Transfer learning methods for the discovery of host-pathogen protein-protein interactions
Meghana Kshirsagar*, Judith Klein-Seetharaman, Jaime Carbonell
Language Technologies Institute, Carnegie Mellon University

Propose techniques for discovering protein-protein interactions (PPIs) in new hosts or pathogens using interactions in known hosts/pathogens

We use known Salmonella-Human PPIs to predict interactions between Salmonella-Arabidopsis proteins

Two approaches are used and predictions from each are combined:
1. Infer interactions using orthology of the host proteins, filter using intra-host PPI network alignment
2. Apply Transductive Support Vector Machines to label interactions in new host using known interactions as labeled data

Challenges:
(a) No labeled data available for Arabidopsis, and a very small labeled dataset available for Human
(b) The two hosts namely Human and Plant have very different features. Building a joint model is tricky!
(c) Difficulty in evaluation of predicted interactions

Supervised learning models (Background)

Transfer learning methods

Approach - 1

(A) Using gold standard PPIs from source host (Human)

Gold standard
Salmonella-Human PPI
(82 interacting gene pairs)

Orthologous
Salmonella-Arabidopsis PPI
(25 interacting gene pairs)

(B) Using predicted interactions from source host (Human)

Transfer using “network alignment” relationship between the host PPI networks

Predictions from
Salmonella-Human PPI model
(115213 interacting gene pairs)

Salmonella-Arabidopsis PPI
where Arabidopsis protein is part of an enriched complex
(23684 interacting gene pairs)

Enriched protein complexes computed using NetworkBlast\textsuperscript{2} algorithm which aligns intra-host PPI network of source with target

Arabidopsis protein complex
Human protein complex

NetworkBlast found 2329 enriched complexes

Approach - 2

(C) Uses unlabeled PPIs from target host (Arabidopsis) in addition to gold standard PPIs from source host (Human)

Transfer using “similarity” between the source and target host proteins

Measure of similarity: defined using host protein properties like Gene-Ontology, Gene expression

\[
\begin{align*}
\text{sim}(x_i, y_j) &= \text{sim}(x_i, x_j) + \text{sim}(y_i, y_j) \\
\text{dot}(x_i, x_j) &= \text{dot}(x_i, y_j)
\end{align*}
\]

\( x_i, y_j = \text{Salmonella protein in source/target} \)

\( x_i, x_j = \text{human/arabidopsis proteins in source/target resp.} \)

Performace on source data using 3-fold cross-validation

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>82.63</td>
<td>58.33</td>
<td>68.14</td>
</tr>
</tbody>
</table>

Experiments and Results

Source data [Salmonella-Human PPIs]
62 positives, 6200 negatives

Target data (Unlabeled Salmonella-Arabidopsis)
150584 interactions

Training: 3 fold CV using all source data + 2000 target examples

Best model applied to remaining unlabeled target examples to get predictions on target

Obtained 1087 interacting gene-pairs

GO enrichment analysis

Applied FuncAssociate\textsuperscript{2} for GO term enrichment analysis on predictions from both approaches. Some top terms are

<table>
<thead>
<tr>
<th>Mass-spectrometry:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binding studies on some predicted plant partners of Salmonella protein spvC show positive results.</td>
</tr>
</tbody>
</table>

Conclusion

Our approaches to build a cross-species model shows very promising results. The techniques can be applied for any new host or pathogen.

Disadvantages:

TSVM solving combinatorial optimization using an approximation, no guarantees on optimality

Future Work:
2. Other ways to transfer knowledge between organisms