

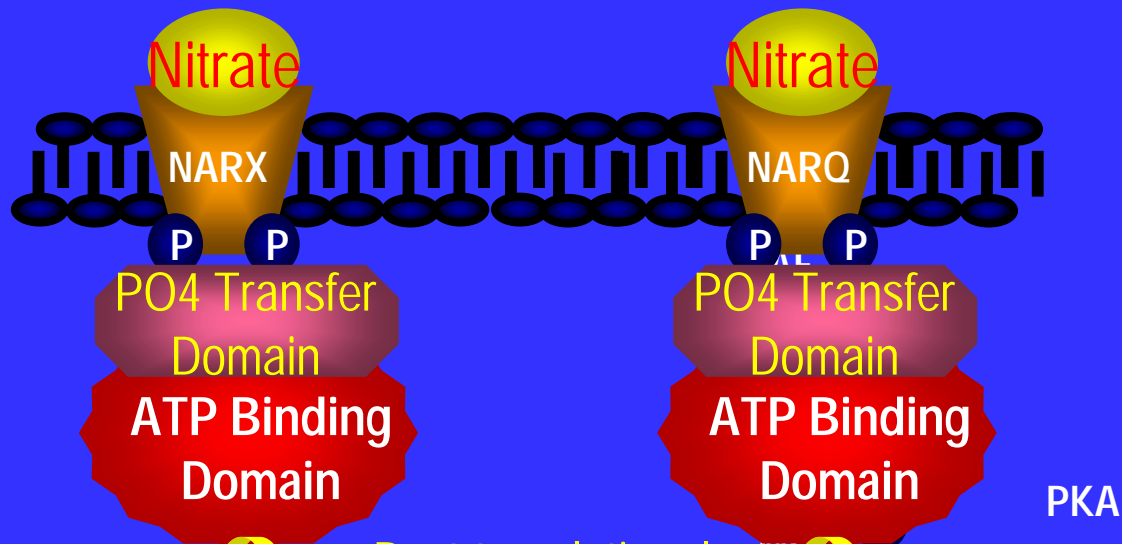
Systems Biology through Data Analysis and Simulation

William Cannon

Computational Biosciences

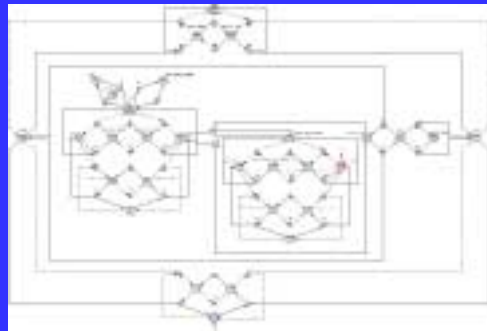
Pacific Northwest National Laboratory

Cellular Dynamics

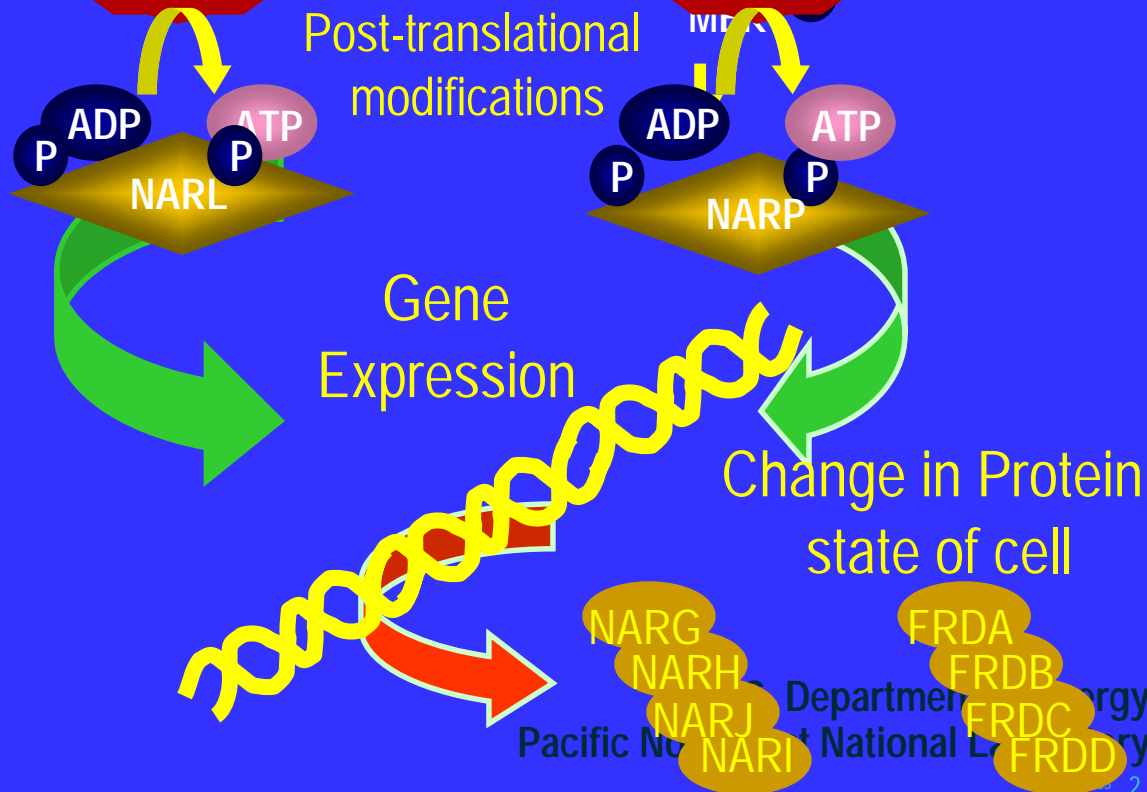


Microbial Cell Dynamics

Data Mining



Battelle



Department of Energy
Pacific Northwest National Laboratory

Deciphering Cell Dynamics

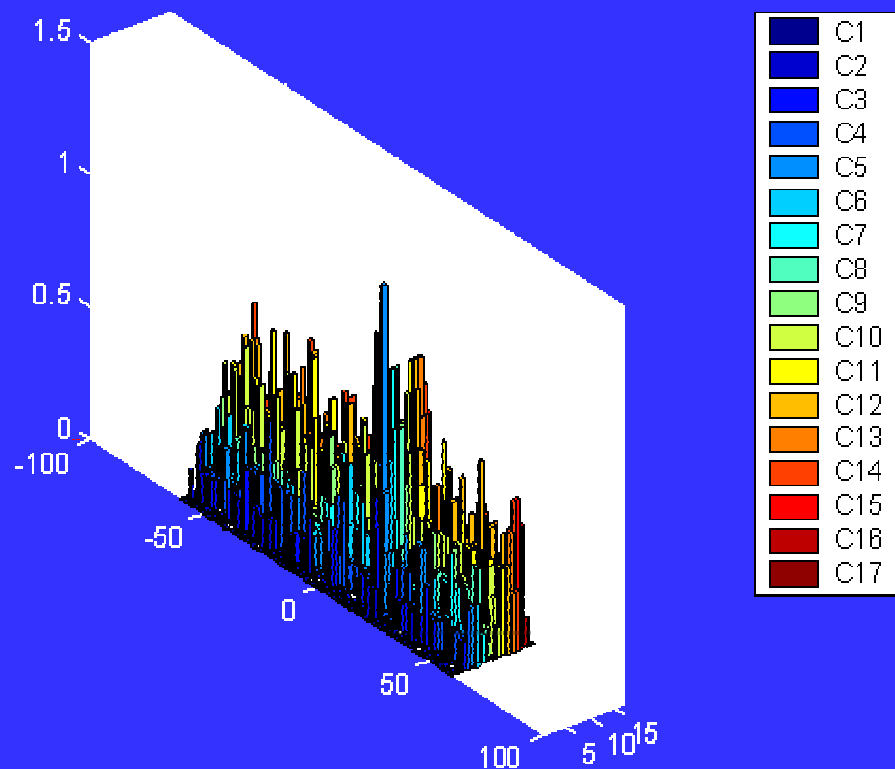
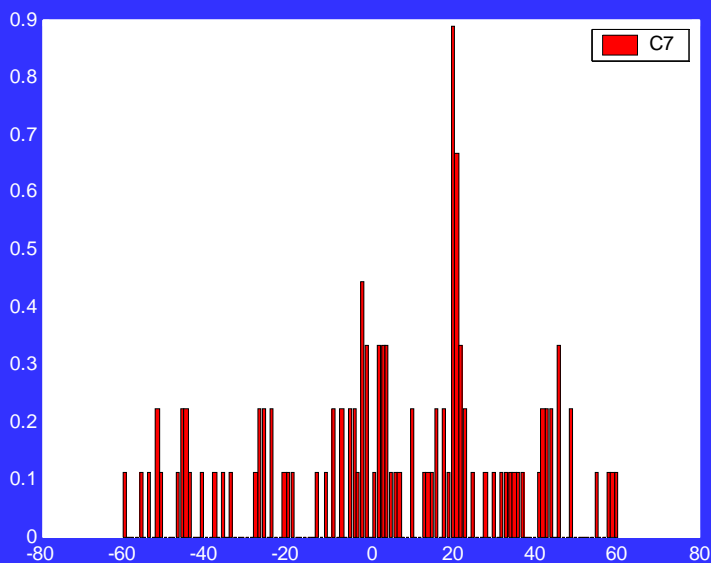
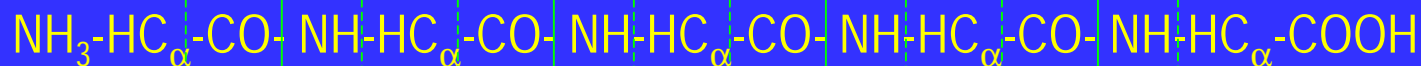
- *Peptide State*: analyze peptide MS/MS data *de novo* with capability to obtain knowledge not captured in sequence databases.
 - Develop a reliable, statistically-based identification method.
 - Account for sequence variations relative to sequence database.

- *Protein State and State of the cell*: High-throughput proteomics must paint a picture of the state of the cell
 - MS data needs to be pieced together into a picture of the state of the protein.
 - All expressed proteins must be analyzed.
 - Microarrays have so far set a bar that must be surpassed if the expense of proteomics is to be justified.

- *Cellular networks*: Determination from high-throughput technologies:
 - Account for hidden variables
 - Make use of existing pathway information
 - Under-determined problem

Statistically-Based Identification Method

Learn the fragmentation patterns and frequency of occurrence from MS/MS spectra of known peptides



Statistically-Based Peptide Identification Method

For peptide in question, construct a MS/MS fingerprint:

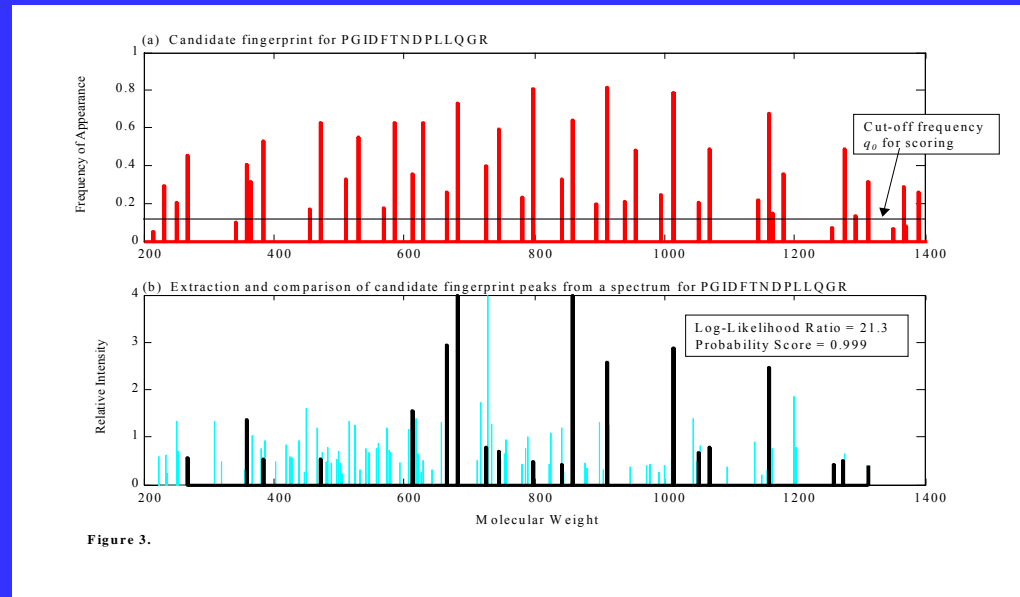
- peak location, l_r, i
- variance in location, s_r, i
- frequency of appearance, p_r, i

Peptide Fingerprint



MS/MS Spectrum:

- fingerprint peaks (black)
- unassigned peaks (aqua)



Statistically-Based Peptide Identification Method

Statistical hypothesis test for identifying a peptide:

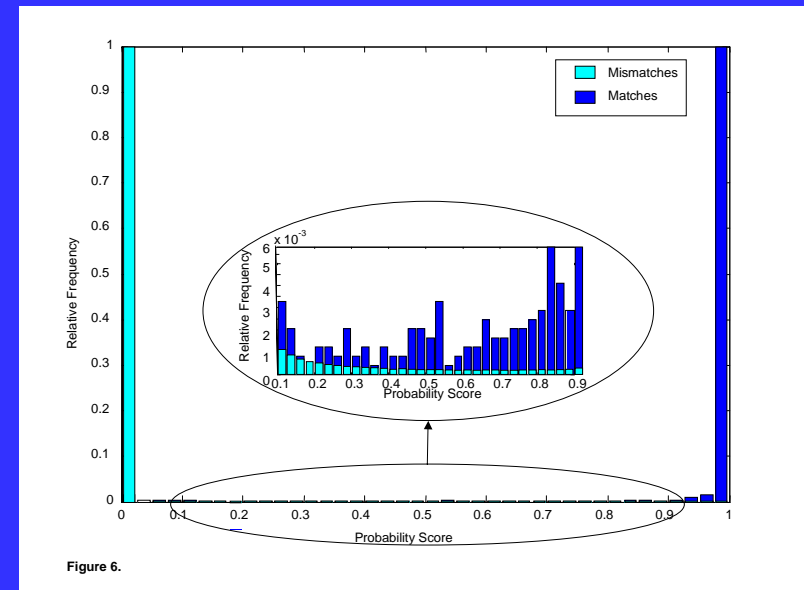
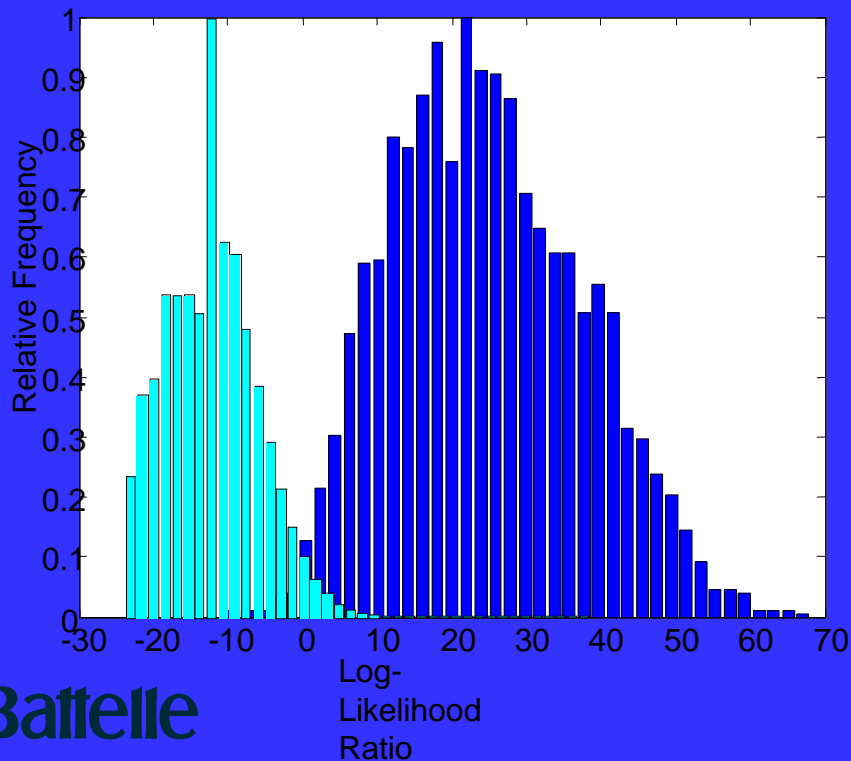
- H_A : the spectrum resulted from a known peptide.
- H_0 : the association between the spectrum and the peptide is no better than random.

$$L = \frac{P\{\text{outcome under } H_A\}}{P\{\text{outcome under } H_0\}}$$
$$= \frac{\prod_{r,i} p_{r,i}^{x_{r,i}} \prod_{r,i} (1 - p_{r,i})^{1-x_{r,i}}}{\prod_{r,i} q_{r,i}^{x_{r,i}} \prod_{r,i} (1 - q_{r,i})^{1-x_{r,i}}}$$

$$P\{H_A \mid \underline{x}\} = \frac{1}{1 + \frac{1}{L} (N_{cand} - 1)}$$

Analysis results

Peptide	LR	P(H _a)	P	chg	mass	Protein
1 VVLEISPYDTSR	15.66	1.00	0.87	1	1378.56	DR2123 initiation factor 1
2 LLLPEHLESDK	2.25	0.01	0.13	1	1380.55	DRB0143 McrB-related protein
3 VVHSMWTPLPGR	-1.07	0.00	-0.06	1	1379.57	DR1838 GTP pyrophosphokinase
4 PELPWGYNGTPF	-1.44	0.00	-0.08	1	1377.53	DR1795 hypothetical protein
5 VVPDFNCQDGGTK	-2.05	0.00	-0.11	1	1379.50	DRC0030 hypothetical protein
6 PQLTVTFDDEGR	-2.51	0.00	-0.14	1	1377.51	DR2194 ribosomal protein S6
7 TTEQTFNVEIAK	-2.57	0.00	-0.14	1	1380.53	DR2429 hypothetical protein
8 PVDLVTLSEHLR	-2.77	0.00	-0.15	1	1378.60	DR0549 DNA helicase (dnaB)
9 QFDSHIEVIHR	-3.12	0.00	-0.17	1	1380.55	DRB0110 thioredoxin-related protein
10 LSLDLALGVGGIPR	-3.52	0.00	-0.20	1	1380.65	DR2340 recA protein



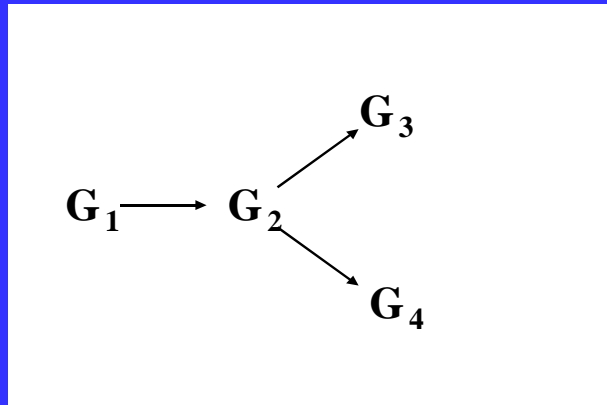
What is a “Genetic Regulatory Network” ?

- What its not:

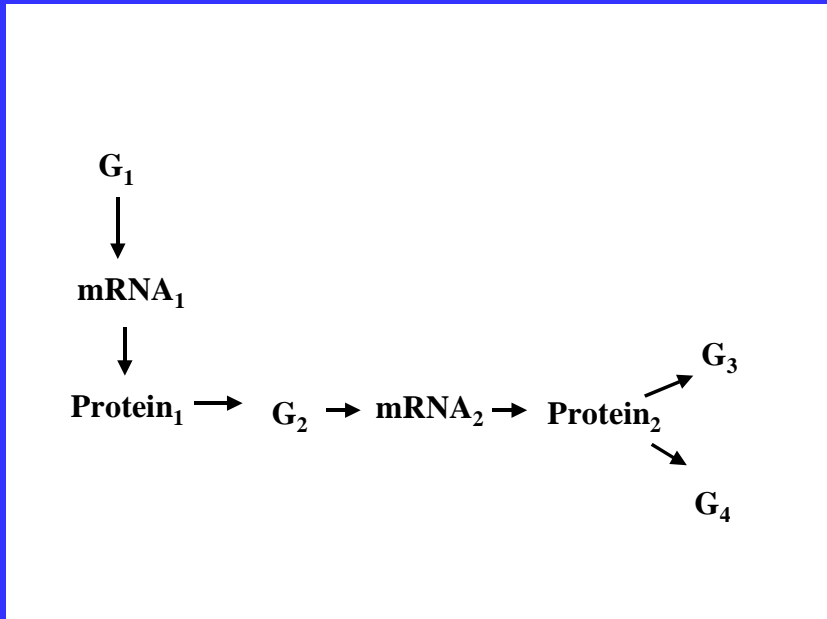
a metabolic or synthetic pathway

- What it is:

- A network inferred from transcriptional data (microarrays); hence the name “genetic”
- The network connections are supposed to show which genes control which other genes; hence the name “regulatory”



Transcription is regulated at multiple levels



- Methylation of DNA
- Regulation by proteins (transcription factors)
- RNA interference
- RNA loop structures
- *etc.*

The statistical inference on microarray data will ideally account for hidden variables

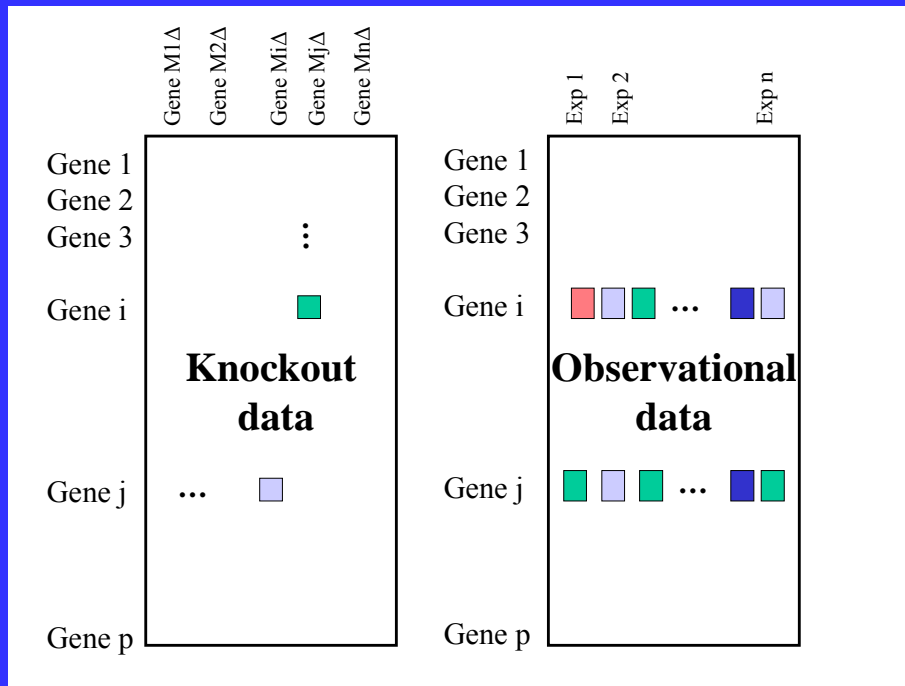
Approaches for network inference

- Cluster analysis: coarse level analysis. Hypothesize function, regulation, etc.
- Relevance network: connect all associated genes – used by some for drug target identification.
- Graphical model: identify 'causal pathways' based on conditional covariation patterns (e.g., Bayesian network)

Using Graphical Models for Analyzing Biological Data

- Resolution:
Inferred networks will be low resolution maps of the cellular networks
→ Fine structure will not be resolved.
- Sample size problem:
Typically trying to infer relationships between 6,000+ genes from only 10-100 microarray experiments.
- Solution: Invert the object/variable space
Instead, infer relations between 10-100 experiments from observations of the behavior of 6,000+ genes.

One approach: Invert the object/variable space



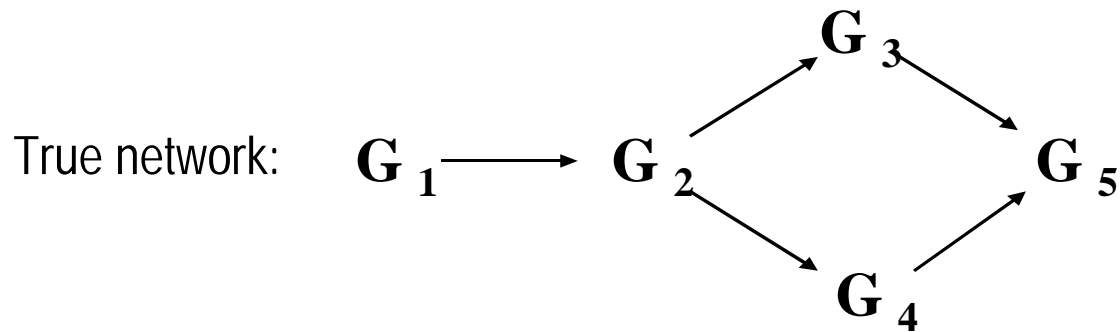
Ideal intervention and Experimental Design:

Each experiment is a gene-knock-out (knock-in, RNAi, etc)

Now relationships between experiments are relationships between targeted genes.

Now have sufficient number of observations, but generally do not observe all genes.

Inferring the connections

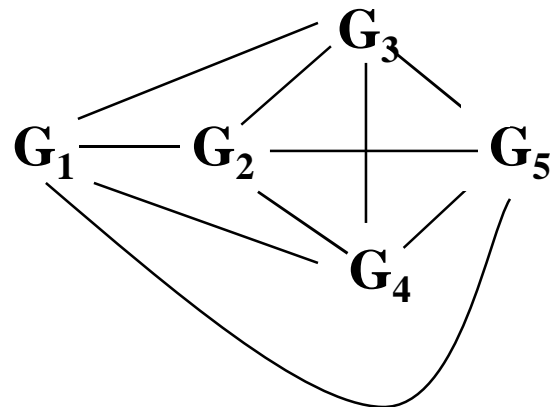


Correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1 & .89 & .80 & .81 & .84 \\ & 1 & .90 & .91 & .94 \\ & & 1 & .82 & .95 \\ & & & 1 & .95 \\ & & & & 1 \end{pmatrix}$$



Resulting graph



Confounding variables: the correlation between genes 1 and 5 may be high, but only in the presence of genes 2,3,4

Partial Correlations

- Partial correlations are correlations between two variables when other variables are controlled for.
- Partial correlations remove the effect of confounding variables

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

Partial Correlations

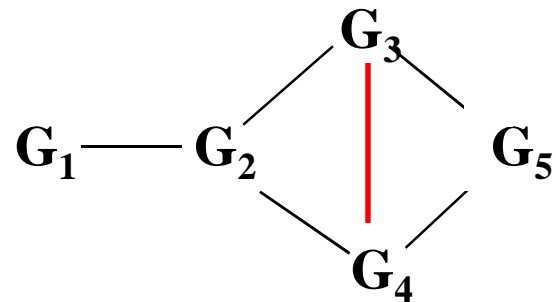
- Pairwise partial correlations when the effect of *all* other genes are removed:

Partial Correlation matrix

$$\tilde{\mathbf{R}}_{1,2} = \begin{pmatrix} 1 & .53 & & & \\ & 1 & .24 & .27 & \\ & & 1 & -.81 & .90 \\ & & & 1 & .90 \\ & & & & 1 \end{pmatrix}$$

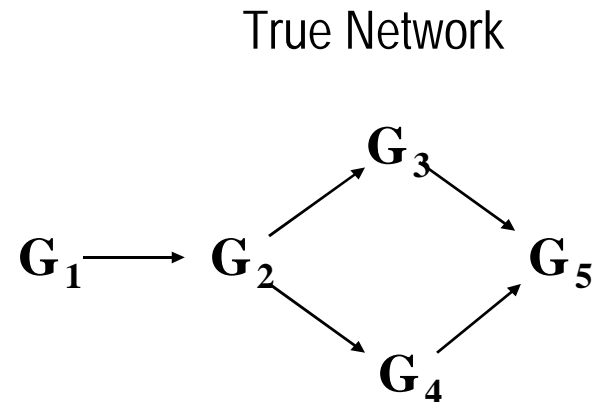
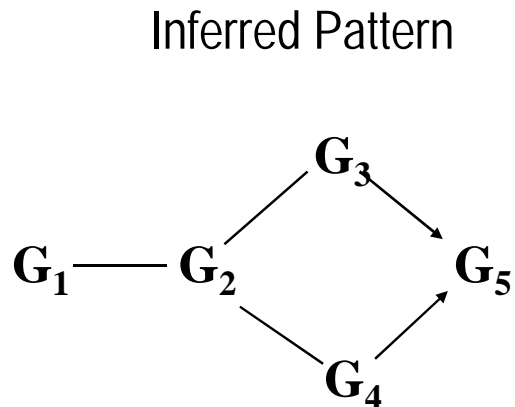


Resulting graph



Partial Correlations using subsets and patterns

- Pairwise partial correlations when the effect of *subsets* of other genes are removed:
- Search for specific correlation patterns when constructing the graphs
 - Directionality

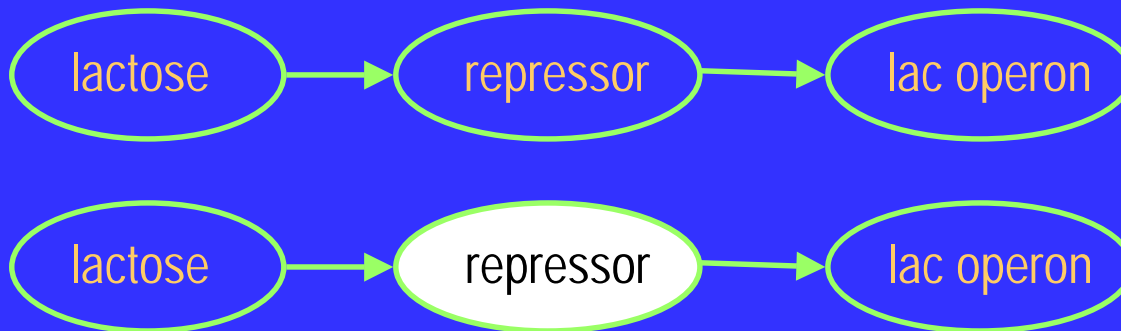


Determining Causal influence

- (1) Gene C causes A and B, or (2) A causes C which causes B:



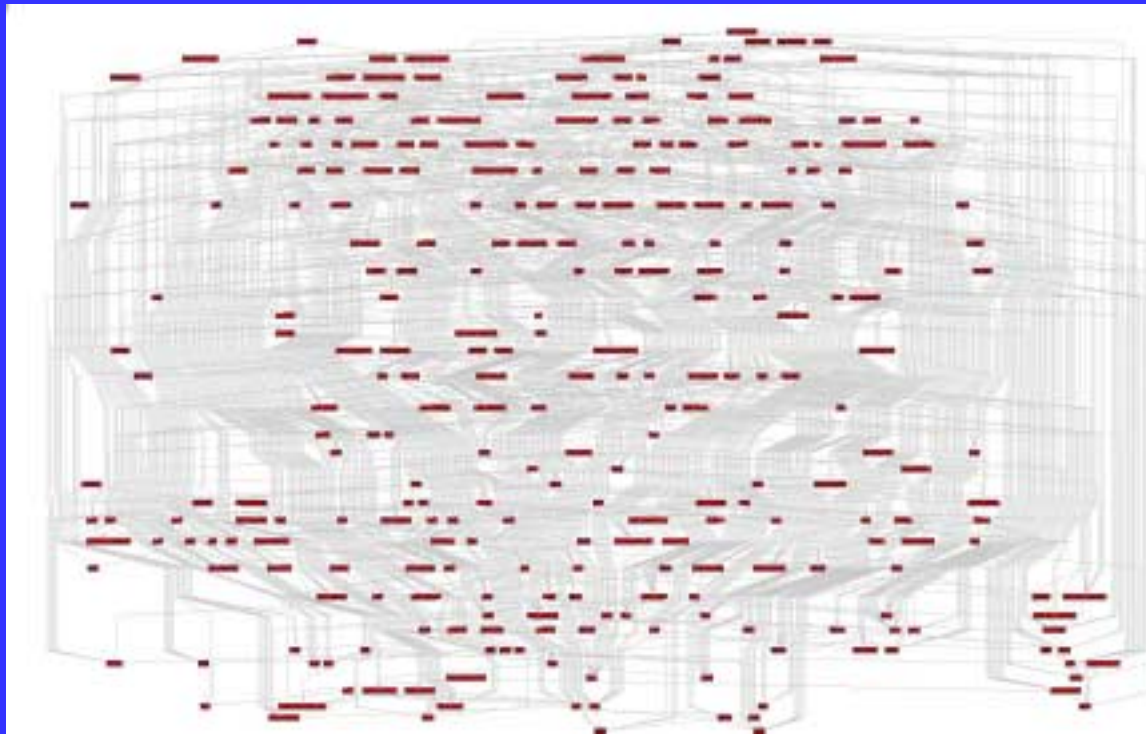
A and B conditionally independent on C



- Inference algorithms work from conditional dependence/independence relationships backwards to causal influence.

Results for 273 yeast knock-out mutants

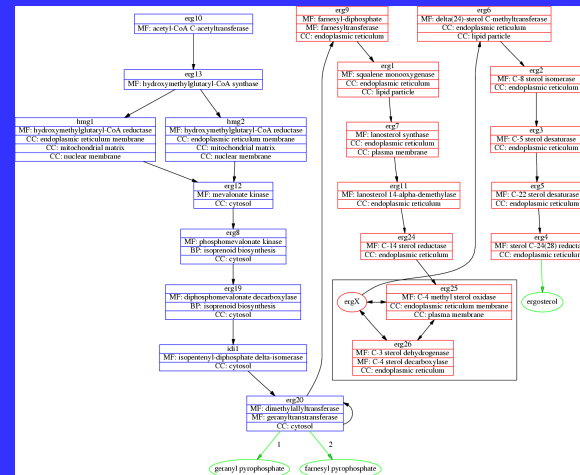
(Hughes, et. al. (2000) Functional discovery via a compendium of expression profiles. *Cell*, 102, 109-126.)



Comparison of inferred Network with Synthetic Pathway

- Genetic Networks are not metabolic/synthetic pathways
- Genes in a pathway can be ordered by their time of appearance.
- Causal networks provide information on order
- Partial order between inferred networks and metabolic pathways may be possible

Erg10 > Erg13 > HMG1, HMG2 > ...
 ... > Erg3 > Erg5 > Erg4.



Partial Order Comparison between synthetic pathway and inferred network

Agreement:

- $IDI1 > ERG4$
- $HMG2 > ERG11 > yer044c \text{ (ERG28)} > ERG2, ERG3$
- $MGG2 > ERG11 > ERG2$
- $ERG6 > ERG2$
- $ERG11 > ERG2$

Disagreement:

- $Erg2 > Erg3$



Regulation of Erg 11 by Swi4

- There has been speculation that Swi4 regulates erg11:
 - Swi4 binding motif: CACGAAA
 - URS1ERG11 motif: CACGAAAAACGAGACAAACGAA
 - Swi4 weakly binds to the ERG11 URS as determined by chromatin immunoprecipitation and microarray analysis (Iyer, et al, Nature 409, p. 533)
- This appears to be the first evidence of functional correlation between erg11 and swi4.

Deciphering Cellular Networks: *Future*

- *Use known relationships as constraints:*
 - *Genes in the same operon*
 - *Regulators and their regulated genes*
 - *Genes in known metabolic pathways*

- *Use protein expression data including post-translational modifications*
- *Use protein-protein interaction data*

Acknowledgments

Statistics and Appl. Math

Kris Jarman

Alan Willse

Sharon Wunschel

Alejandro Heredia-Langner

Macromolecular Structure and Dynamics

Gordon Anderson

Ken Auberry

Dick Smith

Current Funding

DOE OBER

■ Microbial Genome

■ Microbial Cell

■ Genomes-to-Life

Biogeosciences

Jim Fredrickson

Shewanella Federation