Detecting topological patterns in protein networks

Sergei Maslov
Brookhaven National Laboratory
What defines a complex system?

- Complex system has many interacting components (10^{11} neurons, 10^4 types of proteins, 10^6 routers, 10^9 web pages)
- All components are different from each other
- Systems traditionally studied by physics also have many interacting components (10^{23} electrons in a superconductor)
- But they are all the same!
Networks in complex systems

- The simplest question about a complex system: *who interacts with whom?*
- The answer can be visualized as a network
- Network is the *backbone* of a complex system
Why study the topology of complex networks?

- **Lots of easily available data**: that's where the state of the art information is (at least in biology)
- **Large networks** may contain information about basic design principles and/or evolutionary history of the complex system
- This is similar to **paleontology**: learning about an animal from its backbone
Complex networks are the right description when things are interconnected.
Hierarchy of bio-networks

- **Metabolic network**: production of necessary chemical compounds
- **Binding network**: enzymes bind to their substrates in a metabolic network and to other proteins to form complexes
- **Regulatory network**: turns on and off particular groups of proteins in response to signals
- **Higher Levels**: cell-to-cell communication (e.g. neurons in a brain), food webs, social networks, etc.
Protein binding network
Transcription regulatory networks

Prokaryotic bacterium:  
*E. coli*

Single-celled eukaryote:  
*S. cerevisiae*
General properties

- Densely interconnected
- Not very modular: functional modules talk to each other
- Have many random features
- Few proteins (hubs) interact with a lot of neighbors: but most – with just one
in- and out-degree of nodes

In-degree
\( K_{in} = 2 \)

Out-degree
\( K_{out} = 5 \)
How many transcriptional regulators are out there?
Fraction of transcriptional regulators in bacteria

from Stover et al., Nature (2000)
From E. van Nimwegen, Trends in Genetics, 2003
Complexity of regulation grows with complexity of organism

- $N_R <K_{out}> = N <K_{in}> =$ number of edges
- $N_R/N = <K_{in}>/ <K_{out}>$ increases with $N$
- $<K_{in}>$ grows with $N$
  - In bacteria $N_R \sim N^2$ (Stover, et al. 2000)
  - In eucaryots $N_R \sim N^{1.3}$ (van Nimwengen, 2002)
- Networks in more complex organisms are more interconnected than in simpler ones
- Life is not just a bunch of independent modules!
Complexity is manifested in $K_{in}$ distribution

E. coli vs. S. cerevisiae vs. H. sapiens

![Complexity in Kin distribution](image-url)
Beyond degree distributions: How is it all wired together?
Central vs peripheral network architecture

Central (hierarchical)  random  Peripheral (anti-hierarchical)

Correlation profile

- Count \( N(k_0, k_1) \) – the number of links between nodes with connectivities \( k_0 \) and \( k_1 \)
- Compare it to \( N_r(k_0, k_1) \) – the same property in a random network
- Qualitative features are very noise-tolerant with respect to both false positives and false negatives
Correlation profile of the protein interaction network

\[ R(k_0,k_1) = \frac{N(k_0,k_1)}{N_r(k_0,k_1)} \]

Similar profile is seen in the yeast regulatory network

\[ Z(k_0,k_1) = \frac{(N(k_0,k_1) - N_r(k_0,k_1))}{\Delta N_r(k_0,k_1)} \]
Some scale-free networks may appear similar

In both networks the degree distribution is scale-free $P(k) \sim k^{-\gamma}$ with $\gamma \sim 2.2-2.5$
But: correlation profiles give them unique identities

Protein interactions

Internet
How to construct a proper random network?
Null-model of a network

- Distribution of degrees is non-random: the degree of every node has to be conserved in a random network.
- Other topological properties may be also conserved as well:
  - The extent of modularity (by function, sub-cellular localization, etc.)
  - Small motifs (e.g. feed-forward loops)
Randomization

given complex network

random
Edge swapping (rewiring) algorithm

- Randomly select and rewire two edges
- Repeat many times

Metropolis rewiring algorithm

“energy” $E$

- Randomly select two edges
- Calculate change $\Delta E$ in “energy function”
  $E=(N_{actual}-N_{desired})^2/N_{desired}$
- **Rewire** with probability $p=\exp(-\Delta E/T)$

$\Delta E$

How do protein networks evolve?
Gene duplication

Right after duplication

Pair of duplicated proteins

Shared interactions

After some time

Pair of duplicated proteins

Shared interactions
Yeast regulatory network


fraction of shared regulators

amino acid sequence similarity

$10^0$
100 million years ago
Network properties of self-binding proteins

AKA homodimers
There are just **TOO MANY** homodimers

- Null-model
- $P_{\text{self}} \sim \langle k \rangle / N$
- $N_{\text{dimer}} = N \cdot P_{\text{self}} = \langle k \rangle$
- Not surprising as homodimers have many functional roles

<table>
<thead>
<tr>
<th></th>
<th>$N_{\text{dimer}}$</th>
<th>$\langle k \rangle$</th>
</tr>
</thead>
<tbody>
<tr>
<td>yeast</td>
<td>179</td>
<td>6.6 ± 0.2</td>
</tr>
<tr>
<td>worm</td>
<td>89</td>
<td>3.3 ± 0.1</td>
</tr>
<tr>
<td>fly</td>
<td>160</td>
<td>5.9 ± 0.1</td>
</tr>
<tr>
<td>human</td>
<td>1045</td>
<td>5.7 ± 0.1</td>
</tr>
</tbody>
</table>
### Network properties around homodimers

<table>
<thead>
<tr>
<th></th>
<th>$\langle k \rangle$</th>
<th>$\langle k \rangle_{\text{dimer}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>yeast</td>
<td>6.6 ± 0.2</td>
<td>12.4 ± 1.2</td>
</tr>
<tr>
<td>worm</td>
<td>3.3 ± 0.1</td>
<td>13.1 ± 2.2</td>
</tr>
<tr>
<td>fly</td>
<td>5.9 ± 0.1</td>
<td>14.2 ± 1.2</td>
</tr>
<tr>
<td>human</td>
<td>5.7 ± 0.1</td>
<td>14.0 ± 0.6</td>
</tr>
</tbody>
</table>
Likelihood to self-interact vs. $K$

$$P_{dimer}(k) = 1 - (1 - p_{self})^k$$

**Fly:** two-hybrid data

$P_{self} \sim 0.003$, $P_{others} \sim 0.0002$

**Human:** database data

$P_{self} \sim 0.05$, $P_{others} \sim 0.0002$
What we think it means?

- In random networks $p_{dimer}(K) \sim K^2$ not $\sim K$ like our empirical observation.
- $K$ is proportional to the "stickiness" of the protein, which in turn scales with:
  - the area of hydrophobic residues on the surface
  - # copies/cell
  - its’ popularity (in datasets taken from databases)
  - etc.
- Real interacting pair consists of an "active" and "passive" protein, and binding probability scales only with the "stickiness" of the active protein.
- "Stickiness" fully accounts for higher than average connectivity of homodimers.
Summary

- Living cells contain many complex protein networks
- Networks in more complex organisms are more interconnected
- Most have hubs – highly connected proteins
- Hubs often avoid each other (networks are anti-hierarchical)
- Networks evolve by gene duplications
- There are many self-interacting proteins. Probability to self-interact linearly scales with the degree K.
Collaborators:

- Kim Sneppen – U. of Copenhagen
- Kasper Eriksen – U. of Lund
- Koon-Kiu Yan – Stony Brook
- Ilya Mazo, Jaroslav Ispolatov, Anton Yuryev – Ariadne Genomics
THE END
Protective effect of duplicates

Maslov, Sneppen, Eriksen, Yan 2003

Yeast

Worm

lethal fraction out of 14009 tested genes

Maslov, Sneppen, Eriksen, Yan 2003
Protein interaction networks


Graph: Number of common interaction partners against AA sequence identity for Fly, Yeast, and H. pylori.
What shapes the topology of protein networks?

- Duplication-divergence models **CAN** account for their **basic topological features**
- **BUT**: functional organization dominates:
  - Most pairs with many shared interactions are **NOT** homologs
  - **Hubs** are not caused by mult. duplications
- Functional organization is more important than duplication-divergence
Genome-wide protein networks

- **Nodes** - proteins
- **Edges** - interactions between proteins
  - Bindings (physical interactions)
  - Regulations (transcriptional, protein modifications, etc.)
  - Etc, etc, etc.
Correlation profile of the yeast regulatory network

\( R(k_{\text{out}}, k_{\text{in}}) = \frac{N(k_{\text{out}}, k_{\text{in}})}{N_r(k_{\text{out}}, k_{\text{in}})} \)

\( Z(k_{\text{out}}, k_{\text{in}}) = \frac{N(k_{\text{out}}, k_{\text{in}}) - N_r(k_{\text{out}}, k_{\text{in}})}{\Delta N_r(k_{\text{out}}, k_{\text{in}})} \)
YPD full regulatory network

\( O < -2 \) standard deviations

\( O > 2 \) standard deviations
Regulators:
\[ \langle K_{in} \rangle = 2.55 \pm 0.25 \]

Workhorse:
\[ \langle K_{in} \rangle = 2.65 \pm 0.1 \]
Amplification ratios

\[ A^{(dir)} = \frac{\langle K_{in}K_{out} \rangle}{\langle K_{in} \rangle} \quad A^{(undir)} = \frac{\langle K(K - 1) \rangle}{\langle K \rangle} \]

- \( A^{(dir)} \): 1.08 - E. Coli, 0.58 - Yeast
- \( A^{(undir)} \): 10.5 - E. Coli, 13.4 - Yeast
- \( A^{(PPI)} \): ? - E. Coli, 26.3 - Yeast
Two-hybrid experiment

- To test if A interacts with B create two hybrids A* (with Gal4p DNA-binding domain) and B* (with Gal4p activation domain)

Prolexsys 2-hybrid dataset for Human

The image shows a color-coded heatmap with values along the axes labeled $K_0$ and $K_1$. The color scale on the right indicates values ranging from -6 to 6. The data seems to be representing some form of distribution or relationship between the variables $K_0$ and $K_1$.
DIP core correlation profile
Protein binding networks

*S. cerevisiae*

Two-hybrid nuclear

Database (DIP) core set
Pathway vs. network paradigm

Pathway

Network
How to deal with feedback loops?
Propagation of signals often involves feedback loops

- **Feedback** could be used as a **checkpoint** that the signal reached its destination:
  - The main direction of information flow $A \rightarrow B \rightarrow C$
  - “Weak” feedback signal $C \rightarrow A$

- Some closed loops like $A \rightarrow B \rightarrow C \rightarrow A$ are not used to send feedback but to regulate periodic processes:
  - Cell cycle
  - Circadian rhythms
Finding the feedback links

Feedback links go against the main direction of the information flow
Algorithm

- Goal: remove the smallest number of edges to make the network feedback-free

- Solution:
  - Randomly assign hierarchy weights $H_i$ to all nodes. Links $i \rightarrow j$ such that $H_i < H_j$ are candidate feedback links.
  - Try to switch $H_m$ and $H_n$ on randomly selected nodes $m, n$. Accept if this switch reduces the number of feedback links. If increases – accept with a small probability (a la Metropolis).
  - The final network obtained by removing feedback links is feedback-free.
Bow-tie diagram

Human protein modification network: 1100 nodes

wormholes

In: 96 nodes

SCC: 93 nodes

Out: 553 nodes

dangling ends and isolated components
Feedback loops in human protein-modification network

- Removal of only 48 out of 2200 edges makes human protein modification network feedback-free.
- Most links are “reproducible”: rerunning experiment gives almost the same answer.
- Run algorithm 10 times. Each time record all upstream-downstream relations. Look for reproducible relationships: they correspond to main direction of the information flow.
Inhibition of apoptosis

MAPK signaling
Gene disruptions in yeast
Repercussions of gene deletions

Number of proteins affected by a single gene deletion

Total number of proteins in yeast genome
Total: 120,000 interacting protein pairs extracted from PubMed as of 8/2004

Data from Ariadne Genomics