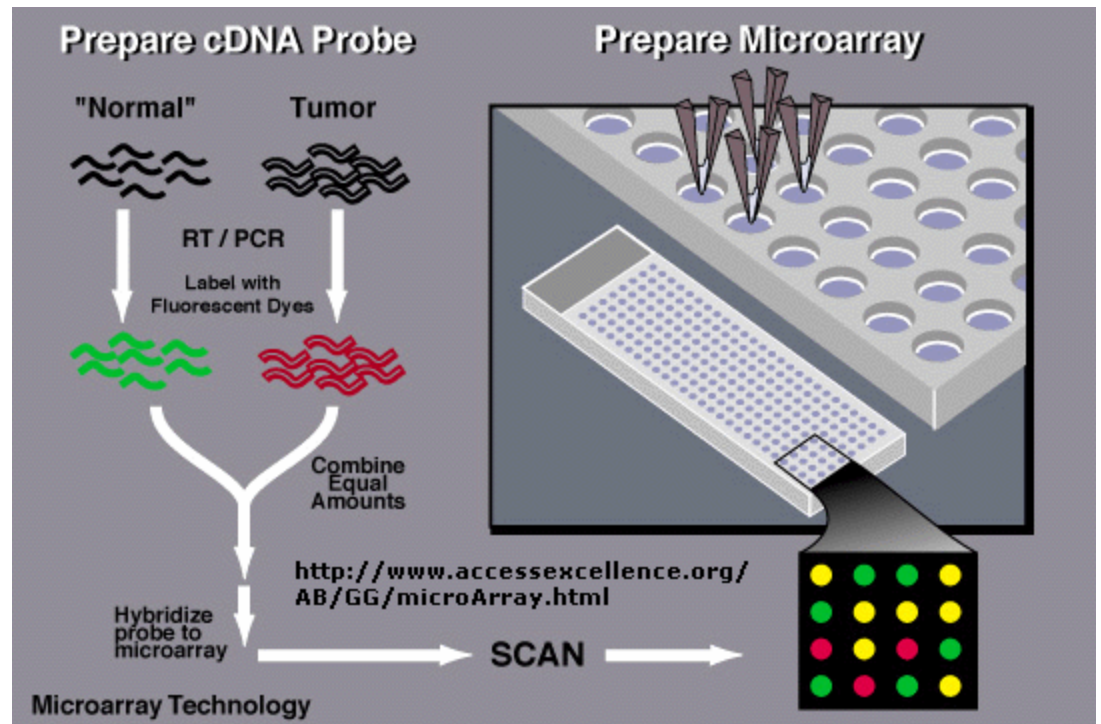
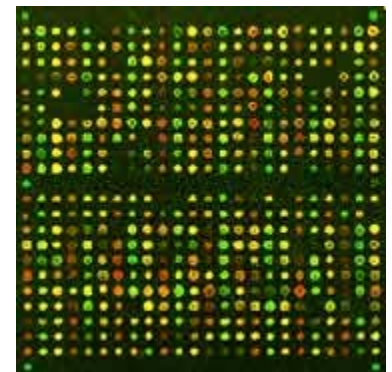


Inference methods

- Probabilistic methods
 - Clustering analysis
 - Data mining
 - Bayesian networks



- Deterministic methods
 - Continuous – Differential equations
 - Discrete - Boolean



Four steps in clustering analysis

- Pair-wise correlation analysis
 - Time-series data
 - spatial data
- Gene co-expression network
- Clustering of genes
- Predicting protein-protein interaction

Drawback: No insight into the causal relationship

Pair-wise similarities between expression profiles:

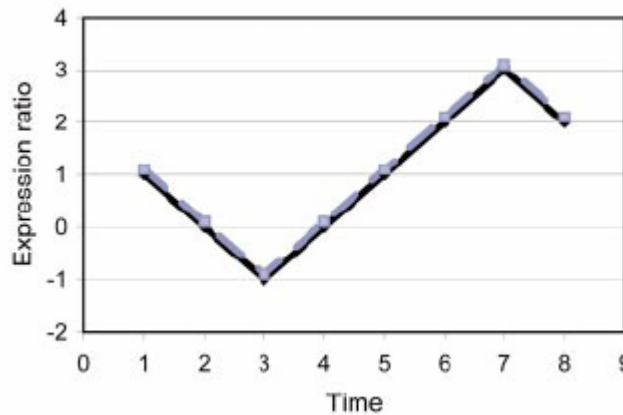
- Pearson correlation
- Squared Pearson correlation coefficient
- Spearman rank correlation
- Jackknife correlation coefficient
- Euclidean distance

Clustering algorithms:

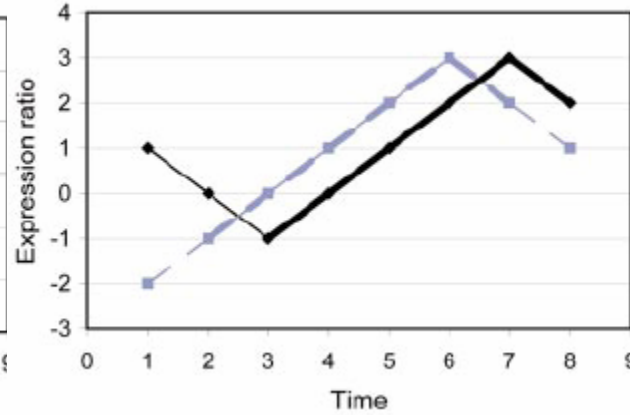
- Bottom-up approach
 - hierarchical clustering
- Top-down approach
 - Self-organizing maps and K-means clustering

Local clustering algorithm

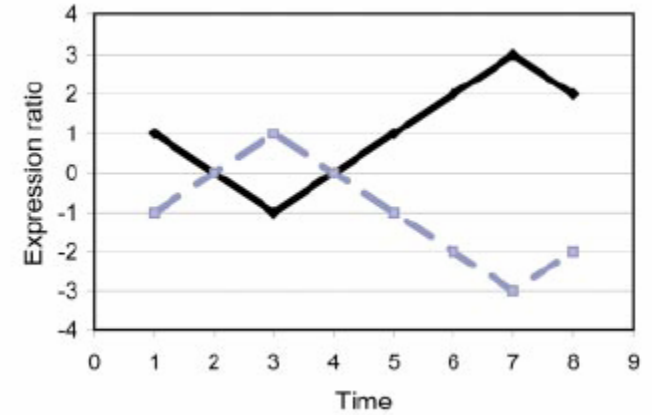
Relationships between expression profiles



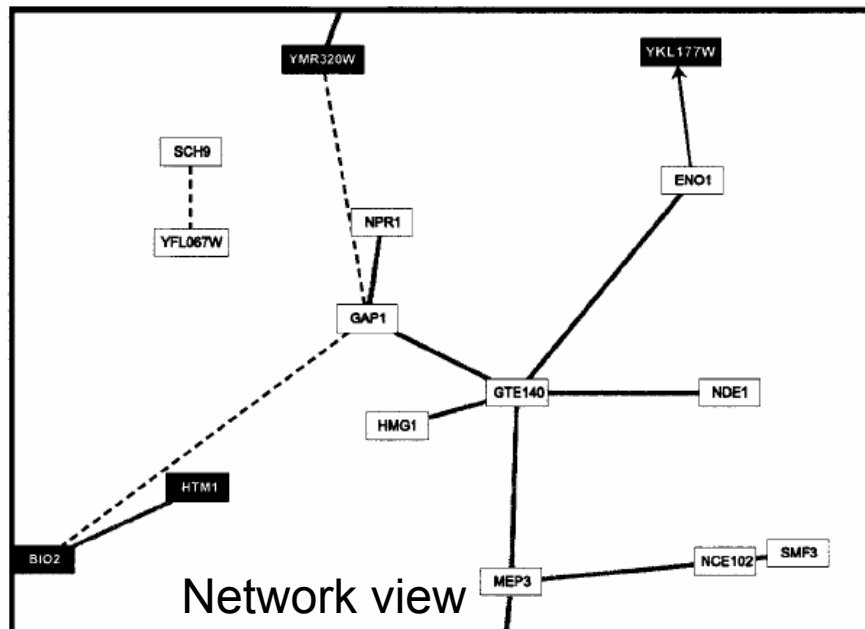
Simultaneous



Time-delayed



Inverted

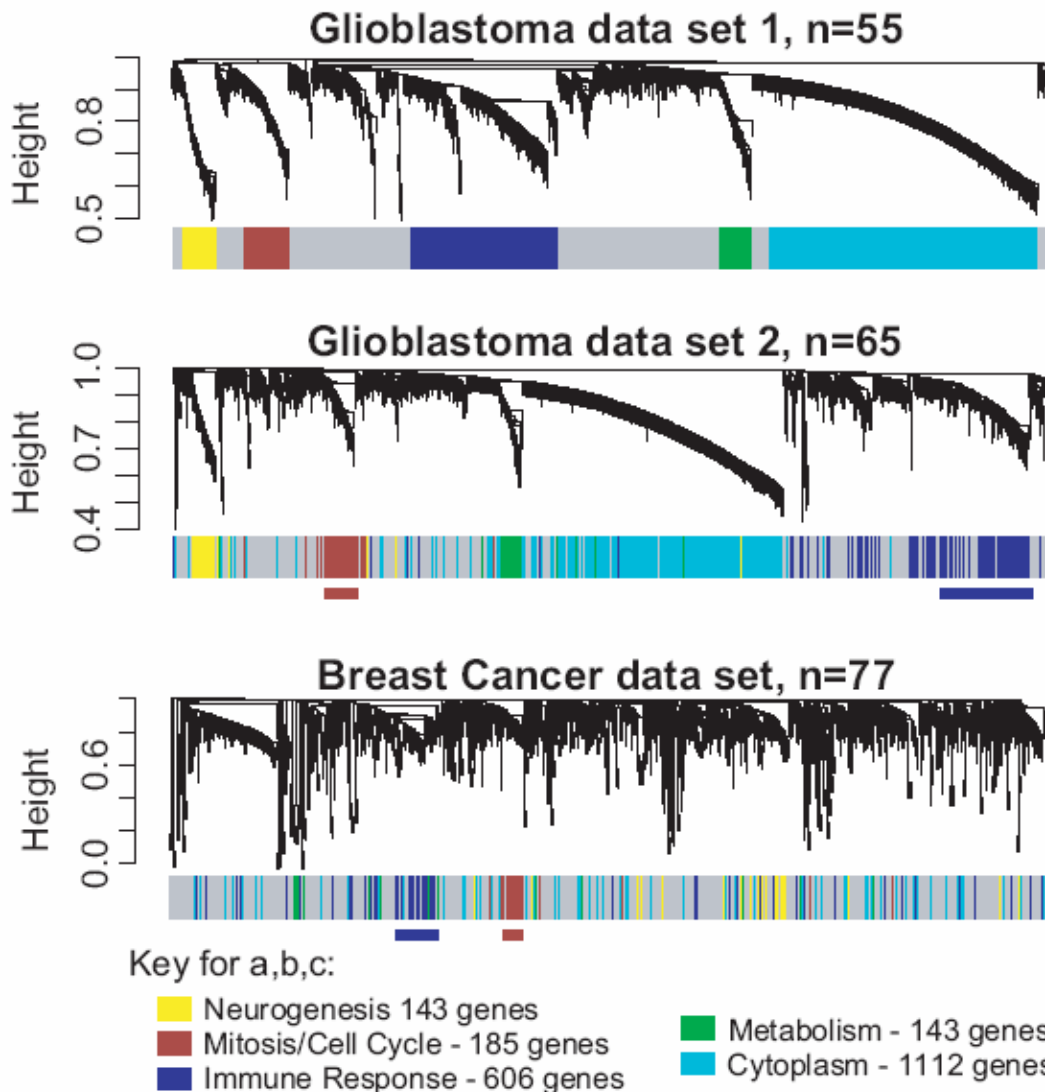


Network view

Predict protein-protein interactions

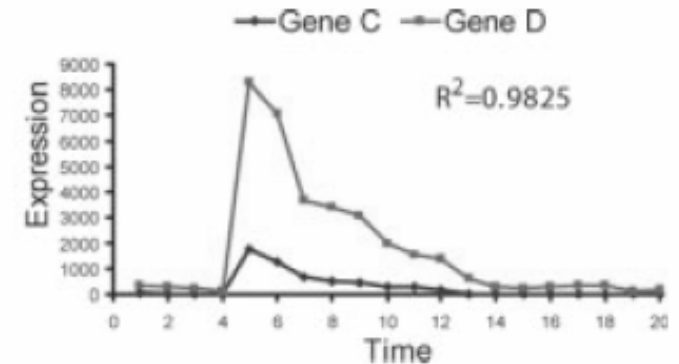
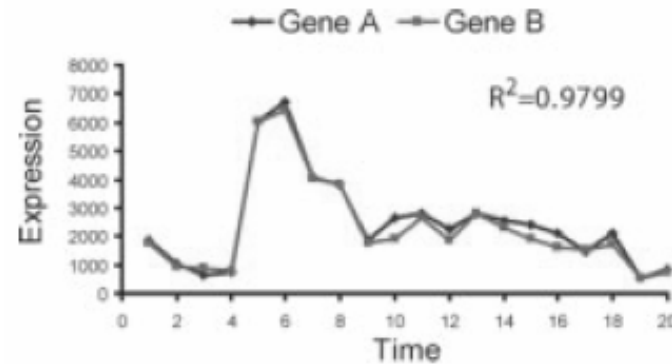
Qian *et al.* (2001) JMB, 314, 1053-1066

Oncogenic signaling network



- Construction of weighted gene co-expression network based on pairwise Pearson correlations
- Hierarchical clustering to detect groups or modules.

Elucidation of directionality



$$SR = \frac{\min(|b_{YX}|, |b_{XY}|)}{\max(|b_{YX}|, |b_{XY}|)}$$

b_{YX}, b_{XY} : regression slopes
e.g. $b_{YX} = 1.004$ and $b_{XY} = 0.976$

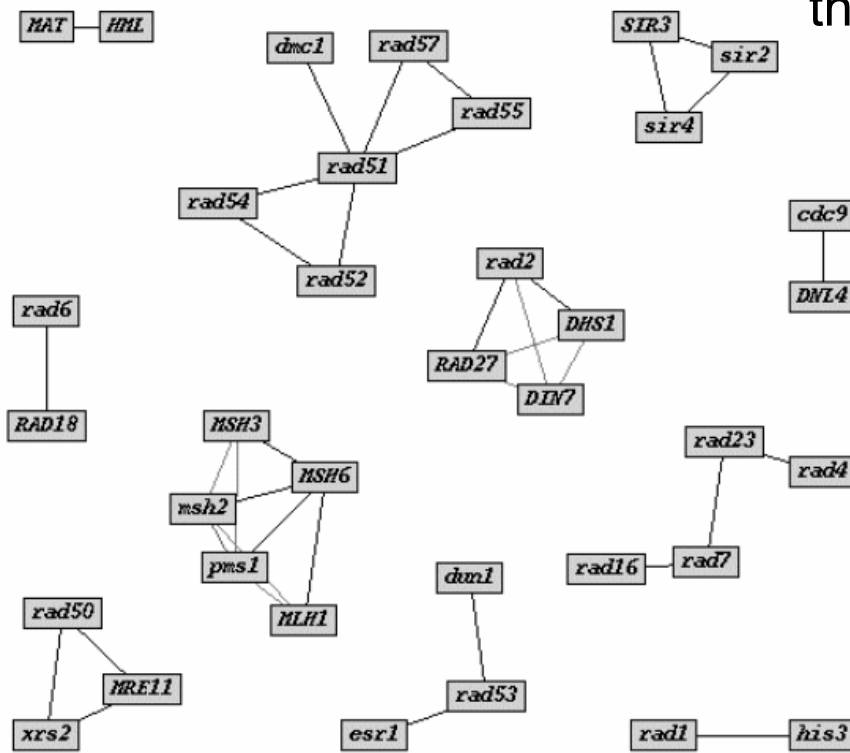
Directionality is assigned to those edges for which $SR \rightarrow 0$

If $SR = \frac{|b_{YX}|}{|b_{XY}|} \Rightarrow Y \rightarrow X$; If $SR = \frac{|b_{XY}|}{|b_{YX}|} \Rightarrow X \rightarrow Y$. $SR_{XY} = 0.97$

Data mining

- Extract information based on the statistical co-occurrence.

Algorithm searched for the co-occurrence of pair of genes resulting in the edge generation according to the user defined threshold.



Network retrieved by the query 'DNA repair'

Bayesian networks

- These protocols are used for sparse datasets
- Probabilistic approach which is capable of handling noise
- The approach is based on the statistical properties of *dependence* and *conditional independence* in the data
- Estimate the confidence in the different features of the network
- Insight into the causal influence

Bayesian analysis

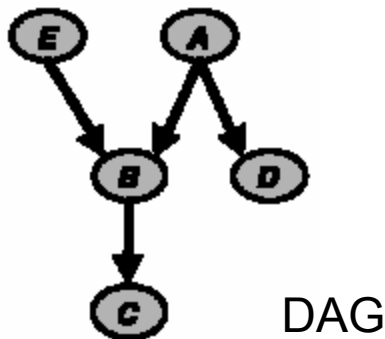
- Bayes' theorem = $P(A|B) = (P(B|A)P(A))/P(B)$
 $\propto L(A|B)P(A)$

$P(B)$ – normalizing constant (NC)

Posterior = (Prior * Likelihood) / NC

Steps in Bayesian analysis

- Define state of the system
 - random variable
 - known information
- Find conditional probability of each node
 - Directed acyclic graph representation of possible causal relationships



Conditional independencies

$$I(A; E)$$

$$I(B; D \mid A, E)$$

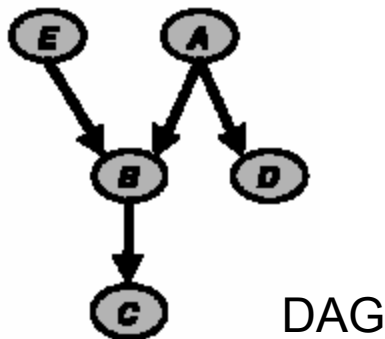
$$I(C; A, D, E \mid B)$$

$$I(D; B, C, E \mid A)$$

$$I(E; A, D)$$

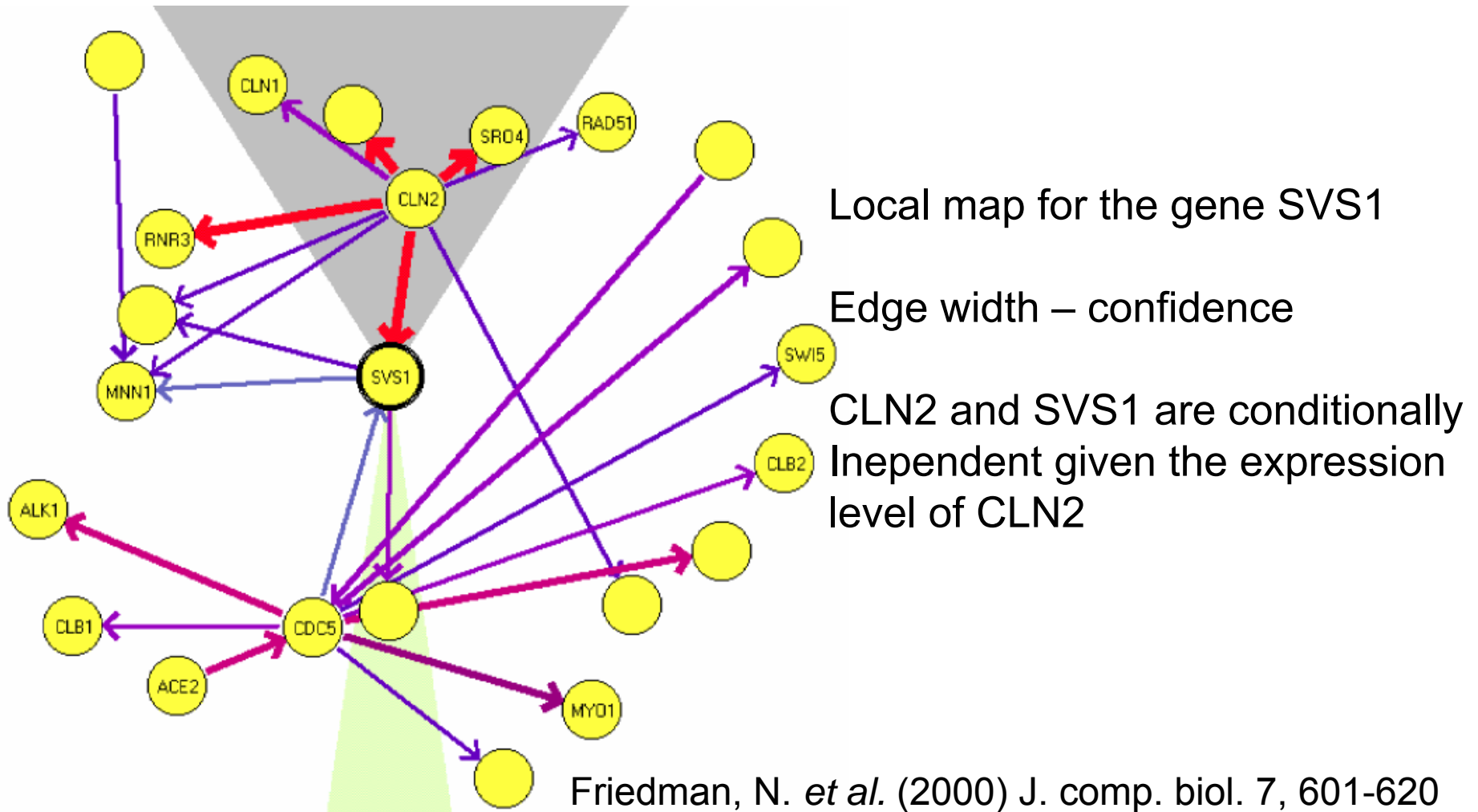
Continued...

- Find joint distribution
 - A set of local joint probability distributions that statistically convey these relationships
- The distribution yielding highest Bayesian score is chosen as the best fit to the data. Benchmarks for weighting are typically obtained from likelihood.



$$P(A, B, C, D, E) = P(A)P(B|A, E)P(C|B)P(D|A)P(E).$$

- Bayesian analysis produces multiple candidate networks
- The links can be established randomly or heuristically
- Iterative search algorithms are employed. e.g. genetic algorithm



Deterministic Methods for Network Inference

A deterministic inference correlates the rate of change in expression level of each gene with the levels of other genes by finding the functional or logical forms of these interdependence relationships.

(Loosely) Two classes of deterministic inference methods:

1) Continuous;

2) Discrete

Continuous Methods

- **Identified as:** systems of linear or nonlinear differential equations in which, for example, rate of change of expression of $X_i(t)$ is a linear combination of concentrations of all other $X(t)$:

$$\frac{dX_i(t)}{dt} = \sum_{j=1}^N w_{ji} X_j(t)$$

- **Pros and cons:**

- can be quite accurate;

- accuracy increases as number of experimental time points increases;

- computational intractability quickly becomes an issue

- **Have been used to infer gene-regulatory networks in:**

- B. subtilis*: Gupta, A., Varner, J. D., and Maranas, C. D.: 'Large-scale inference of the transcriptional regulation of *Bacillus subtilis*', *Computers and Chemical Engineering*, 2005, 29, pp. 565-576.

- Rat*: Chen, T., He, H. L., and Church, G. M.: 'Modeling gene expression with differential equations', *Pac Symp Biocomput*, 1999, pp. 29-40.

An Example: Inferring gene-regulatory networks in *B. subtilis*

A Linear Model of Network Inference

Microarray data

Microarray data

The diagram shows the differential equation $\frac{dX_i(t)}{dt} = \sum_{j=1}^N w_{ji} X_j(t)$. A red arrow points from the left 'Microarray data' text to the derivative term $\frac{dX_i(t)}{dt}$. Another red arrow points from the right 'Microarray data' text to the variable $X_j(t)$. A purple circle highlights the weight parameter w_{ji} , with a purple arrow pointing to it from the text 'To be solved for' below.

$$\frac{dX_i(t)}{dt} = \sum_{j=1}^N w_{ji} X_j(t)$$

To be solved for

$w_{ji} > 0 \Rightarrow$ *activation of i by j*

$w_{ji} < 0 \Rightarrow$ *inhibition of i by j*

Gupta, A., Varner, J. D., and Maranas, C. D.: 'Large-scale inference of the transcriptional regulation of *Bacillus subtilis*', *Computers and Chemical Engineering*, 2005, 29, pp. 565-576.

Optimization

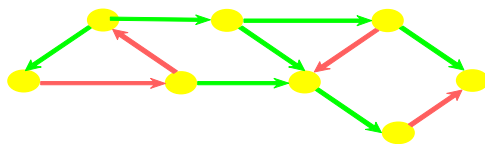
Maximizing Sparseness

$$\text{minimize}_{c_{jk}, w_{ij}^+, w_{ij}^-} \sum_{i,j} (w_{ij}^+ + w_{ij}^-)$$

subject to

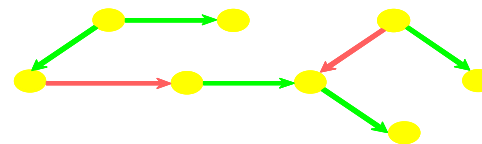
$$\hat{w}_{ij}^+ + \sum_{k=1}^{N-T+1} c_{jk} \hat{v}_{ki} = \hat{w}_{ij}^+ - \hat{w}_{ij}^- \quad \forall i, j = 1, 2, \dots, N$$

$$w_{ij}^+ \geq 0, w_{ij}^- \geq 0 \quad \forall i, j = 1, 2, \dots, N$$



Dense network

Post
Optimization



Sparse network

Discrete Methods

- **Identified as:** Boolean and other logic-based methods that predict discrete regulatory relationships as, for example:

(**Boolean**) A set of nodes $V = \{x_1, \dots, x_n\}$ and a list of Boolean functions $F = (f_1, \dots, f_n)$ where a Boolean function $f = (x_{i1}, \dots, x_{ik})$ with k specified input nodes is assigned to node x_i

(Example to follow in next slide)

- **Pros and cons:**

- More computationally-tractable than continuous methods;
- Less accurate than continuous methods.

- **Much practical application is currently focused on developing and implementing algorithms for large-scale inference:** e.g. REVEAL (REVerse Engineering ALgorithm)

Liang, S., Fuhrman, S., and Somogyi, R.: 'Reveal, a general reverse engineering algorithm for inference of genetic network architectures', Pac Symp Biocomput, 1998, pp. 18-29

An Example: Boolean protocol for network inference

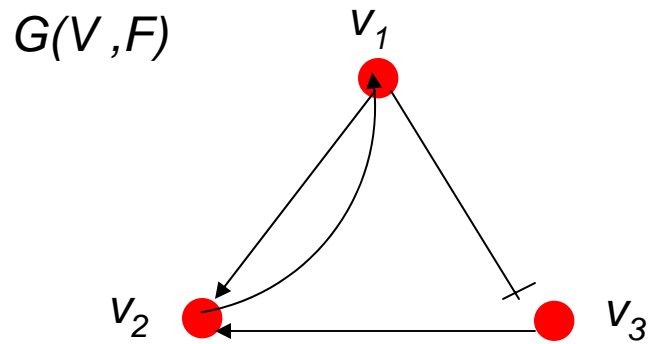
Boolean network basics...

A *Boolean network* $G(V, F)$ consists of a set of nodes, $V = \{v_1, \dots, v_n\}$, representing genes, and a list of Boolean functions, $F = (f_1, \dots, f_n)$.

A Boolean function, $f_j(v_{i_1}, \dots, v_{i_k})$, with inputs from specified nodes, v_{i_1}, \dots, v_{i_k} , is assigned to each node v_j , and this function gives the logical rules (AND, OR, AND NOT, etc...) for the ways in which nodes v_{i_1}, \dots, v_{i_k} will affect the expression of node v_j .

The nodes of a Boolean network can take one of two states: 0 (not-expressed) or 1 (expressed). Thus, the state of each node v_j at time $t+1$ is determined by the states of its input nodes at time t and the Boolean function that dictates how these input nodes affect the expression of v_j .

A picture (no inference yet)



RULES

$v_1' = v_2, v_2' = v_1 \text{ AND } v_3, v_3' = \text{NOT } v_1$

	INPUT			OUTPUT	
v_1	v_2	v_3	v_1'	v_2'	v_3'
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1

An algorithm for inferring a Boolean network

(Akutsu, T. and Miyano, S., Pac. Symp. on Biocomputing 4: 17-28 (1999))

- 1) For each node $v_i \in V$, execute STEP (2).
- 2) If there is a triplet (f_i, v_k, v_h) , satisfying $O_j(v_i) = f_i(I_j(v_k), I_j(v_h))$ for all $j=1, \dots, m$, where O_j and I_j are state outputs and inputs, take f_i as a Boolean function assigned to v_i and take v_k, v_h as input nodes to v_i .
- 2') Enumerate all triplets (f_i, v_k, v_h) satisfying $O_j(v_i) = f_i(I_j(v_k), I_j(v_h))$ for all $j=1, \dots, m$.

What do STEPS 2 and 2' actually mean?!!

An algorithm...continued...

Genes at time t

Genes at time $t+1$

Three Examples

	v_1	v_2	v_3	v_1'	v_2'	v_3'	
I_1	1	0	0	0	0	1	O_1
I_2	0	1	0	0	1	1	O_2
I_3	0	1	1	1	0	1	O_3

G_1

$$v_1' = v_3$$

$$v_2' = v_2 \text{ AND } (\text{NOT } v_3)$$

$$v_3' = \text{NOT } v_3$$

Not consistent, so we'd reject it

G_2

$$v_1' = v_3$$

$$v_2' = v_2 \text{ XOR } v_3$$

$$v_3' = v_1 \text{ OR } v_3$$

Conduct an exhaustive search to find these (f_i, v_k, v_h) triplets. All possible combinations of the 16 Boolean functions (x AND y, x OR y, etc.) for the three node pairs $((v_1, v_2), (v_1, v_3), (v_2, v_3))$ must be checked for each node.

This assumes that only two (or less) nodes determine the next state of each node.

Hybrid Methods

- **Inference methods that bridge the gap between probabilistic and deterministic approaches, usually by incorporating some type of stochastic process (variability, uncertainty) into the inference algorithm.**
- **Pros and cons:**
 - Arguably most accurate and realistic network inference methods;
 - Amount of training data and computational time make methods prohibitive for large networks;

For Example: Probabilistic Boolean Inference

- N Boolean functions are assigned to each node, each with some probability of being selected to advance the state of the node; a machine-learning algorithm must be used to update the state of each node at each time point.