Static network inference (Barcelona variant)

BCH 5101: Analysis of -Omics Data

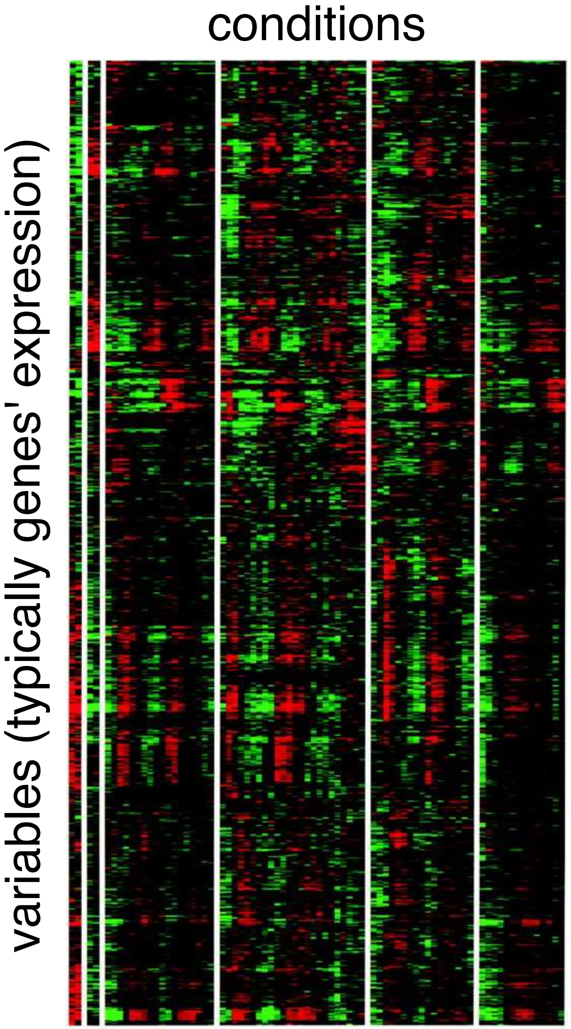
Why network inference?

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?

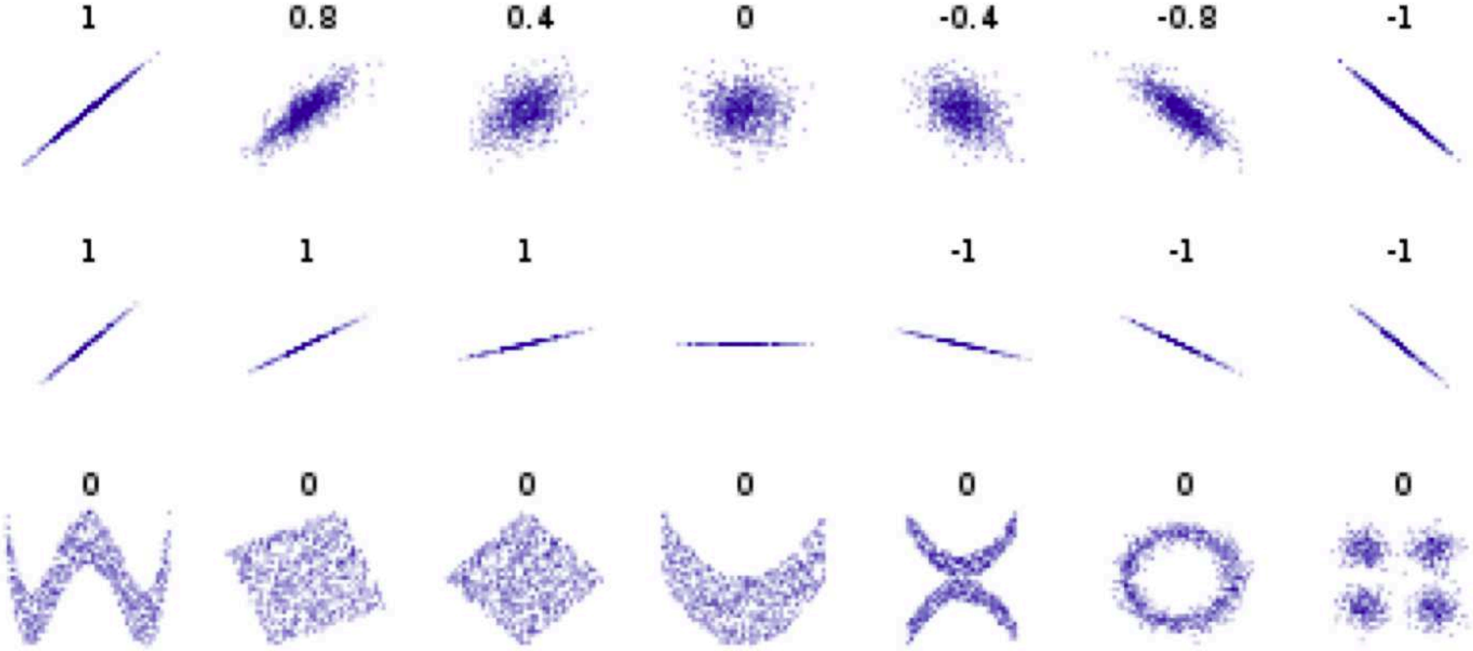
- When are two variables similar enough?

Pearson's linear correlation coefficient

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

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

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

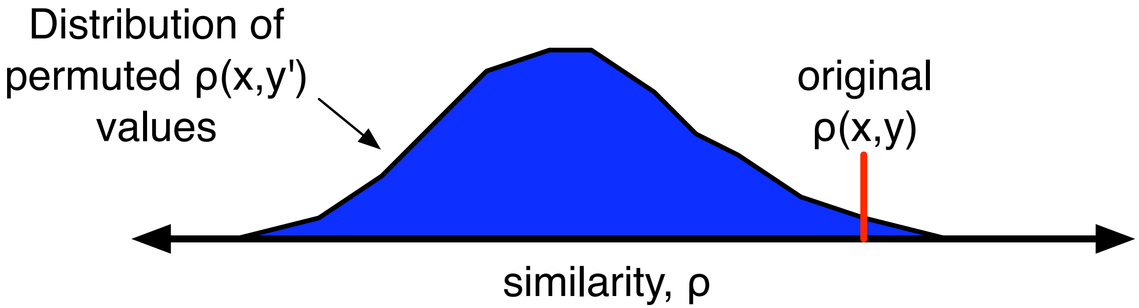
- Suppose you've got paired data on two variables, x and y:

$$\begin{array}{c|cccccccc} x & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\ y & y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 & y_8 \end{array}$$

- 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.:

$$\begin{array}{c|cccccccc} x & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\ y' & y_3 & y_7 & y_6 & y_1 & y_5 & y_4 & y_2 & y_8 \end{array}$$

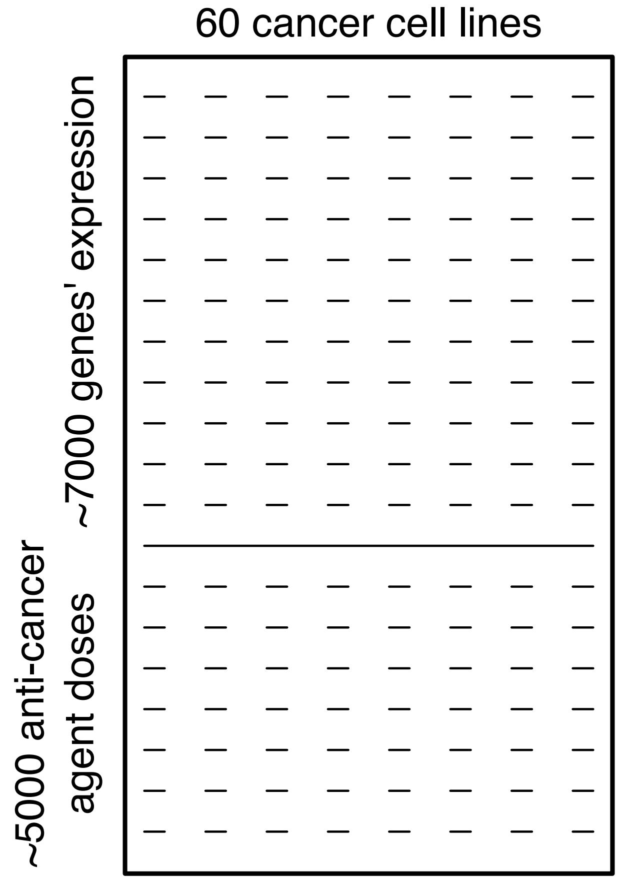
- The location of the original $\rho(x, y)$ with respect to the permuted ρ values gives a p-value.



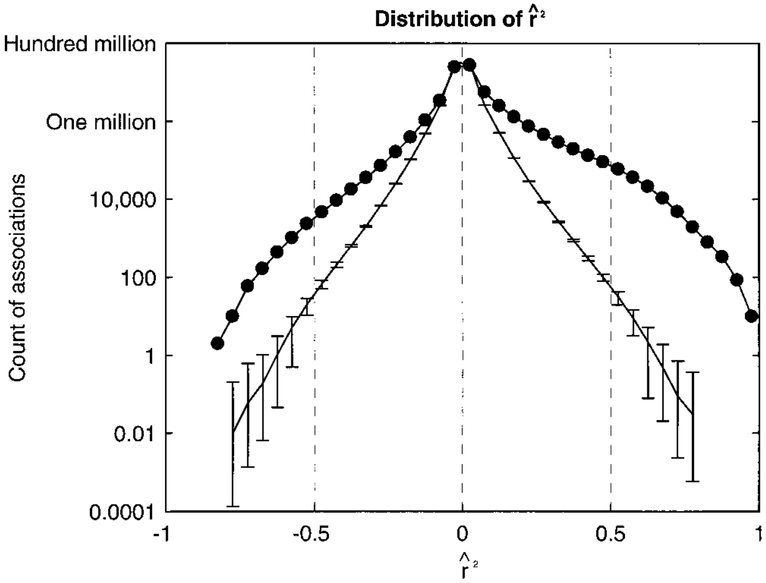
- Approach is distribution and similarity-measure neutral! Still need to choose a p-value threshold (or FDR)...

Back to Butte et al.

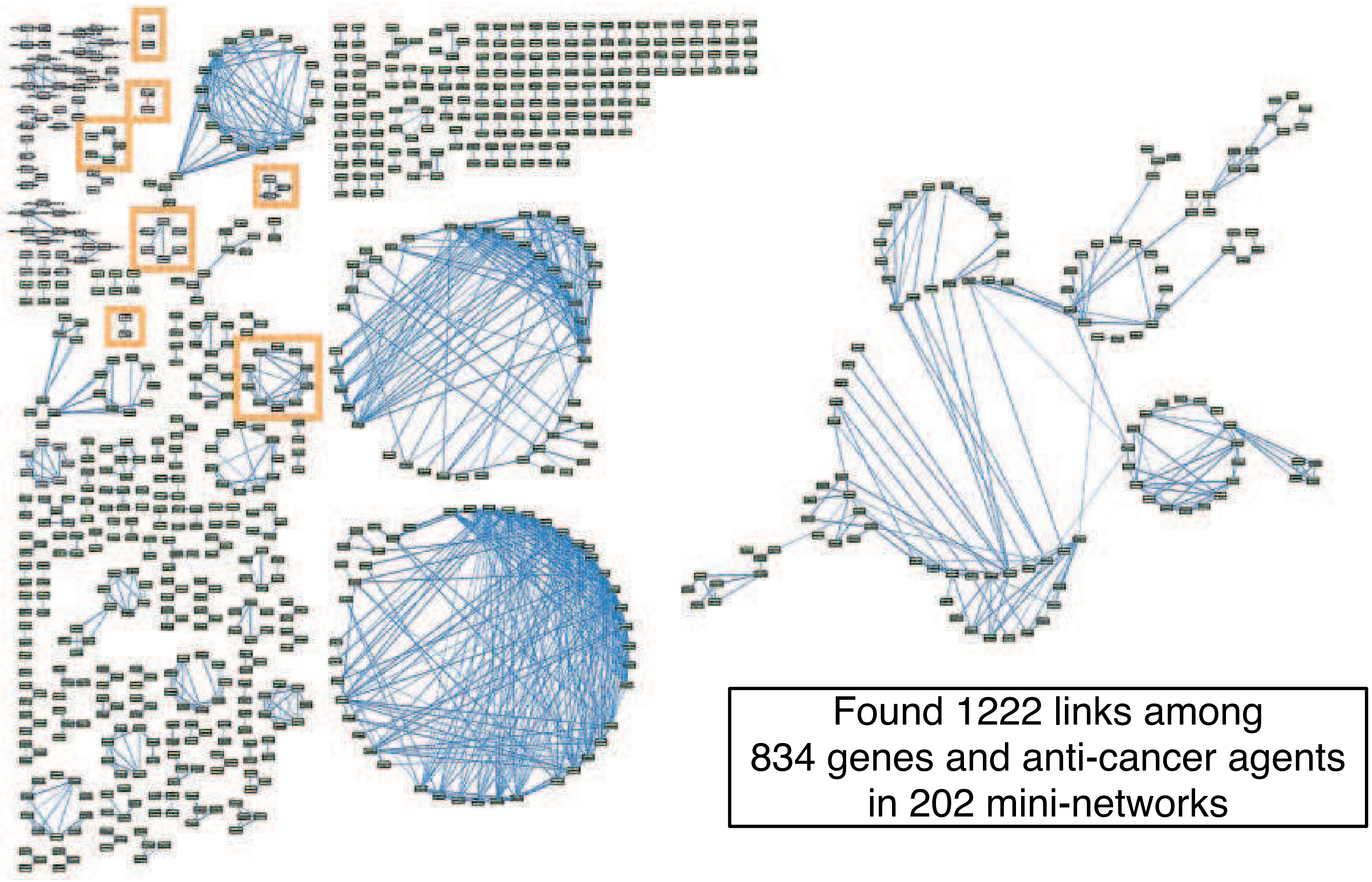
Butte *et al.* took 60 cancer cell lines as conditions, and used as variables microarray expression data for 6,701 genes and susceptibility of those lines to 4,991 anti-cancer agents.



- 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



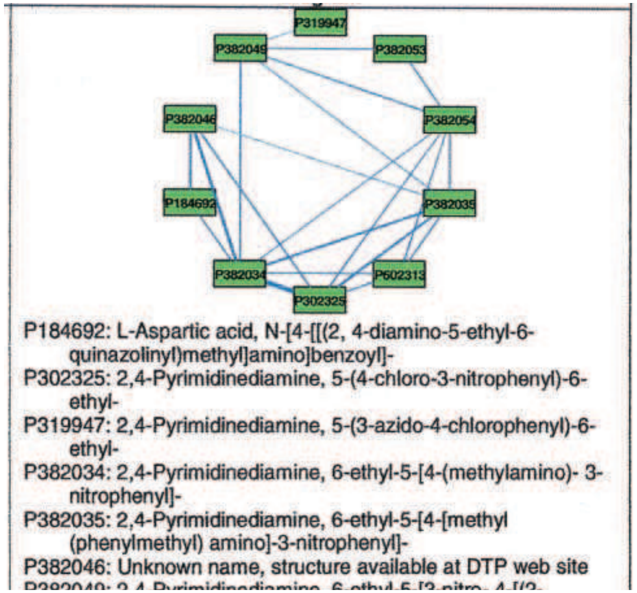
Network found



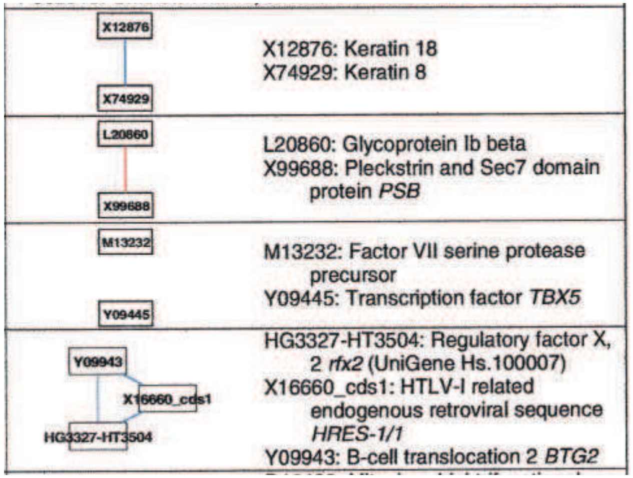
Found 1222 links among
834 genes and anti-cancer agents
in 202 mini-networks

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 ± 0.80 were drawn with line thickness proportional to $\hat{r}^2$. Features without an association at ± 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



P184692: L-Aspartic acid, N-[4-[[(2, 4-diamino-5-ethyl-6-quinazolinyl)methyl]amino]benzoyl]-
P302325: 2,4-Pyrimidinediamine, 5-(4-chloro-3-nitrophenyl)-6-ethyl-
P319947: 2,4-Pyrimidinediamine, 5-(3-azido-4-chlorophenyl)-6-ethyl-
P382034: 2,4-Pyrimidinediamine, 6-ethyl-5-[4-(methylamino)- 3-nitrophenyl]-
P382035: 2,4-Pyrimidinediamine, 6-ethyl-5-[4-[methyl (phenylmethyl) amino]-3-nitrophenyl]-
P382046: Unknown name, structure available at DTP web site
P382049: 2,4-Pyrimidinediamine, 6-ethyl-5-[3-nitro- 4-[(2-

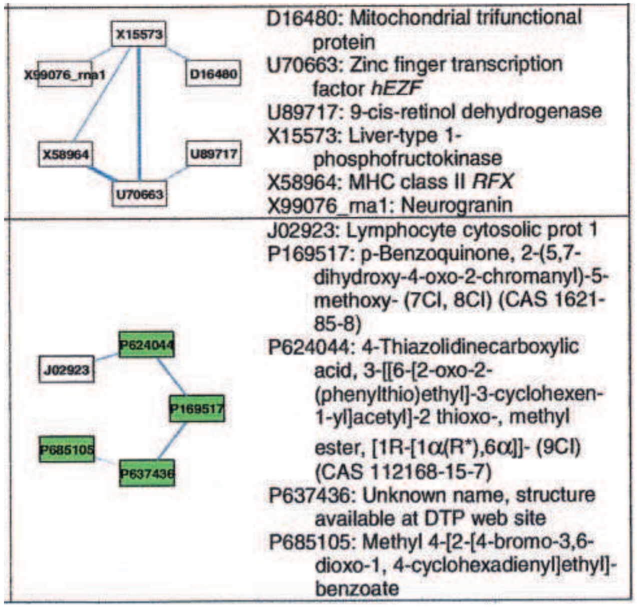


X12876: Keratin 18
X74929: Keratin 8

L20860: Glycoprotein Ib beta
X99688: Pleckstrin and Sec7 domain protein *PSB*

M13232: Factor VII serine protease precursor
Y09445: Transcription factor *TBX5*

HG3327-HT3504: Regulatory factor X, 2 *rfx2* (UniGene Hs.100007)
X16660_cds1: HTLV-I related endogenous retroviral sequence *HRES-1/1*
Y09943: B-cell translocation 2 *BTG2*



D16480: Mitochondrial trifunctional protein
U70663: Zinc finger transcription factor *hEZF*
U89717: 9-cis-retinol dehydrogenase
X15573: Liver-type 1-phosphofructokinase
X58964: MHC class II *RFX*
X99076_rna1: Neurogranin

J02923: Lymphocyte cytosolic prot 1
P169517: p-Benzoquinone, 2-(5,7-dihydroxy-4-oxo-2-chromanyl)-5-methoxy- (7CI, 8CI) (CAS 1621-85-8)
P624044: 4-Thiazolidinecarboxylic acid, 3-[[6-[2-oxo-2-(phenylthio)ethyl]-3-cyclohexen-1-yl]acetyl]-2 thioxo-, methyl ester, [1R-[1α(R*),6α]]- (9CI) (CAS 112168-15-7)
P637436: Unknown name, structure available at DTP web site
P685105: Methyl 4-[2-[4-bromo-3,6-dioxo-1, 4-cyclohexadienyl]ethyl]-benzoate

- 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?

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