Network inference from dynamic (state) information

Input: components; states of components (in time)
Hypotheses: regulatory framework
Output: proposed regulatory network
Validation: capture known interactions

For inference of gene regulatory networks, the most frequently used state information comes from gene expression arrays (microarrays)

There are several microarray types and methods, for our purposes it suffices to say that a microarray provides a readout of the relative or (semi)absolute expression level of each gene in the array.

Inference methods

- Need expression snapshots:
  - Clustering analysis
  - Bayesian networks
- Need expression timecourse:
  - Continuous – Differential equations
  - Discrete - Boolean
- Need other types of information:
  - Data mining

Clustering analysis

- Pairwise correlation of expression levels of two genes across time or conditions, e.g. by Pearson correlation or Euclidean distance

- These correlations are then clustered
  - Hierarchical clustering - forms a dendogram
    - Successive clusters are formed by aggregation of existing clusters.
    - Difficult to decide cut-off of similarity
  - K-clustering
    - K - predetermined number of groups
    - Decide criteria to group the genes – e.g. Group together genes with similar correlation coefficient.
    - Often used as an exploratory data analysis tool
    - More computationally intensive than hierarchical clustering but optimization can be performed.

Drawback: No insight into the causal relationship

Example of clustering analysis

Qian *et al*. (2001) J Mol. Bio, 314, 1053-1066

- Data set – yeast cell cycle expression, expression ratios
- Score matrix $M_{i,j}$ – matrix of similarities between expression ratio of each pair of genes.
- Two aggregate matrices

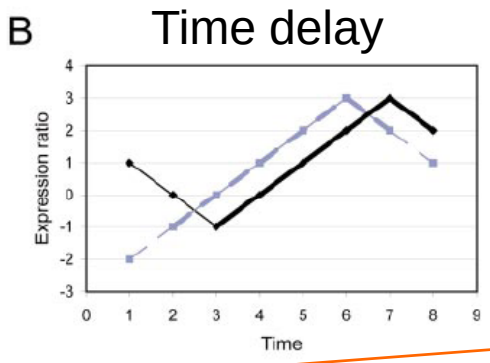
$$E_{i,j} = \max(E_{i-1,\,j-1} + M_{i,j},\, 0)$$

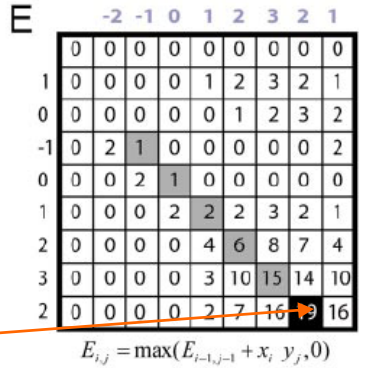
$$D_{i,j} = \max(D_{i-1,\,j-1} - M_{i,j},\, 0)$$

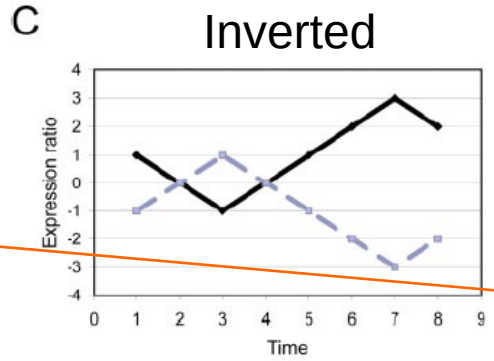
The central idea is to find the maximal aggregated score

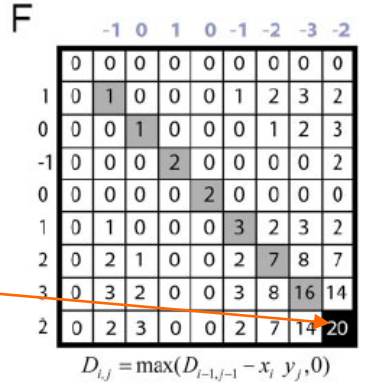
Max(E) off-diagonal – time shifted relationship

Max(D) diagonal – inverted relationship

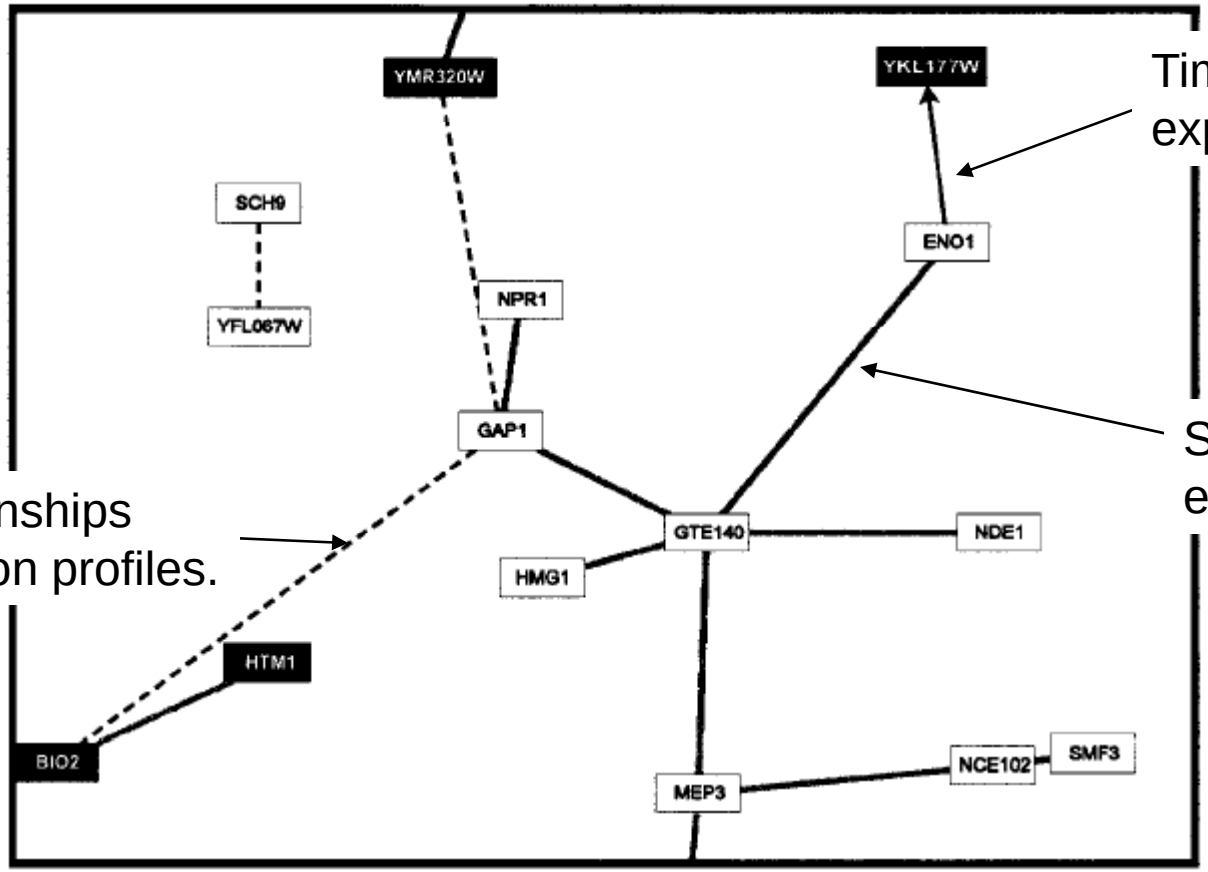








Network is constructed using clustering methods



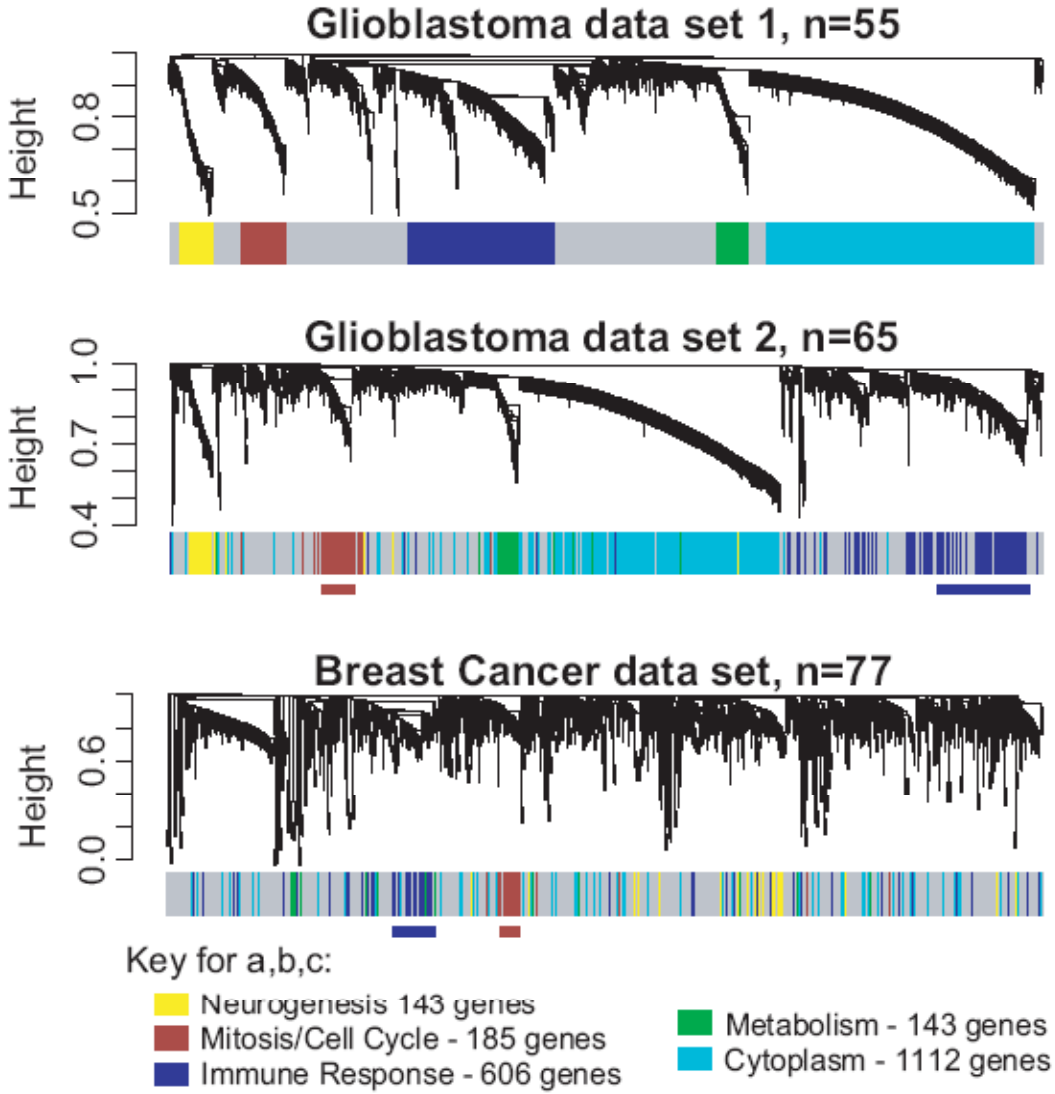
Time delayed expression

Simultaneous expression

Inverted relationships in the expression profiles.

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

Oncogenic signaling network



- Constructed weighted gene co-expression network based on pairwise Pearson correlations
- Hierarchical clustering to detect groups or modules of co-expressed genes.
- Five co-expressed groups in data set 1 are mapped to the other data sets.

Horvath, S. *et al*. (2006) 103, 17402-7

Discussion

1. If genes A and B are co-expressed can we say whether A→B or B→A?

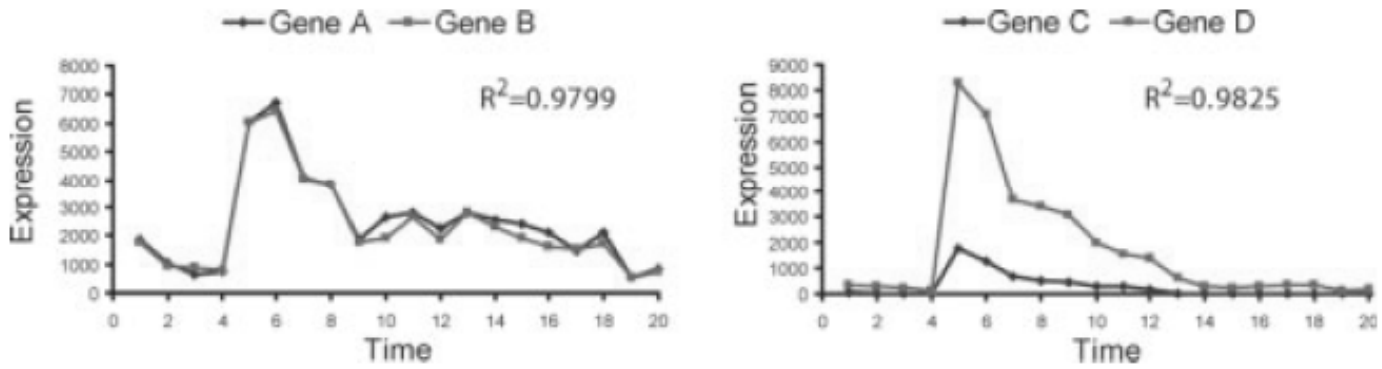
2. What can we infer from time delayed versus simultaneous relationships?

Elucidation of directionality

B. subtilis time series data

Measure of correlation: $R^2 = b_{YX} * b_{XY}$
b_{YX}, b_{XY} *:* regression slopes (regressing Y on X and X on Y)

R^2 values cannot differentiate between expression levels



Slope ratio metric -

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

e.g. b_{YX} = 1.004 *and* b_{XY} = *0.976* SR_{XY} = 0.97

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

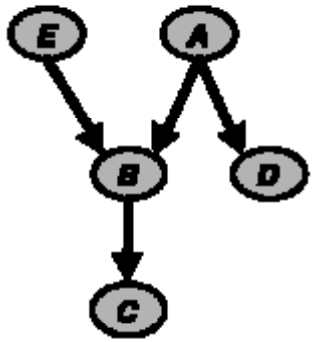
$$\text{If} \quad \mathrm{SR} = \frac{|b_{YX}|}{|b_{XY}|} \Rightarrow Y \rightarrow X; \quad \text{If} \quad \mathrm{SR} = \frac{|b_{XY}|}{|b_{YX}|} \Rightarrow X \rightarrow Y.$$

Gupta, A. *et al*. (2006) Bioinformatics 22, 209-214

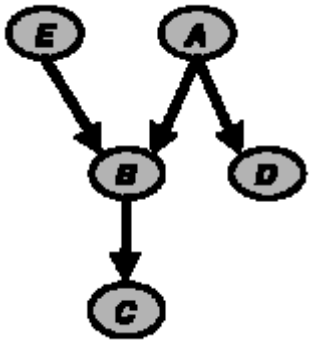
Bayesian networks

- Probabilistic approach usable for sparse datasets which is capable of handling noise
- The approach is based on the statistical properties of *dependence* and *conditional independence* in the data
- Estimates the confidence in the different features of the network
- Insight into the causal influence
- Main step: construct a directed acyclic graph indicating dependencies
- In this graph each node is independent

of its non-descendents (nodes that are not reachable from node), given its parents (upstream nodes).



Steps in Bayesian analysis



Conditional independencies

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

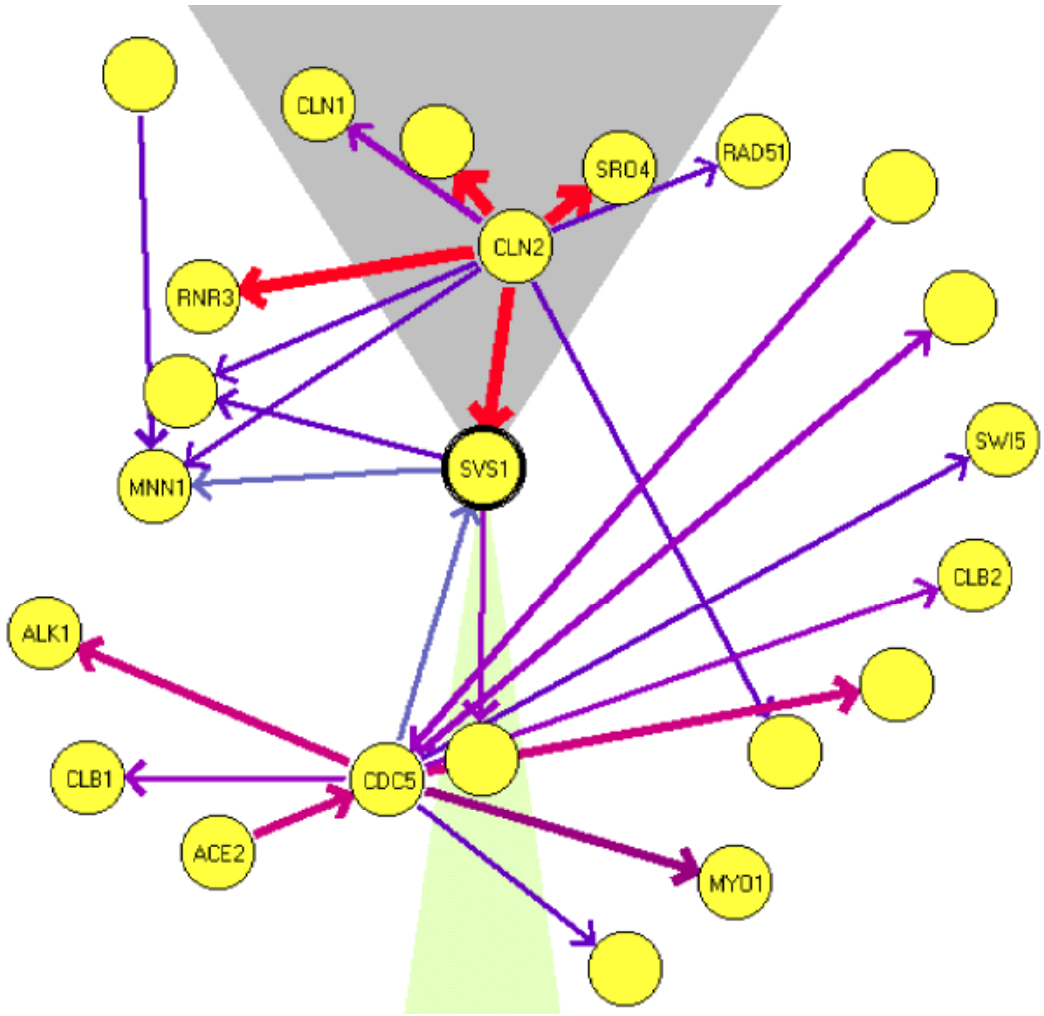
; - independency
| - conditionality
, - or

- Joint probability can be given by a product form,

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

- The data set is used to describe the independencies between different components.
- More than one DAG can imply exactly same independencies.
- A scoring function is designed to evaluate each of this network with respect to the training data to search for the optimal network.
- Start with random or heuristic graph, change edges iteratively. The graph yielding highest Bayesian score is chosen as the best fit to the data

Example



Local map for the gene SVS1

Edge width – confidence

Genes connected to CLN2 are conditionally independent of each other.

Friedman, N. *et al.* (2000) J. Comp. Biol. 7, 601-620

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.

Can only be applied if time-course expression data is available

(Loosely) Two classes of deterministic inference methods:

1) **Continuous;**

2) **Discrete**

Continuous Methods

- Systems of linear or nonlinear differential equations in which, for example, the 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-sale 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.

Example: Inferring gene-regulatory networks in *B. subtilis* using a linear model

Microarray data

Microarray data

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

To be solved for regulatory coefficient

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

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

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

Network filtering

The general solution of this equation accounts for all possible alternate network configurations that are consistent with the experimental data.
Goal: find the sparsest network

Used linear programming (LP)

- To get sparse network we want to maximize number of zero weights
- The objective function minimizes deviation of weights from zero

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

The model is able to identify hub regulators; interactions of highly expressed genes e.g. genes involved in information processing, energy metabolism and signal transduction

Nonlinear model

Mass balance type of equation -

Max. rate

Control input

Growth rate

Degradation rate

$$\frac{\mathrm{d}z_j(t)}{\mathrm{d}t} = r_{T,z_j}(t)u_{z_j}(t) - (\hat{r}_{\mathrm{g}}(t) + \beta_{z_j})z_j(t),$$

$$y^{\mathrm{M}}_{z_j}(t) = f_{z_j}(\mathbf{z}(t), \mathbf{k}), \qquad j = 1, 2, \ldots, N$$

Relationship between intracellular concentrations and signals

- Non linear and linear models are able to identify global regulators with high level confidence
- Non-linear approach captures development regulation, carbon and nitrogen specific interactions

Discrete Methods

- Boolean and other logic-based methods that predict discrete regulatory relationships

  - **Pros and cons:**

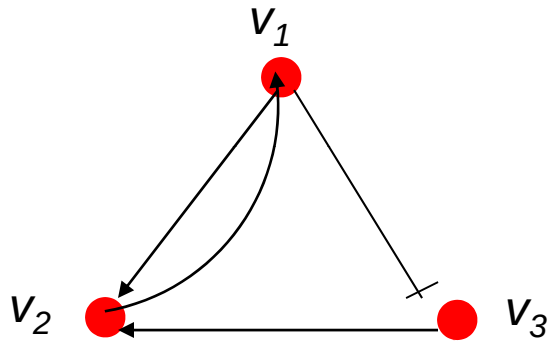
    -More computationally-tractable than continuous methods;

    -Less accurate than continuous methods.

  - **Implemented 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

Example of Boolean model



RULES

$v_1' = v_2,\ v_2' = v_1\ AND\ v_3,\ \ v_3' = NOT\ v_1$

	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

Example of Boolean inference

	Genes at time *t*			Genes at time *t+1*			
	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	0	O_3

G_1

$$v_1' = v_3$$
$$v_2' = v_2 \; AND \; (NOT \, v_3)$$
$$v_3' = NOT \, v_3$$

Consistent

G_2

$$v_1' = v_3$$
$$v_2' = v_2$$
$$v_3' = v_1 \; OR \; v_3$$

Not consistent

Conduct an exhaustive search through all Boolean rules for the nodes. Start with one input rules, then go to two input rules, until a cutoff.

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.

- For example: Probabilistic Boolean Inference

- Pros and cons:
  - Arguably most accurate and realistic network inference methods;
  - Amount of training data and computational time make methods prohibitive for large networks;

Probabilistic Boolean Networks

- Boolean network are completely deterministic. Randomness can be included in the selection of initial states.
- *N* Boolean functions are assigned to each node, each with some probability of being selected to advance the state of the node.
- This can be captured by considering joint probability distribution of all Boolean functions corresponding to all nodes.
- Joint probability for the next time step can be calculated based on present time step.
- A machine-learning algorithm must be used to update the state of each node at each time point.

Shmulevich, et.al.(2002) Bioinformatics, 18(2): 261-274.

Data mining

- Data-mining can be used to infer protein-protein interactions, gene-regulatory relationships, and even metabolic pathways.
- Extract information based on the statistical co-occurrence of features of interest e.g. their inclusion in databases and biomedical journals .
- In this case, by correlating the frequencies of keywords with the probability that a given interaction is addressed in a paper (estimated from a training set), machine learning algorithms can determine whether or not a particular paper is likely to discuss a specific interaction.
- Search tools such as STRING (http://string.embl.de/) employ similar data-mining methods for the inference of both direct and indirect protein-protein interactions in eukaryotes and prokaryotes.
- Example - Algorithm searched for the co-occurrence of pair of genes resulting in the edge generation according to the user defined threshold.

Marcotte, et.al. (2001) Bioinformatics, 17(4), 359-363