Static Properties of Biological Networks

*Ron Y. Pinter*

Dept. of Computer Science
Technion – Israel Institute of Technology

Haifa, Israel

December 2, 2015
Biological Networks

- **Networks** are a powerful method for describing and analyzing the relations among entities.
- Used in Computer Science, Operations Research, Electrical Engineering, etc.
- We use them to **model** the complexities of large biological systems.
A network of interactions can be built from the individual P-P interactions using Pajek software.

A large network of 8184 interactions among 4140 *S. Cerevisiae* proteins.
Likewise, a regulatory network can be built from the individual TF-target relationships.

DATA TYPE
Gal4 → gal1
MCM1 → swi4
Ste12 → fus3
.....

6254 interactions between 160 TFs and 2698 target genes
Global Network Properties

- **Degree distribution** — counting the number of nodes in the network that have a specific number of immediate neighbors.

- **Clustering coefficient** — the average density of nodes’ neighborhoods; measures the tendency of the network to form highly interconnected regions called clusters.

- **Diameter** — the (length of the) longest shortest path between any pair of nodes.
Global Properties

<table>
<thead>
<tr>
<th>k</th>
<th>P(k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ C_v = \frac{2e_1}{n_1(n_1-1)} \]

C=0

C=1/4

D=8

8
What are they good for?

• Used to characterize the type of network we are dealing with
• Is it small-world, scale-free, or geometric?
  – most interact with few, few interact with many
  – hubs — essential, conserved
  – cliques — complexes
• Problematic with networks obtained by high-throughput experiments — contaminated by artifacts (sticky proteins, auto-activators, etc.)
Small-World vs. Scale-Free Networks

- The Erdös-Rényi model of random graphs
  - Edges inserted with equal probability $p$
  - Tend to generate small-world networks (high clustering coeff, small diameter)

- Barabási & Albert scale-free networks
  - Edges inserted using preferential attachment
  - $P(k) \sim k^{-\gamma}$, where $k$ is node degree and $2<\gamma<3$
Taking a Closer Look

- **Graphlets** – a graphlet $g$ is a small (e.g. 2-4 nodes), connected graph pattern whose occurrence in a larger graph $G=(V,E)$ as a subgraph is a mapping $m$ of the nodes of $g$ to a subset of the nodes in $G$ such that if $(v_1,v_2)$ is an edge in $g$ then $(m(v_1),m(v_2))$ is an edge in $E$.

- **Motifs** – subgraphs that occur in a network at frequencies much higher than expected at random.
Integrative Analyses of Interaction Networks Underlying the Cellular Circuitry in Yeast

- Combined two types of network data in one framework
- Found *motifs* in yeast data
- Many interesting biological observations

The Cellular Circuitry

- Involves several types of interactions, including PP interaction and TR interactions

- Integrating them allows better understanding of the underlying cellular processes

- Challenge: devise integrative approaches and methods to analyze this circuitry
Analysis of transcription regulation networks

Feed-forward loop

Single input module

Dense overlapping regulons

(Shen-Orr S. et al., 2002)
Many molecular pathways in the cell involve both protein-protein (PP) interactions and transcription regulation (TR).

Control the activity of proteins and the expression levels of genes.

**Galactose Metabolism**

- Galactose (out) → HXTs → Galactose (in) → Gal2 → Gal1 → Galactose-1-P → Gal7 → Glucose-1-P → Gal10 → Glucose-6-P

**Glycolysis**

- UDP-Galactose → UDP-Glucose

**Pathway Regulation**

- Gal3, Gal4, Gal80

Legend:
- Red: Transcriptional activation
- Yellow: Protein inhibition
- Green: Protein activation
Network Motifs
Connected patterns of interactions that recur in the integrated cellular network statistically significantly more often than at random

Computational Challenge
Generate random networks that maintain the characteristics of the original network with multiple types of connections
Generation of Random Networks

Definitions:

**The extended degree of a node:** #, type, direction of edges near a node

\[ P_1: \leftrightarrow \leftrightarrow \quad P_2: \rightarrow \]

**The edge profile of two nodes:** types and direction of the edges connecting two nodes

\[ (P_1, P_2): \leftarrow \quad (P_1, P_4): \leftrightarrow \]
Switchability Condition

Randomization: retain both the extended degree of each node and the edge profiles

Use iterative edge switching:
If $\text{profile}(s1,t1) = \text{profile}(s2,t2)$ and $\text{profile}(s1,t2) = \text{profile}(s2,t1)$ then switch
Switchability Condition

Randomization: retain both the extended degree of each node and the edge profiles

Use iterative edge switching:
If \( \text{profile}(s_1,t_1) = \text{profile}(s_2,t_2) \) and \( \text{profile}(s_1,t_2) = \text{profile}(s_2,t_1) \), then switch

Only for 2-protein motifs:
PP and TR interactions are independent of each other
Identifying $k$-protein motifs

Computational procedure:
- Find all connected patterns $g$ over $k$ nodes in the cellular network
- Create 1000 random networks; in each network count # of occurrences of $g$

$p$-value($g$) : fraction of random networks in which $g$ is at least as frequent as in cellular network

$g$ is a network motif if $p$-value($g$) $\leq$ 0.05 and appears $\geq$ 5 in cellular network
Overcoming Noisy Data

A *stringent* network – based on a more reliable set of interactions

- PP interactions discovered by at least 2 methods
- TR interactions discovered by methods other than genome-wide DNA location analysis

<table>
<thead>
<tr>
<th>Network Type</th>
<th># of PP Interactions</th>
<th># of TR Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stringent Network</td>
<td>1832 interactions between 1385 proteins</td>
<td>1351 interactions between 128 TFs and 591 target genes</td>
</tr>
<tr>
<td>Non-stringent Network</td>
<td>6159 interactions between 3617 proteins</td>
<td>6254 interactions between 160 TFs and 2698 target genes</td>
</tr>
</tbody>
</table>
Only the mixed feedback loop is a motif (P<0.001 in stringent and non-stringent networks).
Preference for Mixed Feedback

Mixed feedback has an advantage over pure transcriptional feedback

May stem from its faster response time:
In negative feedback long delays can lead to instability and noisy oscillations

May be a design principle in other organisms
Three-protein Network Motifs: Cliques, Co-regulation, Complex Formation

The stringent network:
5 network motifs out of 29 connected patterns
(p < 0.001)

All 5 are also motifs in the non-stringent network
(p < 0.001)
<table>
<thead>
<tr>
<th>Motif</th>
<th>Stringent</th>
<th>Non-stringent</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Protein clique</td>
<td>1293</td>
<td>2016</td>
<td>1198 occurrences in experimentally identified complexes.</td>
</tr>
<tr>
<td>B. Interacting TFs co-regulating a gene</td>
<td>243</td>
<td>476</td>
<td>21 pairs of co-regulating proteins, most act in concert.</td>
</tr>
<tr>
<td>C. Feed-forward loop</td>
<td>83</td>
<td>994</td>
<td>Analyzed in Shen-Orr et al. 2002.</td>
</tr>
<tr>
<td>D. Co-regulated interacting proteins</td>
<td>66</td>
<td>285</td>
<td>25 sets of co-regulated interacting proteins, most act in concert or participate in a common complex.</td>
</tr>
<tr>
<td>E. Mixed feedback loop between TFs that co-regulate a gene</td>
<td>46</td>
<td>118</td>
<td>4 distinct pairs of TFs, also involved in a mixed feedback loop.</td>
</tr>
</tbody>
</table>
4-Protein network motifs

Largely composed of 3-protein motifs

- 63 network motifs out of 201 patterns in network.
  - 36 motifs are 3-protein motifs with a dangling fourth node
  - 21 motifs are combinations of two or more 3-protein motifs.
### 4-protein motifs as combinations of 3-Protein Motifs

<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><img src="image1.png" alt="Diagram A" /></td>
<td><img src="image2.png" alt="Diagram B" /></td>
<td><img src="image3.png" alt="Diagram C" /></td>
<td><img src="image4.png" alt="Diagram D" /></td>
</tr>
<tr>
<td>B</td>
<td><img src="image1.png" alt="Diagram A" /></td>
<td><img src="image2.png" alt="Diagram B" /></td>
<td><img src="image3.png" alt="Diagram C" /></td>
<td><img src="image4.png" alt="Diagram D" /></td>
</tr>
<tr>
<td>C</td>
<td><img src="image1.png" alt="Diagram A" /></td>
<td><img src="image2.png" alt="Diagram B" /></td>
<td><img src="image3.png" alt="Diagram C" /></td>
<td><img src="image4.png" alt="Diagram D" /></td>
</tr>
<tr>
<td>D</td>
<td><img src="image1.png" alt="Diagram A" /></td>
<td><img src="image2.png" alt="Diagram B" /></td>
<td><img src="image3.png" alt="Diagram C" /></td>
<td><img src="image4.png" alt="Diagram D" /></td>
</tr>
</tbody>
</table>
Remaining 4-Protein Motifs
4-Protein Network Motifs

Motifs contain higher-order hubs – pairs/triplets that recur in most motif occurrences.

The non-stringent network
- Some stringent network motifs are missing
- Overall, largely combined of smaller motifs
Summary

- **Composite network motifs:**
  - Mixed feedback loop
  - Co-regulation and complex formation
  - Four-protein motifs are composed of three-protein motifs

Basic framework for detecting network motifs in networks with multiple types of connections
Looking at Specific Nodes using Graphlet Degree Vectors

- Characterizing individual nodes in a network by their neighborhoods [T. Milenkovic and N. Przulj, Cancer Informatics 2008]
- Leveraging operations on vector spaces + enrichment with GO terms
- Application to PPI in human, yeast, and *E. coli*

Michal Gordon, Erez A. Livneh, Ron Y. Pinter, and Eitan Rubin: "Elucidating Protein Function using Graphlet Degree Vectors in Protein-Protein Interactions Networks"; under review.
Occurrences of Graphlets
Graphlets of sizes 2-5 and internal positions (orbits)

from N. Przulj, *Bioinformatics* 2007
Graphlet Degree Vectors

• For each node in the network, count how many times it appears in each of the graphlet positions (roles, orbits)
• Tally results in vectors
  – resolution
  – length
• Generalize the notion of node degrees

from T. Milenkovic and N. Przulj,
Cancer Informatics 2008
2-4 size graphlets arranged at three levels of granularity
High $Q_{4,1}$ count - Two nodes away from a highly connected node (hub)

<table>
<thead>
<tr>
<th>Index</th>
<th>Q2,1</th>
<th>Q3,1</th>
<th>Q3,2</th>
<th>Q3,3</th>
<th>Q4,1</th>
<th>Q4,2</th>
<th>Q4,3</th>
<th>Q4,4</th>
<th>Q4,5</th>
<th>Q4,6</th>
<th>Q4,7</th>
<th>Q4,8</th>
<th>Q4,9</th>
<th>Q4,10</th>
<th>Q4,11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>NUP43</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>133</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
High $Q_{4,3}$ count - One node away from a hub

<table>
<thead>
<tr>
<th>Index</th>
<th>Q2.1</th>
<th>Q3.1</th>
<th>Q3.2</th>
<th>Q3.3</th>
<th>Q4.1</th>
<th>Q4.2</th>
<th>Q4.3</th>
<th>Q4.4</th>
<th>Q4.5</th>
<th>Q4.6</th>
<th>Q4.7</th>
<th>Q4.8</th>
<th>Q4.9</th>
<th>Q4.10</th>
<th>Q4.11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>CYP2C19</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

![Network Diagram](Image)
High $Q_{4,6}$ count - One node away from a clique or a highly connected sub-graph
### Full $Q$ vector

<table>
<thead>
<tr>
<th>Index</th>
<th>Q2,1</th>
<th>Q3,1</th>
<th>Q3,2</th>
<th>Q3,3</th>
<th>Q4,1</th>
<th>Q4,2</th>
<th>Q4,3</th>
<th>Q4,4</th>
<th>Q4,5</th>
<th>Q4,6</th>
<th>Q4,7</th>
<th>Q4,8</th>
<th>Q4,9</th>
<th>Q4,10</th>
<th>Q4,11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>P53</td>
<td>140</td>
<td>1887</td>
<td>9591</td>
<td>139</td>
<td>26739</td>
<td>252064</td>
<td>28563</td>
<td>428968</td>
<td>1396</td>
<td>1008</td>
<td>7027</td>
<td>18070</td>
<td>205</td>
<td>514</td>
<td>28</td>
</tr>
</tbody>
</table>
Generate Graphlet Degree Vectors (GDVs) to characterize each node in the network by its local topological environment.

Cluster the nodes in human PPI network according to their GDVs.

Find biological meaning of the GDV-clusters using GO enrichment analysis

Show enrichments are better with the full vector compared to the degree alone
Clusters of proteins in the human PPI network according to their network local topology
Clusters of proteins in the human PPI network according to their network local topology
Yeast

Set 1 (1811) Set 2 (1153) Set 3 (866)

Set 4 (247) Set 5 (77) Set 6 (70)

Set 7 (69) Set 8 (57) Set 9 (41)

Set 10 (40) Set 11 (39) Set 12 (26)

Set 13 (24) Set 14 (17)
E. Coli

Set 1 (648)  
Set 2 (102)  
Set 3 (91)  
Set 4 (72)  
Set 5 (57)  
Set 6 (25)  
Set 7 (22)  
Set 8 (18)
“Orthologous” Clusters across Species

Human

Yeast

E. coli

Cluster 3 (523)

Cluster 3 (866)

Cluster 7 (22)

Similarity between clusters were measured using Kendall correlation and found to be the most correlated between Clusters 3 in human and in yeast and Cluster 7 in E. coli
Evidence that Orthologous Genes tend to be in Orthologous clusters

• If a human gene is in human Cluster 3, its yeast ortholog is 1.5 times more likely to be in yeast Cluster 3 ($p = 0.0001$, $\chi^2$ test)

• This difference cannot be explained by similarity in the connectivity of orthologous gene pairs, since no correlation is observed between orthologs ($R^2=0.04$, $p=0.104$)
**GO enrichment analysis**

Cluster 3 – Cell Organization

<table>
<thead>
<tr>
<th>GO term</th>
<th>GO id</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>apoptosis</td>
<td>GO:0006915</td>
<td>0.001</td>
</tr>
<tr>
<td>cell cycle</td>
<td>GO:0007049</td>
<td>0.002</td>
</tr>
<tr>
<td>cell organization and biogenesis</td>
<td>GO:0016043</td>
<td>0.001</td>
</tr>
<tr>
<td>cytoskeletal protein binding</td>
<td>GO:0008092</td>
<td>0.001</td>
</tr>
<tr>
<td>cytoskeleton organization and biogenesis</td>
<td>GO:0007010</td>
<td>0.023</td>
</tr>
<tr>
<td>hemostasis</td>
<td>GO:0007599</td>
<td>0.034</td>
</tr>
<tr>
<td>regulation of apoptosis</td>
<td>GO:0042981</td>
<td>0.001</td>
</tr>
<tr>
<td>signal transduction</td>
<td>GO:0007165</td>
<td>0.009</td>
</tr>
<tr>
<td>structural constituent of cytoskeleton</td>
<td>GO:0005200</td>
<td>0.001</td>
</tr>
</tbody>
</table>
GO enrichment analysis

Cluster 5 –
Transcription / Kinases

<table>
<thead>
<tr>
<th>GO term</th>
<th>GO id</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA binding</td>
<td>GO:0003677</td>
<td>0.022</td>
</tr>
<tr>
<td>intracellular signaling cascade</td>
<td>GO:0007242</td>
<td>0.001</td>
</tr>
<tr>
<td>protein amino acid phosphorylation</td>
<td>GO:0006468</td>
<td>0.017</td>
</tr>
<tr>
<td>protein kinase cascade</td>
<td>GO:0007243</td>
<td>0.006</td>
</tr>
<tr>
<td>Protein tyrosine kinase activity</td>
<td>GO:0004713</td>
<td>0.001</td>
</tr>
<tr>
<td>regulation of cellular process</td>
<td>GO:0050794</td>
<td>0.001</td>
</tr>
<tr>
<td>regulation of transcription, DNA dependent</td>
<td>GO:0006355</td>
<td>0.001</td>
</tr>
<tr>
<td>SH3/SH2 adaptor activity</td>
<td>GO:0005070</td>
<td>0.001</td>
</tr>
<tr>
<td>signal transduction</td>
<td>GO:0007165</td>
<td>0.001</td>
</tr>
<tr>
<td>transcription factor binding</td>
<td>GO:0008134</td>
<td>0.008</td>
</tr>
<tr>
<td>transcription from RNA polymerase II promoter</td>
<td>GO:0006366</td>
<td>0.001</td>
</tr>
<tr>
<td>transcriptional repressor activity</td>
<td>GO:0016564</td>
<td>0.036</td>
</tr>
<tr>
<td>transmembrane receptor protein tyrosine kinase activity</td>
<td>GO:0004714</td>
<td>0.001</td>
</tr>
<tr>
<td>transmembrane receptor protein tyrosine kinase signaling pathway</td>
<td>GO:0007169</td>
<td>0.001</td>
</tr>
</tbody>
</table>
GO enrichment analysis

\( \vec{Q} \) Cluster 13 –

Chemokines; many-to-many relationship
Common GO enrichment in “orthologous” clusters

**Human**

Cluster 3 (523)

**Yeast**

Cluster 3 (866)

cytoskeleton organization and biogenesis GO:0007010 in Human and Yeast - Cluster 3
Shuffled labels – keeping the degree of proteins in clusters

Cluster 1 (2658)
Cluster 2 (1714)
Cluster 3 (523)
Cluster 4 (361)
Cluster 5 (280)
Cluster 6 (217)
Cluster 7 (112)
Cluster 8 (101)
Cluster 9 (93)
Cluster 10 (72)
Cluster 11 (62)
Cluster 12 (62)
Cluster 13 (49)
Cluster 14 (45)
Cluster 15 (34)

protein kinase activity
transcription coactivator

regulation of cellularphysiological process
Summary

• Local topological properties are meaningful
• Functional enrichment is important
• Could be useful in other domains
Future Directions

• Extend to motifs and graphlets of size 5 and larger

• Analyze motif and graphlet hierarchy and define algebra for motif combinations

• Turn the network from static to dynamic

• Find causal relations
  – direct the PP interactions
  – assign +/- regulation

• Insert additional relations