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

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)]
Hierarchies & DAGs of controlled-vocab terms but still have issues...

MIPS (Mewes et al.)

GO (Ashburner et al.)

[Seringhaus & Gerstein, Am. Sci. '08]
Towards Developing Standardized Descriptions of Function

- Subjecting each gene to standardized expt. and cataloging effect
  - KOs of each gene in a variety of std. conditions => phenotypes
  - Std. binding expts for each gene (e.g. prot. chip)

- Function as a vector

  | nucleic acids | small molecules | proteins |
  | DNA | RNA | ATP | Metal | CoA | NAD | G protein | CDC28 | Calmodulin |...... |
  | protein 1 | 1.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |...... |
  | protein 2 | 0 | 0.9 | 0 | 0 | 0 | 0 | 0 | 0 |...... |
  | protein 3 | 1.0 | 0 | 1.0 | 0 | 0 | 0 | 0 | 0 |...... |
  | protein 4 | 0 | 0 | 0 | 0 | 0.8 | 0 | 0 | 0 | 1.0 |...... |
  | protein 5 | 1.0 | 0 | 0 | 0 | 0 | 0 | 0.9 | 0 |...... |
  | protein 6 | 0.9 | 0 |...... |
  | protein 7 | 0 | 0.8 |...... |
  |...... |...... |...... |...... |...... |...... |...... |...... |...... |...... |

Interaction Vectors [Lan et al, IEEE 90:1848]
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

[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
Using the position in networks to describe function

Guilt by association

Finding the causal regulator (the "Blame Game")

[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

Part of the TCA cycle

Genetic interaction (synthetic lethal)
Signaling pathways
Outline: Molecular Networks

• Why Networks?
• Predicting Networks (yeast ppi)
  ◊ Propagating known information
• Central Points in Networks
  ◊ Hubs & Bottlenecks (yeast ppi & reg. net)
  ◊ Tops of Heirarchies (yeast reg. net)
  ◊ Identified by score (human miRNA-targ. net)
• Dynamics of Networks
  ◊ Across environments (in prokaryote metab. pathways)
• Protein Networks & Variation (human ppi & miRNA-targ. net)
Example: yeast PPI network

Actual size:

◊ ~6,000 nodes
→ Computational cost: ~18M pairs
◊ Estimated ~15,000 edges
→ Sparseness: 0.08% of all pairs (Yu et al., 2008)

Known interactions:

◊ Small-scale experiments: accurate but few
→ Overfitting: ~5,000 in BioGRID, involving ~2,300 proteins
◊ Large-scale experiments: abundant but noisy
→ Noise: false +ve/-ve for yeast two-hybrid data up to 45% and 90% (Huang et al., 2007)
Different Types of Molecular Networks

- Protein-protein Interaction networks
- TF-target-gene Regulatory networks
- Metabolic pathway networks
- miRNA-target networks

Predicting Networks

How do we construct large molecular networks? From extrapolating correlations between functional genomics data with fairly small sets of known interactions, making best use of the known training data.
Training sets

- **Known interactions**
- **Known non-interactions**
- **Unknown**

![Diagram with nodes and edges indicating known interactions, known non-interactions, and unknowns.](image-url)
Network prediction: features

• Example 1: gene expression

\[ x_1 = (0.2, 2.4, 1.5, \ldots) \]
\[ x_2 = (0.8, 2.2, 1.5, \ldots) \]
\[ x_3 = (4.3, 0.1, 7.5, \ldots) \]

\[ \sim(x_1, x_2) = 0.62 \]
\[ \sim(x_1, x_3) = -0.58 \]

Gasch et al., 2000
Network prediction: features

• Example 2: sub-cellular localization

\[ x_1 = (1, 1, 0, 0, \ldots) \]
\[ x_2 = (1, 1, 1, 0, \ldots) \]
\[ x_3 = (1, 0, 1, 0, \ldots) \]
\[ \ldots \]
\[ \text{sim}(x_1, x_2) = 0.81 \]
\[ \text{sim}(x_1, x_3) = 0.12 \]
\[ \ldots \]
Data integration & Similarity Matrix

![Graph with nodes and edges]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.57</td>
<td>0.55</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>0.57</td>
<td>1.00</td>
<td>0.66</td>
<td>0.89</td>
</tr>
<tr>
<td>3</td>
<td>0.55</td>
<td>0.66</td>
<td>1.00</td>
<td>0.79</td>
</tr>
<tr>
<td>4</td>
<td>0.40</td>
<td>0.89</td>
<td>0.79</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Learning methods

An endless list:

- Docking (e.g. Schoichet and Kuntz 1991)
- Evolutionary (e.g. Ramani and Marcotte, 2003)
- Topological (e.g. Yu et al., 2006)
- Bayesian (e.g. Jansen et al., 2003)
- **Kernel methods**
  - ◊ Global modeling:
    - em (Tsuda et al., 2003)
    - kCCA (Yamanishi et al., 2004)
    - kML (Vert and Yamanishi, 2005)
    - Pairwise kernel (Pkernel) (Ben-Hur and Noble, 2005)
  - ◊ Local modeling:
    - Local modeling (Bleakley et al., 2007)

**Let’s compare in a public challenge!**

*(DREAM: Dialogue for Reverse Engineering Assessment and Methods)*
DREAM3: *in silico* regulatory network reconstruction

<table>
<thead>
<tr>
<th>Actual network</th>
<th>Expression data</th>
<th>Modeling</th>
<th>Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Nodes and Connections]</td>
<td>![Graph with Time Series]</td>
<td>![Graph with Formula]</td>
<td>![Graph with Expression Rates]</td>
</tr>
</tbody>
</table>

**Deletion strains**

**Expression rate models**

\[
\text{Prob(}\text{signal} | \text{point}) = 2\Phi((\text{point} - \text{ref}) / s) - 1
\]

**Noise models**

\[
\frac{dy_j}{dt} = a_{j0} - a_{j1}y_j + \sum_{k \neq j} a_{j,k}y_k
\]

\[
\frac{dy_j}{dt} = \frac{b_{j0}}{1 + \exp\left(a_{j0} + \sum_{k \neq j} a_{j,k}y_k\right)} - b_{j1}y_j
\]

\[
\frac{dy_j}{dt} = a_{j0} \prod_{k \neq j} \left(\frac{b_{j,k}}{y_k + b_{j,k}}\right) \prod_{k \neq j} \left(\frac{y_k + b_{j,k}}{y_k + b_{j,k}}\right) - a_{j1}y_j
\]

**Accuracy (AUC)**

<table>
<thead>
<tr>
<th>Size</th>
<th>E. Coli 1</th>
<th>E. Coli 2</th>
<th>Yeast 1</th>
<th>Yeast 2</th>
<th>Yeast 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size-10</td>
<td>0.928</td>
<td>0.912</td>
<td>0.949</td>
<td>0.747</td>
<td>0.714</td>
</tr>
<tr>
<td>Size-50</td>
<td>0.930</td>
<td>0.924</td>
<td>0.917</td>
<td>0.792</td>
<td>0.805</td>
</tr>
<tr>
<td>Size-100</td>
<td>0.948</td>
<td>0.960</td>
<td>0.915</td>
<td>0.856</td>
<td>0.783</td>
</tr>
</tbody>
</table>

[Yip et al., DREAM3]
Our work: efficiently propagating known information

Training set expansion
- Motivation: lack of training examples
- Expand training sets horizontally

Multi-level learning
- Motivation: hierarchical nature of interaction
- Expand training sets vertically

DREAM3 in silico regulatory network reconstruction challenge
Kernels

Kernel: a similarity matrix that is positive semi-definite (p.s.d.)

![Diagram showing objects in a feature space and a similarity matrix]

- Objects in a feature space
- Compute inner products
- p.s.d. implies

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.72</td>
<td>0.45</td>
<td>-0.56</td>
</tr>
<tr>
<td>2</td>
<td>0.72</td>
<td>1.00</td>
<td>-0.30</td>
<td>-0.98</td>
</tr>
<tr>
<td>3</td>
<td>0.45</td>
<td>-0.30</td>
<td>1.00</td>
<td>0.49</td>
</tr>
<tr>
<td>4</td>
<td>-0.56</td>
<td>-0.98</td>
<td>0.49</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Good for integrating heterogeneous datasets (protein sequences, PSSM, gene expression, …)
– no need to explicitly place them in a common feature space
Kernel methods

Use the kernel as proxy to work in the feature space

Example: SVM (finding the best separating hyperplane)

Maximize $\sum \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle$

Subject to $\lambda \geq 0$

$\sum_i \lambda_i y_i = 0$

The only thing that we need to know about the objects: their similarity values (inner products)
Kernel methods for predicting networks: local vs. global modeling

Global modeling: build one model for the whole network

Pairwise kernel: consider object pairs instead of individual objects
Problem: $O(n^2)$ instances, $O(n^4)$ kernel elements

Direct methods: threshold the kernel to make predictions
Problem: One single global model, may not be able to handle subclasses
Kernel methods for predicting networks: local vs. global modeling

Local modeling: build one model for each node

Problem: insufficient and unevenly distributed training data (what if node 3 has no known interactions at all?)
Our work: training set expansion

• Goal:
  ◊ Utilize the flexibility of local modeling
  ◊ Tackle the problem of insufficient training data

• Idea: generate auxiliary training data
  ◊ Prediction propagation
  ◊ Kernel initialization

[Yip and Gerstein, Bioinformatics ('09, in press)]
Prediction propagation

- Motivation: some objects have more examples than others
- Our approach:
  ◊ Learn models for objects with more examples first
  ◊ Propagate the most confident predictions as auxiliary examples of other objects

[Yip and Gerstein, Bioinformatics ('09, in press)]
Kernel initialization

- Motivation: what if most objects have very few examples?
- Our approach (inspired by the direct method):
  - Add the most similar pairs in the kernel as positive examples
  - Add the most dissimilar pairs in the kernel as negative examples

[Yip and Gerstein, Bioinformatics ('09, in press)]
Remarks

• Can be used in combination
• Prediction propagation theoretically related to co-training (Blum and Mitchell, 1998)
  ◊ Semi-supervised
    • Similarity with PSI-BLAST
• Algorithm complexity $O(nf(n))$ of local modeling vs. $O(f(n^2))$ of global modeling

[Yip and Gerstein, Bioinformatics ('09, in press)]
# Prediction accuracy (AUC)

<table>
<thead>
<tr>
<th>Mode 1</th>
<th>phy</th>
<th>loc</th>
<th>exp-gasch</th>
<th>exp-spellman</th>
<th>y2h-ito</th>
<th>y2h-uetz</th>
<th>tap-gavin</th>
<th>tap-krogan</th>
<th>int</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct</td>
<td>58.04</td>
<td>66.55</td>
<td>64.61</td>
<td>57.41</td>
<td>51.52</td>
<td>52.13</td>
<td>59.37</td>
<td>61.62</td>
<td>70.91</td>
</tr>
<tr>
<td>kCCA</td>
<td>65.80</td>
<td>63.86</td>
<td>68.98</td>
<td>65.10</td>
<td>50.89</td>
<td>50.48</td>
<td>57.56</td>
<td>51.85</td>
<td>80.98</td>
</tr>
<tr>
<td>kML</td>
<td>63.87</td>
<td>68.10</td>
<td>69.67</td>
<td>68.99</td>
<td>52.76</td>
<td>53.86</td>
<td>60.86</td>
<td>57.69</td>
<td>73.47</td>
</tr>
<tr>
<td>em</td>
<td>71.22</td>
<td>75.14</td>
<td>67.53</td>
<td>64.96</td>
<td>55.90</td>
<td>53.13</td>
<td>63.74</td>
<td>58.20</td>
<td>81.55</td>
</tr>
<tr>
<td>local</td>
<td>71.67</td>
<td>71.41</td>
<td>72.66</td>
<td>70.63</td>
<td>67.27</td>
<td>67.27</td>
<td>64.60</td>
<td>67.48</td>
<td>75.65</td>
</tr>
<tr>
<td>local+pp</td>
<td>73.89</td>
<td>75.25</td>
<td>77.43</td>
<td>75.35</td>
<td>71.60</td>
<td>71.51</td>
<td>74.62</td>
<td>71.39</td>
<td>83.63</td>
</tr>
<tr>
<td>local+ki</td>
<td>71.68</td>
<td>71.42</td>
<td>75.89</td>
<td>70.96</td>
<td>69.40</td>
<td>69.05</td>
<td>70.53</td>
<td>72.03</td>
<td>81.74</td>
</tr>
<tr>
<td>local+pp+ki</td>
<td>72.40</td>
<td>75.19</td>
<td>77.41</td>
<td>73.81</td>
<td>70.44</td>
<td>70.57</td>
<td>73.59</td>
<td>72.64</td>
<td>83.59</td>
</tr>
</tbody>
</table>

**Observations:**

- Highest accuracy by training set expansion
- Over fitting of local modeling without training set expansion
- Prediction propagation theoretically related to co-training (Blum and Mitchell, 1998)
  - ◊ Semi-supervised (Similarity with PSI-BLAST)

[Yip and Gerstein, Bioinformatics ('09)]
Complementarity of the two methods

[Yip and Gerstein, Bioinformatics (‘09, in press)]
From horizontal to vertical

Training set expansion
• Motivation: lack of training examples
• Expand training sets horizontally

Multi-level learning
• Motivation: hierarchical nature of interaction
• Expand training sets vertically
Yeast NADP-dependent alcohol dehydrogenase 6 (PDB: 1piw)

Protein-level features for interaction prediction: functional genomic information

[Yip and Gerstein, BMC Bioinfo. ('09, press)]
Domain interaction

Pfam domains: PF00107 (inner) and PF08240 (outer)

Domain-level features for interaction prediction: evolutionary information

[Yip and Gerstein, BMC Bioinfo. ('09, press)]
Residue interaction

Interacting residues: 283 (yellow) with 287 (cyan), and 285 (purple) with 285

Residue-level features for interaction prediction: physical-chemical information

[Yip and Gerstein, BMC Bioinfo. ('09, press)]
Combining the three problems

- Protein interactions
- Domain interactions
- Residue interactions

i. Independent levels
ii. Unidirectional flow
iii. Bidirectional flow

[Yip and Gerstein, BMC Bioinfo. ('09, press)]
Empirical results (AUCs)

<table>
<thead>
<tr>
<th></th>
<th>Ind. levels</th>
<th>Unidirectional flow</th>
<th>Bidirectional flow</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PD</td>
<td>PR</td>
</tr>
<tr>
<td>Level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proteins</td>
<td>71.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domains</td>
<td>53.18</td>
<td>61.51</td>
<td></td>
</tr>
<tr>
<td>Residues</td>
<td>57.36</td>
<td></td>
<td>54.89</td>
</tr>
</tbody>
</table>

- Highest accuracy by bidirectional flow
- Additive effect: 2 vs. 3 levels

[Yip and Gerstein, BMC Bioinfo. ('09, press)]
Finding Central Points in Networks: Hubs & Bottlenecks

Where are key points networks? How do we locate them?
Global topological measures

Indicate the gross topological structure of the network

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]
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)]
Bottlenecks are what matters in regulatory networks

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

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

Finding Central Points in Networks #2: Tops of the Hierarchy

Where are key points networks? How do we locate them?
Social Hierarchy

THE GOVERNMENT OF THE UNITED STATES

LEGISLATIVE BRANCH

THE CONGRESS
SENATE
HOUSE
ARCHITECT OF THE CAPITOL
UNITED STATES BOTANIC GARDEN
GENERAL ACCOUNTING OFFICE
GOVERNMENT PRINTING OFFICE
LIBRARY OF CONGRESS
CONGRESSIONAL BUDGET OFFICE

EXECUTIVE BRANCH

PRESIDENT
WHITE HOUSE OFFICE
OFFICE OF THE VICE PRESIDENT
COUNCIL OF ECONOMIC ADVISORS
COUNCIL ON ENVIRONMENTAL QUALITY
NATIONAL SECURITY COUNCIL
OFFICE OF ADMINISTRATION

JUDICIAL BRANCH

THE SUPREME COURT OF THE UNITED STATES
UNITED STATES COURTS OF APPEALS
UNITED STATES DISTRICT COURTS
TERITORIAL COURTS
UNITED STATES COURT OF INTERNATIONAL TRADE
UNITED STATES COURT OF FEDERAL CLAIMS
UNITED STATES COURT OF APPEALS FOR THE ARMED FORCES
UNITED STATES TAX COURT
UNITED STATES COURT OF APPEALS FOR VETERANS CLAIMS
ADMINISTRATIVE OFFICE OF THE UNITED STATES COURTS
FEDERAL JUDICIAL CENTER
UNITED STATES SENTENCING COMMISSION

INDEPENDENT ESTABLISHMENTS AND GOVERNMENT CORPORATIONS

AFRICAN DEVELOPMENT FOUNDATION
CENTRAL INTELLIGENCE AGENCY
COMMODITY FUTURES TRADING COMMISSION
CONSUMER PRODUCT SAFETY COMMISSION
CORPORATION FOR NATIONAL AND COMMUNITY SERVICE
DEFENSE NUCLEAR FACILITIES SAFETY BOARD
ENVIRONMENTAL PROTECTION AGENCY
EQUAL EMPLOYMENT OPPORTUNITY COMMISSION
EXECUTIVE OFFICE OF THE PRESIDENT
FARM CREDIT ADMINISTRATION
FEDERAL COMMUNICATIONS COMMISSION
FEDERAL DEPOSIT INSURANCE CORPORATION
FEDERAL ELECTION COMMISSION
FEDERAL HOME LOAN BANK BOARD
FEDERAL LABOR RELATIONS AUTHORITY
FEDERAL MARITIME COMMISSION
FEDERAL MEDICAL AND CONCILIATION SERVICE
FEDERAL MINERAL SAFETY AND HEALTH REVIEW COMMISSION
FEDERAL RESERVE SYSTEM
FEDERAL RETIREMENT THRIFT INVESTMENT BOARD
FEDERAL TRADE COMMISSION
GENERAL SERVICES ADMINISTRATION
INTERAMERICAN FOUNDATION
MERIT SYSTEMS PROTECTION BOARD
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
NATIONAL ARCHIVES AND RECORDS ADMINISTRATION
NATIONAL CAPITAL PLANNING COMMISSION
NATIONAL CREDIT UNION ADMINISTRATION
NATIONAL FOUNDATION ON THE ARTS AND THE HUMANITIES
NATIONAL LABOR RELATIONS BOARD
NATIONAL MEDIATION BOARD
NATIONAL RAILROAD PASSENGER CORPORATION (AMTRAK)
NATIONAL SCIENCE FOUNDATION
NATIONAL TRANSPORTATION SAFETY BOARD
NUCLEAR REGULATORY COMMISSION
OCCUPATIONAL SAFETY AND HEALTH REVIEW COMMISSION
OFFICE OF GOVERNMENT ETHICS
OFFICE OF PERSONNEL MANAGEMENT
OFFICE OF SPECIAL COUNSEL
OVERSEAS PRIVATE INVESTMENT CORPORATION
PEACE CORPS
PENSION BENEFIT GUARANTY CORPORATION
POSTAL RATE COMMISSION
RAILROAD RETIREMENT BOARD
SECURITIES AND EXCHANGE COMMISSION
SELECTIVE SERVICE SYSTEM
SMALL BUSINESS ADMINISTRATION
SOCIAL SECURITY ADMINISTRATION
TENNESSEE VALLEY AUTHORITY
TRADE AND DEVELOPMENT AGENCY
U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT
U.S. COMMISSION ON CIVIL RIGHTS
U.S. INTERNATIONAL TRADE COMMISSION
U.S. POSTAL SERVICE
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)

Level 1

Level 2

Level 3

[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)]
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.
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)]
Finding Central Points in Networks #3: Points of Maximal Regulatory Effect
• How much does a regulator influence its targets?
• For micro-RNA-target networks easy to calculate, as all influence is down-regulation
  ◊ target prediction methods: TargetScan, PITA, PicTar, miRanda, …
• Look at down-reg. genes in a sample & compare with targets of a specific micro-RNA
  ◊ more down-reg genes => stronger regulatory effect

**RE-score: Another way to measure "importance" in networks**
Application of RE-score to measure changing miRNA effect in different conditions (ER- and ER+ breast cancer)

Cheng et al., Genome Biology, 2009
miRNA RE-scores can be used to classify cancers

RE-score profiles for 8 miRNAs
Differential expression of miRNA processing genes

The majority of miRNAs have higher RE-score in ER- than in ER+

Distribution of ER-/ER+ T-scores for all miRNAs

(a)

(b)
Network Dynamics #2: Environments

How do molecular networks change across environments?
What pathways are used more?
Used as a biosensor?
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
ATCGTGTAGATGATGATGAG
ATCGTGTAGATGATGATGAG
ATCGTGTAGATGATGATGAG
ATCGTGTAGATGATGATGAG
ATCGTGTAGATGATGATGAG
ATCGTGTAGATGATGATGAG
ATCGTGTAGATGATGATGAG
ATCGTGTAGATGATGATGAG
ATCGTGTAGATGATGATGAG
ATCGTGTAGATGATGATGAG

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

\[ UPI = a \cdot \text{GRE} + b \cdot \text{GPA} + c \]

\[ GPI = a' \cdot \text{GRE} + b' \cdot \text{GPA} + c' \]

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

Score | # of papers published
---|---
GRE | 

Undergraduate Performance Index (UPI) | Graduate School Performance Index (GPI)
---|---
GRE | GPA

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 $Y = \{y_1, y_2, \ldots, y_m\}$

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

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

[ 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) ]
Conclusion #3: Covariation of AA biosynthesis and Import

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) ]
Networks & Variation

Which parts of the network vary most in sequence? Which are under selection, either positive or negative?
METHODOLOGY: MAP SNP AND CNV DATA ONTO ENSEMBL GENES, AND THEN MAP ENSEMBL GENES TO THE KNOWN INTERACTOME


Source: PMK
ADAPTIVE EVOLUTION CAN BE SEEN ON TWO DIFFERENT LEVELS

Intra-species variation

- Single-basepair
  - Positive Selection

Single-Nucleotide Polymorphisms

Structural variation

- Positive Selection
- Copy Number Variants
- Segmental Duplications

Fixed mutations
(differences to other species)

- Fixed Differences

Source: PMK
POSITIVE SELECTION LARGELY TAKES PLACE AT THE NETWORK PERIPHERY

Positive selection in the human interactome

CENTRAL PROTEINS ARE LESS LIKELY TO BE UNDER POSITIVE SELECTION

- Peripheral genes are likely to under positive selection, whereas hubs aren’t
- This is likely due to the following reasons:
  - Hubs have stronger structural constraints, the network periphery doesn’t
  - Most recently evolved functions (e.g. “environmental interaction genes” such as sensory perception genes etc.) would probably lie in the network periphery
- Effect is independent of any bias due to gene expression differences

*With a probability of over 80% to be positively selected as determined by Ka/Ks. Other tests of positive selection (McDonald Kreitmann and LDD) corroborate this result.

CENTRAL NODES ARE LESS LIKELY TO LIE INSIDE OF SDs

• This result also confirms our initial hypothesis – peripheral nodes tend to lie in regions rich in SDs.

• Since segmental duplications are a different mechanism of ongoing evolution, the less constrained peripheral proteins are enriched in them.

• Note that despite the small size of our dataset for known SD’s we get significant correlations. It is to be expected that the correlations will get clearer as more data emerges.*

* Specifically, a number of the SDs are likely not fixed, but rather common CNVs in the reference genome
Why do we observe this? Perhaps central hub proteins are involved in more interactions & have more surface buried.

**BURIED SITES ARE CONSERVED AND MUCH LESS LIKELY TO HARBOR NON-SYNONYMOUS MUTATIONS**

<table>
<thead>
<tr>
<th></th>
<th>Exposed sites</th>
<th>Buried sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>dN/dS Ratio</td>
<td>0.49</td>
<td>0.35</td>
</tr>
<tr>
<td>p</td>
<td>&lt;&lt;0.01</td>
<td></td>
</tr>
</tbody>
</table>

- **Average Relative Surface Exposure**

<table>
<thead>
<tr>
<th></th>
<th>Site with Synonymous Mutations only</th>
<th>Sites with Non-synonymous Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.26</td>
<td>2.66</td>
</tr>
<tr>
<td>p</td>
<td>&lt;&lt;0.01</td>
<td></td>
</tr>
</tbody>
</table>

Source: Kim et al. PNAS (2007)
**Another explanation: THE NETWORK PERIPHERY CORRESPONDS TO THE CELLULAR PERIPHERY**

<table>
<thead>
<tr>
<th>Region</th>
<th>Betweenness Centrality ($x 10^4$)</th>
<th>Degree Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome</td>
<td>5.5</td>
<td>10</td>
</tr>
<tr>
<td>Nucleus</td>
<td>5.0</td>
<td>8.6</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>5.2</td>
<td>8.1</td>
</tr>
<tr>
<td>Membrane</td>
<td>4.0</td>
<td>6.5</td>
</tr>
<tr>
<td>Extracellular Region</td>
<td>3.8</td>
<td>5.9</td>
</tr>
</tbody>
</table>

Source: Gandhi et al. (*Nature Genetics* 2006), Kim et al. PNAS (2007)
IS RELAXED CONSTRAINT OR ADAPTIVE EVOLUTION THE REASON FOR THE PREVALENCE OF BOTH SELECTED GENES AND SDs AT THE NETWORK PERIPHERY?

**ILLUSTRATIVE**

<table>
<thead>
<tr>
<th>Relaxed Constraint</th>
<th>Adaptive Evolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-Species Variation (Fixed differences)</td>
<td></td>
</tr>
<tr>
<td>• Increases inter-species variation – more variable loci are under less negative selection</td>
<td>• Increases inter-species variation – more variable loci are under less negative selection</td>
</tr>
<tr>
<td>• Can be seen in higher Ka/Ks ratio or SD occurrence</td>
<td>• Can be seen in higher Ka/Ks ratio or SD occurrence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intra-Species Variation (Polymorphisms)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Increases intra-species variation – for the very same reason</td>
<td>• Should not have effects on intra-species variation</td>
</tr>
<tr>
<td>• Can be seen in both SNPs or CNVs</td>
<td></td>
</tr>
</tbody>
</table>

Source: Kim et al. PNAS (2007)
SOME, BUT NOT ALL OF THE SINGLE-BASEPAIR SELECTION AT THE PERIPHERY IS DUE TO RELAXED CONSTRAINT

Inter vs. Intra-Species Variation in Networks

Inter-Species (Fixed differences)

Betweenness Centrality ($x \times 10^4$)

- Genes with $dN/dS>1$
- Genes with $dN/dS<1$

Betweenness Centrality ($x \times 10^4$)

- Genes with $pN/pS>1$
- Genes with $pN/pS<1$

Reasoning

- There is a difference in **variability** (in terms of SNPs) between the network periphery and the center

- However, this difference is much smaller than the difference in **selection**

- This most likely means, that part of the effect we’re seeing is due to relaxed constraint (and higher variability)

- But, not the entire effect*

* But it’s hard to quantify

Source: Kim et al. (2007) PNAS
Similar Results for Large-scale Genomic Changes (CNVs and SDs)

Inter vs. Intra-Species Variation in Networks

**Inter-Species (SDs)**

Betweenness Centrality ($x 10^4$)

<table>
<thead>
<tr>
<th>Genes intersecting SDs</th>
<th>All others</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.61</td>
<td>4.18</td>
</tr>
</tbody>
</table>

p < 0.01

**Intra-Species (CNVs) [ Variability ]**

Betweenness Centrality ($x 10^4$)

<table>
<thead>
<tr>
<th>Genes intersecting CNVs</th>
<th>All others</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.25</td>
<td>4.20</td>
</tr>
</tbody>
</table>

p < 0.01

Reasoning

- There a small difference in **variability** (in terms of CNVs) between the network periphery and the center.

- But, there is a (as shown before) marked difference in fixed (and hence, presumably, selected) SDs at the network periphery and center.

Source: Kim et al. (2007) PNAS
Networks & Variation 2

Variation in the miRNA network
Analyze Regulation in microRNA-target Network

- Relationship between target in degree (number of micro-RNAs that regulate gene) & evolutionary rate of gene?
  ◊ In deg. related 3' UTR size

- Expectation: more regulation, more constraint
Relationship between microRNA regulation and protein evolution

Important genes are regulated more intensively regulated by the microRNAs

<table>
<thead>
<tr>
<th>Human vs.</th>
<th>Number of genes</th>
<th>Correlation</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>chimpanzee</td>
<td>11326</td>
<td>-0.11</td>
<td>2.E-32</td>
</tr>
<tr>
<td>mouse</td>
<td>13280</td>
<td>-0.21</td>
<td>7.E-128</td>
</tr>
<tr>
<td>rat</td>
<td>12270</td>
<td>-0.20</td>
<td>4.E-107</td>
</tr>
<tr>
<td>cow</td>
<td>11683</td>
<td>-0.21</td>
<td>8.E-115</td>
</tr>
<tr>
<td>chicken</td>
<td>8061</td>
<td>-0.18</td>
<td>1.E-57</td>
</tr>
</tbody>
</table>

[Cheng et al., BMC Genomics, 2009 (in press)]
MicroRNA regulation: a two-way strategy

For non-housekeeping genes, functionally critical genes are intensively regulated by miRNAs and prefer long 3'UTR.

Housekeeping genes, however conserved, are selected to have shorter 3'UTRs to avoid miRNA regulation.

[Cheng et al., BMC Genomics, 2009 (in press)]
Network dynamics constrain evolution

Hypothesis: Nodes in a molecular network with the strongest impact on dynamic behavior should be under strong purifying selection and thus exhibit the least genetic variation.

Network dynamics constrain evolution

Hypothesis: Nodes in a molecular network with the strongest impact on dynamic behavior should be under strong purifying selection and thus exhibit the least genetic variation.

Algorithm:
1) Reconstruct families of molecular networks from genomic data.
2) Map some kind of genetic variation onto the networks.
3) Analyze sensitivity of dynamical model of the generic network.

Speculation: Why more tightly regulated gene might have less variation

Example: MAP Kinase signaling pathway

Dynamic model:
- ODE model with Michaelis-Menten kinetics
- parameters fit to time series data of protein activities in response to EGF and NGF from rat PC12 cell line

In sensitivity analysis, stiff parameters cluster around Ras and Raf.

Population study in fruit flies:
- allele variation based on PCR of pathway genes

Ras and Raf have less allele variation than other proteins in the network.


Analogies show it reasonable for more variable part of network to be periphery

• **Computer Networks**
  – Servers in center have much depending on them; thus, can't be frequently updated & patched
  – Servers on periphery often attacked and so need frequent patches

• **Social Networks**
  – Individuals at center under more constraint (to conform), whereas those at periphery have more freedom to experiment
Outline: Molecular Networks

• Why Networks?
• Predicting Networks (yeast ppi)
  ◊ Propagating known information
• Central Points in Networks
  ◊ Hubs & Bottlenecks (yeast ppi & reg. net)
  ◊ Tops of Hierarchies (yeast reg. net)
  ◊ Identified by score (human miRNA-targ. net)
• Dynamics of Networks
  ◊ Across environments (in prokaryote metab. pathways)
• Protein Networks & Variation (human ppi & miRNA-targ. net)
Conclusions on Networks: Predictions

• Predicting Networks
  ◊ Extrapolating from the Training Set
  ◊ Principled ways of using known information in the fullest possible fashion
    • Prediction Propagation
    • Multi-level learning
Conclusions: Analysis of Network Structure

• Centrality Measures in Protein Network
  ◊ Hubs & Bottlenecks
  ◊ Importance of later 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: Points of Network Centrality

- RE-score measures degree of (down) regulation of targets vs. non-targets
- Application to miRNA network
- Different RE-score of miRNAs can be used in cancer classification
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.
Conclusions: Connecting Networks & Variation

• We find ongoing evolution (positive selection) at the network periphery.
  ◊ This trend is present on two levels:
    • On a sequence level, it can be seen as positive selection of peripheral nodes
    • On a structural level, it can be seen as the pattern of SDs that display significantly higher allele frequencies in non-central genes
  ◊ 2 possible mechanisms for this: adaptive evolution at cellular periphery & relaxation of structural constraints at the network periphery
  • We show that the latter can only explain part of the increased variability
Conclusions: Connecting Networks & Variation 2

- More highly regulated genes are under more constraint in miRNA-target networks
- Exception for housekeeping genes
- Speculation as to why variation at periphery is quite reasonable
TopNet – 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

H Yu
P Kim
K Yip
T Gianoulis
C Cheng

A Paccanaro
P Alves
T Emonet
P Cayting
M Seringhaus
Y Xia
J Korbel
A Sboner
P Patel
P Bork
J Raes
E Franzosa
M Snyder
N Bhardwaj
R Alexander

Networks.GersteinLab.org
Job opportunities currently for postdocs & students
More Information on this Talk

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

SUBJECT: Networks

DESCRIPTION:
Network Biology: Understanding metabolic and protein interactions, Mathematical Biosciences Institute, Columbus, OH; 2009.09.14, 13:30-14:30; [I:MBINETS] (Long networks talk, adding in for the first time: rescore*, mirnatargevolrate* & netdynamicsrev*. Fits easily into 55’ w. 5’ questions. PPT works on mac & PC and has many photos.)

(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.