAIR V4 prediction model. The coverage of newly predicted protein interactions is expected to be xxxx% of the entire Arabidopsis interactome, with 43% reliability. The predicted interactions overlap xxxx (xxxx%) experimental physical protein interactions collected form Intact, BioGRID BIND, TAIR, as of December 21rd, 2014.

(\*Web application coming soon)

2. PAIR V4 Prediction Method

2.1. Introduction

The core data mining method of PAIR V4 is SVM (Support Vector Machine). To train an SVM prediction mode, a training data set is required assembling by pairs of proteins which are marked as positive or negative according to direct evidence and their feature vectors extracted from indirect evidence.

2.2. Direct Evidence

2.2.1. Introduction

Direct evidence shall be collected to mark a protein pair as positive, which means we have strong evidences to believe that this protein pair is truly ‘Functional Interacting’. However, it is hard to define and describe ‘Functional Interacting’ based on any biological knowledge base we see nowadays. A resulting compromise is to infer ‘Functional Interacting’ protein pair by ‘Physical Interacting’ protein pair whose supporting data sets can be widely obtained from the Internet (##table1##). Consequently, several filters are needed to qualify physical interacting protein pairs that can be seen in chapter

2.2.2. Data Collection

We retrieved and integrated experimentally reported physical protein interaction data from Intact (#cite#), BioGRID (#cite#), BIND (#cite#) and TAIR (#cite#). (Table1) The original data is curated by their evidence, which provided by the database to determine whether it is an experimentally reported physical protein interactions instead of any predicted one. (Table2)

|  |  |  |  |
| --- | --- | --- | --- |
| Number of Interactions and Proteins of Original Data | | | |
| Database | Number of interactions | Number of Proteins | Date |
| BIND | 1224 | 673 | 14-Jan-14 |
| BioGRID | 25340 | 7804 | 1-Dec-14 |
| TAIR | 2656 | 1333 | 18-Nov-10 |
| IntAct | 17818 | 5657 | 1-Dec-14 |

Table1

2.2.3. Data Filter (Why and How, the mysterious Train-Test Score Drop)

As discussed before, in order to mark positive examples of ‘Functional Interacting’ proteins by experimentally reported physical protein interactions, we have to filter the original interaction data by some standards including:

a) The number of bibliographies reporting this interaction with different experiment methods. (If a bibliography reported an interaction in N different experiment methods, the number counts as N)

b) The throughput of the experiment method.

Defining the low-throughput experiment threshold as 30, with these two filters, we assemble the final dataset with two parts:

Part. 1) Interactions reported by at least 2 bibliographies in different experiment methods, meanwhile these methods are High-throughput. (Number: 1137)

Part. 2) All interactions reported by Low-throughput experiment methods (Number: 4648)

|  |  |  |
| --- | --- | --- |
| Number of Interactions and Proteins of Curated Data | | |
| Database | Number of interactions | Number of Proteins |
| BIND | 782 | 404 |
| BioGRID | 23057 | 6844 |
| TAIR | 2604 | 1293 |
| IntAct | 5208 | 1551 |
| Total | 26013 | 9047 |
| Total Filtered | 5785 | 3071 |

Table2

2.2.4. Negative Examples Generation

Not as that our filtered experimentally reported physical protein interactions as positive examples, there is barely any research, which focuses on the investigation and detecting of non-interacting proteins. The resulting situation leads to the demand of the simulation or prediction of non-interacting proteins required as negative examples. Several studies have been conducted to choose certain pairs of proteins to represent the non-interacting proteins such as the proteins whose most common sub-cellular locations are different.

Unfortunately, this approach seems not to help improve the model, moreover, there are studies indicating a potential bias of the estimation of model accuracy caused by the unexpected reduced difficulty when the prediction were carried out based on this simulated negative example. (##cite Ben-Hur and Noble, 2006##)  
 Therefore, to obtain a negative example dataset as unbiased as possible, we choose to follow the approach demonstrated in (##cite Zhang el al., 2004##) by selecting negative examples as random protein pairs without overlap between known positive example

2.2.5. Protein Interactions Distribution Analysis by GO-Enrichment