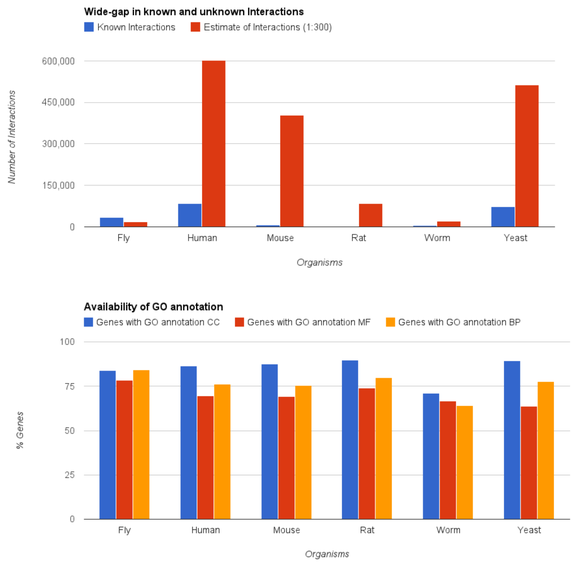
Transfer Learning for protein-protein interaction

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA USA

Introduction

Since experimental methods for discovering/validating protein-protein interactions
(PPIs), including high-throughput techniques, are highly resource intensive, **machine learning algorithms** have been suggested to accelerate discovery. By training models with known interactions and features of proteomes, the most-promising PPIs can be predicted. To predict PPIs more accurately, sufficient training data is required. However, except for only a few model organisms like human and yeast, **interactomes are unknown**, and even the model organisms still



To solve insufficient training dataset (known PPIs and GO annotation)

➔

Transferring knowledge from other well-known organisms using *computational transfer* and *bioinformatic transfer*

Methods

Here, we explore the feasibility of discovering PPIs of one organisms using two different knowledge transfer methods between organisms to improve the performance of learning; **transferring a model** and **transferring features.**

Approach 1 : Transferring a model developed a random forest classifier with known interactions in multiple species (i.e. fly, mouse, rat, worm, and yeast) to predict interactions in other species (i.e. human and mouse) while keeping the feature representation uniform across different domains.

1) Trained a model with multiple species (i.e. fly, mouse, human, rat, worm, and yeast)
2) Test mouse or human interactome

Approach 2 : Transferring features in the absence of feature information, we transferred missing Gene Ontology feature values from orthologues into the human data. For obtaining GO features from orthologous genes, we follow the below flow:

1) For a human interactome to be predicted, first we check whether it has annotations present in GO.
2) If GO annotations are available, features are computed for protein pairs and a random forest classifier is built.
3) If not available, for each protein which has a missing GO annotation, find the GO annotations

Features are generic and not organism specific

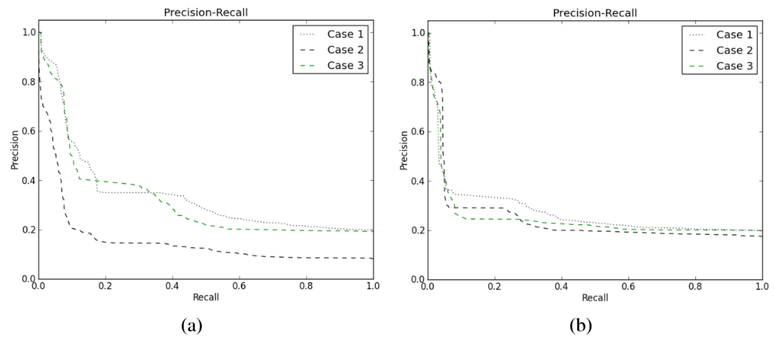
➔

Transferring a model and GO features

Results

Transferring a model

The left figure shows the results of mouse interactome prediction and the right figure shows the results of human interactome prediction. Case 2 and Case 3 perform similarly to Case 1 in human, and even at 70% precision, Case 2 outperformed Case 1.



Mouse Interactome Prediction

Training Set
1. Case I: Mouse interactions as training data:
2. Case II: Mouse + Fly, Human, Rat, Worm and Yeast as training data:
3. Case III: Fly, Human, Rat, Worm and Yeast as training data:

Test Set
Mouse

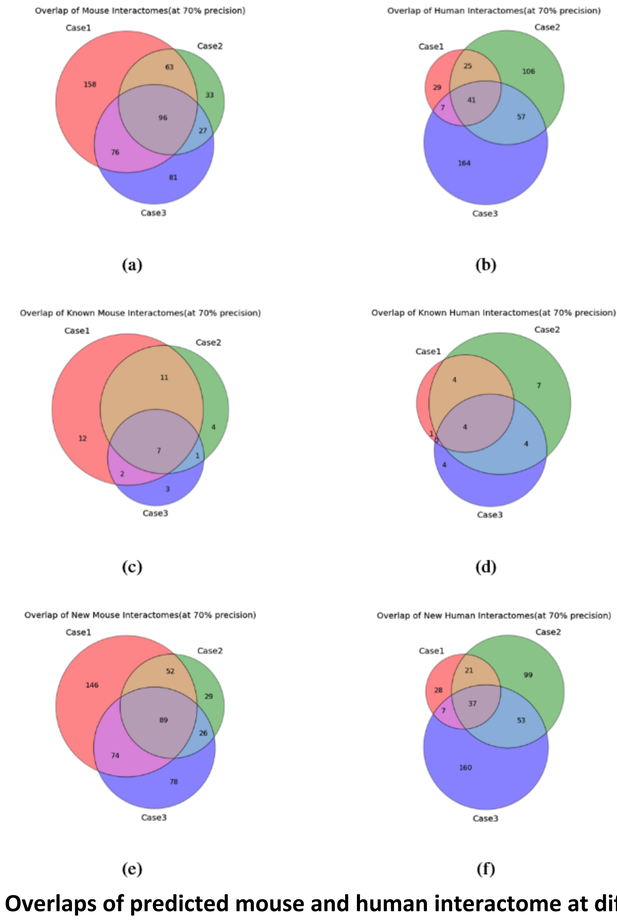
Human Interactome Prediction

Training Set
1. Case I: Human interactions as training data:
2. Case II: Human + Fly, Mouse, Rat, Worm and Yeast as training data:
3. Case III: Fly, Human, Rat, Worm and Yeast as training data:

Test Set
Human

(1:4)

	Train Positive	Train Negative	Test Positive	Test Negative
Ratio	1	4	1	4
Models	5,000	20,000	833	3,332

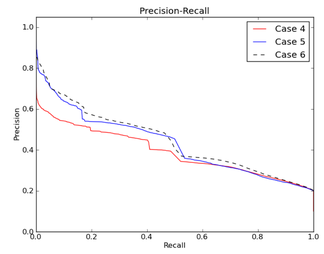


Overlaps of predicted mouse and human interactome at different cases (at 70% precision). Figure 6(a), Figure 6(c), Figure 6(e) show the overlap of the predicted mouse interactome, and Figure 6(b), Figure 6(d), Figure 6(f) show the overlap of the predicted human interactome. Compared to mouse, less overlaps are observed in human interactome. with Case 2 and Case 3

Transferring features

The Case 6 (missing GO features are filled with transferred GO features from five organisms, fly, mouse, rat, worm, yeast) performed better than Case 4 (original dataset) and Case 5 (missing GO features are filled with transferred GO features from the only one organisms)

The number of missing GO features at each test cases, and the percentage of GO feature coverage. After applying the transferring features method, the feature coverage increased by 25%. However, there are still missing features for two main reasons: some of genes do not have any orthologues , and/or some of orthologues do not have GO annotations.



Case 4	Original human dataset
Case 5	Missing GO annotations are filled with transferred GO annotations from orthologousgenes of one organism
Case 6	Missing GO annotations are filled with transferred GO annotations from orthologuous genes of five organisms: fly, mouse, rat, worm, and yeast

	Case4		Case 5& 6	
GO categories	# missing	% coverage	# missing	% coverage
CC	10,660	78.68%	9,552	80.90%
MF	18,101	63.80%	15,784	68.43%
BP	16,663	66.67%	15,536	68.93%

Features

We use GO annotations and protein domain information in building features of protein-pairs, and are described in detail below:
Domain Domain Interaction based features (4 features) from three
different databases, namely 3DID, IPFAM , DIMA, and InterPRO
Gene Ontology (3 features) Cellular Component, Biological Process, Molecular Function:Co-occurrence of each pair of GO terms that occur in known interactions and divided by the product of the individual frequency of each GO term amongst all proteins in training data

Our methods are feasible to discover new PPIs and improve precision

Conclusions

Transferring a model

Our results show that at 80 % precision, our model uncovers **104 mouse interactions** and **150 human interactions** in 249,167 protein-pairs. The results reveal that transferred GO features contribute to improved performance of PPI prediction.

Transferring a features

From our results, it was clear that incorporating information from other organisms helps in giving better results and new interactions in both mouse and humans which would have been missed otherwise. These findings should help to predict PPIs in other organisms that have insufficient training data.

References

Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." Knowledge and Data Engineering, IEEE Transactions on 22.10 (2010): 1345-1359.

Ananthasubramanian, Seshan, et al. "Mycobacterium tuberculosis and Clostridium difficille interactomes: demonstration of rapid development of computational system for bacterial interactome prediction." *Microbial informatics and experimentation* 2.1 (2012): 1-12.