Why network inference?

Overview of the day

- Lecture 1: Static network inference
- Practical 1: Practicing on Yogi's gap gene data
- Lecture 2: Dynamical network inference
- Practical 2: Practicing on Yogi's gap gene data
- Lecture 3: Stochastic network inference

- Mostly discussing *computational* methods, rather than *experimental*
- General principles interleaved with literature examples
- Mostly practical techniques, some abstruse theory
- There's no substitute for your intelligence!
- Interrupt, ask questions, clarify, ...

Static Network Inference

CRG Summer Course – Modeling for Systems Biology

Theodore J. Perkins

www.perkinslab.ca

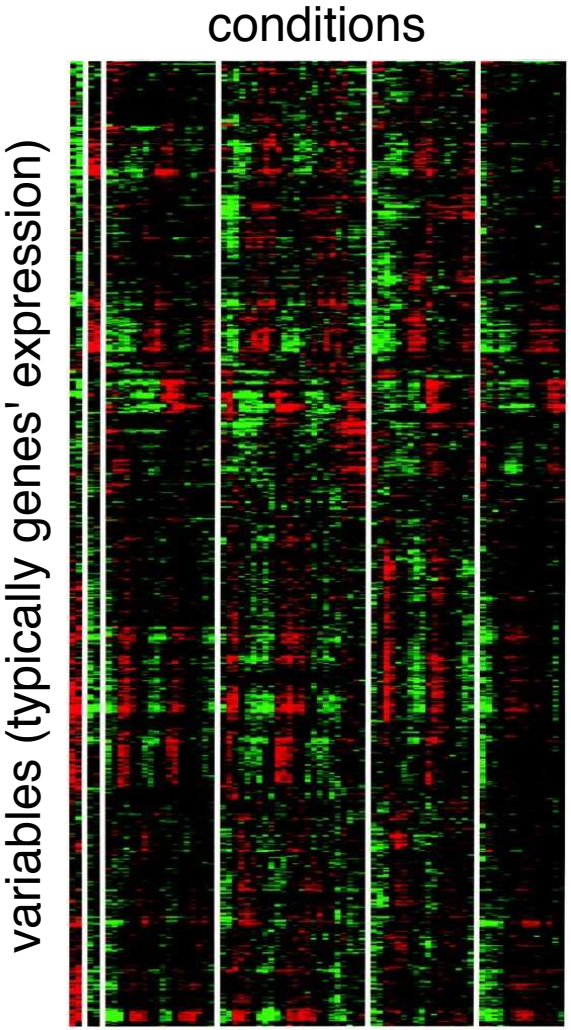
Ottawa Hospital Research Institute &
University of Ottawa

Overview

- "Correlation" networks from expression – similar genes are connected
  - Correlation
  - Information theory & mutual information
  - Permutation testing
  - Relevance Networks
  - ARACNE
- Genetic interaction networks – when deleting two genes doesn't give what you'd expect
  - Synthetic interaction lethality
  - Avery & Wasserman's qualitative framework for epistasis analysis
  - A quantitative statistical model for epistasis analysis

"Correlation" networks

Typically, we start with a data matrix measuring the expression of genes under different conditions.



- Main idea: Make a big graph in which "similarly" expressed genes are connected.

- Could represent one TF regulating another, or co-regulated genes in a complex / pathway, or any number of other things . . . .

- The resulting graph can then be inspected / analyzed to extract biological meaning.

- What does "similar" mean?

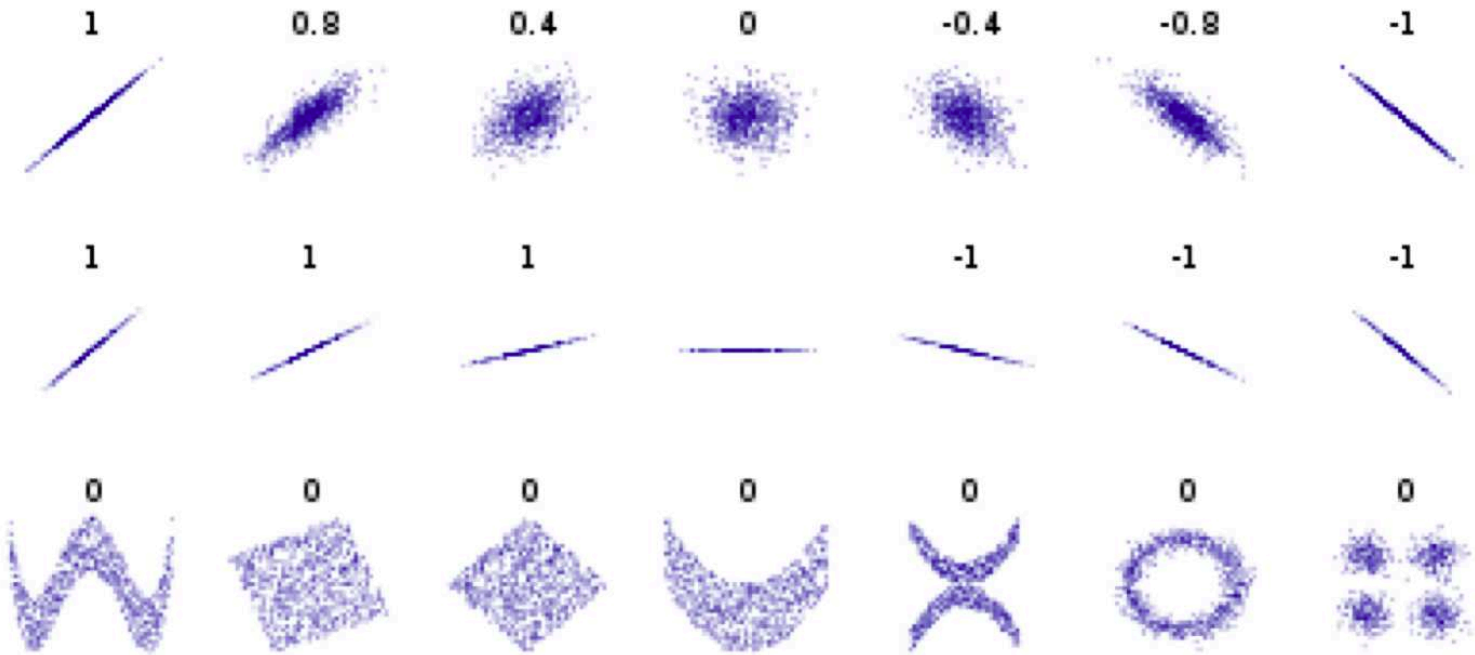
- When are two variables similar enough?

Pearson's linear correlation coefficient

Linear correlation between $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$:

$$r(x,y) = \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}} = \frac{\sum_{i=1}^{n} \frac{1}{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} \frac{1}{n}(x_i - \bar{x})^2 \sum_{i=1}^{n} \frac{1}{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are sample means.



Lit. Example: Relevance Networks (Butte et al. PNAS 2000)

1. Remove variables with "low information content" – e.g., genes that are always on or always off.
2. For every pair of variables (i, j) compute the Pearson's (linear) correlation coefficient across the conditions, r_{ij}, and then compute:

$$\hat{r}_{ij}^2 = \frac{r_{ij}}{|r_{ij}|} r_{ij}^2 = sgn(r_{ij}) r_{ij}^2$$

3. Choose a threshold, τ, to determine statistically significant values of $\hat{r}_{ij}^2$
4. Connect nodes i and j with an undirected edge, if $\hat{r}_{ij}^2 > \tau$

How to choose τ? Permutation testing for significance

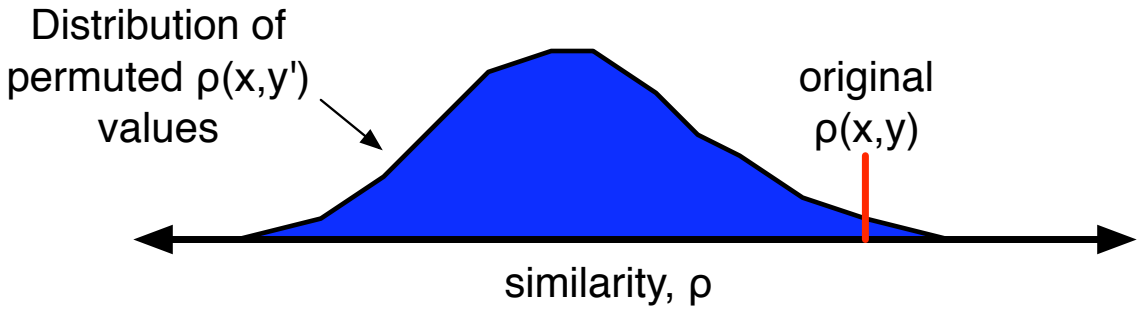
- Suppose you've got paired data on two variables, x and y:

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
y	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8

- Suppose you've got any measure of similarity ρ, which assigns a score to such paired data, $\rho(x, y)$.
- N times, randomly permute the y values and recompute ρ. E.g.:

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
y'	y_3	y_7	y_6	y_1	y_5	y_4	y_2	y_8

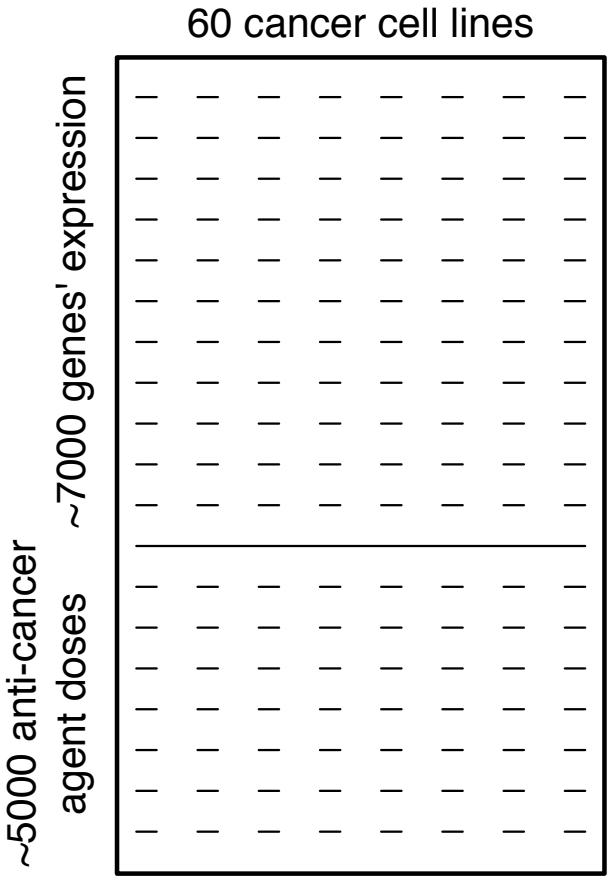
- The location of the original $\rho(x, y)$ with respect to the permuted ρ values gives a p-value.



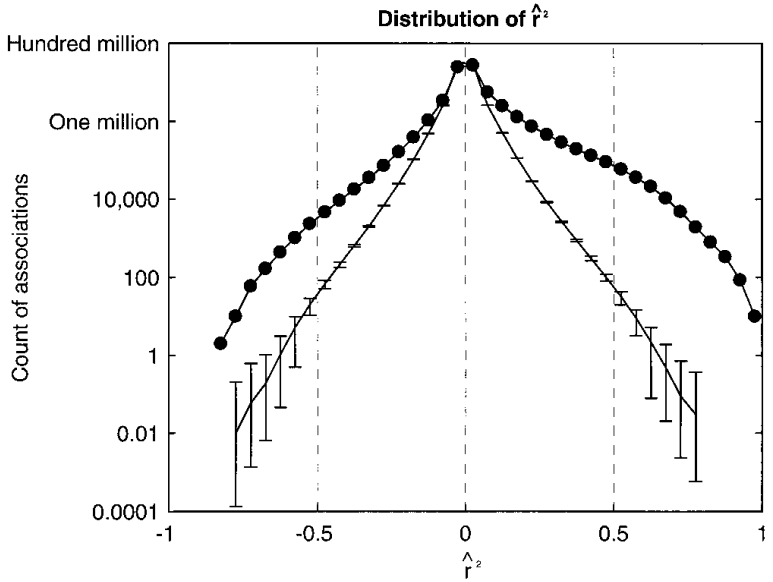
- Approach is distribution and similarity-measure neutral! Still need to choose a p-value threshold (or FDR)...

Back to Butte et al.

Butte *et al.* took 60 cancer cell lines as conditions, and used as variables microarray expression data for 6,701 genes and susceptibility of those lines to 4,991 anti-cancer agents.



- 544 genes and 93 anti-cancer agents discarded as being low information
- 68,345,586 pairwise correlations computed
- $\tau = 0.8$ was deemed statistically significant



Network found

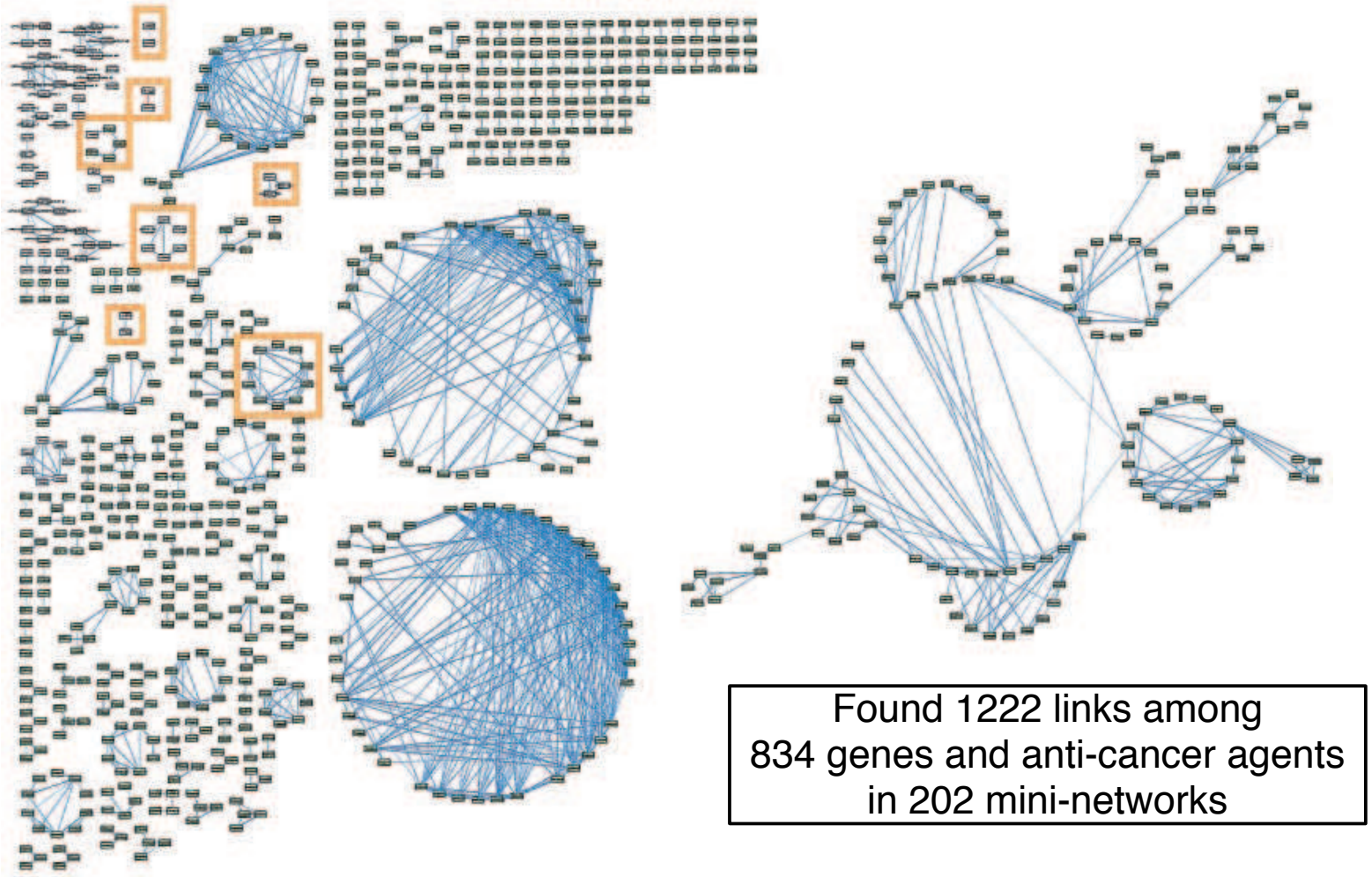
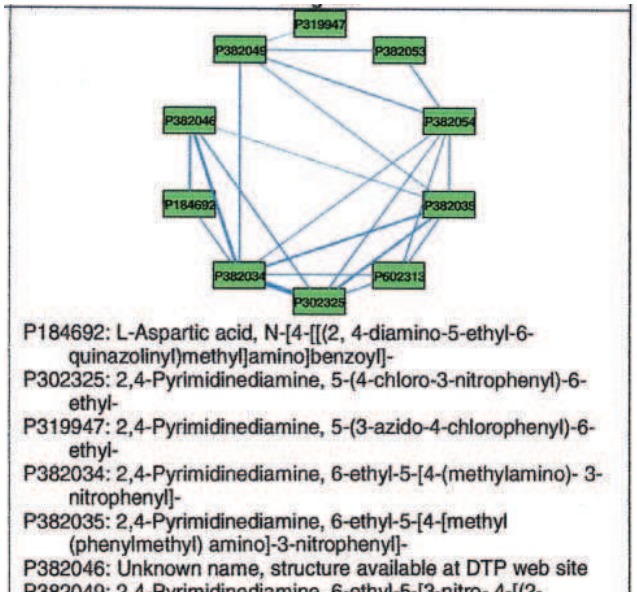
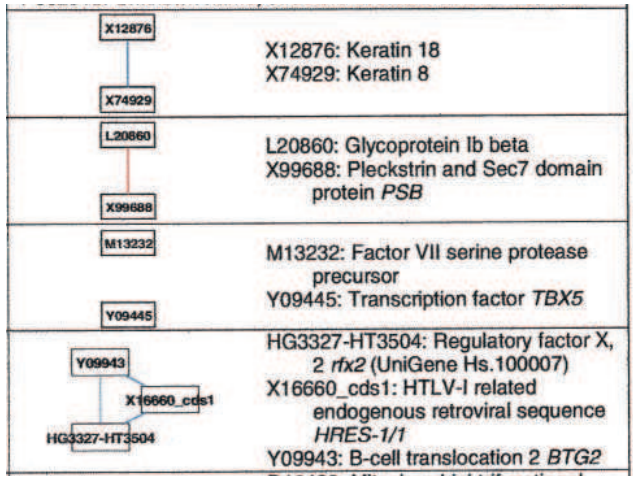
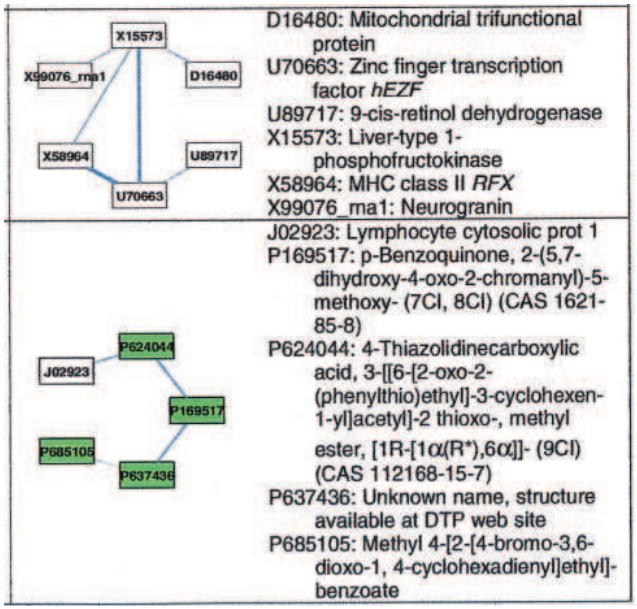


Fig. 2. Relevance networks constructed from the joined databases of baseline gene expression in 60 cancer cell lines and measures of susceptibility of the same cell lines to anticancer agents. The pairs of features (anticancer agents in green boxes, genes in white boxes) with $\hat{r}^2$ at or greater than $\pm$ 0.80 were drawn with line thickness proportional to $\hat{r}^2$. Features without an association at $\pm$ 0.80 were removed. Associations with negative $\hat{r}^2$ are in red. Seven networks are highlighted in orange and are in Table 1. Large versions of all figures and descriptions for each accession number may be found at http://www.chip.org/genomics.

Some subnetworks



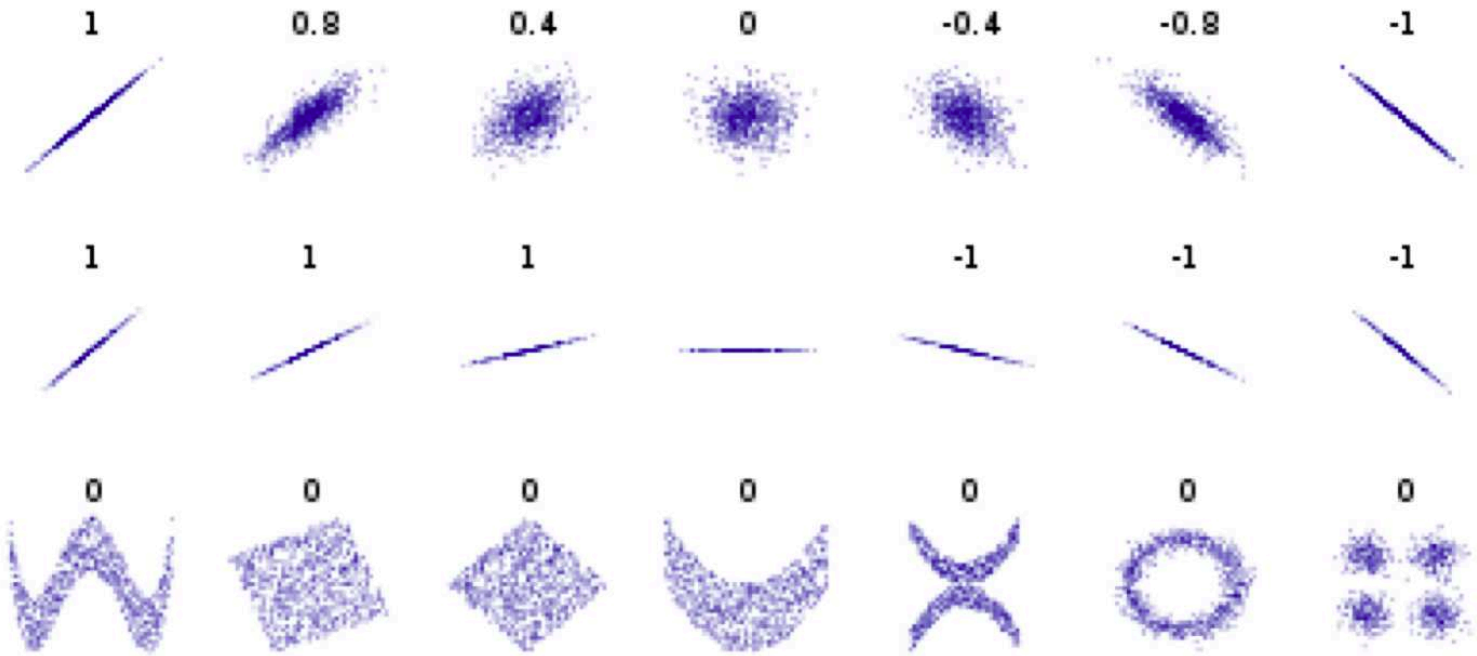




- Found many links between genes and between anti-cancer agents
- Only one gene was linked to an anti-cancer agent – possibly meaningful
- Their τ is very conservative (IMHO)
- Should they compute a different τ for each pair of variables?

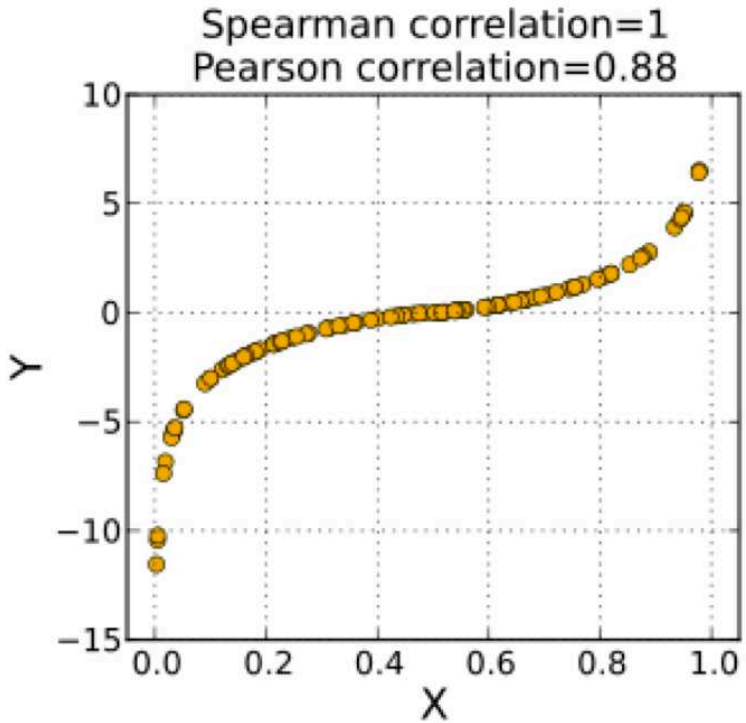
$\Rightarrow$ Analysis of subnetworks leads to generation of hypothesis; the Relevance Networks algorithm has been used in many subsequent studies.

Correlation is limited in the kinds of relationships it detects



Spearman's rank correlation not limited to linearity

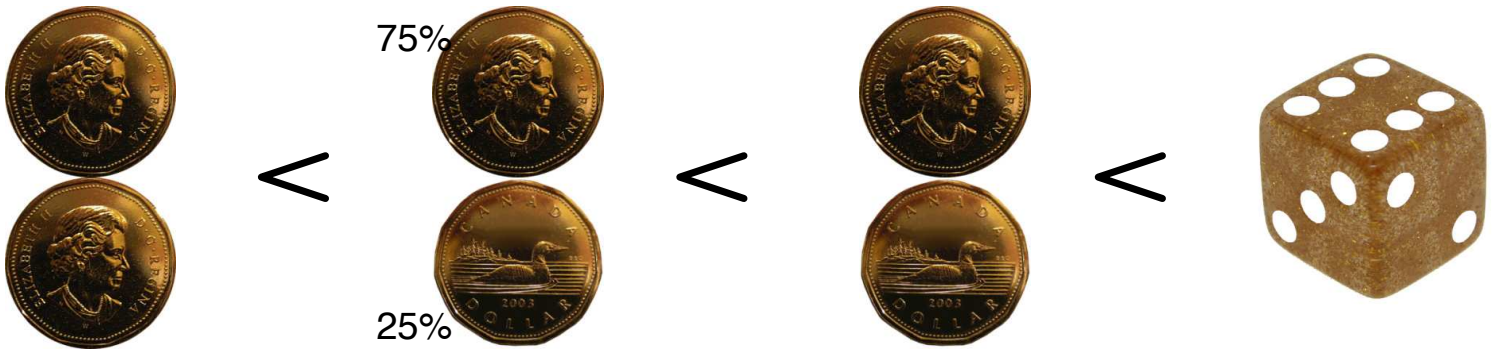
- Replace each x_i by its rank in sorted-increasing order.
  (E.g. 10, 3, 14, 200, 5 $\rightarrow$ 3, 1, 4, 5, 2)
- Do the same for the y_i.
- Compute Pearson's correlation on the ranks.



Captures monotone but nonlinear relationships; more robust to outliers.

Some rudiments of information theory

- The "information" or "entropy" of a random variable $\approx$ how unpredictable that variable is.

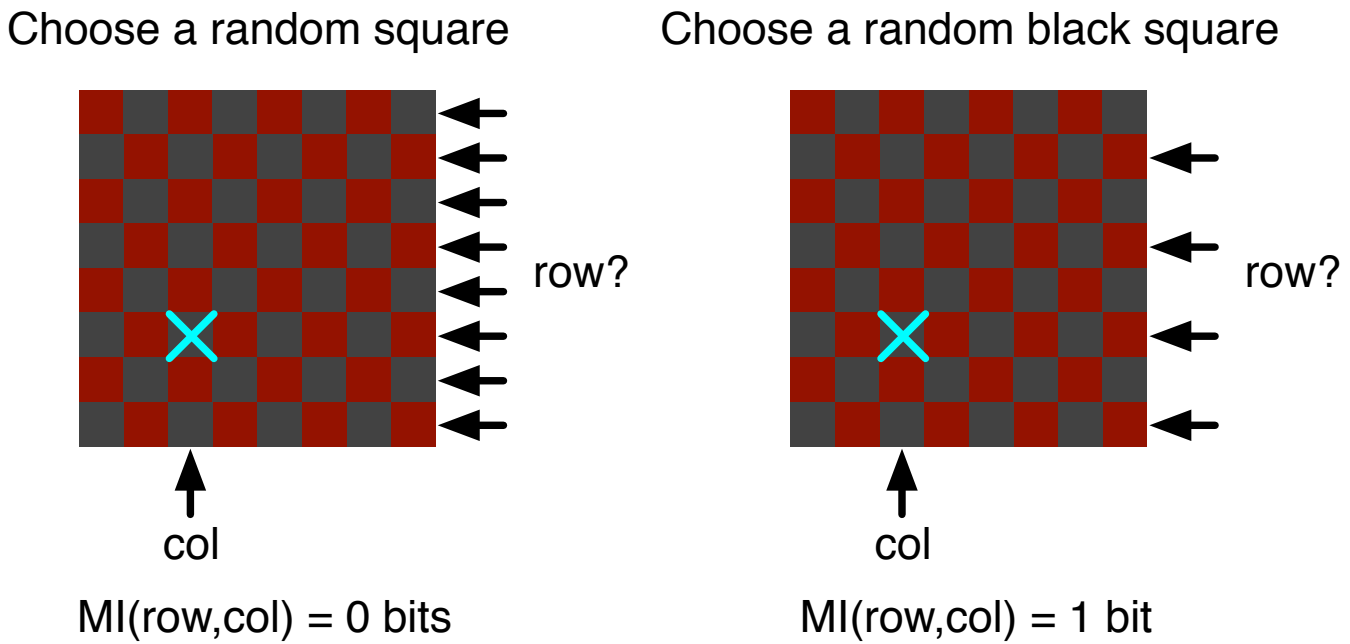


0 bits 0.8113 bits 1 bit 2.585 bits

- The more uniform the outcome and the larger the number of outcomes, the greater the information.
  (If "more random"="more information" seems confusing, think in terms of sending a message.)
- Formally, it is $H(X) = -\sum_x p(x) \log_2 p(x)$ for a discrete r.v. X, or $-\int_x p(x) \log_2 p(x) dx$ for a continuous r.v. X.

Mutual Information

- The "mutual information" between two variables $\approx$ how much better we can predict each variable if we know the other



$$MI(X,Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \text{ or } \int_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} dxdy$$

$$= H(X) - H(X|Y) = H(Y) - H(Y|X)$$

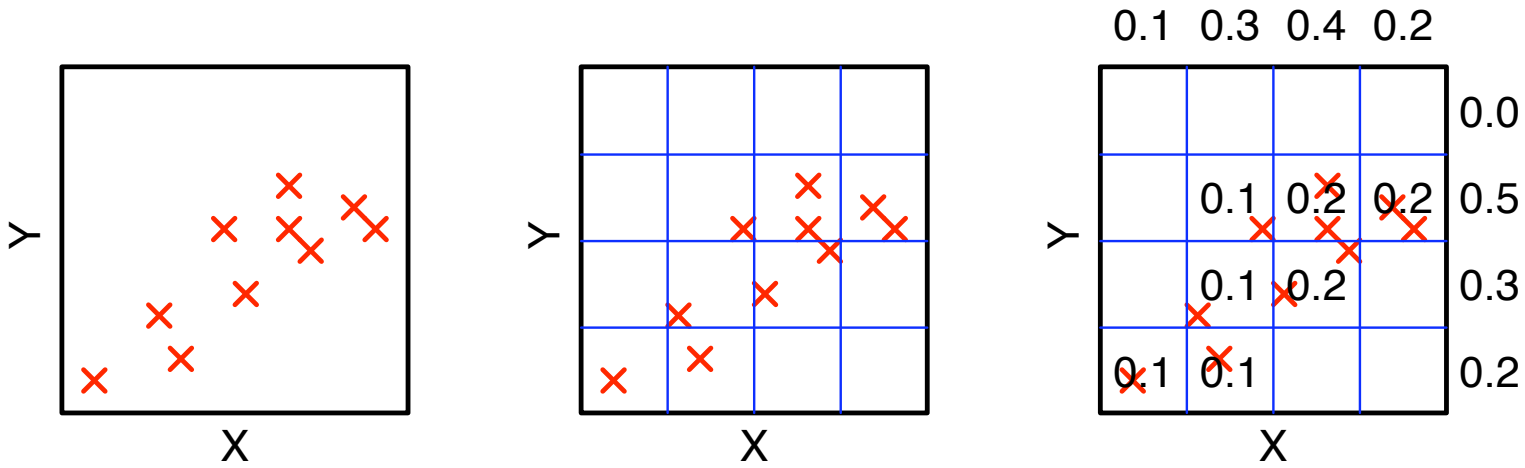
- Note that $MI(X,Y) = MI(Y,X)$.

Literature example: Mutual information relevance networks

- Mutual information relevance networks (see Butte & Kohane, PSB, 2000) is highly analogous to Relevance Networks
- Mutual information was used instead of Pearson correlation to associate genes
- But how to compute (estimate) mutual information for gene expression data – where we don't know $p(x, y)$, but only have samples?

Estimating mutual information from real-valued data

- The simplest approach is to discretize:



. . . and then apply the definition from the previous pages.

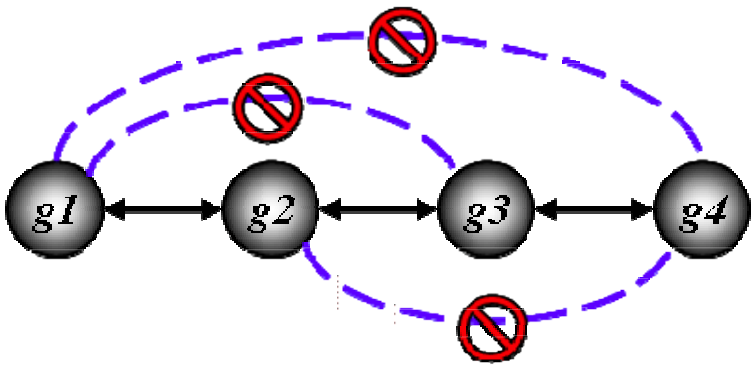
- How many bins? Where to draw lines? In Mutual Information Relevance Networks, they discretized each gene's real-valued expression range into 10 equal-sized bins – but this is not the only plausible choice!

The ARACNE algorithm (Margolin et al., BMC Bioinfo, 2006)

Motivation: Many correlations between variables may be the result of indirect effects

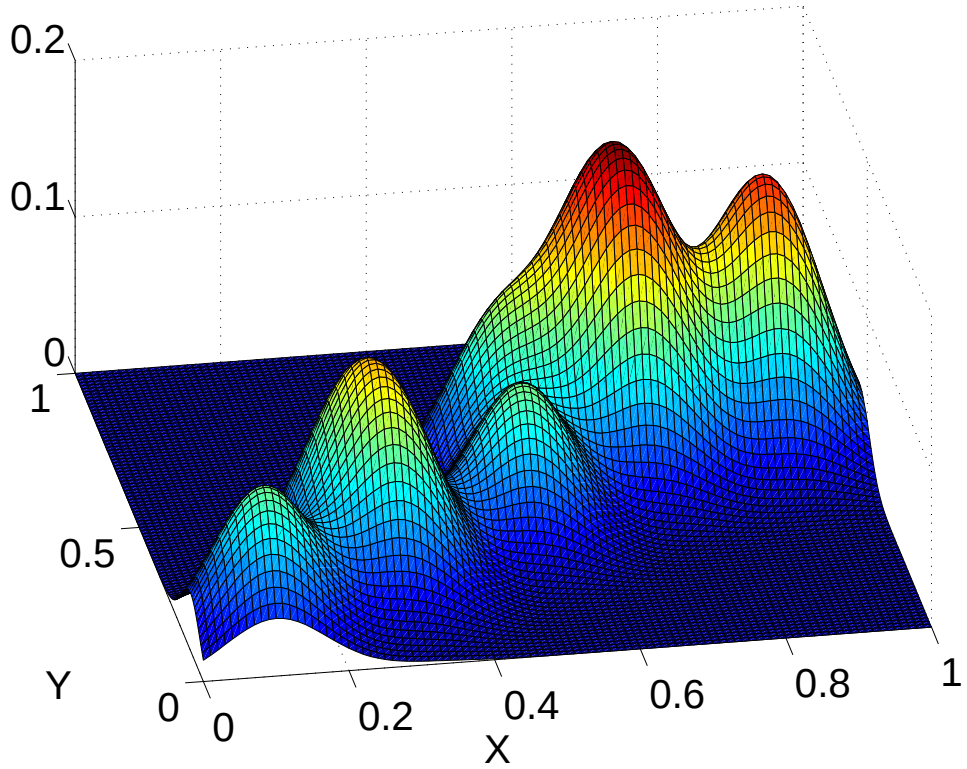
1. Estimate mutual information between all pairs of variables.
   (Using a smoothing method.)
2. Link variables with mutual information $\geq \tau$.
   (Using a permutation test to choose τ, as in Relevance Networks.)
3. Remove $X \leftrightarrow Y$ if $\tau \leq MI(X,Y) < \min(MI(X,Z), MI(Z,Y))$

The last condition is called the "Data Processing Inequality", and it removes the weakest link in every triplet of connected variables.



Estimating mutual information from real-valued data (II)

- In ARACNE they approximate a smooth density over X-Y space using a mixture of Gaussians.

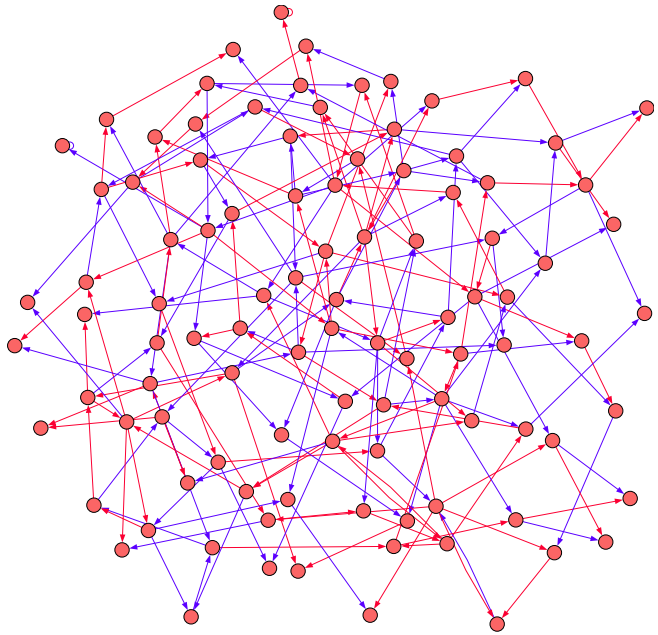


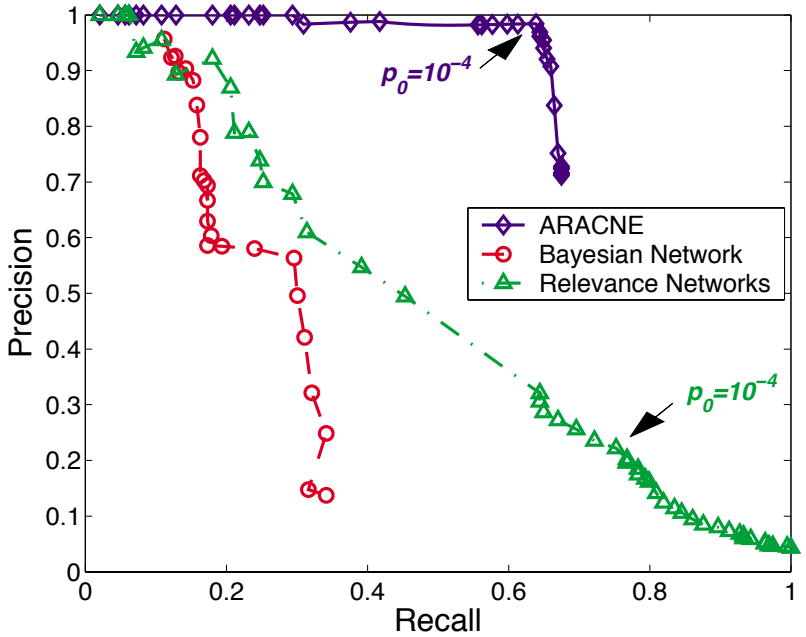
. . . and then apply continuous mutual information definition.

- How wide are the Gaussian? Do they have non-zero covariance? Why Gaussian at all?

ARACNE results on simulated, random networks

- They tested the approach on randomly generated networks (with 100 nodes, 200 links, and simulated steady state expression data):





- ARACNE was more accurate than competitors at finding correct links.
  - Precision = what fraction of the top K estimated links are really in the network?
  - Recall = what fraction of the true links are among the top K estimated links?

ARACNE results on B lymphocyte data

- Applied to expression data from B lymphocytes ($\approx$ 340 conditions; normal, tumor, manipulated)
- Found 29 of 56 known transcriptional targets of the proto-oncogene c-MYC, which was a hub in the network
- Neighbors of those 29 were often highly correlated to c-MYC as well, but had many fewer known targets

$\Rightarrow$ Some evidence that the algorithm can pull out real transcriptional relationships.

"Correlation" networks summary

- In "correlation" networks, the genes whose expression is most strongly correlated over a set of conditions are linked
- "Correlation" can be assessed in several different ways

\+ Correlation networks can be computed efficiently

\+ Can find true / known relationships, as well as many new ones

\+ Subnetwork inspection leads to new hypotheses

– Links are directionless, and of unclear meaning
(Though some directional proposals have been made.)

– The networks are not predictive. What would happen if gene X were deleted?

Questions?

Overview

- "Correlation" networks from expression – similar genes are connected
  - Correlation
  - Information theory & mutual information
  - Permutation testing
  - Relevance Networks
  - ARACNE
- Genetic interaction networks – when deleting two genes doesn't give what you'd expect
  - Synthetic interaction lethality
  - Avery & Wasserman's qualitative framework for epistasis analysis
  - A quantitative statistical model for epistasis analysis

Genetic Interaction Networks

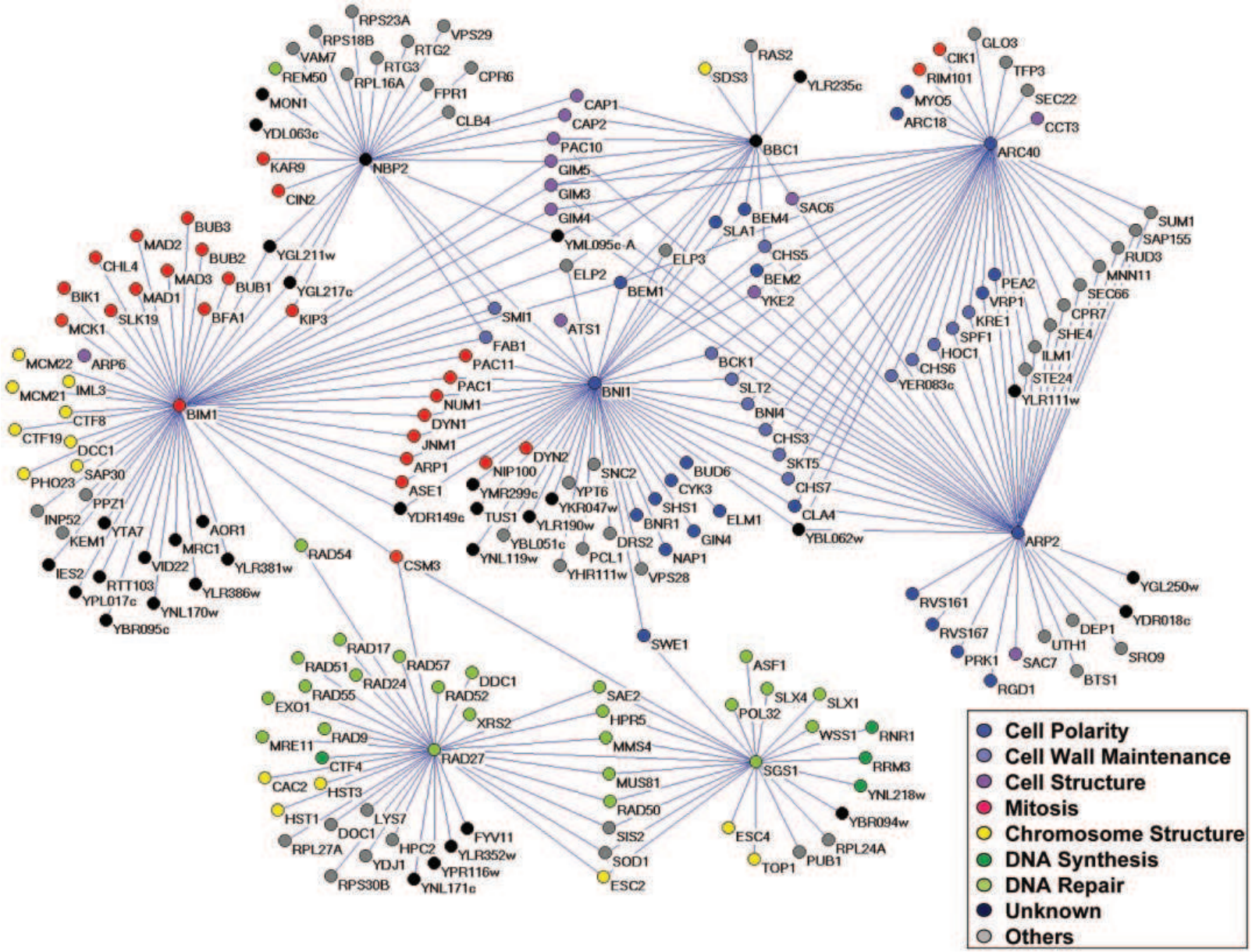
- Biologists have long used genetic perturbations (including knock-out, knock-down, knock-in, overexpression, ectopic expression) to test regulatory network theories.
- When knocking out two genes, X and Y, two obvious outcomes of interest are:
  - $\Delta X \Delta Y$ has "stronger" effect than expected (e.g. Synthetic lethality / sickness)
  - $\Delta X \Delta Y$ has "weaker" effect than expected (e.g. Epistasis / Masking / Suppression)

Literature example: Tong et al., Science, 2001

Investigated synthetic lethality in yeast.

- Crossed a library of 4700 non-essential single-gene deletion strains with 10 "query" strains, generating 47,000 double-deletion mutants
  (Robotics allowed such large numbers for the first time.)
- Assays each for liveness/growth under "standard conditions"
- Made a network by connecting genes whose pairwise deletion resulted in dead / very sick yeast

Tong et al. results, graphically



Tong et al. selected results

- The recovered 8 of 11 previously-known ΔBNI1 interactors, plus 43 new ones (all verified with low-throughput testing)
- The interactors of a gene tended to be member of a common pathway / function. E.g. ΔBNI1 interactors:
  BEM1, BEM2 BEM4 (bud emergence genes), CHS3, SKT5, CHS5, CHS7, BNI4 (chitin sythase III activity-related), etc.
- Genes of unknown function could be assigned putative functions based on interactions with query genes

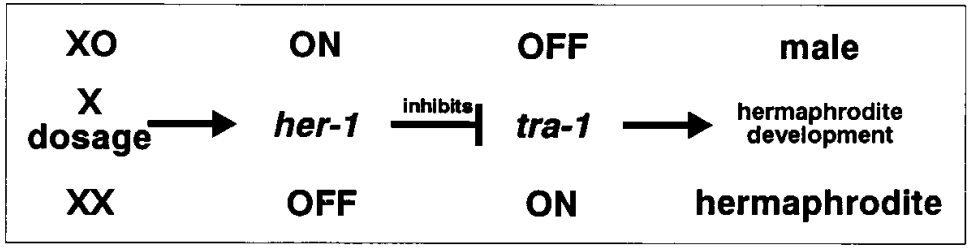
$\Rightarrow$ Synthetic lethal interactions tend to reveal pathways parellel to / dependent on query gene

- Subsequent studies have increased scale of screens, and are looking at more quantitative/multidimensional phenotypes

Epistasis analysis of masking phenotypes

Two examples from Avery & Wasserman (TIG, 1992):

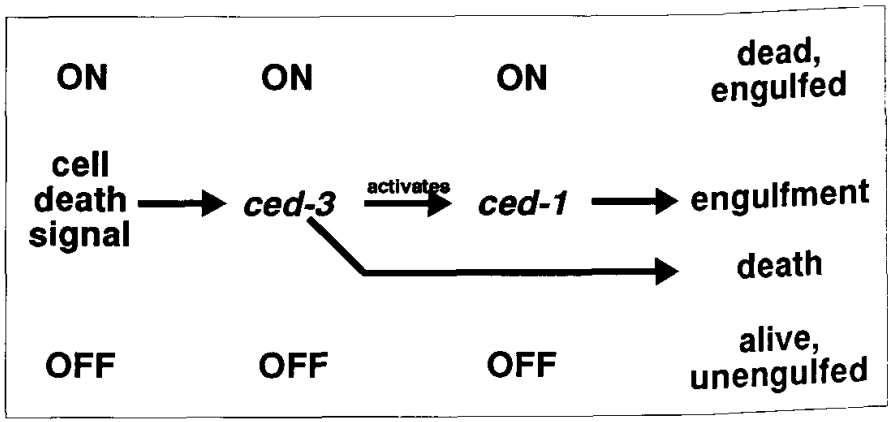
Sex determination in *C. elegans*



⇒ Tra-1 is epistatic to her-1

(because Δher1Δtra1 = Δtra1)

Apoptosis in *C. elegans*



⇒ Ced-3 is epistatic to ced-1

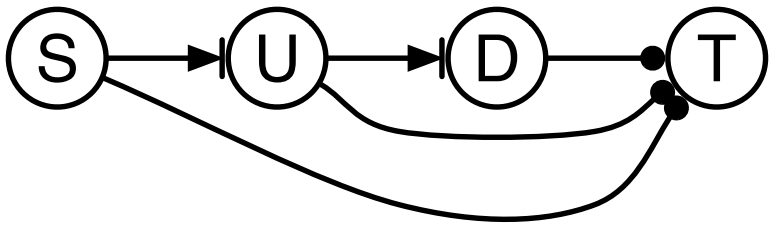
(because Δced3Δced1 = Δced3)

Question: If we don't know the pathway relationship, but only observe the phenotype for wild-type, ΔX, ΔY and ΔXΔY, what can we conclude?

Avery and Wasserman's model

Avery & Wasserman created a simple Boolean model to formalize / explain how to reason about epistasis.

- Consider a pathway that takes an input signal S, through an upstream gene U, a downstream gene D, and on to the phenotypic trait T.

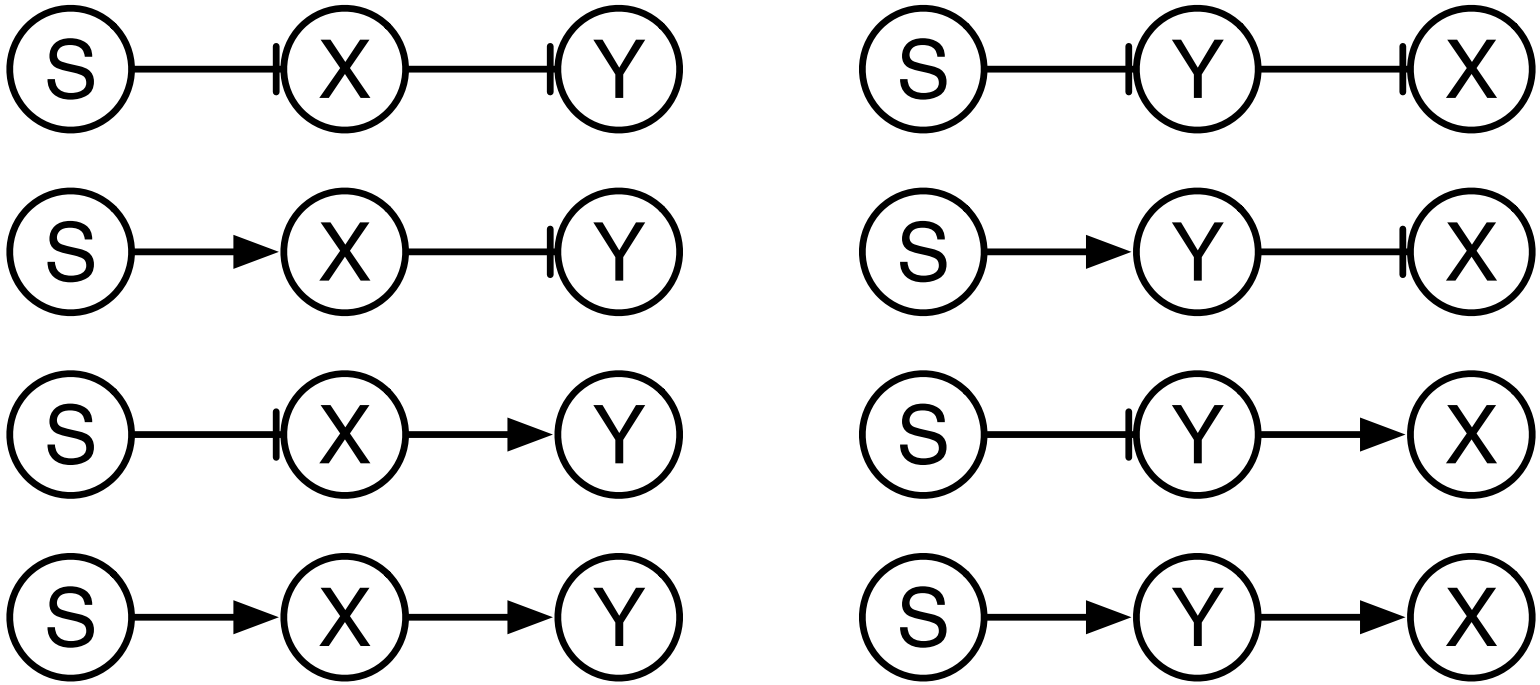


- Suppose S, U and D are all Boolean (ON/OFF).
- Suppose that S either activates or represses U, and U either activates or repressed D. (Directly or indirectly).
- Suppose all three can influence the trait T.

(Note, this is a slight re-interpretation of what they say.)

Eight possibilities structures

For two genes X and Y, there are eight possible pathway structures, depending on which is upstream, and whether links are activating or repressing.



Analyzing a particular pathway

We can construct a truth table to see the behavior of any pathway, and verify the rules. E.g., for $S \to X \to Y$:

S	ΔX	ΔY	X	Y	T	Description
0	0	0	0	0	T_0	signal off, wild-type
0	0	1	0	0	T_0	signal off, Y deleted
0	1	0	0	0	T_0	signal off, X deleted
0	1	1	0	0	T_0	signal off, both deleted
1	0	0	1	1	T_{SXY}	signal on, wild-type
1	0	1	1	0	T_{SX}	signal on, Y deleted
1	1	0	0	0	T_S	signal on, X deleted
1	1	1	0	0	T_S	signal on, both deleted

Avery and Wasserman's rules

Going through this process for all 8 possible pathways establishes the following rules:

1. Each single deletion only affects trait T either when signal S is ON or when it is OFF, but not both.
2. If two single deletions affect T in opposite signal states, and one is epistatic to the other, then the epistatic gene is downstream.
3. If two single deletions affect T in the same signal state, and one is epistatic to the other, then the epistatic gene is upstream.

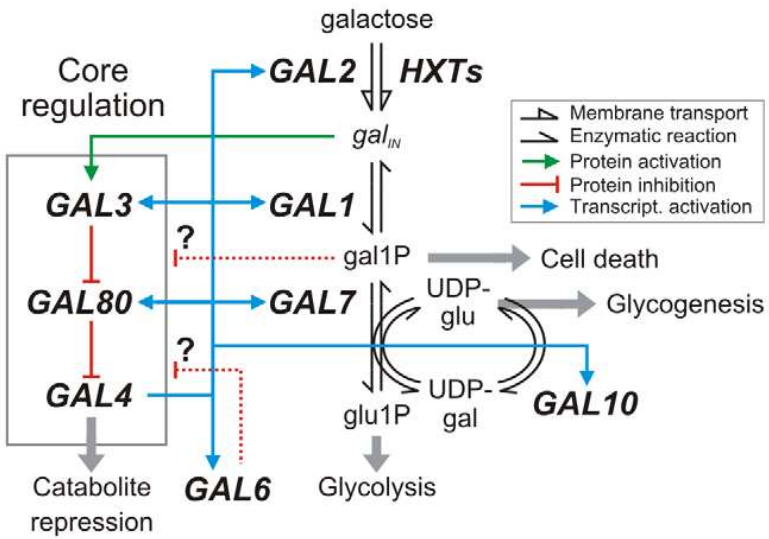
Here X epistatic to Y means that $\Delta X \Delta Y = \Delta X \neq \Delta Y$.

- These rules give us a test for whether the model applies (rule 1)
- And allow us to make conclusions about network structure just by observing changes in phenotype when genes are deleted (rules 2 and 3)

A statistical framework for quantitative epistasis analysis

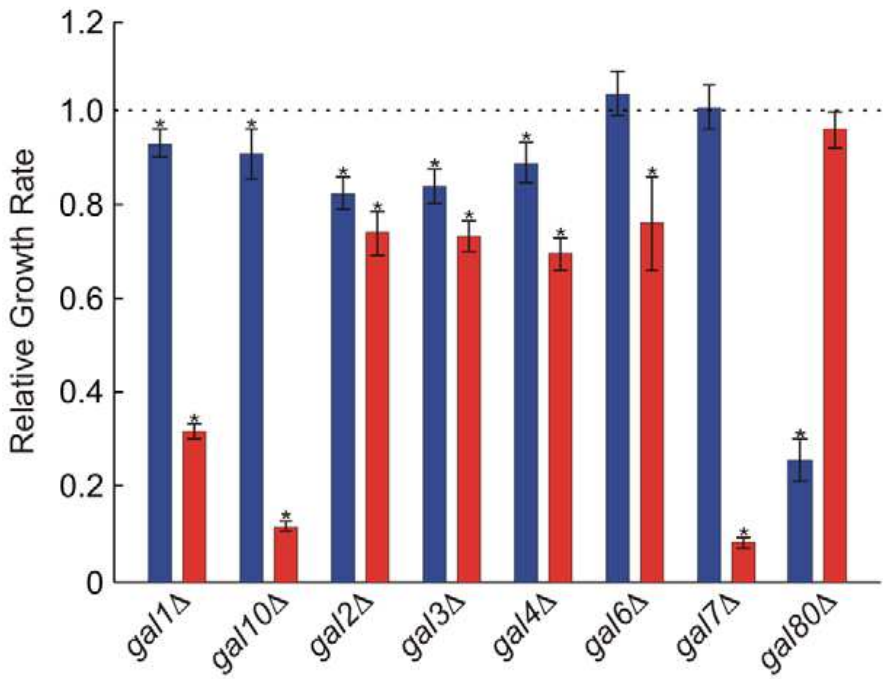
[Phenix *et al.*, PLoS CB, 2011]

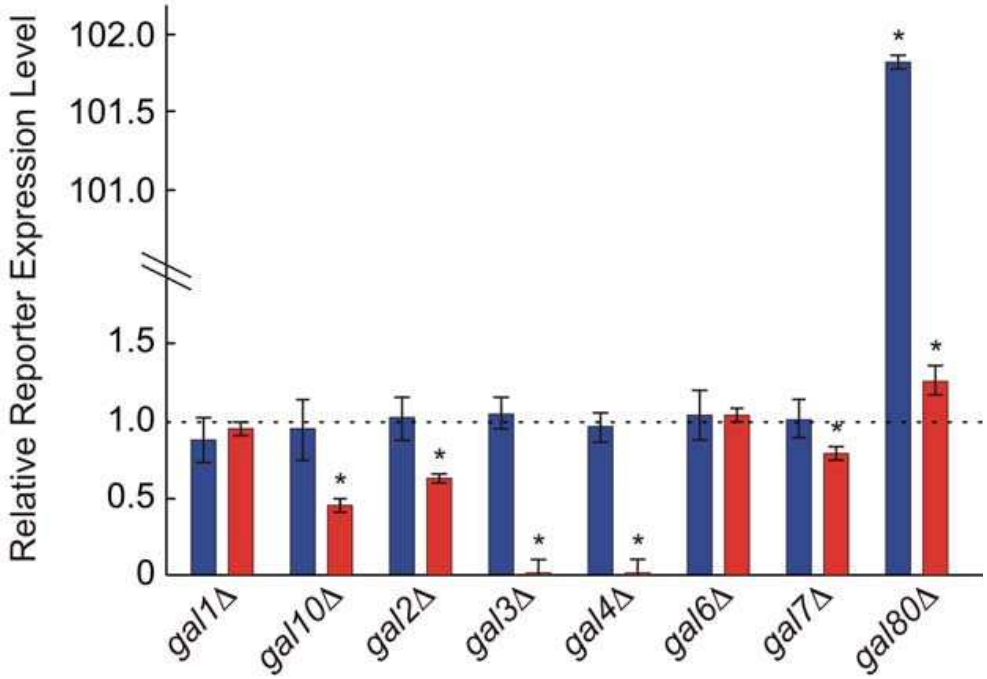
The Avery & Wasserman model is fundamental for qualitative reasoning about epistasis and a key tool for reconstructing pathways. But it doesn't generalize well to quantitative data: How to handle noise? How to handle off-signal effects / pathway crosstalk?



- We had single-cell growth data and Gal10p-reporter data, under WT and all single- and double-knockouts
- Avery & Wasserman, because its assumptions (and conclusions) were violated.

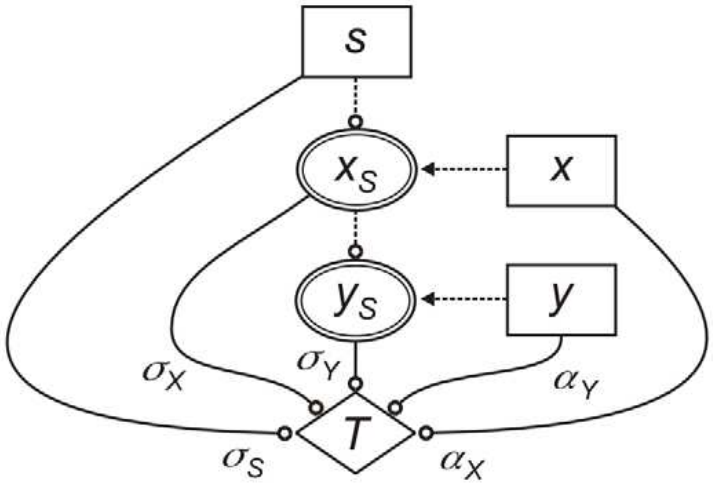
Single knockout data





- Many gene knockouts affect the trait (growth or reporter expression) in *both* signal states – i.e. there are signal-independent effects to deletion.
- There are some strong feedforward / off-path effects – in some cases stronger than the on-path effects.
- Variability in data requires statistical framework for determining epistasis.

A statistical model



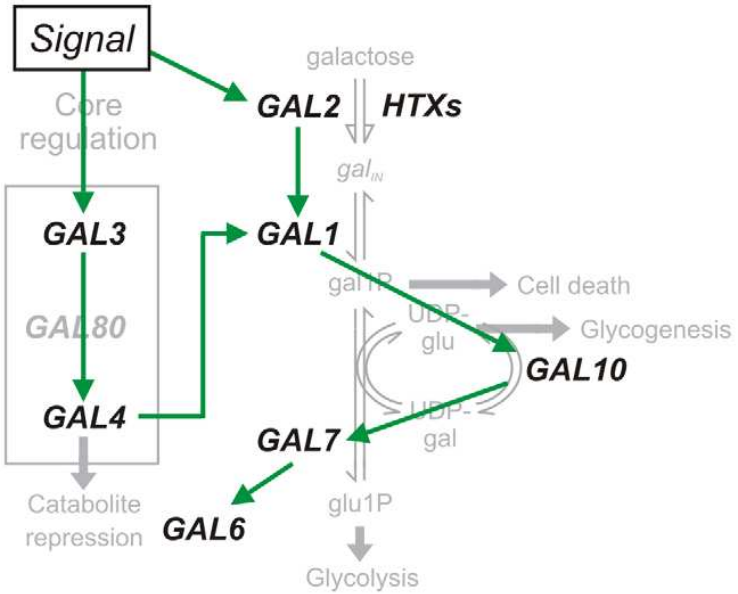
- S, X, X_S, Y, Y_S binary variables, T is the trait
- X, Y represent whether or not genes X and Y have been deleted
- X_S, Y_S denote the signal/pathway-specific effect of X and Y
- All five variables can influence the trait through additive weights

- This allows us to separate signal-dependent influences (σ_S, σ_X, σ_Y) from signal-independent influences (α_X, α_Y).

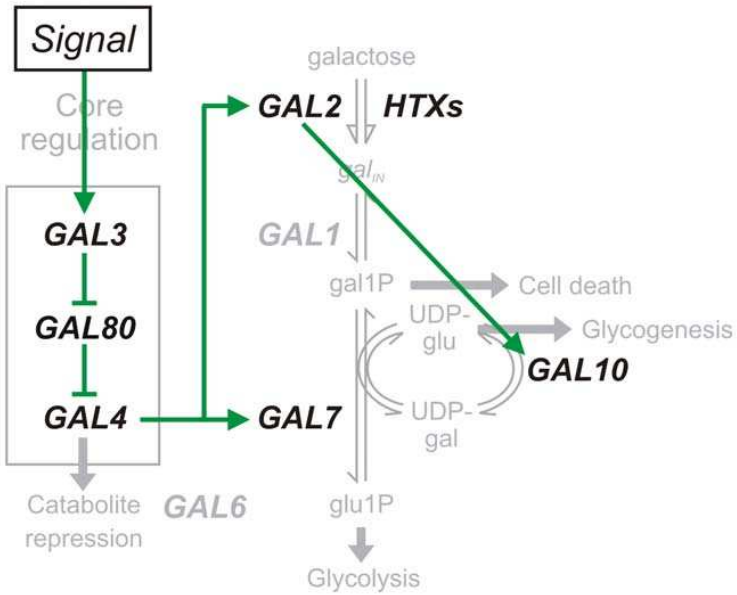
$\Rightarrow$ When the signal-dependent effect of deleting X masks the signal-dependent effect of deleting Y–up to our ability to statistically discriminate–then Y is downstream.

Results

Reconstruction from growth rate



Reconstruction from Gal10p-expression



- Growth rate data helped find metabolic links, some regulatory links
- Expression data helped find regulatory links
- Together, the two recreated virtually the whole network

Genetic interaction networks

- Genetic interaction networks are built based observing a phenotype in WT, single- and pairwise-deletion conditions.

+ Computationally easy to construct

+ May reveal pathways (e.g., synthetic lethality)

+ May be directed, and predictive (e.g. epistasis)

− They require deletion experiments!

Questions?