Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

Mark B Gerstein
Yale

Slides at
Lectures.GersteinLab.org

(See Last Slide for References & More Info.)
The problem: Grappling with Function on a Genome Scale?

• 250 of ~530 originally characterized on chr. 22 [Dunham et al. Nature (1999)]

• >25K Proteins in Entire Human Genome (with alt. splicing)
Traditional single molecule way to integrate evidence & describe function

**EF2_YEAST**

**Descriptive Name:** Elongation Factor 2

**Lots of references to papers**

**Summary sentence describing function:**
This protein promotes the GTP-dependent translocation of the nascent protein chain from the A-site to the P-site of the ribosome.
Some obvious issues in scaling single molecule definition to a genomic scale

- Fundamental complexities
  - Often >2 proteins/function
  - Multi-functionality:
    - 2 functions/protein
  - Role Conflation:
    - molecular, cellular, phenotypic
Some obvious issues in scaling single molecule definition to a genomic scale

- Fundamental complexities
  - Often >2 proteins/function
  - Multi-functionality:
    - 2 functions/protein
  - Role Conflation:
    - molecular, cellular, phenotypic

- Fun terms… but do they scale?....
  - **Starry night** (P Adler, ’94)

[Seringhaus et al. GenomeBiology (2008)]
An Ontology of Naming Pathologies

**Single**

- **M** - Explicit meaning
  - **M-scientific** - SEMA5A
  - **M-literal** - drop dead
  - **M-embed** - Inherent meaning of words is sufficient to describe gene function in some way; no cultural knowledge is required

**Multi**

- **T** - Transferred naming system
  - **T-relation** - kryptonite and superman
  - **T-norelation** - arleekin valiet tungus...

- **P** - Problematic relationships
  - **P-clash** - PKD1 and lov-1
  - **P-confusion** - MT-1
  - **P-defunct** - BAF45 and BAF47

**~M**

- **~M-outside** - kuzbanian
- **~M-irrel** - ring
- **~M-nr** - yippee

Silly or funny names. No relevance to underlying gene function

---

[Seringhaus et al. GenomeBiology (2008)]
Gene Name Skew

Seringhaus et al. Genome Biology (2008)
Hierarchies & DAGs of controlled-vocab terms but still have issues...

MIPS (Mewes et al.)

GO (Ashburner et al.)

[Seringhaus & Gerstein, Am. Sci. '08]
Networks (Old & New)

Classical KEGG pathway

- **Fringe**: Vital in boundary formation in developing fly wing.
- **Numb**: mutations impair sensory organs in flies
- **Notch**: with defects, flies develop notches in wings

Same Genes in High-throughput Network

- **Itch**: linked to itchy skin in mice
- **DLK1**
- **Fringe**
- **Delta**
- **Numb**

[Seringhaus & Gerstein, Am. Sci. '08]
Networks occupy a midway point in terms of level of understanding

1D: Complete Genetic Partslist

2D: Bio-molecular Network Wiring Diagram

3D: Detailed structural understanding of cellular machinery

Networks as a universal language

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

Breast Cancer
Alzheimer's Disease
Parkinson's Disease
Multiple Sclerosis

Interactome networks

[Adapted from H Yu]
Using the position in networks to describe function

[NY Times, 2-Oct-05, 9-Dec-08]
Combining networks forms an ideal way of integrating diverse information

- Metabolic pathway
- Transcriptional regulatory network
- Physical protein-protein Interaction
- Co-expression Relationship
- Genetic interaction (synthetic lethal)
- Signaling pathways

Part of the TCA cycle
Outline: Molecular Networks

• Why Networks?
• Network Structure: Key Positions
  ◊ Hubs & Bottlenecks
  ◊ Tops of a Hierarchy
• Networks, Variation & the Environment
  ◊ Which pathways change most with the environment
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]
Global topological measures for directed networks

Regulatory and metabolic networks are directed
**Scale-free networks**

*Hubs* dictate the structure of the network

[Barabasi]
Hubs tend to be Essential

Integrate gene essentiality data with protein interaction 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".

[Yue et al., 2003, TIG]
Another measure of Centrality: Betweenness centrality

Betweenness of a node is the number of shortest paths of pairs of vertices that run through it -- a measure of information flow.


Betweenness centrality -- Bottlenecks

Proteins with high betweenness are defined as *Bottlenecks* (top 20%), in analogy to the traffic system.
Bottlenecks & Hubs

- **Hub-bottleneck node**
- **Non-hub-bottleneck node**
- **Hub-non-bottleneck node**
- **Non-hub-non-bottleneck node**

[Yu et al., PLOS CB (2007)]
Different Interactome networks

Interaction networks

[Jeong et al, Nature (2001)]

Regulatory networks

Undirected

Directed
Bottlenecks are what matters in regulatory networks

\[ P < 10^{-20} \]

\[ P < 10^{-4} \]

Signaling transduction pathways are directed

[Xianglin Shi]

Bottlenecks in signaling pathways are important

*Yu et al., PLoS Comput Biol (2007)*
Outline: Molecular Networks

• Why Networks?
• Network Structure: Key Positions
  ◊ Hubs & Bottlenecks
  ◊ Tops of a Hierachy
• Networks, Variation & the Environment
  ◊ Which pathways change most with the environment
Determination of "Level" in Regulatory Network Hierarchy with Breadth-first Search

I. Example network with all 4 motifs

II. Finding terminal nodes (Red)

III. Finding mid-level nodes (Green)

IV. Finding top-most nodes (Blue)

[Yu et al., PNAS (2006)]
Regulatory Networks have similar hierarchical structures

[S. cerevisiae]

[E. coli]

[Yu et al., *Proc Natl Acad Sci U S A* (2006)]
Example of Path Through Regulatory Network

Expression of MOT3 is activated by heme and oxygen. Mot3 in turn activates the expression of NOT5 and GCN4, mid-level hubs. GCN4 activates two specific bottom-level TFs, Put3 and Uga3, which trigger the expression of enzymes in proline and nitrogen utilization.

[Yu et al., PNAS (2006)]
Yeast Regulatory Hierarchy: the Middle-managers Rule

A. Regulatory hierarchy in *S. cerevisiae*

[Yu et al., PNAS (2006)]
Yeast Network Similar in Structure to Government Hierarchy with Respect to Middle-managers

B. Governmental hierarchy of a representative city (Macao)
Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks

Average betweenness at each level

[Yu et al., PNAS (2006)]
Characteristics of Regulatory Hierarchy: The Paradox of Influence and Essentiality

[Yu et al., PNAS (2006)]
Outline: Molecular Networks

• Why Networks?
• Network Structure: Key Positions
  ◊ Hubs & Bottlenecks
  ◊ Tops of a Hierachy
• Networks, Variation & the Environment
  ◊ Which pathways change most with the environment
What is metagenomics?

**Genomics Approach**

- **Culture Microbes**
- **Extract DNA**
- **Sequence**
  - ATCGTATA
  - CGCGAAG
  - ACGTCTGA
  - AGTCTGCT

**Assemble and Annotate**

**Problem:** Estimated that less than 1% can be cultured in the lab.

---

**Metagenomics Approach**

- **Collect Sample**
- **Extract DNA**
- **Sequence**
  - ATCGTAGATAGATAGATAGA
  - ATCGTAGATAGATAGATAGA
  - ATCGTAGATAGATAGATAGA
  - ATCGTAGATAGATAGATAGA
  - ATCGTAGATAGATAGATAGA
  - ATCGTAGATAGATAGATAGA
  - ATCGTAGATAGATAGATAGA
  - ATCGTAGATAGATAGATAGA
  - ATCGTAGATAGATAGATAGA
  - ATCGTAGATAGATAGATAGA

**Partially Assemble and Annotate**

**Problem:** Lose information about which gene belongs to which microbe.
Global Ocean Survey Statistics (GOS)

6.25 GB of data
7.7M Reads
1 million CPU hours to process

Rusch, et al., PLOS Biology 2007
Expressing data as matrices indexed by site, env. var., and pathway usage

[Rusch et. al., (2007) PLOS Biology; Gianoulis et al., PNAS (in press, 2009)]
Simple Relationships: Pairwise Correlations

[ Gianoulis et al., PNAS (in press, 2009)]
Canonical Correlation Analysis: Simultaneous weighting

\[ UPI = a \cdot GRE + b \cdot \text{# of papers published} + c \cdot GPA \]

\[ GPI = a' \cdot GRE + b' \cdot \text{Undergraduate Performance Index (UPI)} + c' \cdot \text{Graduate School Performance Index (GPI)} \]

[ Gianoulis et al., PNAS (in press, 2009) ]
Canonical Correlation Analysis: Simultaneous weighting

<table>
<thead>
<tr>
<th>Score</th>
<th># of papers published</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Undergraduate Performance Index (UPI)</th>
<th>Graduate School Performance Index (GPI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE</td>
<td>GPA</td>
</tr>
</tbody>
</table>

Environmental Features
- Temp
- Chlorophyll

Metabolic Pathways
- Photosynthesis
- Lipid Metabolism

[ Gianoulis et al., PNAS (in press, 2009) ]
The goal of this technique is to interpret cross-variance matrices. We do this by defining a change of basis.

Given $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$

$$C = \frac{\sum_X \sum_{X,Y} \sum_{Y,X}}{\sum_X \sum_{Y,X}}$$

$$\text{max } \text{Corr}(U,V) = \frac{a' \sum_{12} \sum_{12} b}{a \sqrt{a' \sum_{11} a} \sqrt{b' \sum_{22} b}}$$

[ Gianoulis et al., PNAS (in press, 2009) ]
Strength of Pathway co-variation with environment

[ Gianoulis et al., PNAS (in press, 2009) ]
Conclusion #1: energy conversion strategy, temp and depth

[ Gianoulis et al., PNAS (in press, 2009) ]
Conclusion #2: Outer Membrane components vary the environment

[ Gianoulis et al., PNAS (in press, 2009) ]
Why is their fluctuation in amino acid metabolism? Is there a feature(s) that underlies those that are environmentally-variant as opposed to those which are not?

[ Gianoulis et al., PNAS (in press, 2009) ]
Biosensors:
Beyond Canaries in a Coal Mine

[ Gianoulis et al., PNAS (in press, 2009) ]
Conclusions

• Developing Standardized Descriptions of Protein Function
• Gene Naming
• Betweenness is an important global network statistic
  ◊ Bottlenecks are more correlated with essentiality than hubs in regulatory networks
• Regulatory Network Hierarchies
  ◊ Middle managers dominate, sitting at info. flow bottlenecks
  ◊ Paradox of influence and essentiality
  ◊ Topmost proteins sit at center of interaction network
Conclusions: Networks Dynamics across Environments

- Developed and adapted techniques to connect quantitative features of environment to metabolism.
- Applied to available aquatic datasets, we identified footprints that were predictive of their environment (potentially could be used as biosensor).
- Strong correlation exists between a community’s energy conversion strategies and its environmental parameters (e.g. temperature and chlorophyll).
- Suggest that limiting amounts of cofactor can (partially) explain increased import of amino acids in nutrient-limited conditions.
- an automated web tool

Normal website + Downloaded code (JAVA) + Web service (SOAP) with Cytoscape plugin

[Yu et al., NAR (2004); Yip et al. Bioinfo. (2006);
Similar tools include Cytoscape.org, Idekar, Sander et al]
Acknowledgements
TopNet.GersteinLab.org
Acknowledgements

TopNet.GersteinLab.org
RNAi: 
Birth of a 
Field in 
the 
Literature 
Culminating in the 2006 Nobel
Extra
Types of Networks

Interaction networks

Nodes: proteins or genes
Edges: interactions

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

Regulatory networks

Metabolic networks
More Information on this Talk

**TITLE:** Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

**SUBJECT:** Networks

**DESCRIPTION:**
National Academy of Engineering, Meeting at Columbia U, 2009.04.14, 14:00-14:30; [I: NAECU] (Short networks talk, incl. the following topics:
why networks w. amsci*, funnygene*, bottleneck*, nethierarchy*, metagenomics*, tyna* + topnet*, & pubnet*. Fits into 30’ w. 5’ questions. PPT works on mac & PC and has many photos w. EXIF tag kwtimewemet.)

(Paper references in the talk were mostly from Papers.GersteinLab.org. The above topic list can be easily cross-referenced against this website. Each topic abbrev. which is starred is actually a papers “ID” on the site. For instance, the topic pubnet* can be looked up at http://papers.gersteinlab.org/papers/pubnet)

**PERMISSIONS:** This Presentation is copyright Mark Gerstein, Yale University, 2008. Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html. Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).

**PHOTOS & IMAGES.** For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info. In particular, many of the images have particular EXIF tags, such as kwpotppt, that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt.