Analysis of biological networks on a genome scale

Haiyuan Yu
Gerstein lab
Department of MB&B
Yale University
March 11th, 2005
Outline

1. Why networks?
Outline

1. Why networks?
2. Generation of biological networks
Outline

1. Why networks?
2. Generation of biological networks
3. Topological analysis of individual biological networks
Outline

1. Why networks?
2. Generation of biological networks
3. Topological analysis of individual biological networks
4. Comprehensive comparison of biological networks
Outline

1. Why networks?
2. Generation of biological networks
3. Topological analysis of individual biological networks
4. Comprehensive comparison of biological networks
Networks are mathematical models of complex systems

A network (a.k.a. graph) is a pair $G = (V, E)$ of sets satisfying $E \subseteq V \times V$

Vertices: $V = \{1, 2, 3, 4, 5, 6, 7\}$

Edges: $E = \{\{1, 2\}, \{1, 5\}, \{2, 5\}, \{3, 4\}, \{5, 7\}\}$
i. Many biological processes can be modeled as networks

Protein-protein Interactions

[Barabasi]
i. Many biological processes can be modeled as networks

[Horak, et al, Genes & Development, 16:3017-3033]
[DeRisi, Iyer, and Brown, Science, 278:680-686]
[Jeong et al, Nature, 41:411]
ii. Many information systems in biology are networks

Trees are special networks

Hierarchies & DAGs
[Enzyme, Bairoch; GO, Ashburner; MIPS, Mewes, Frishman]
iii. Networks as a universal language

- Internet [Burch & Cheswick]
- Food Web
- Electronic Circuit
- Disease Spread [Krebs]
- Protein Interactions [Barabasi]
- Neural Network [Cajal]
- Social Network
Outline

1. Why networks?
   i. Many biological processes are networks;
   ii. Many information systems are networks;
   iii. Network is a universal language in many scientific fields;

2. Generation of biological networks
3. Topological analysis of biological networks
4. Comprehensive comparison of biological networks
Outline

1. Why networks?
2. Generation of biological networks
   3. Topological analysis of individual biological networks
   4. Comprehensive comparison of biological networks
2. Generation of biological networks?

◊ **Manually-curated small-scale experiments**
  - MIPS, BIND, DIP, KEGG, etc.

◊ **High-throughput experiments**
  1. Interaction networks:
     a) TAP-tag (Ho et al., Gavin et al., J Greenblatt & A Emili, H Yu)
     b) Two hybrid (Ito et al., Uetz et al.; M Vidal, H Yu)
  
  2. Expression networks:
     a) Microarray (Spellman et al., M Snyder, J Qian, H Yu, et al.)
  
  3. Regulatory networks
     a) ChIP-chip (Horak et al.; A Borneman, M Snyder, H Yu)
2. Generation of biological networks?

- Manually-curated small-scale experiments
  MIPS, BIND, DIP, KEGG, etc.

- High-throughput experiments
  1. Interaction networks:
     a) TAP-tag (Ho et al., Gavin et al., J Greenblatt & A Emili, H Yu)
     b) Two hybrid (Ito et al., Uetz et al.; M Vidal, H Yu)

  2. Expression networks:
     a) Microarray (Spellman et al., M Snyder, J Qian, H Yu, et al.)

  3. Regulatory networks
     a) ChIP-chip (Lee et al.; A Borneman, M Snyder, H Yu)

- Computational prediction
  Bayesian integration (Jansen, Yu, et al., Science; Yu, et al., Genome Res.)
Basic idea of Bayesian integration
Basic idea of Bayesian integration

In essence, Bayesian integration is similar to a weighted vote of different voices!
Prediction of protein interactions: Bayesian integration

Network

Gold-Standards

Feature 1, e.g. co-expression
Feature 2, e.g. same function
Gold-standard +
Gold-standard –

[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]
Prediction of protein interactions: Bayesian integration

Network

Gold- Standards

Likelihood Ratio for Feature $i$:

$$L_i = \frac{p(x_i | +)}{p(x_i | -)}$$

[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]
Prediction of protein interactions: Bayesian integration

Likelihood Ratio for Feature $i$:

$$L_i = \frac{p(x_i | +)}{p(x_i | -)}$$

$$L_1 = \frac{4/4}{3/6} = 2$$

Feature 1, e.g. co-expression
Feature 2, e.g. same function
Gold-standard +
Gold-standard –

[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]
Prediction of protein interactions: Bayesian integration

Likelihood Ratio for Feature $i$:

$$L_i = \frac{p(x_i | +)}{p(x_i | -)}$$

Gold-Standards

$$L_1 = \frac{4/4}{3/6} = 2$$

[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]
Prediction of protein interactions: Bayesian integration

Likelihood Ratio for Feature $i$:

$$L_i = \frac{p(x_i | +)}{p(x_i | -)}$$

Gold-Standards

Network

$$L_1 = \frac{4/4}{3/6} = 2$$

[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]
Prediction of protein interactions: Bayesian integration

Network

Gold-Stands

Likelihood Ratio for Feature $i$:

$$L_i = \frac{p(x_i | +)}{p(x_i | -)}$$

Gold-standards

$$L_1 = \frac{4/4}{3/6} = 2$$

[Jaansen, Yu, et al., Science; Yu, et al., Genome Res.]
Prediction of protein interactions: Bayesian integration

Likelihood Ratio for Feature $i$:

$$L_i = \frac{p(x_i | +)}{p(x_i | -)}$$

Gold-Standards

$L_1 = \frac{4/4}{3/6} = 2$

[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]
Prediction of protein interactions: Bayesian integration

Likelihood Ratio for Feature $i$:

$$L_i = \frac{p(x_i | +)}{p(x_i | -)}$$

Gold-Standards:

- $L_1 = \frac{4/4}{3/6} = 2$
- $L_2 = \frac{3/4}{3/6} = 1.5$

For each protein pair:

$$LR = L_1 \times L_2$$

$$\log(LR) = \log(L_1) + \log(L_2)$$

[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]
Prediction of protein interactions: Bayesian integration

For each protein pair:
\[ LR = \frac{L_1}{L_2} \]
\[ \log(LR) = \log(L_1) + \log(L_2) \]

Likelihood Ratio for Feature \( i \):
\[ L_i = \frac{p(x_i | +)}{p(x_i | -)} \]

Gold-Standards
\[ L_1 = \frac{4/4}{3/6} = 2 \]
\[ L_2 = \frac{3/4}{3/6} = 1.5 \]

Network

[Jansen, Yu, et al., Science; Yu, et al., Genome Res.]
Prediction of protein interactions: Bayesian integration

1. Individual features are weak predictors,
   \[ LR \sim 10; \]

2. Bayesian integration is much more powerful,
   \[ LR_{cutoff} = 600 \quad \sim 9000 \text{ interactions} \]
Prediction of protein interactions: Across species mapping

E. coli  S. cerevisiae  C. elegans  D. melanogaster  H. Sapiens

Regulation

Expression

Interaction

Metabolism
Across species -- Mapping

\[\text{E. coli} \rightarrow \text{S. cerevisiae} \rightarrow \text{C. elegans} \rightarrow \text{D. melanogaster} \rightarrow \text{H. Sapiens}\]

- **Regulation**
- **Expression**
- **Interaction**
- **Metabolism**

[Yu et al, Genome Res 14: 1107-18]
Interologs

[ Yu et al, Genome Res 14: 1107-18 ]
Likelihood ratios of interolog mapping

More likely to interact

More similar

LR_{cutoff} \sim 1000

[Yu et al, Genome Res 14: 1107-18]
Performance of interolog mapping is comparable to large-scale experiments

ROC-like curve:

TP: True positives
FP: False positives
P: Total positives

[Yu et al, Genome Res 14: 1107-18]
Performance of interolog mapping is comparable to large-scale experiments.

ROC-like curve:

TP: True positives
FP: False positives
P: Total positives

Coverage

Accuracy

[Yu et al, Genome Res 14: 1107-18]
Interolog mapping

- 20470 interactions in *C. albicans*
- 91224 interactions in *C. elegans*
- 101920 interactions in *D. melanogaster*
- 201754 interactions in *A. thaliana*

8250 high-quality interactions

*[Yu et al, Genome Res 14: 1107-18]*
Interolog database

http://interolog.gersteinlab.org
Outline

1. Why networks?

2. Generation of biological networks
   i. Bayesian integration
      ~9000 yeast interactions
   iii. Interolog mapping
      Interaction networks in other organisms

3. Topological analysis of biological networks

4. Comprehensive comparison of biological networks
Outline

1. Why networks?
2. Generation of biological networks
3. Topological analysis of individual biological networks
4. Comprehensive comparison of biological networks
Global topological measures

Indicate the gross topological structure of the network

Degree ($K$) 5
Path length ($L$) 2
Clustering coefficient ($C$) 1/6

Interaction and expression networks are **undirected**

[Barabasi]
Scale-free networks

**Hubs** dictate the structure of the network

[Barabasi]
Hubs tend to be Essential

Integrate gene essentiality data with network. Perhaps hubs represent vulnerable points?

[Lauffenburger, Barabasi]
Relationships extends to "Marginal Essentiality"

Marginal essentiality measures relative importance of each gene (e.g. in growth-rate and condition-specific essentiality experiments) and scales continuously with "hubbiness".

[Yu et al., 2004, TIG]
More important genes are better candidates for drug targets

Gotcha!!!
Global topological measures

- In-degree: 3
- Out-degree: 5
- Path length: 2
- Clustering coefficient: 1/6

Regulatory and metabolic networks are **directed**
Scale-free (power law) Network Structure

Most TFs have few target genes
Few TFs have many target genes

[Yu et al., 2004, TIG]
Regulatory hubs tend to be essential, too!

[Yu et al., 2004, TIG]
Target hubs tend not to be essential!

[Hubs]

Non-hubs

Fraction of Essential genes (%)

[Yu et al., 2004, TIG]
Local Network Motifs

Regulatory modules within the network

[Alon]
TopNet – an automated web tool

[Yu et al., NAR, 2004]
Outline

1. Why networks?
2. Generation of biological networks
3. Topological analysis of individual biological networks
   i. Global topology
      a) Marginal essentiality correlates with the degree
         Proteins with more interaction are more important for the cell
      c) Regulatory networks are scale-free
      d) Regulatory hubs are essential, while target hubs not
   iii. TopNet - an web tool for topological analysis of networks
4. Comprehensive comparison of biological networks
Outline

1. Why networks?
2. Generation of biological networks
3. Topological analysis of individual biological networks

4. Comprehensive comparison of biological networks
Outline

1. Why networks?
2. Generation of biological networks
3. Topological analysis of individual biological networks

4. Comprehensive comparison of biological networks
   i. Pair-wise comparisons
   ii. Multi-way comparison
Regulatory vs. Expression

Composite motifs:

- SIM
- FFL
- MIM

[Yu et al., 2003, TIG]
Co-regulated proteins are co-expressed

Occurrence of motifs relative to random expectation (log)

[Yu et al., 2003, TIG]
Outline

1. Why networks?
2. Generation of biological networks
3. Topological analysis of individual biological networks

4. Global comparison of biological networks
   i. Pair-wise comparisons
   ii. Multi-way comparison
Four networks at the same time

[Horak, et al, Genes & Development, 16:3017-3033]
[DeRisi, Iyer, and Brown, Science, 278:680-686]
[Jeong et al, Nature, 41:411]
Combined networks

- Metabolic pathway
- Transcriptional regulatory network
- Physical protein-protein Interaction
- Co-expression Relationship

Part of the TCA cycle
All networks are related

Neighbors in one network tend to be close in another network

1. Correlation coefficient
2. Mutual information
3. Chi-2 tests

... We are all related as a family!

All P-values in chi-2 tests are smaller than $10^{-100}$
Hub overlap ???

Action networks

Hub overlap relative to random expectation

<table>
<thead>
<tr>
<th></th>
<th>P &lt; 10^{-9}</th>
<th>P &lt; 10^{-12}</th>
<th>P &lt; 0.02</th>
<th>P = 0.62</th>
<th>P = 0.42</th>
</tr>
</thead>
<tbody>
<tr>
<td>Met-Int</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp-Int</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp-Met</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp-Reg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Int-Reg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Composite motifs – co-regulation

SIM

FFL

TF-target

Regulation

Distance K

Adjacent

Far-away

1

2

3

K
Immediate pairs are co-regulated; Distant pairs are not

Enrichment of composite motifs

Fraction of all pairs at k that are co-regulated by the same SIM

Adjacent  $K$  Far-away

Less
Immediate pairs are co-regulated; Distant pairs are not.
Composite motifs – inter-regulation

SIM

FFL

Bridge

TF-target
Regulation

Distance K
Composite motifs – inter-regulation

SIM

FFL

Bridge

TF-target

Regulation

Distance $K$
Metabolic networks are long-range networks; Interaction and expression networks are short-range.
Metabolic networks are assembly lines in the cell

[DeRisi, Iyer, and Brown, Science, 278:680-686]
Metabolic networks are assembly lines in the cell

Henry Ford’s auto assembly plan

[David Kimble]
Metabolic networks are assembly lines in the cell
Metabolic networks are assembly lines in the cell

[DeRisi, Iyer, and Brown, Science, 278:680-686]
Bridges

TF-target Regulation

Distance $K$

Co-expressed

Delayed
Metabolic networks are long-range networks; Interaction and expression networks are short-range.
Classification of biological networks

Regulation

Metabolism

Expression

Interaction
Classification of biological networks

Directed

- Regulation
- Metabolism

Undirected

- Expression
- Interaction
Classification of biological networks

Directed
- Regulation
- Metabolism

Undirected
- Expression
- Interaction

Regulation ➔ Action
Classification of biological networks

Directed
- Regulation
- Metabolism

Undirected
- Expression
- Interaction

Long-range
- Regulation

Short-range
- Action
Summary

- Generation of biological networks
  - Large-scale experiments (ChIPchip, TAP-tag, Y2H)
  - Bayesian integration
  - Interolog database

- Topological analysis of individual biological networks
  - Marginal essentiality correlates with the degree
  - Regulatory networks are scale-free
  - Regulatory hubs are essential, while target hubs not
  - TopNet - an web tool for topological analysis of networks

- Comprehensive comparison of biological networks
  - Co-expressed, interacting pairs, as well as co-enzymes, are co-regulated
  - TFs targeting the same target tend to interact
  - Regulation vs. Action networks
  - Metabolic networks have long-range regulatory relationships
Acknowledgment

Mark Gerstein
Acknowledgment

Michael Snyder

Perry Miller
Acknowledgment

Jiang Qian, Nick, Alberto, Brandon, Valery, Ronald, Paul, Jason, David,
Dov, Duncan, Joel, Tom, Zhaolei, and everyone in the Gerstein group
Acknowledgment