Combining biological networks to predict genetic interactions

Sharyl L. Wong*, Lan V. Zhang*, Amy H. Y. Tong†, Zhijian Li†, Debra S. Goldberg*, Oliver D. King*, Guillaume Lesage‡, Marc Vidal§, Brenda Andrews‡, Howard Bussey‡, Charles Boone‡, and Frederick P. Roth*‡

*Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, MA 02115; †Banting and Best Department of Medical Research and Department of Medical Genetics and Microbiology, University of Toronto, Toronto, ON, Canada M5G 1L6; §Department of Biology, McGill University, Montreal, QC, Canada H3A 1B1; and ‡Dana–Farber Cancer Institute and Department of Genetics, Harvard Medical School, Smith B58, 1 Jimmy Fund Way, Boston, MA 02115
Motivation and overview.

- Tong, et al. (Science, 2001) identified ~3800 Synthetic Sick and Lethal (SSL) interactions in yeast.
- This represents a small fraction of the total possible SSLs.
- Classic and genomic methods very time consuming and labor intensive.
- Wong et al. (this paper) report on a method for predicting SSL interactions.
Synthetic Sick and Lethal (SSL) Interactions - Common Pathway

Gene 1

Gene 2

Gene 3

Gene 4

...
Synthetic Sick and Lethal (SSL) Interactions - Complementary

Gene 1

Gene 3

Gene 4

Gene 5

Gene 6
Why do we care about SSL interactions?

- Describe Gene-Gene Interaction networks (distinct from PPI networks).
- Reveal genetic redundancy, generally used to protect critical cell functions.
- Assign functions to proteins in redundant systems.
- Identify multi-genic characteristics of human diseases.
- Identify potential combination-therapy drug targets.
How are SSLs identified? - Gene replacement.

- Integrate plasmid with mutant gene into chromosome.
- Select for strains that have the URA3 marker.
- Sub-select for strains that lack the URA3 marker (FOA).
- Mate two mutant strains to make double-mutant.
How are SSLs identified? - Genomics style.

Synthetic Genetic Array (SGA; a) and Synthetic Lethality Analyzed by Microarray (SLAM; b).

SGA analysis

- SGA first introduced by Tong, et al. (Science 2001).
- Used 130 query strains to identify ~3800 SSLs.
- They estimated 80k SSLs in yeast (~5800 genes total).
What can SGA tell us?

Fig. 3. Genetic interaction network representing the synthetic lethal/sick interactions determined by SGA analysis. Genes are represented as nodes, and interactions are represented as edges that connect the nodes; 291 interactions and 204 genes are shown. All of the interactions were confirmed by tetrad analysis, with 8 to 14 tetrads examined in each case. The genes are colored according to their YPD cellular roles (18). For genes assigned multiple cellular roles, we chose one that we considered the most probable on the basis of a review of published abstracts for studies concerning the gene. Synthetic lethal relationships can also be represented by two-dimensional hierarchical clustering, as is used for analysis of DNA microarray experiments (16).

What can we say about SSL networks?

- SSLs in yeast form a small-world, scale-free network.
  - 'hub' genes with increased importance in fitness;
  - dense local networks with short path lengths, connected by longer paths.

- SSL connectivity follows a power-law distribution.
  - many genes with few interactions;
  - few genes with many interactions.
So what do we know about SSLs now?

- SSLs are enriched between genes that give rise to the same mutant phenotype.
- SSLs are enriched between genes with the same GO function.
- SSLs are enriched between paralogs.
- SSLs are enriched between genes that code for proteins in the same complex.
Where does the Wong study fit in?

- SSLs are important, crucial interactions.
- Classical genetics (identify one at a time) is too slow.
- Arrays and genomic techniques help, but not much.
  - Only 5% of total potential SSLs identified so far.
  - Single organism, single condition.

Wong et al.: Can we predict SSL interactions in yeast?
Wong's group wants to predict SSLs based on available yeast data.

- Compile known data of different types from multiple sources.
- Use this integrated dataset to predict SSLs based on particular characteristics.
26 categories.
123 characteristics.
18 sources.
2hop relationships.

<table>
<thead>
<tr>
<th>Major category</th>
<th>No. of characteristics</th>
<th>Appears in trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common upstream regulator</td>
<td>3</td>
<td>38</td>
</tr>
<tr>
<td>Gene cooccurrence</td>
<td>1</td>
<td>18-20</td>
</tr>
<tr>
<td>Chromosomal distance</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Gene fusion</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Conserved gene neighborhood</td>
<td>1</td>
<td>16-18</td>
</tr>
<tr>
<td>Physical interaction</td>
<td>15</td>
<td>39-42</td>
</tr>
<tr>
<td>mRNA coexpression</td>
<td>17</td>
<td>43, 44</td>
</tr>
<tr>
<td>Same predicted physical complex</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>Same MIPS function</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>Same MIPs protein class</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>Same subcellular localization</td>
<td>42</td>
<td>15</td>
</tr>
<tr>
<td>Same phenotype</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>Sequence homology</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Mutual clustering coefficient in physical interaction</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>Posterior probability of physical interaction</td>
<td>4</td>
<td>21</td>
</tr>
</tbody>
</table>

**H:** Sequence homology.

**P:** Physical interaction.

**R:** Common regulator.

**X:** Correlated expression.

**S:** SSL interaction.

- 2hop relationships will play an important role in SSLs.

Summary of methods.

- What can all of the characteristics of a gene pair tell us about the likelihood of an SSL interaction in that pair?

- Used Conditional Information Gain to rank characteristics.

- Used Probabilistic Decision Trees to identify the characteristics most likely to indicate an SSL.

- Assigned a score to each leaf in the tree, corresponding to the probability that a gene pair on that leaf will have an SSL.
Training the Decision Tree.

Start with all T gene pairs in the root node.

T (training set) = all gene pairs analyzed by SGA in Tong et al. (692, 865 gene pairs, including 3868 SSLs.)
• Find the characteristic that yields the highest *Conditional Information Gain* wrt SSL.

• Branch the tree on that characteristic.

• Recurse until a minimum Gain is reached, then stop. This forms a leaf.

• Score = the fraction of gene pairs in the leaf with SSL.

• Define a threshold score, above which considered an SSL.
What is Conditional Information Gain?

The amount of information we learn about A if we first know B.

\[
\text{Cond. Info. Gain} = H_N(X) - \sum_{a=0,1} \frac{|N_a|}{|N|} H_{N_a}(X)
\]

Entropy of the parent Node
Entropy of the child Node
Relative size of the child Node

\[
\text{Entropy (H)} = -p_N \log(p_N) - (1 - p_N) \log(1 - p_N)
\]

Probability that a gene pair is SSL
How well does this work? - 4-fold cross-validation.

- Training set: 692,865 gene pairs, including 3868 SSLs, from Tong.
- Randomly split training set into four equal-sized sets.
- Train the tree on three of the sets, and score the fourth.

Sensitivity = true positive rate = \( \frac{\text{number correctly predicted SSLs}}{\text{number total SSLs in test set}} \)


False positive rate = \( \frac{\text{number incorrectly predicted SSLs}}{\text{number total SSLs in test set}} \)

How well does this work? - experimental validation.

- Built a tree using all data from the cross-validation.
- Scored test set of new SSL interactions, determined by SGA (8 query genes; 36k gene pairs).
- Test set had same 'biological bias' as training set:
  - actin-base cell polarity
  - cell wall biosynthesis
  - microtubule-based chromosome segregation
  - DNA synthesis and repair.
Validation results: Sensitivity vs. FPR.

A decision tree from cross-validation.

Cross-validation SSL decision tree. Top five scoring leaves labeled by rank.
Top predictors of SSL in cross-validation.

- 2hop SSL-SSL.
  - SSL partners of a gene tend to interact with each other.

- 2hop physical-SSL.
  - compensatory: pathway with A compensates for pathway in which B-C physically interact.

- Surprise: few SSL genes share sequence homology.
What combinations of characteristics are good predictors of SSLs?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Characteristics</th>
<th>SSL +/-</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2hop S–S; 2hop P–S; MIPS complex; GO function; ER localization</td>
<td>4/?</td>
</tr>
<tr>
<td>2</td>
<td>2hop S–S; 2hop P–S; MIPS complex; 2hop H–P</td>
<td>13/4</td>
</tr>
<tr>
<td>3</td>
<td>2hop S–S; 2hop P–S; 2hop S–X; mutant phenotype</td>
<td>206/90</td>
</tr>
<tr>
<td>4</td>
<td>2hop S–S; 2hop P–S; 2hop H–S; 2hop S–X; mutant phenotype; seq. homology p&lt;e–3</td>
<td>4/1</td>
</tr>
<tr>
<td>5</td>
<td>2hop S–S; 2hop P–S; 2hop H–S; mutant phenotype; 2hop S–X; GO function</td>
<td>9/7</td>
</tr>
</tbody>
</table>
Special note: Compensatory pathways.

![Diagram of compensatory pathways](image)

Fig. 8. SSL gene pairs from the highest-scoring leaves of the decision tree may belong to compensating pathways. When gene 1 and gene 2 are lost, synthetic sickness or lethality may result, because both compensating pathways are impaired (a) or because two of three (or more) compensating pathways are impaired (b). Blue circles represent genes. "gene 1" and "gene 2" represent a query gene pair from the first (a) or third (b) leaf. H indicates sequence homology; P indicates physical interaction; S indicates SSL interaction; X indicates correlated mRNA expression.

New SSL Predictions

- Used same tree to predict new SSLs.
- No way to assess widespread validity until tested (i.e., by SGA analysis) but specific examples appear reasonable.

![Gene-pair relationships](image)

**Fig. 3.** Gene-pair relationships. (a and b) Known (a) and predicted (b) SSL gene pairs from the highest-scoring leaf of the decision tree. (c and d) Known (c) and predicted (d) SSL gene pairs from the third-highest-scoring leaf. P, physical interaction; S, synthetic sick or lethal interaction; X, correlated mRNA expression.

Conclusions.

- SSLs can be predicted from integrated proteomic and genomic information.
- Relatively few SSL pairs share sequence homology.
- Many SSL pairs share GO function.
- Strongest predictors of SSL are local network topology (2hop) characteristics.
- SSL prediction offers insights into biological relationships that may be difficult to visualize otherwise.