The society of genes: networks of functional links between genes from comparative genomics

Itai Yanai
and
Charles DeLisi

Genome Biology October 25, 2002
Volume 3, No. 11
Introduction

The authors of this paper use three comparative genomics methods to identify functional correlations between genes (methods other than traditional sequence similarity):

• the examination of common phylogenetic distributions

• the analysis of conserved proximity along the chromosomes of multiple genomes

• observations of fusions of genes into a multidomain gene in another organism
Introduction

• The authors of this paper previously used these three methods individually to generate the links for 43 known microbial genomes.

• In this paper they combine these three methods to construct networks of functional associations.
Background

Overview of the three methods

• The three comparative genomics methods that were applied use correlations in the properties and occurrence of genes across known genomes.
Background

The First Method: **Phylogenetic Profiling**

- Infers functional links between genes A and B based on the same phyletic distribution across the known genomes. (here arbitrarily labeled W, X, Y, Z, etc.)
Background

The Second Method: **Chromosomal Proximity**

- Infers functional links between genes based on their proximity on chromosomes from different genomes.
Background

The Third Method: **Domain Fusion**

- This method is based on the observation that distinct non-homologous genes are functionally related if their orthologs are fused in another organism.
Previous Work

After combining the total sets of links generated by these three methods, the authors came up with the following findings:

• The sets of functional links provided by the three methods are largely additive (they mostly do not overlap).

• The reliability of the links is approximately 70% (which is well above the background noise which they estimated to be around 10-15%)
Networks

The relationships uncovered by each of the three methods form networks whose structures are method dependant.

- Phylogenetic profile links, which they refer to as “phylo links”, are transitive; meaning that two genes linked to a common gene are also linked to each other.

- Phylo networks are characterized by the number and size distributions of their cliques.
Networks

• Gene networks uncovered by chromosomal proximity links, which they refer to as “chromo links”, typically exhibit a chain structure.

• One of the long chains in this network of *Helicobacter pylori* genes corresponds to the chromosomal region containing highly conserved ribosomal genes.

• This network formation results from the fact that the conservation of gene order tends to involve more than two genes.
Networks

- The domain fusion links, which they refer to as “fusion links”, have complicated network relations, including the appearance of a major cluster.

- Even for small networks there are a few nodes that have a large number of links, and a large number that have few links.

*Found in H. pylori*
Using these previous methods

- These three methods for constructing networks all produce distinctive structures.

- The authors discovered that by combining these three different representations of genes onto a single grid, they generated a largely non-overlapping map of functional linkages that provided more information than any of the three maps alone.

- They found that the resulting combined networks have many important structural and functional properties.
Combined Networks

- This combined network captures between 30 and 80% of the genes in a genome and 57% on average.

- The chromo, fusion, and phylo networks individually capture 48%, 29%, and 19% of the genes on average.

- This implies that the graphs overlap significantly in terms of nodes:
• Of the three methods, the chromo networks have the greatest coverage.

• The fusion and phylo networks share 72% and 75% of their nodes with the chromo networks.

• On the other hand, only 20% and 14% of the functional links in fusion and phylo networks are found in chromo networks.

• The least overlapping of all are the phylo and fusion graphs, with only 41% of the nodes in a phylo network found in a fusion network and 6% of the links in a phylo network found in a fusion network.

### Overlap of nodes and edges between the four types of network

<table>
<thead>
<tr>
<th></th>
<th>Nodes</th>
<th>Chromo</th>
<th>Fusion</th>
<th>Phylo</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromo</td>
<td>41 (6)</td>
<td>31 (3)</td>
<td>100 (0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fusion</td>
<td>72 (6)</td>
<td>30 (3)</td>
<td>100 (0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phylo</td>
<td>75 (10)</td>
<td>41 (6)</td>
<td>100 (0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>80 (6)</td>
<td>46 (6)</td>
<td>34 (4)</td>
<td>100 (0)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Edges</th>
<th>Chromo</th>
<th>Fusion</th>
<th>Phylo</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromo</td>
<td>14 (2)</td>
<td>11 (1)</td>
<td>100 (0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fusion</td>
<td>20 (4)</td>
<td>7 (2)</td>
<td>100 (0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phylo</td>
<td>14 (7)</td>
<td>6 (2)</td>
<td>100 (0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>43 (10)</td>
<td>30 (7)</td>
<td>38 (11)</td>
<td>100 (0)</td>
<td></td>
</tr>
</tbody>
</table>

Element in row i column j is the percentage of average nodes (a) or average edges (b) in network of type i that are also found in network of type j, with the standard deviation in parentheses. For example, 75% of the nodes found by phylogenetic profiling are contained in the chromo networks. This should, however, be contrasted with the finding that only 14% of the functional links in the phylo networks are in chromo networks.
• These results indicate that, although the three methods capture overlapping sets of genes, the links generated by the different methods individually show much less overlap than the nodes.

• Therefore, in their study of combined networks, the authors construct the union of the links generated by each method, instead of the intersection.

<table>
<thead>
<tr>
<th></th>
<th>Nodes</th>
<th>Chromo</th>
<th>Fusion</th>
<th>Phylo</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromo</td>
<td></td>
<td>41 (6)</td>
<td></td>
<td>31 (3)</td>
<td>100 (0)</td>
</tr>
<tr>
<td>Fusion</td>
<td>72 (6)</td>
<td></td>
<td>30 (3)</td>
<td></td>
<td>100 (0)</td>
</tr>
<tr>
<td>Phylo</td>
<td>75 (10)</td>
<td>41 (6)</td>
<td></td>
<td></td>
<td>100 (0)</td>
</tr>
<tr>
<td>Combined</td>
<td>80 (6)</td>
<td>46 (6)</td>
<td>34 (4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Edges</th>
<th>Chromo</th>
<th>Fusion</th>
<th>Phylo</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromo</td>
<td></td>
<td>14 (2)</td>
<td>11 (1)</td>
<td>100 (0)</td>
<td></td>
</tr>
<tr>
<td>Fusion</td>
<td>20 (4)</td>
<td></td>
<td>7 (2)</td>
<td></td>
<td>100 (0)</td>
</tr>
<tr>
<td>Phylo</td>
<td>14 (7)</td>
<td>6 (2)</td>
<td></td>
<td></td>
<td>100 (0)</td>
</tr>
<tr>
<td>Combined</td>
<td>43 (10)</td>
<td>30 (7)</td>
<td>38 (11)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Element in row $i$ column $j$ is the percentage of average nodes (a) or average edges (b) in network of type $i$ that are also found in network of type $j$, with the standard deviation in parentheses. For example, 75% of the nodes found by phylogenetic profiling are contained in the chromo networks. This should, however, be contrasted with the finding that only 14% of the functional links in the phylo networks are in chromo networks.
Quality of the links

• An important question is the quality of the links in terms of a direct functional correlation between the linked members.

• This quality can be estimated by referencing databases that classify genes into broad functional categories:
  • Clusters of orthologous groups (COGs).
  • Biological pathways (Kyoto Encyclopedia of Genes and Genomes (KEGG)).

• 72%, 68%, 64% of the fusion, chromo and phylo links are in the same COGs functional category.
Shuffling Networks

To evaluate the significance of a given network, the authors produced 100 shuffled versions of that network.

They wanted to preserve the unique degree distribution of the observed network in its random counterpart.

They shuffled the edges of the observed network according to the following algorithm:
Shuffled Networks

hey began with the observed network and randomly chose two links in that network:

\[ X_1 = Y_1 \]

And rewired them to:

\[ X_2 = Y_2 \]

- They did this over and over again, about 10,000 times.
Shuffled Networks

or each of these 100 shuffled versions of that network, the authors:

- calculated the percentage of linked pairs whose members were present in the same functional category (according to COGs) or pathway (according to KEGG)

- then estimated the average and standard deviation for the population of random networks.

- They found significant differences between the observed networks and the shuffled ones.
In average, 33 standard deviations of separation distinguished an observed network from its 100 shuffled versions.

Each circle corresponds to a network of one of the 4 types of the observed networks for one of the 43 genomes. Each triangle corresponds to the mean of 100 shuffled versions of each of the observed networks.

- The observed networks had high functional correlations centered around 60% agreement with COGs and 70% with KEGG.

- The shuffled networks formed a tight distant cluster with a low background functional correlation - a noise level in the range of 10 to 20%.
E. Coli example

In E. coli, 26 clusters of seven or more genes corresponding to distinct KEGG pathways were identified in the network (where a cluster is a minimally connected subgraph).

- The ability to reconstruct pathways depends on an integration of the three methods.

- In other words, the clusters shown would become fragmented if the methods were used individually.

- To illustrate this, for each of the 26 E. coli clusters, the authors asked what fraction is obtained solely by the individual methods:
The blue spheres correspond to clusters of genes ascribed to a particular functional pathway. The three-dimensional coordinates of the spheres correspond to the fraction of the clusters (in terms of nodes) that could have been recovered by each of the methods (the axes). Only the clusters of seven or more genes are shown, and the names of a few of the pathways are shown.

They conclude that combined networks reconstruct portions of known pathways that cannot be obtained by applying the methods independently.
Some of the pathway clusters can be completely recovered by just one of the methods:

- The ribosomal system, for example, can be completely accounted for by chromo links.
- Overall, most pathway clusters can only be found by the integration of all three methods.
Another example

New functional information on particular genes can be predicted on the basis of the combined networks.

The most fundamental unit of prediction in the network corresponds to a link between two nodes.

For example, *M. jannaschii* genes MJ1313 and MJ1407, whose functions are completely uncharacterized, are linked by fusion, chromo and phylo links.

- From this example the authors conclude that, although the actual functions of the genes in question remain unknown, a functional link probably exists between the two genes and they need to be experimentally studied together.
Structural Properties

surprising property that emerges when the networks are combined is the formation of a giant cluster.

• This occurs for all but three genomes (*Mycoplasma pneumoniae, Mycoplasma genitalium, and Ureaplasma urealyticum*) which might be due to the fact that the number of links in these short genomes is very small.
Structural Properties

Although the number of nodes in the combined networks range from 400 to 1,600:

The percentage of nodes contained in a genome's largest cluster is relatively invariant, ranging from 80 to 90%.
hey conclude from these observations that the properties of the combined networks have stabilized.

Iso, their general characteristics should be relatively invariant against further increases in the number of functional links discovered.
Minimum Path Length

- In general, many paths of different lengths connect each pair of nodes in the giant cluster.

- A useful measure of the global characteristics of the cluster is the minimum path length between each pair of nodes.

- The authors state that such minimum paths are the most biologically relevant of all paths connecting a pair.
Characteristic Path Length

or each genome, the set of shortest path lengths has a Gaussian distribution.

ts average is referred to as the characteristic path length.

he characteristic path length averaged over the 43 genomes that have a giant cluster is 7, and the standard deviation about this average is 3.

o, on average, seven comparative genomic links separate any two genes in a giant cluster.
Local Environments

One descriptive characteristic of a local environment is the **clustering coefficient**: the average probability that two genes linked to a common gene are also linked to each other.

His number is 1 for systems that are fully transitive and becomes very small for random networks.
Local Environments

Not surprisingly, the combined networks for all 43 genomes showed a significant amount of local clustering - they called this ‘cliquishness’.

The chromo and fusion networks both had a similar clustering coefficient of 0.24 and 0.25 respectively, as calculated for the giant component of the networks.

When they shuffled the networks the clustering coefficients for these random graphs were 0.02 and 0.06 for chromo and fusion networks.

The clustering coefficient increased to 0.36 for the combined network (0.02 for the shuffled networks), reflecting the high coefficient of the
Universality

Quantitative measures of the local (clustering coefficient) and global (characteristic path distance) properties of networks allowed the authors to analyze the similarities and differences among all of the networks.

This figure shows the properties of the network mapped onto a two-dimensional space defined by the characteristic path distance and clustering coefficient:

- chromosome proximity (red)
- domain fusion (blue)
- phylogenetic profiling (yellow)
- combined networks (black)
Universality

• They found that networks of the same type (phylo, chromo, fusion or combined) cluster together, demonstrating the universality of network structures of each method.
• The combined networks had a characteristic path that was roughly the average of the fusion and chromo networks and a clustering coefficient that was greater than both.

chromosome proximity (red)
domain fusion (blue)
phylogenetic profiling (yellow)
combined networks (black)
Conclusion

Uncovering functional links between genes is a major step towards deducing the function of individual genes.

Here the authors described the properties of networks generated by the combination of three comparative genomics methods for 43 microbial genomes representing the three domains of life.
**Conclusion**

Hey found that a giant 'small-world' cluster consistently includes 80% to 90% of the nodes in these networks, so that the average minimum path between any two genes is small, but local cliquishness is frequent relative to random networks.

His structure for a society of genes supports the notion that the network of relationships among a cell's genes is a set of highly cliquish functional systems interlinked by genes that are common to multiple pathways.