Biological Network Analysis

Mark B Gerstein
Yale

slides at

Lectures.GersteinLab.org

(See Last Slide for References & More Info.)
GersteinLab.org Research Overview: Bioinformatics

• **Genome Annotation**
  ◦ Characterizing non-coding regions of the genome, focusing on protein fossils and novel RNAs (Pseudogene.org + GenomeTech.GersteinLab.org)
  ◦ Personal Genomics – esp. related to SVs

• **Molecular Networks**
  ◦ Using molecular networks to integrate & mine functional genomics information and describe gene function on a large-scale (Networks.GersteinLab.org)

• **Macromolecular Motions**
  ◦ Analyzing select populations of 3D-structures in detail, trying to understand their flexibility in terms of packing (MolMovDB.org)
The problem: Grappling with Function on a Genome Scale?

sequence of human chr. 7

~1,200 protein-coding genes
(~950 pseudogenes)

Traditional single molecule way to integrate evidence & describe function

Descriptive Name: Elongation Factor 2

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.

Lots of references to papers
Some obvious issues in scaling single molecule definition to a genomic scale

• Fundamental complexities
  ◊ Role Conflation: molecular, cellular, phenotypic
  ◊ Often >2 proteins/function
  ◊ Also Multi-functionality:
    2 functions/protein
    • phenotypically – e.g. Pleiotropic effects such as human PKU being involved in retardation & eczema
    • cellular role – e.g. Depending on the molecule it interacts with HSP70 is involved with protein folding, translocation of proteins into mitochondria, biogenesis of certain subunits..

[HSP from Craig et al, Rev Physiol Biochem Pharmacol (2006) 156:1 ; Terms from Seringhaus et al. GenomeBiology (2008)]
Some obvious issues in scaling single molecule definition to a genomic scale

• Fundamental complexities
  ◊ Role Conflation:
    molecular, cellular, phenotypic
  ◊ Often >2 proteins/function
  ◊ Also Multi-functionality:
    2 functions/protein
    • phenotypically – e.g. Pleiotropic effects such as human PKU being involved in retardation & eczema
    • cellular role – e.g. Depending on the molecule it interacts with HSP70 is involved with protein folding, translocation of proteins into mitochondria, biogenesis of certain subunits..

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

[HSP from Craig et al, Rev Physiol Biochem Pharmacol (2006) 156:1 ; Terms from 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]
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**
- **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 and 4D: Detailed structural understanding of cellular machinery (e.g. ribosome in different functional states)

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
- Genetic interaction (synthetic lethal)
- Signaling pathways

Part of the TCA cycle
• Why Networks?

• Generating Networks
  ◦ Processing Protein Chips
    (yeast & human nets)
  ◦ Propagating Known Information
    (yeast ppi)

• 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
    (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**

Generating Networks

How do we construct large molecular networks. From processing high-throughput protein array data?
Protein Networks from Processing Protein Chip Data

- Array functional proteins on a chip
- Readout can show presence of proteins in sera (via autoantibodies), small mol. interactions, enzymatic activity, & protein interactions

- Technical issues in processing protein chips similar but not identical to those for DNA chips
  ◇ Hybridization v protein binding
  ◇ Background correction & denoising,
    Normalizing across chips & replicates,
    Calling "hits"
  ◇ ProCAT (Zhu et al., GenomeBiology, '06)
    &
    RLM (Sboner et al., J Proteome Sci. '09)

~6000 yeast proteins on a chip, Zhu et al. Science ('01)

4200 phosphorylations involving 1325 proteins, Ptacek et al. Nature ('05)
Signal Distribution & Metrics

Goal:
Decreasing variation betw. replicates (both inter- & intra-array), measured by CV, & increasing separability (S) betw. known positive & negative samples

Protein Chip Sig. Intensity Distribution from different applied sera (NEGative & with POSitive sera)

CV = \frac{\sigma}{\mu}

S = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}

Representative DNA Chip Sig. Dist.

[Sboner et al., J Proteome Sci. ’09]
RLM Normalization, how it compares?

**NORMALIZATION**

- **Global**
  - A single scaling factor

- **Quantile**
  - Signals are normalized robustly according to the quantiles of a reference distribution

- **Robust Linear Model (RLM)**

\[ y_{ijkr} = \alpha_i + \beta_j + \tau_k + \varepsilon_{ijkr} \]

- \( \alpha_i \): Slide-effect (inter slide)
- \( \beta_j \): Sub-array effect (intra slide)
- \( \tau_k \): Signal
- \( \varepsilon_{ijkr} \): Random error

**SAMPLE SEPARABILITY**

[Sboner et al., J Proteome Sci. '09]
Check #2: How Signal Intensity Correlates over a Titration

**Expectation**

“Positive” protein signal should positively correlate with “Positive” serum dilution

Higher number of “hits” for the “Positive” serum

[Sboner et al., J Proteome Sci. '09]
Generating Networks #2

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
Network prediction: features

• Example 1: gene expression

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

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

Similarity scale:

1 \rightarrow -1

Gasch et al., 2000
Network prediction: features

- Example 2: sub-cellular localization

\[ x_1 = (1, 1, 0, 0, \ldots) \]
\[ x_2 = (1, 1, 1, 0, \ldots) \]
\[ x_3 = (1, 0, 1, 0, \ldots) \]
\[
\begin{align*}
\text{sim}(x_1, x_2) &= 0.81 \\
\text{sim}(x_1, x_3) &= 0.12 \\
&\ldots
\end{align*}
\]

Similarity scale:

1 \quad 1
Data integration & Similarity Matrix

<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)
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
Protein interaction

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

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

Domain interactions

Residue interactions

[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>Level</td>
<td></td>
<td>PD</td>
<td>PR</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>54.89</td>
<td>53.81</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]
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]
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

[ Yu et al., PLOS CB (2007) ]
Bottlenecks are what matters in regulatory networks

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

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

[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.
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

[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 miRNA-target networks easy to calculate, as all influence is down-regulation
  ◊ target prediction via: 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 identify "important" network nodes**

Cheng et al., Genome Biology, 2009
Application of RE-score to measure changing miRNA effect in different conditions (ER- and ER+ breast cancer)

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

(1) RE-score profile for diff. miRNA in 1 cancer sample.
(2) Tabulate over many different breast cancer samples

(3) Clustering based on RE score divides samples into 2 main types of cancer

(4) Clustering better than based on indiv. gene expression levels

Cheng et al., Genome Biology, 2009
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
- AGTGTGCT

Assemble and Annotate

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

Metagenomics Approach

Collect Sample → Extract DNA → Sequence

- ATCGTATAGATAGATAGA
- ATCGTATAGATAGA
- ATCGTATAGATAGA
- ATCGTATAGATAGA
- ATCGTATAGATAGA
- ATCGTATAGATAGA
- ATCGTATAGATAGA
- ATCGTATAGATAGA

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
Pathway Sequences (Community Function) → Environmental Features

Expressing data as matrices indexed by site, env. var., and pathway usage

[Expressing data as matrices indexed by site, env. var., and pathway usage are discussed.]

[Expressing data as matrices indexed by site, env. var., and pathway usage are discussed.]

[Rusch et. al., (2007) PLOS Biology; Gianouulis et al., PNAS (in press, 2009)]
Simple Relationships: Pairwise Correlations
Canonical Correlation Analysis: Simultaneous weighting

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

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

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

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

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

[ 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)]
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)

Fixed Differences

Structural variation

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 undergo 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.

IS RELAXED CONSTRAINT OR ADAPTIVE EVOLUTION THE REASON FOR THE PREVALENCE OF BOTH SELECTED GENES AND SDs AT THE NETWORK PERIPHERY?

**Relaxed Constraint**

- Increases inter-species variation – more variable loci are under less negative selection
- Can be seen in higher Ka/Ks ratio or SD occurrence

**Adaptive Evolution**

- Increases inter-species variation – more variable loci are under less negative selection
- Can be seen in higher Ka/Ks ratio or SD occurrence
- Should not have effects on intra-species variation

---

**Inter-Species Variation (Fixed differences)**

- Increases inter-species variation – more variable loci are under less negative selection
- Can be seen in higher Ka/Ks ratio or SD occurrence

**Intra-Species Variation (Polymorphisms)**

- Increases intra-species variation – for the very same reason
- Can be seen in both SNPs or CNVs

Source: Kim et al. PNAS (2007)
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)**
  - Betweenness Centrality ($\times 10^4$)
  - Genes intersecting CNVs: 3.25, All others: 4.20
  - $p < 0.01$

**Reasoning**

- There is 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.

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.


• Why Networks?

• Generating Networks
  ◊ Processing Protein Chips
    (yeast & human nets)
  ◊ Propagating Known Information
    (yeast ppi)

• 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
    (prokaryote metab. pathways)

• Protein Networks & Variation
  (human ppi & miRNA-targ. net)
Conclusions on Networks: Generation

• Networks from processing protein chip data
  ◊ RLM normalization suppresses quantile

• Predicting Networks
  ◊ Extrapolating from the Training Set
  ◊ Principled ways of using known information in the fullest possible fashion
    • 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 & essentiality
Conclusions:
Points of Network Centrality

- RE-score measures degree of (down) regulation of targets v. non-targets
- Application to miRNA network
- Use 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

• Positive selection (adaptive evolution) at the network periphery
  ◊ 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

• miRNA network
  ◊ More highly regulated genes are under more constraint in miRNA-target networks
  ◊ Exception for housekeeping genes
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

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

SUBJECT: Networks

DESCRIPTION:
CSHL, Cold Spring Harbor, NY; 2010.01.06, 12:00-13:00; [I:CSHL2]
(Long networks talk, derived from [I:MBINETS], including rlm* & new intro. for 1st time)

(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 .