Using Graph Theory to Analyze Gene Network Coherence

Francisco A. Gómez-Vela
fgomez@upo.es

Norberto Díaz-Díaz
ndiaz@upo.es

José A. Lagares
José A. Sánchez
Jesús S. Aguilar
Outlines

- Introduction
- Proposed Methodology
- Experiments
- Conclusions
Outlines

- Introduction
- Proposed Methodology
- Experiments
- Conclusions
There is a need to generate patterns of expression, and behavioral influences between genes from microarray.

GNs arise as a visual and intuitive solution for gene-gene interaction.

They are presented as a graph:
- Nodes: are made up of genes.
- Edges: relationships among these genes.
Introduction
Gene Network
Introduction

Gene Network

- Many GN inference algorithms have been developed as techniques for extracting biological knowledge
  - Ponzoni et al., 2007.
  - Gallo et al., 2011.

- They can be broadly classified as (Hecker M, 2009):
  - Boolean Network
  - Information Theory Model
  - Bayesian Networks
Introduction

Gene Network Validation in Bioinformatics

Once the network has been generated, it is very important to assure network reliability in order to illustrate the quality of the generated model.

- **Synthetic data based validation**
  - This approach is normally used to validate new methodologies or algorithms.

- **Well-Known data based validation**
  - The literature prior knowledge is used to validate gene networks.
Introduction

Well-Known Biological data based Validation

- The quality of a GN can be measured by a direct comparison between the obtained GN and prior biological knowledge (Wei and Li, 2007; Zhou and Wong, 2011).

- However, these approaches are not entirely accurate as they only take direct gene–gene interactions into account for the validation task, leaving aside the weak (indirect) relationships (Poyatos, 2011).
Outlines

- Introduction
- Proposed Methodology
- Experiments
- Conclusions
Proposed Methodology

- The main features of our method:
  - Evaluate the similarities and differences between gene networks and biological database.
  - Take into account the indirect gene-gene relationships for the validation process.
  - Using Graph Theory to evaluate with gene networks and obtain different measures.
Proposed Methodology

Input Network

Biological Database

Floyd Warshall Algorithm

Distance Matrices

$DM_{IN}$

$DM_{DB}$
Proposed Methodology

Input Network

\( DM_{IN} \)

\[ CM = |DM_i - DM_j| \]

Coherence Matrix

\[ CM = |DM_{IN} - DM_{DB}| \]

Biological Database

\( DM_{DB} \)
Proposed Methodology
Floyd-Warshall Algorithm

- This approach is a graph analysis method that solves the shortest path between nodes.

Network

Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Proposed Methodology
Distance Threshold

- Distance threshold (δ)
  - It is used to exclude relationships that lack biological meaning.
  - This threshold denotes the maximum distance to be considered as relevant in the Distance Matrix generation process.
  - If the minimum distance between two genes is greater than δ, then no path between the genes will be assumed.
Proposed Methodology
Distance Threshold

\[ \delta = 1 \]

Network

Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Proposed Methodology
Distance Threshold

Network

\[ \delta = 1 \]

Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>(\infty)</td>
<td>1</td>
<td>1</td>
<td>(\infty)</td>
</tr>
<tr>
<td>B</td>
<td>(\infty)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>(\infty)</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>(\infty)</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>(\infty)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>(\infty)</td>
<td>(\infty)</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
**Proposed Methodology**

**DM\text{$_{IN}$}**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>∞</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>∞</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**CM** = |DM$_i$ − DM$_j$|

**DM\text{$_{DB}$}**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Coherence Matrix (CM)**
Proposed Methodology

Obtaining Measures

- Coherence Level threshold ($\theta$)
  - This threshold denotes the maximum coherence level to be considered as relevant in the Coherence Matrix.

  - It is used to obtain well-Known indices by using the elements of the coherence matrix:

$$\begin{align*}
CM_{i,j} & \begin{cases} 
|v-y| \leq \theta & \rightarrow TP \\
|v-y| > \theta & \rightarrow FP \\
|\infty - y| (\alpha) & \rightarrow FN \\
|\infty - \infty| (\beta) & \rightarrow TN \\
\end{cases} \\
0 < v, y < \infty
\end{align*}$$
## Proposed Methodology

\[ \theta = 3 \]

### Coherence Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>1</td>
<td>(\alpha)</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>-</td>
<td>(\beta)</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>(\alpha)</td>
<td>(\beta)</td>
<td>-</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>7</td>
<td>5</td>
<td>8</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>
Proposed Methodology

$\theta = 3$

Coherence Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>TP</td>
<td>FN</td>
<td>FP</td>
<td>FP</td>
</tr>
<tr>
<td>B</td>
<td>TP</td>
<td>-</td>
<td>TN</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>C</td>
<td>FN</td>
<td>TN</td>
<td>-</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>D</td>
<td>FP</td>
<td>TP</td>
<td>TP</td>
<td>-</td>
<td>TP</td>
</tr>
<tr>
<td>E</td>
<td>FP</td>
<td>FP</td>
<td>FP</td>
<td>TP</td>
<td>-</td>
</tr>
</tbody>
</table>
Outlines

- Introduction
- Proposed Methodology
- Experiments
- Conclusions
Results

Real data experiment

- Input networks were obtained by applying four inference network techniques on the well-known yeast cell cycle expression data set (Spellman et al., 1998).
  - Soinov et al., 2003.
  - Bulashevska et al., 2005.
  - Ponzoni (GRNCOP) et al., 2007

- Comparison with Well-Known data:
  - BioGrid
  - KEGG
  - SGD
  - YeastNet
Results

Real data experiment

- Several studies were carried out using different threshold value combinations:
  - Distance threshold ($\delta$) and Coherence level threshold ($\theta$) have been modified from one to five, generating 25 different combinations.

- The results show that the higher $\delta$ and $\theta$ values, the greater is the noise introduced.
  - The most representative result, was obtained for $\delta=4$ and $\theta=1$. 
## Results

<table>
<thead>
<tr>
<th></th>
<th>Soinov</th>
<th></th>
<th></th>
<th>Ponzoni</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>F-measure</td>
<td>Accuracy</td>
<td>F-measure</td>
<td>Accuracy</td>
<td>F-measure</td>
</tr>
<tr>
<td>Biogrid</td>
<td>0.27</td>
<td>0.42</td>
<td>0.65</td>
<td>0.79</td>
<td>0.82</td>
<td>0.90</td>
</tr>
<tr>
<td>KEGG</td>
<td>0.58</td>
<td>0.48</td>
<td>0.34</td>
<td>0.50</td>
<td>0.28</td>
<td>0.43</td>
</tr>
<tr>
<td>SGD</td>
<td>0.31</td>
<td>0.47</td>
<td>0.53</td>
<td>0.69</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YeastNet</td>
<td>0.29</td>
<td>0.45</td>
<td>0.50</td>
<td>0.66</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Results

- These results are consistent with the experiment carried out in Ponzoni et al., 2007.

- Ponzoni was successfully compared with Soinov and Bulashevska approaches.
Outlines

- Introduction
- Proposed Methodology
- Experiments
- Conclusions
Conclusions

- A new approach of a gene network validation framework is presented:
  - The methodology not only takes into account the direct relationships, but also the indirect ones.
  - Graph theory has been used to perform validation task.
Conclusions

- Experiments with Real Data.
  - These results are consistent with the experiment carried out in Ponzoni et al., 2007.
  - Ponzoni was successfully compared with Soinov and Bulashevska approaches.
  - These behaviours are also found in the obtained results. Ponzoni presents better coherence values than Soinov and Bulashevska in BioGrid, SGD and YeastNet.
Future Works

- The methodology has been improved:
  - The elements in coherence matrix will be weighted based on the gene-gene relationships distance.
  - A new measure, based on different databases will be generated.

- Moreover, a Cytoscape plugin will be implemented.
Some References


Using Graph Theory to Analyze Gene Network Coherence

Thanks for your attention