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)
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

<table>
<thead>
<tr>
<th></th>
<th>nucleic acids</th>
<th>small molecules</th>
<th>proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>RNA</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>ATP</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Metal</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>CoA</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>NAD</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>G protein</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>CDC28</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Calmodulin</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

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
- Delta
- Serrate
- PSE2
- PSEN
- NCSTN
- APH-1
- TACE
- CSL

Same Genes in High-throughput Network

- Itch: linked to itchy skin in mice
- DLK1
- Dvl
- GSK3B
- ITCH
- Numb
- LCK
- CNTN1
- BXW7
- CTNNB1
- APP
- SMAD3
- PCAF
- YY1
- RELA
- APBA1
- G22P1
- CSL
- RBPMS
- CSNK2A1
- CSNK2A2
- EPS8
- AP2A
- TP53
- MDM2
- GRB2
- MAML

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

1D: Complete Genetic Partslis

~2D: Bio-molecular Network Wiring Diagram

3D: Detailed structural understanding of cellular machinery

Networks as a universal language
Networks as a Central Theme in Systems Biology

Reductionist Approach

Integrative Approach

[Adapted from H Yu]
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
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
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

• Predicting Networks
  ◊ Training set expansion

• Properties of Protein Networks
  ◊ Hubs

• Dynamics of Networks
  ◊ Dynamics across cellular states
  ◊ Dynamics across environments

• Protein Networks and Human Variation
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.
**Network Prediction**

- Only small portions are already known
- Many other kinds of data available

→ Use them to learn models for predicting the unknown portions

**Ex. of Predicted Membrane Protein Interactome in Xia et al. JMB (2006)**

**Figure 6:** A map of known and a subset of predicted interactions among helical membrane proteins. Nodes represent helical membrane proteins, and edges represent interactions among them. Red edges represent known interactions that are also predicted.
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)
Learning

Concepts in machine learning:

• Training sets:
  ◊ Positive set: known interactions
  ◊ Negative set: known non-interactions

• Features:
  ◊ Data describing the objects

• Model:
  ◊ A function that takes two objects as input and predicts whether they interact
# Training sets

![Diagram of known interactions and non-interactions]

<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>0</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>?</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
<td>0</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

- **Known interactions**
- **Known non-interactions**
- **Unknown**
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
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 and table showing data integration and similarity matrix.](image)
Evaluation

• Computational:
  ◊ Cross-validation
  ◊ Indirect evidence (e.g. same GO category)

• Experimental:
  ◊ Validation of de novo predictions
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 fairly in a public challenge! (DREAM)
Kernels

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

Objects in an feature space

Compute inner products

p.s.d. implies

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 \[ \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle \]

Subject to \[ \lambda_i \geq 0 \]
\[ \sum_i \lambda_i y_i = 0 \]

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

Local modeling: build one model for each node

Model for node 3:

Problem: insufficient and unevenly distributed training data (what if node 3 has no known interactions at all?)
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
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)]
Experiments

- Gold-standard interactions: BioGRID, from studies that report less than 10 interactions
- Features:

<table>
<thead>
<tr>
<th>Code</th>
<th>Data type</th>
<th>Source</th>
<th>Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>phy</td>
<td>Phylogenetic profiles</td>
<td>COG v7 (Tatusov et al., 1997)</td>
<td>RBF ($\sigma=3,8$)</td>
</tr>
<tr>
<td>loc</td>
<td>Sub-cellular localization</td>
<td>(Huh et al., 2003)</td>
<td>Linear</td>
</tr>
<tr>
<td>exp-gasch</td>
<td>Gene expression (environmental response)</td>
<td>(Gasch et al., 2000)</td>
<td>RBF ($\sigma=3,8$)</td>
</tr>
<tr>
<td>exp-spellman</td>
<td>Gene expression (cell-cycle)</td>
<td>(Spellman et al., 1998)</td>
<td>RBF ($\sigma=3,8$)</td>
</tr>
<tr>
<td>y2h-ito</td>
<td>Yeast two-hybrid</td>
<td>(Ito et al., 2000)</td>
<td>Diffusion ($\beta=0.01$)</td>
</tr>
<tr>
<td>y2h-uetz</td>
<td>Yeast two-hybrid</td>
<td>(Uetz et al., 2000)</td>
<td>Diffusion ($\beta=0.01$)</td>
</tr>
<tr>
<td>tap-gavin</td>
<td>Tandem affinity purification</td>
<td>(Gavin et al., 2006)</td>
<td>Diffusion ($\beta=0.01$)</td>
</tr>
<tr>
<td>tap-krogan</td>
<td>Tandem affinity purification</td>
<td>(Krogan et al., 2006)</td>
<td>Diffusion ($\beta=0.01$)</td>
</tr>
<tr>
<td>int</td>
<td>Integration</td>
<td></td>
<td>Summation</td>
</tr>
</tbody>
</table>

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

<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>53.85</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>68.20</td>
<td>81.65</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
- Overfitting of local modeling without training set expansion
- Comparing prediction propagation and kernel initialization

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

[Yip and Gerstein, Bioinformatics ('09, in press)]
Network Dynamics #1: Cellular States

How do networks change across different cellular states?
How can this be used to assign function to a protein?
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".

[Yu et al., 2003, TIG]
Dynamic Yeast TF network

- Analyzed network as a static entity
- But network is *dynamic*
  - Different sections of the network are active under different cellular conditions
- Integrate gene expression data

Luscombe et al. Nature 431: 308
Gene expression data for five cellular conditions in yeast

<table>
<thead>
<tr>
<th>Cellular condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell cycle</td>
</tr>
<tr>
<td>Sporulation</td>
</tr>
<tr>
<td>Diauxic shift</td>
</tr>
<tr>
<td>DNA damage</td>
</tr>
<tr>
<td>Stress response</td>
</tr>
</tbody>
</table>

[Brown, Botstein, Davis....]
Backtracking to find active sub-network

- Define differentially expressed genes
- Identify TFs that regulate these genes
- Identify further TFs that regulate these TFs

Active regulatory sub-network
Network usage under different conditions

static

Luscombe et al. Nature 431: 308
Network usage under different conditions

cell cycle
Network usage under different conditions

sporulation
Network usage under different conditions

diauxic shift
Network usage under different conditions

DNA damage
Network usage under different conditions

stress response
Network usage under different conditions

SANDY:

1. Standard graph-theoretic statistics:
   - Global topological measures
   - Local network motifs

2. Newly derived follow-on statistics:
   - Hub usage
   - Interaction rewiring

3. Statistical validation of results

Luscombe et al. Nature 431: 308
Network usage under different conditions

Cell cycle  Sporulation  Diauxic shift  DNA damage  Stress

SANDY:
1. Standard graph-theoretic statistics:
   - Global topological measures
     - Local network motifs

2. Newly derived follow-on statistics:
   - Hub usage
     - Interaction rewiring

3. Statistical validation of results
Analysis of condition-specific subnetworks in terms of global topological statistics

<table>
<thead>
<tr>
<th></th>
<th>Cell cycle</th>
<th>Sporulation</th>
<th>Diauxic shift</th>
<th>DNA damage</th>
<th>Stress response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outdegree</td>
<td>7.9</td>
<td>6.5</td>
<td>17.1</td>
<td>15.0</td>
<td>9.0</td>
</tr>
<tr>
<td>Indegree</td>
<td>2.0</td>
<td>1.9</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>Pathlength</td>
<td>4.5</td>
<td>3.4</td>
<td>2.1</td>
<td>2.0</td>
<td>2.2</td>
</tr>
<tr>
<td>Clustering coefficient</td>
<td>0.15</td>
<td>0.14</td>
<td>0.09</td>
<td>0.09</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Multi-stage
Controlled, ticking over of genes at different stages

Binary
Quick, large-scale turnover of genes

Luscombe et al. Nature 431: 308
Analysis of condition-specific subnetworks in terms of occurrence of local motifs

- **Single-input module**
  - 32% Sporulation
  - 39% Cell cycle

- **Multi-input module**
  - 24% Sporulation
  - 17% Cell cycle

- **Feed-forward loop**
  - 44% Sporulation
  - 45% Cell cycle

- **Sporulation**
  - 57%
  - 56%
  - 59%

- **Cell cycle**
  - 24%
  - 27%
  - 20%

- **Diauxic shift**
  - 19%

- **DNA damage**
  - 17%

- **Stress response**
  - 21%

Multi-stage controlled, ticking over of genes at different stages

Binary quick, large-scale turnover of genes

Luscombe et al. Nature 431: 308
Summary

- multi-stage conditions: less pronounced, longer, more complex loops (FFLs)
- binary conditions: more pronounced, shorter, less simpler (SIMs)
Transient Hubs

Questions:
◊ Do hubs stay the same or do they change over between conditions?
◊ Do different TFs become important?

Our Expectations
◊ Literature:
  • Hubs are permanent features of the network regardless of condition
◊ Random networks (sampled from complete regulatory network)
  • Random networks converge on same TFs
  • 76-97% overlap in TFs classified as hubs (ie hubs are permanent)

Luscombe et al. Nature 431: 308
- Some permanent hubs
  ◊ house-keeping functions

- Most are transient hubs
  ◊ Different TFs become key regulators in the network

- Implications for condition-dependent vulnerability of network

Luscombe et al. Nature 431: 308
transient hubs
permanent hubs

Swi4, Mbp1
Ime1, Ume6
Msn2, Msn4

cell cycle
sporulation
diauxic shift
DNA damage
stress response
all conditions

Luscombe et al. Nature 431: 308
Unknown functions

- Cell cycle
- Sporulation
- Diauxic shift
- DNA damage
- Stress response
- All conditions

Luscombe et al. Nature 431: 308
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

PROBLEM: Estimated that less than 1% can be cultured in the lab

Metagenomics Approach

Collect Sample → Extract DNA → Sequence

PROBLEM: Lose information about which gene belongs to which microbe.
Do the proportions of pathways represented in these two samples differ?

Dinsdale et. al., Nature 2008
Do the proportions of pathways represented in these two samples CHANGE as a function of their environments?
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
Canonical Correlation Analysis: Simultaneous weighting

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

\[
\text{GPI} = a' \text{GRE} + b' \text{PowerPoint} + c' \text{money}
\]

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

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

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

[ 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) ]
**Conclusion #4: Cofactor (Metal) Optimization**

**IS DEPENDENT-ON**
- Methionine synthesis
  - Cobalamin biosynthesis
  - Cobalt transporters

**IS NEEDED FOR**
- Methionine degradation
  - S-adenosyl Methionine Biosynthesis
    (synthesize SAM one of the most important methyl donors)
  - Polyamine biosynthesis

**RELIES ON**
- Methionine Salvage
  - Spermidine/Putrescine transporters
  - Arg/His/Ornithine transporters

---

[ 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
- Fixed mutations (differences to other species)
  - Copy Number Variants
  - Segmental Duplications

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

Centrality vs. SD occurrence

- 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

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

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?

Source: Kim et al. PNAS (2007)

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

<table>
<thead>
<tr>
<th>Intra-Species Variation (Polymorphisms)</th>
<th>Relaxed Constraint</th>
<th>Adaptive Evolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Increases intra-species variation – for the very same reason</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Can be seen in both SNPs or CNVs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Should not have effects on intra-species variation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Some, but not all of the single-basepair selection at the periphery is due to relaxed constraint.

- 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 ($\times 10^4$)
  - Genes intersecting SDs: 2.61
  - All others: 4.18
  - $p < 0.01$

**Intra-Species (CNVs) [Variability]**
- Betweenness Centrality ($\times 10^4$)
  - Genes intersecting CNVs: 3.25
  - All others: 4.20
  - $p < 0.01$

**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
Conclusions: Net Intro. + Predicting Networks

• Developing Standardized Descriptions of Protein Function
  ◊ Gene Naming

• Predicting Networks
  ◊ Extrapolating from the Training Set
  ◊ Principled ways of using the training set data in the fullest possible fashion
    • Prediction Propagation
    • Kernel Initialization
Conclusions: Network Dynamics across Cellular States

- Merge expression data with Networks
- Active network markedly different in different conditions
- Identify transient hubs associated with particular conditions
- Use these to annotate genes of unknown function
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 & Human 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,
TopNet – an automated web tool

Similar tools include Cytoscape.org, Idekar, Sander et al

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
Acknowledgements
TopNet.GersteinLab.org
Acknowledgements
TopNet.GersteinLab.org

Job opportunities currently for postdocs & students

P Bork, J Raes

A Sboner
MS

J Korbel

A Paccanaro

M Seringhaus

MG

S Teichmann

M Babu

H Yu

N Luscombe

T Gianoulis

P Kim

J Lu

S Douglas

P Cayting

P Patel

Y Xia

K Yip
Extra
DPM: Discriminative Partition Matching

Cluster (Partition)

Environment

Site-Set 1
B1
B3

Site-Set 2
B2
B4
B5

Test

DPM FOOTPRINT

B1
B3
B2
B4
B5

Functional class | pval
---|---
InfoStorage & Processing | .07
Cellular Process | .08
Metabolism | 4x10^-14

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

Taurine biosynthesis
Heme biosynthesis
Asparagine degradation
Nitrogen fixation
Acylglycerol degradation
Asparagine biosynthesis
Cysteine Metabolism
protein-DNA interactions

protein-protein interactions

protein-small molecule interactions

[From H Yu]
Networks help us understand biological processes

[From H Yu]
More Information on this Talk

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

SUBJECT: Networks

DESCRIPTION:
Cornell Medical School, Physiology, Biophysics and Systems Biology (PBSB) graduate program, 2009.01.26, 16:00-17:00; [I: CORNELL-PBSB]
(Long networks talk, incl. the following topics:
why networks w. amsci*, funnygene*, net. prediction intro, memint*,
tse*, essen*, sandy*, metagenomics*, netpossel*, tyna*+ topnet*, & pubnet*. Fits easily into 60’ w. 10’ questions. PPT works on mac &
PC and has many photos w. EXIF tag kwcornellpbsb.)

(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