10-810 /02-710

Computational Genomics

Protein interactions
Protein Interactions
Domains of a Protein

• While predicting the structure from the sequence is still an open problem, we can identify several domains within the protein.
• Domains are compactly folded structures.
• In many cases these domains are associated with specific biological function.
• Domains are also associated with interactions.
Assigning Function to Proteins

• While almost 30000 genes have been identified in the human genome, relatively few have known functional annotation.

• Determining the function of the protein can be done in several ways.
  - Sequence similarity to other (known) proteins
  - Using domain information
  - Using three dimensional structure
  - Based on high throughput experiments (when does it functions and who it interacts with)
Protein Interaction

In order to fulfill their function, proteins interact with other proteins in a number of ways including:

• Pathways, for example A -> B -> C

• Post translational modifications

• Forming protein complexes
Protein interaction

- Traditionally protein interactions were studied individually
- Many new proteins from complete genome sequences
- New methods for genome wide interaction data
- Comparatively evaluate accuracy, biases, overlaps and complementarities
PPI Lab Experiments

• **Small-scale** PPI experiments
  • One protein or several proteins at a time
  • Small amount of available data
  • Expensive and slow lab process

• **High-throughput** PPI experiments
  • Hundreds / thousands of proteins at a time
  • Highly noisy and incomplete data
  • Surprisingly little overlap among different sets
Methods

- Yeast two-hybrid screens
- Protein complex purification techniques using mass spectrometry
- Correlated messenger RNA expression profiles
- Genetic interaction data
- 'in silico' interactions
- Focus on the yeast proteome
Yeast two-hybrid assay

- Pairs of proteins to be tested for interaction are expressed as fusion proteins ('hybrids') in yeast:
- One protein is fused to a DNA-binding domain, the other to a transcriptional activator domain.
- Any interaction between them is detected by the formation of a functional transcription factor.
Yeast 2 Hybrid Technique

- **a plasmid**
  - target DNA-binding domain
  - fusion
  - precocious promoter
  - promoter
  - lac-Z gene
  - β-galactosidase

- **α plasmid**
  - bait activation domain

- Bait protein
  - LexA operators
  - Reporter genes (LacZ and HIS3)

- Prey Activation domain
  - LexA operators
  - Reporter genes (LacZ and HIS3)

- Interaction of bait and target results in transcription ON
Mass spectrometry of purified complexes

- Individual proteins are tagged and used as 'hooks' to biochemically purify whole protein complexes. These are then separated and their components identified by mass spectrometry.
mRNA Expression
Genetic interactions (synthetic lethality).

- Two nonessential genes that cause lethality when mutated at the same time form a synthetic **lethal interaction**.
- Such genes are often functionally associated and their encoded proteins may also interact physically.
In silico predictions through genome analysis.

- Whole genomes can be screened for three types of interaction evidence:
- In prokaryotic genomes, interacting proteins are often encoded by conserved operons
- Interacting proteins have a tendency to be either present or absent together from fully sequenced genomes, that is, to have a similar 'phylogenetic profile';
- Proteins are sometimes found fused into one polypeptide chain. This is an indication for a physical interaction.
functional associations between proteins:

high-throughput data in yeast (2002)

<table>
<thead>
<tr>
<th>Method</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>yeast two-hybrid</td>
<td>5125</td>
</tr>
<tr>
<td>complex-purification*</td>
<td>18027 (TAP)</td>
</tr>
<tr>
<td>(analysis by mass spectrometry)</td>
<td></td>
</tr>
<tr>
<td>33014 (HMS-PCI)</td>
<td></td>
</tr>
<tr>
<td>mRNA synexpression</td>
<td>~ 15000</td>
</tr>
<tr>
<td>(cell-cycle + Rosetta data)</td>
<td></td>
</tr>
<tr>
<td>in silico predictions</td>
<td>~ 7000</td>
</tr>
<tr>
<td>(neighborhood, fusion, co-occurrence)</td>
<td></td>
</tr>
<tr>
<td>genetic interactions</td>
<td>886</td>
</tr>
<tr>
<td>(synthetic lethality, Tong et.al. + MIPS)</td>
<td></td>
</tr>
</tbody>
</table>

* Two new large scale datasets published in 2006
Coverage and overlap

- In total ~80,000 interactions, but only 2400 in the overlap; why?
- the methods have not have reached saturation
- methods produce a significant fraction of false positives
- methods have strength/weaknesses for certain types of interactions/proteins → complementarities between the methods
distribution of interacting proteins (e.g. TAP complexes)

- energy production
- aminoacid metabolism
- other metabolism
- translation
- transcription
- transcriptional control
- protein fate
- cellular organization
- transport and sensing
- stress and defense
- genome maintenance
- cellular fate/organization
- uncharacterized

interaction density

0
10

(actual interactions per 1000 possible pairs)
reference interactions

manually annotated protein complexes: MIPS / YPD

high-throughput interaction data: OVERLAP OF 2+ METHODS

10907 interactions

2455 interactions
protein interaction datasets

- **purified complexes (TAP):** 18027 interactions
- **purified complexes (HMS-PCI):** 33014 interactions
- **genomic associations:** 7446 interactions
- **mRNA synexpression:** 16496 interactions
- **yeast two-hybrid:** 5125 interactions
- **synthetic lethals:** 886 interactions
Benchmarking

• Comparing the data with a reference set of trusted interactions allows the estimation of **lower limits** for accuracy and coverage.
• The highest accuracy is achieved for interactions supported by more than one method.
Benchmarking high-throughput interaction data

- Coverage: fraction of reference set covered by data (log %)
- Accuracy: fraction of data confirmed by reference set (log %)

- Purified complexes
- TAP
- mRNA synexpression
- HMS-PCI
- Genomic associations
- Synthetic lethality
- Yeast two-hybrid
- Combined evidence
- Two methods
- Three methods

Legend:
- ▲ raw data
- ■ filtered data
- ◆ parameter choices
Biases in coverage

• Most protein interaction data (including the curetted complexes) are biased towards proteins of high abundance.
• The two “genetic” approaches (two-hybrid and synthetic lethality) appear relatively unbiased.
• Data sets are biased towards particular cellular localizations. For example mitochondrial proteins in the case of the *in silico* predictions. (such protein are of bacterial descent)
Protein interaction as a classification problem

- Given these direct and indirect datasets, we can design a classifier which will take as an input high throughput data for a pair of proteins.
Challenges

- Feature are heterogeneous
- Most features are noisy
- Most features have missing values
- Highly skewed class distribution
  - Much more non-interacting pairs than interacting pairs
  - No negative (not interacting) set available
- Only a small positive (interacting) set available

<table>
<thead>
<tr>
<th>Species</th>
<th>Database (Small-scale PPI)</th>
<th>Genome Size</th>
<th>Predicted # of Interactions</th>
<th>Estimated Ave. Num. Partners Per Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>DIP (3867 interactions ; 1773 proteins)</td>
<td>~6300</td>
<td>~30,000</td>
<td>~10</td>
</tr>
<tr>
<td>Human</td>
<td>HPRD (14608 interactions; 5712 proteins)</td>
<td>~30,000</td>
<td>~90,000</td>
<td>~6</td>
</tr>
</tbody>
</table>
Method

1. Determine a distance measure for the pair-wise difference between protein pairs
   - This method will overcome missing values and address the relationships between feature
2. Use a classifier (kNN) to calculate a confidence score for each pair
   - This method can take into account the skewed distribution
3. Rank the pairs by the above score and use training set to select a cutoff
# Features List

<table>
<thead>
<tr>
<th>Source Index</th>
<th>Feature Category of Yeast</th>
<th>Coverage in Yeast (Percentage)</th>
<th>Feature Category of Human</th>
<th>Coverage in Human (Percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HMS_PCI Mass</td>
<td>8.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>TAP Mass</td>
<td>8.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Yeast-2-Hybrid</td>
<td>3.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>GO Molecular Function</td>
<td>80.7</td>
<td>GO Function</td>
<td>39.1</td>
</tr>
<tr>
<td>5</td>
<td>GO Biological Process</td>
<td>76.1</td>
<td>GO Process</td>
<td>37.6</td>
</tr>
<tr>
<td>6</td>
<td>GO Component</td>
<td>81.5</td>
<td>GO Component</td>
<td>36.3</td>
</tr>
<tr>
<td>7</td>
<td>Essentiality</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>MIPS Protein Class</td>
<td>4.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MIPS Mutant Phenotype</td>
<td>9.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Gene Neighborhood / Gene Fusion / Gene Co-occur</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Sequence Similarity</td>
<td>100</td>
<td>BlastP E-value</td>
<td>100</td>
</tr>
<tr>
<td>12</td>
<td>Homology based PPI</td>
<td>100</td>
<td>Homology PPI Yeast</td>
<td>100</td>
</tr>
<tr>
<td>13</td>
<td>Domain-Domain Interaction</td>
<td>100</td>
<td>Domain Interaction</td>
<td>37.7</td>
</tr>
<tr>
<td>14</td>
<td>Gene Expression</td>
<td>88.9</td>
<td>Gene Expression</td>
<td>34.0</td>
</tr>
<tr>
<td>15</td>
<td>Protein Expression</td>
<td>42.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Protein-DNA TF group binding</td>
<td>98.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Synthetic Lethal</td>
<td>7.6</td>
<td>Co-Tissue</td>
<td>57.1</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Approach

- Random Forest Classifier
  - A collection of independent decision trees (ensemble classifier)
  - Each tree is grown on a bootstrap sample of the training set
  - Within each tree’s training, for each node, the split is chosen from a bootstrap sample of the attributes

- Robust to noisy feature
- Can handle different types of features
- Able to estimate missing values
Performance Comparison (Yeast)

- **AUC R100**: the area under the ROC curve until reaching 100 negative predictions
- **Precision**: the fraction of interacting pairs predicted by the classifier that are truly interacting (“true positives”).
- **Recall**: the fraction of the known positive pairs of interacting proteins have been identified by the learning model (true positive rate)
Performance Comparison (Human – Receptor Related)
**EGFR Predicted Partners List**

<table>
<thead>
<tr>
<th>Index</th>
<th>geneID</th>
<th>geneSymbol</th>
<th>RFpScore</th>
<th>hprdLabel</th>
<th>GeneticDisorder</th>
<th>2Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5159</td>
<td>PDGFRB</td>
<td>4.9346204</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5744</td>
<td>PRR2</td>
<td>4.4444236</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>760</td>
<td>DDR1</td>
<td>4.4444236</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2441</td>
<td>FER</td>
<td>4.4444236</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2654</td>
<td>ERBB2</td>
<td>4.4444236</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5706</td>
<td>PIPR1</td>
<td>4.1094236</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>5707</td>
<td>PIPR2</td>
<td>4.0281508</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>674</td>
<td>SRC</td>
<td>3.8257427</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>5775</td>
<td>PIPRN1</td>
<td>3.8257427</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2863</td>
<td>GRB10</td>
<td>3.9656405</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>10186</td>
<td>TNK2</td>
<td>3.9656405</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2963</td>
<td>PRKCI</td>
<td>3.7484237</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2744</td>
<td>EPHB3</td>
<td>3.7484237</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>2568</td>
<td>EPHB4</td>
<td>3.3078974</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>1866</td>
<td>DNM2</td>
<td>3.1000216</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>7042</td>
<td>TGFEB2</td>
<td>3.0262983</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>5596</td>
<td>PRKCA</td>
<td>3.0262983</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>7417</td>
<td>VAV2</td>
<td>3.0945073</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>6654</td>
<td>SOS1</td>
<td>3.0945073</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>681</td>
<td>BTC</td>
<td>3.0945073</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>3791</td>
<td>KDR</td>
<td>2.9707281</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>3078</td>
<td>IGAS</td>
<td>2.9707281</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>2821</td>
<td>DDR1</td>
<td>2.9707281</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>2821</td>
<td>FGFR1</td>
<td>2.9707281</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>6650</td>
<td>SYK</td>
<td>2.9707281</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>2586</td>
<td>HCK</td>
<td>2.9707281</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>374</td>
<td>AREG</td>
<td>2.7002083</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>8603</td>
<td>IRF5</td>
<td>2.7002083</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>4860</td>
<td>ROR2</td>
<td>2.7002083</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>9807</td>
<td>CD44</td>
<td>2.7002083</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **EGFR has 91 validated partners in HPRD**
- **72 are in the top 200 computationally predicted partners for EGFR**
Interaction of EGFR with Hck

**Hck Domain structure**

**Functions:**

Binds and regulates Nef during HIV infection

Function in signal transduction, but not well defined

**Scheme of pull-down assay experiment**
Interaction of EGFR with Dynamin-2

Dynamin-2 Domain structure

**Function:** Receptor internalization

Scheme of pull-down assay experiment

Anti-GFP blot of Dyn-C+Juxta EGFR and untransfected Cos-1 cells negative control
Mixture of Feature Experts Method

- Make protein interaction prediction by
  - **Weighted voting** from the four roughly homogeneous feature categories
  - Treat each feature category as a prediction expert
  - The **weights are also dependent** on the input example
Mixture of Four Feature Experts: Details

- Parameters \((w_i, \nu)\) are trained using EM
- Experts and root gate use logistic regression (ridge estimator)

\[
p(y^{(n)} | x^{(n)}) = \sum_{i=1}^{4} p(m_i^{(n)} = 1 | x^{(n)}, \nu) \times p(y^{(n)} | x^{(n)}, m_i^{(n)} = 1, w_i)
\]
Mixture of Feature Experts Method

• Selecting features:
  – Useful for experimental biologists to know which features contributed to specific predictions
  – Researchers may have various opinions regarding the reliability of diverse features
Predicting interactions for the yeast pheromone pathway

- 25 proteins involved in this pathway
- Test all possible 300 protein pairs
- 51 predicted interactions
  - 33 known
  - 18 new

The frequency at which each of the four experts has maximum contribution among validated and predicted pairs

- 33 Correct Predictions
- 18 New Predictions

Proteins involved:
- Ste2
- Cdc42
- Ste3
- Ste4
- Ste18
- Ste11
- Ste7
- Fus3
- Ste12
- Macl
- Dig1.2
- Polarization
- Cell cycle arrest
- Mating
What you should know

• New methods are enabling researchers to look at protein-protein interactions on a much larger scales

• These methods are highly inaccurate containing many false positives and false negatives

• Computational tools (classification algorithms) can be used to combine direct and indirect information to improve the accuracy