Microarray Data Analysis

Gene Regulatory Networks:

Bayesian Networks

國立台灣大學資訊所

Course: 生物資訊與計算分子生物學

2006/11/07

吳漢銘
hmwu@stat.sinica.edu.tw
http://www.sinica.edu.tw/~hmwu

中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica

Outlines

- What are Gene Regulatory Networks?
- Bayesian Networks
- Bayesian Statistics
- Binary Case for Gene Expression Data
- Learning Bayesian Network from Data
- Inference Given a Network
- Application to Microarray Data of Yeast Cell Cycle
- Software

What are Gene Regulatory Networks?

Wikipedia: A ***gene regulatory network*** (also called a **GRN** or genetic regulatory network) is a collection of DNA segments in a cell which interact with each other and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed into mRNA.

Question?

- Is gene *a* regulating gene *b* or vice versa?
- Is the regulation direct
- or indirect where there is a mediating gene c so that a regulates c and then c regulates b?

- Inspecting the finer structure, which are called ***regulatory network***, give us a more intricate view of molecular interactions offering further possibilities for medical interventions.

- Inferring regulatory networks from gene expression data, a process which is called ***reverse-engineering*** of gene regulatory network.

From Gene Expression Data to Gene Regulatory Networks

First Step

Supervised/unsupervised expression profile learning, or extensive data visualization.

From finding gene clusters to finding the functional roles of the respective genes, and moreover, to understanding the underlying biological process.

Next Step

Find potential regulatory sequence elements in genomes. (e.g., transcription factor binding sites, promoter regions,…)

Gene expression data permits us to study finer structure of molecular pathways exposing causal regulation relations between genes.

Hypothesis

- genes with similar expression profiles (i.e. genes that are co-expressed) may have common regulatory mechanisms (i.e. they may be co-regulated), and hence have similar transcription factor binding sites.

Bayesian Approach

Mathematical modeling of regulation inside a network:

- Bayesian network, Boolean network and its generalization,
- ordinary and partial differential equations, qualitative differential equations, stochastic master equations, Petri nets, transform grammars, process algebra, and rule-based formalisms.

Classical Probability: true or physical probability of an event, measured by repeated trials.

Bayesian Probability: the degree of belief in that event, measured by arbitrary techniques for sensible choice.

Bayesian approach: offers a clear separation of structure and parameter optimization, and adding predefined rules and information is easy, widely used for microarray data.

Bayesian Network Modeling

A Bayesian Network for $\boldsymbol{X}=\{X_1, \ldots X_n\}$ consists of

Qualitative part

(a Network Structure):

Directed acyclic graph (DAG) (***G***)

- Nodes - random variables (***V***)
- Edges - direct influence (***E(i, j)***)

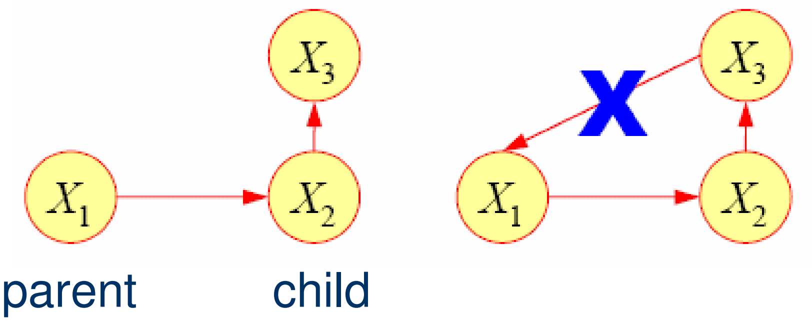
(no cycles allowed)

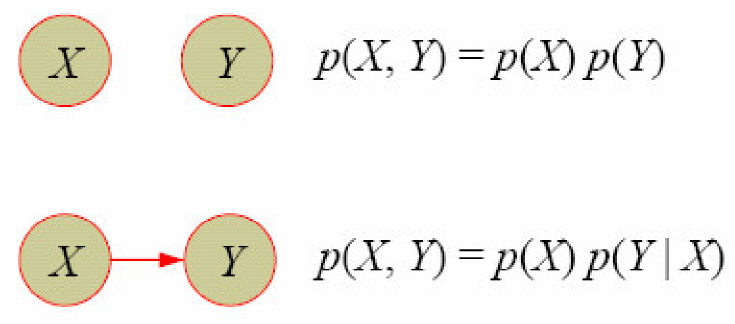
Quantitative part

(a set of probability distribution):

local conditional probability distributions are attached to nodes of graph.

$$\theta = P(X_i \mid Pa_i)$$

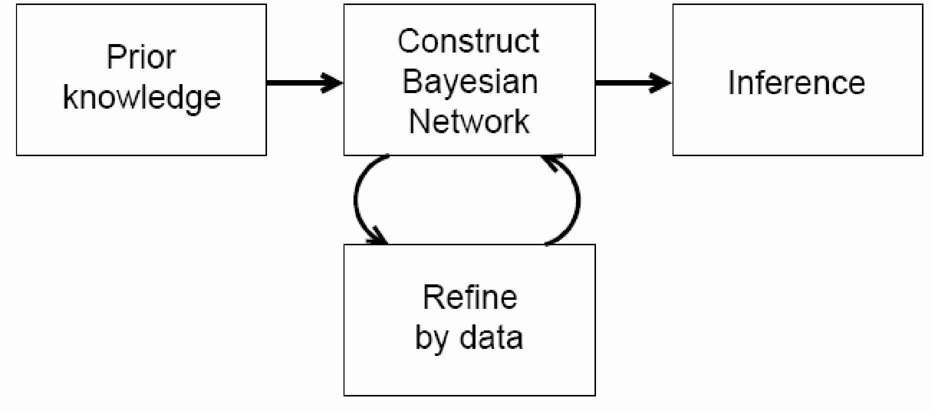




Together: Define a unique distribution in a factored form.

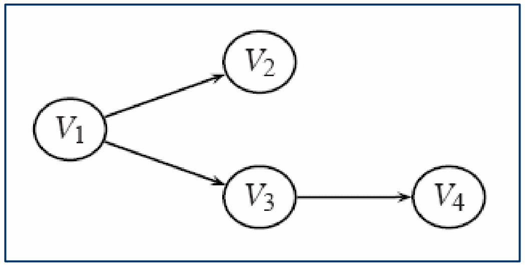
- Arcs represent probabilistic dependence between variables.
- Conditional probabilities encode the strength of dependencies.

Bayesian Networks



Learning Bayesian Networks

- Given a training set $X=\{X_1, \ldots X_n\}$, find a network $B = <G, \theta>$ that best matches X.



Construction

- Determine the variables to model.
- Build DAG that encodes conditional independence edge: cause -> effect
- Assess local probability distribution $\theta = P(X_i | Pa_i)$

Probabilistic Inference

- Compute a probability of interest given a model.
- Use Bayes theorem and simplify by conditional independence.

Bayesian Networks

- *Markov Condition*: Each variable X_i is independent of its non-descendants given its parents.

  ➜ Local probability in X_i depends only on the parents.

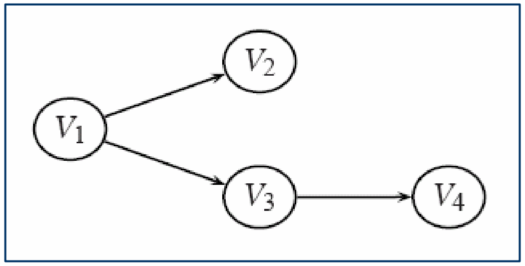
- *Conditional Independence*: Given its parents, X_i is independent from the other nodes in the graph.

- Judea Pearl. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, 1988.
- Heckerman, David: A Tutorial on Learning with Bayesian Networks, MSR-TR-95-06

Gene Expression

- Expression levels X_i of nodes V_i are considered as random variables and the edges represent conditional dependencies between distributions of the random variables.
- $V_i \rightarrow V_j$: gene V_i regulates gene V_j (up or down-regulation).

Learn the network structure from gene expression data.
Problem: Noise, sparse data



Bayesian Statistics

expression level X_i of gene or node V_i at point j ➡ $\{X_i[j] \mid 1 \le i \le n, 1 \le j \le m\}$

joint probability ➡ $\mathrm{P}(X \cap Y) = \mathrm{P}(X, Y) = \mathrm{P}(Y \mid X)\ \mathrm{P}(X) = \mathrm{P}(X \mid Y)\ \mathrm{P}(Y)$

conditional probability ➡ $\mathrm{P}(X \mid Y) = \dfrac{\mathrm{P}(Y \mid X)\ \mathrm{P}(X)}{\mathrm{P}(Y)}$ Bayes' rule

generalization of it forms the chain rule

➡
$$\begin{aligned}
\mathrm{P}(K, X, Y, Z) &= \mathrm{P}(Z \mid K, X, Y)\ \mathrm{P}(K, X, Y) \\
&= \mathrm{P}(Z \mid K, X, Y)\ \mathrm{P}(Y \mid K, X)\ \mathrm{P}(K, X) \\
&= \mathrm{P}(Z \mid K, X, Y)\ \mathrm{P}(Y \mid K, X)\ \mathrm{P}(X \mid K)\ \mathrm{P}(K).
\end{aligned}$$

X and Y are independent ➡ $\mathrm{P}(X \cap Y) = \mathrm{P}(X)\mathrm{Pr}(Y)$

X and Y are independent for given a value of Z ➡ $\mathrm{P}(X \mid Z, Y) = \mathrm{P}(X \mid Z)$

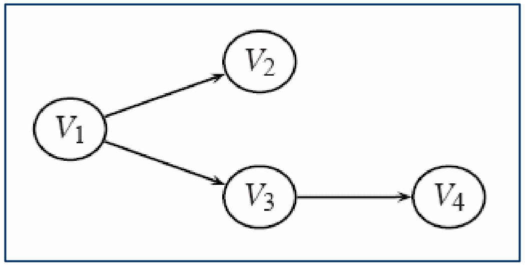
Bayesian Statistics

Bayes Formula:

$$\underbrace{P(\ \vartheta \mid D\ )}_{\text{Posterior}} = \frac{\overbrace{P(\vartheta)}^{\text{Prior}}\ \overbrace{P(D \mid \vartheta)}^{\text{likelihood}}}{P(D)}$$

Joint distribution:

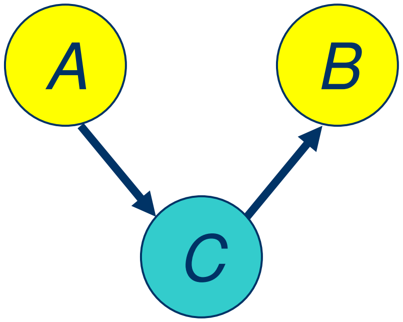
$$P(\vartheta, D) \;=\; P(\vartheta)\ P(D \mid \vartheta)$$



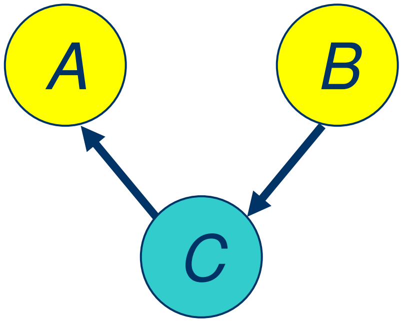
$$\mathrm{P}(X_1, X_2, X_3, X_4) = \ \mathrm{P}(X_1)\ \mathrm{P}(X_2 \mid X_1)\ \mathrm{P}(X_3 \mid X_1)\ \mathrm{P}(X_4 \mid X_3)$$

$$\mathrm{P}(X_4 \mid X_1, X_2, X_3) = \ \mathrm{P}(X_4 \mid X_3)$$

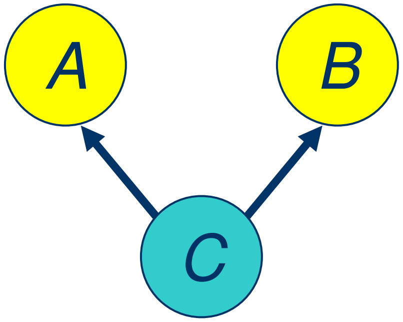
Exercise



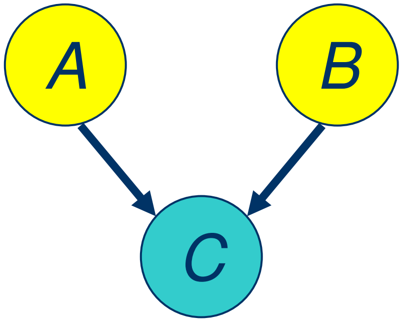
$P(A,B,C) = P(B \mid C)P(C \mid A)P(A)$



$P(A,B,C) = P(A \mid C)P(C \mid B)P(B)$

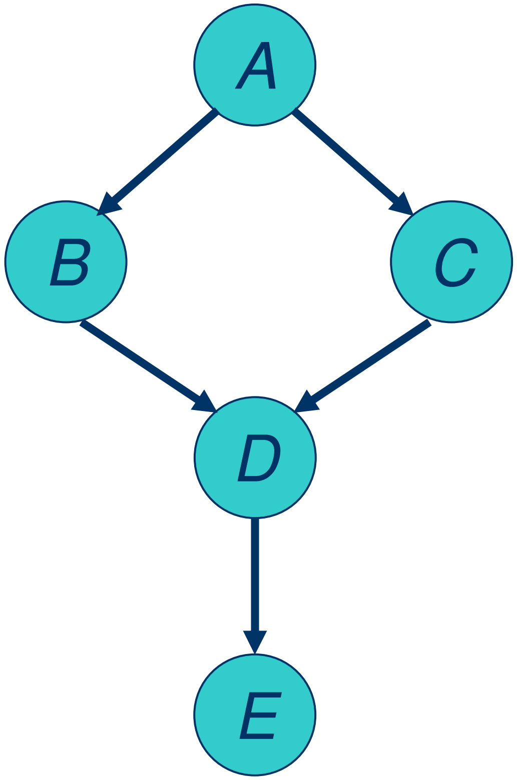


$P(A,B,C) = P(A \mid C)P(B \mid C)P(C)$



$P(A,B,C) = P(C \mid A,B)P(A)P(B)$

More Complex Example



$$P(A, B, C, D, E) = \prod_i P(\text{node}_i | \text{parents}_i)$$

Conditional independence

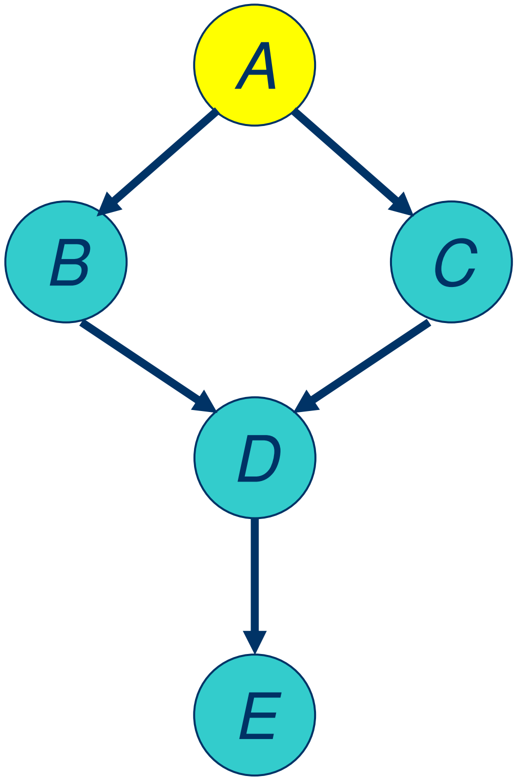
I(A; E),
I(B;D | A, E),
I(C;A,D,E |B),
I(D;B,C,E|A)
I(E;A,D)

Joint distribution

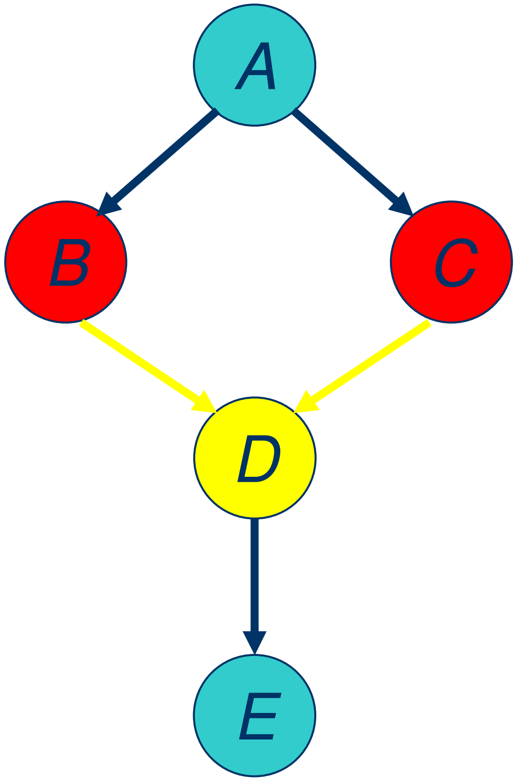
P(A, B, C, D, E) = P(A)P(B|A, E) P(C|B)P(D|A) P(E)

Biological interpretation

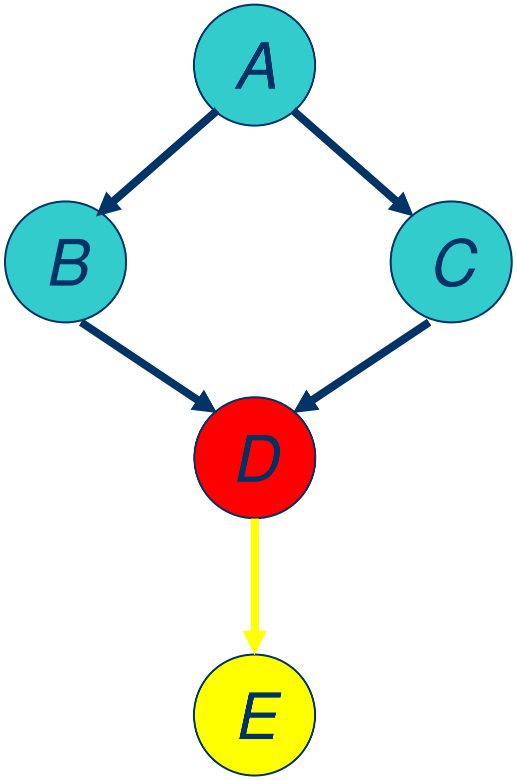
- Is the structure right? Is the order of regulation correct?
- If there are several possible structures with respect to the experiments done so far, which one is right? so a graph represents hypothesis based on current knowledge.



Initiation of cell (sub-)cycle



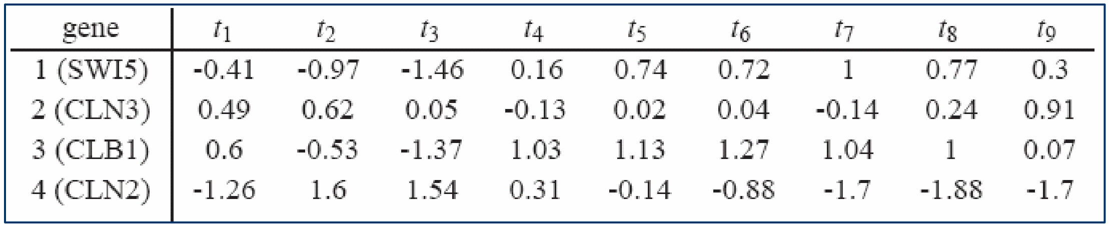
Co-regulation



Mediation

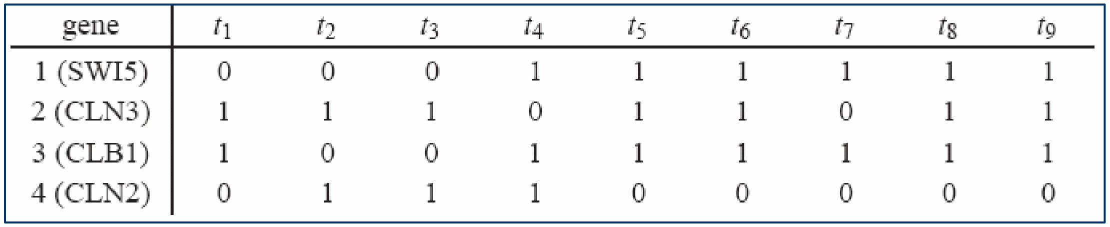
Binary Case

- Binary case: gene can be "off" (0) or "on" (1), but not both.



gene	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
1 (SWI5)	-0.41	-0.97	-1.46	0.16	0.74	0.72	1	0.77	0.3
2 (CLN3)	0.49	0.62	0.05	-0.13	0.02	0.04	-0.14	0.24	0.91
3 (CLB1)	0.6	-0.53	-1.37	1.03	1.13	1.27	1.04	1	0.07
4 (CLN2)	-1.26	1.6	1.54	0.31	-0.14	-0.88	-1.7	-1.88	-1.7

- Having a fixed structure, the conditional probabilities are easy to calculate.



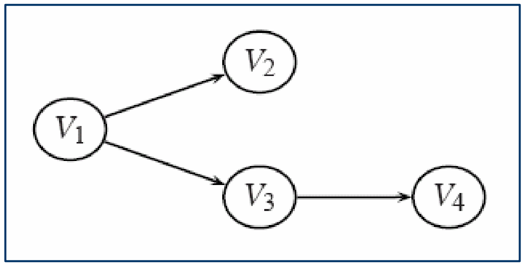
gene	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
1 (SWI5)	0	0	0	1	1	1	1	1	1
2 (CLN3)	1	1	1	0	1	1	0	1	1
3 (CLB1)	1	0	0	1	1	1	1	1	1
4 (CLN2)	0	1	1	1	0	0	0	0	0

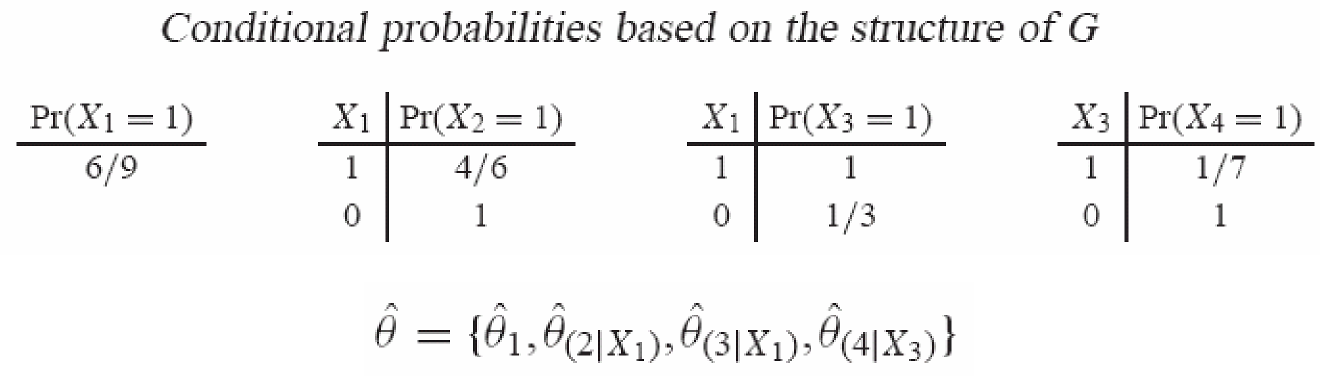
Binomial Distribution

***X* ~ *B*(*n*, *p*)**

The probability of getting exactly k successes is given by the probability mass function:

$$f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$





Conditional probabilities based on the structure of G

$\Pr(X_1 = 1)$
6/9

X_1	$\Pr(X_2 = 1)$
1	4/6
0	1

X_1	$\Pr(X_3 = 1)$
1	1
0	1/3

X_3	$\Pr(X_4 = 1)$
1	1/7
0	1

$$\hat{\theta} = \{\hat{\theta}_1, \hat{\theta}_{(2|X_1)}, \hat{\theta}_{(3|X_1)}, \hat{\theta}_{(4|X_3)}\}$$

Type of Variables

- **Discrete variables**: Usage of more states for genes like "low", "medium", and "high" follows the multinomial theory conveniently generalizing the binomial theory.

The probabilities are given by

$$P(X_1 = x_1, \ldots, X_k = x_k) = \begin{cases} \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} & \text{when } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise.} \end{cases}$$

- **Continuous variables**: Linear Gaussian model

$$P(X \mid u_1, \ldots, u_k) \sim N(a_0 + \sum_i a_i \cdot u_i, \sigma^2).$$

parents $U_1, \ldots, U_k$

- **Hybrid Networks**

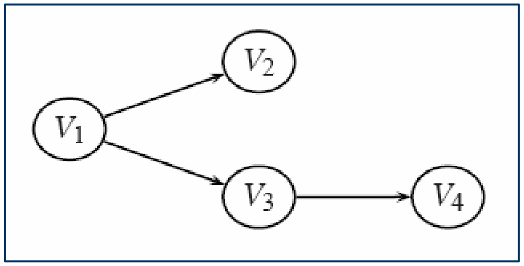
Calculating BN Parameters

- Having graph G and the expression matrix D, our aim is to obtain distribution dependency parameters θ ={ θ 1, θ 2,…} that are fitted best to the structure of G and data D.
- Maximal likelihood method

$$\mathrm{L}(\theta : D) = \mathrm{P}(D : \theta) = \prod_{j=1}^{m} \mathrm{P}(X_1[j], X_2[j], \ldots X_n[j] : \theta).$$

$$\mathrm{L}(\theta : D) = \prod_{j=1}^{m} \mathrm{P}(X_1[j] : \theta_1) \cdot \prod_{j=1}^{m} \mathrm{P}(X_2[j] \mid X_1[j] : \theta_2)$$

$$\cdot \prod_{j=1}^{m} \mathrm{P}(X_3[j] \mid X_1[j] : \theta_3) \cdot \prod_{j=1}^{m} \mathrm{P}(X_4[j] \mid X_3[j] : \theta_4)$$



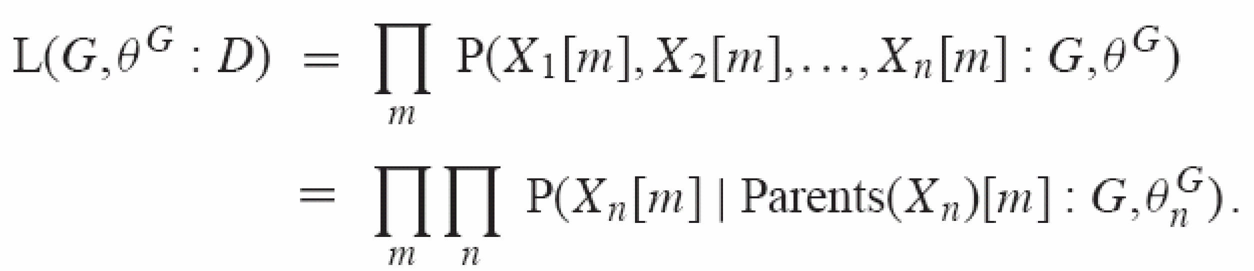
$$\mathrm{L}(\theta : D) = \prod_{i} \mathrm{L}_i(\theta_i : D)$$

$$\mathrm{L}_1(\theta_1 : D) = \mathrm{P}(D : \theta_1) = \prod_{j=1}^{m} \mathrm{P}(X_i[j] : \theta_1) = (\theta_1)^{N(X_1=1)}(1-\theta_1)^{m-N(X_1=1)}$$

$$\Pr(X_1 = 1) = \hat{\theta}_1 = \frac{N(X_1 = 1)}{N(X_1 = 1) + N(X_1 = 0)} = 6/9$$

$$\begin{aligned}
\hat{\theta}_{(2|X_1=1)} &= 4/6, & \hat{\theta}_{(2|X_1=0)} &= 1, \\
\hat{\theta}_{(3|X_1=1)} &= 1, & \hat{\theta}_{(3|X_1=0)} &= 1/3, \\
\hat{\theta}_{(4|X_3=1)} &= 1/7, & \hat{\theta}_{(4|X_3=0)} &= 1.
\end{aligned}$$

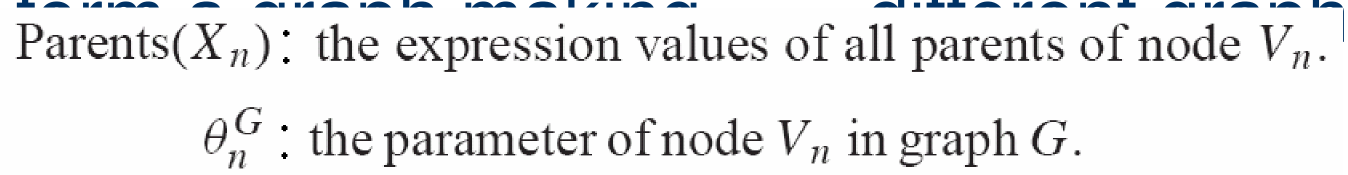
Searching Bayesian Network Structure

- **Learning**: find network structure which fits the prior knowledge and data.
- Given a graph G, we know now how to calculate the par$\binom{4}{2}$neter set θ G maximizing the likelihood score $L(G, \theta: D)$.$2^{\binom{4}{2}}$
- 

$$\mathrm{L}(G,\theta^G : D) = \prod_m \mathrm{P}(X_1[m], X_2[m], \ldots, X_n[m] : G, \theta^G)$$

$$= \prod_m \prod_n \mathrm{P}(X_n[m] \mid \mathrm{Parents}(X_n)[m] : G, \theta_n^G).$$

sting of four ble be taken to form a graph making different graph in total.



$\mathrm{Parents}(X_n)$: the expression values of all parents of node V_n.

θ_n^G : the parameter of node V_n in graph G.

Given a random sample D compute the posterior probability

Searching Bayesian Network Structure

$$\log(\mathrm{L}(G,\theta^{G} : D)) = m\sum_{i=1}^{n}(\mathrm{I}(V_i,\mathrm{Parents}(V_i)) - \mathrm{H}(V_i))$$

$$\mathrm{I}(V_i,\mathrm{Parents}(V_i)) = \sum_{\substack{x=0,1\\(\forall i):x_i=0,1}} \Pr(X_i = x \textstyle\bigcap_i P_i = x_i)\,\mathrm{lg}\,\dfrac{\Pr(X_i = x \bigcap_i P_i = x_i)}{\Pr(X_i = x)\prod_i \Pr(P_i = x_i)}\,.$$

mutual expression information between node V_i and its parent nodes $\mathrm{Parents}(V_i) = \{P_1, P_2, \ldots\}$

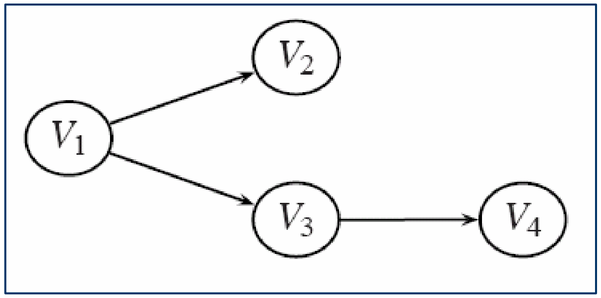
$$\xcancel{\mathrm{H}(V_i) = \sum_{x=0,1} \Pr(X_1 = x)\,\mathrm{lg}\,\frac{1}{\Pr(X_1 = x)}}$$

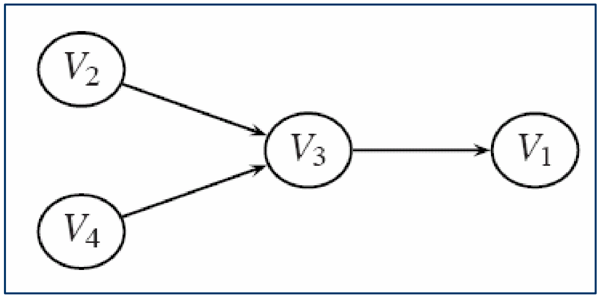
~~entropy of expression values X_i of node V_i~~

- Function $I(V_i, Parents(V_i)) \geq 0$ measures how much information the expression values of nodes $Parents(V_i)$ provide about V_i.
- If V_i is independent of parent nodes, then $I(V_i, Parents(V_i))$ has the value of zero.
- If V_i is totally predictable for given values of $Parents(V_i)$, then $I(V_i, Parents(V_i))$ reduces into the entropy function H(V_i).
- It should be noted that in general I(X,Y) = I(Y , X) so the direction of edges matters.

Searching Bayesian Network Structure

- Is the structure optimal with respect to the data?
- Search for high scoring structure by greedy search, simulated annealing



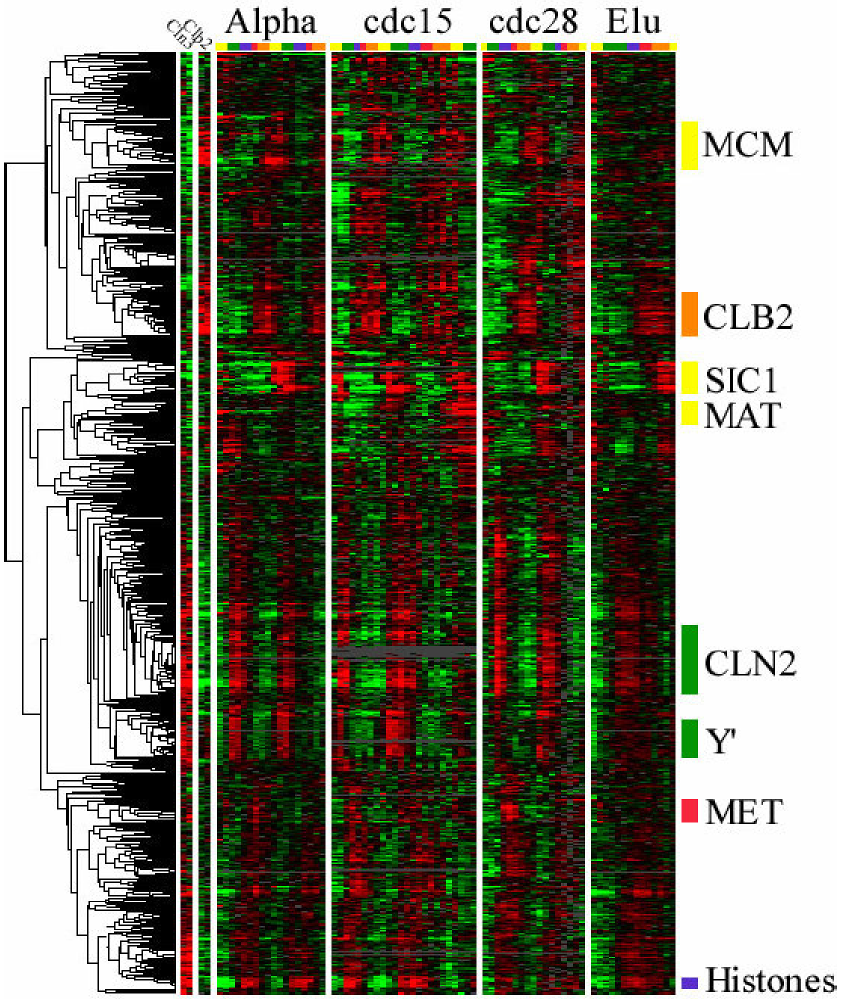


$$
\begin{array}{llll}
\hat{\theta}_2 & = 7/9\,, & \hat{\theta}_4 & = 3/9\,, \\
\hat{\theta}_{(3|X_2=0,X_4=0)} & = 1\,, & \hat{\theta}_{(3|X_2=0,X_4=1)} & = 1\,, \\
\hat{\theta}_{(3|X_2=1,X_4=0)} & = 1\,, & \hat{\theta}_{(3|X_2=1,X_4=1)} & = 0\,, \\
\hat{\theta}_{(1|X_3=0)} & = 1/3\,, & \hat{\theta}_{(1|X_3=1)} & = 1\,.
\end{array}
$$

Yeast Cell Cycle

Spellman et. al. (1998): Microarray data of yeast cell cycle

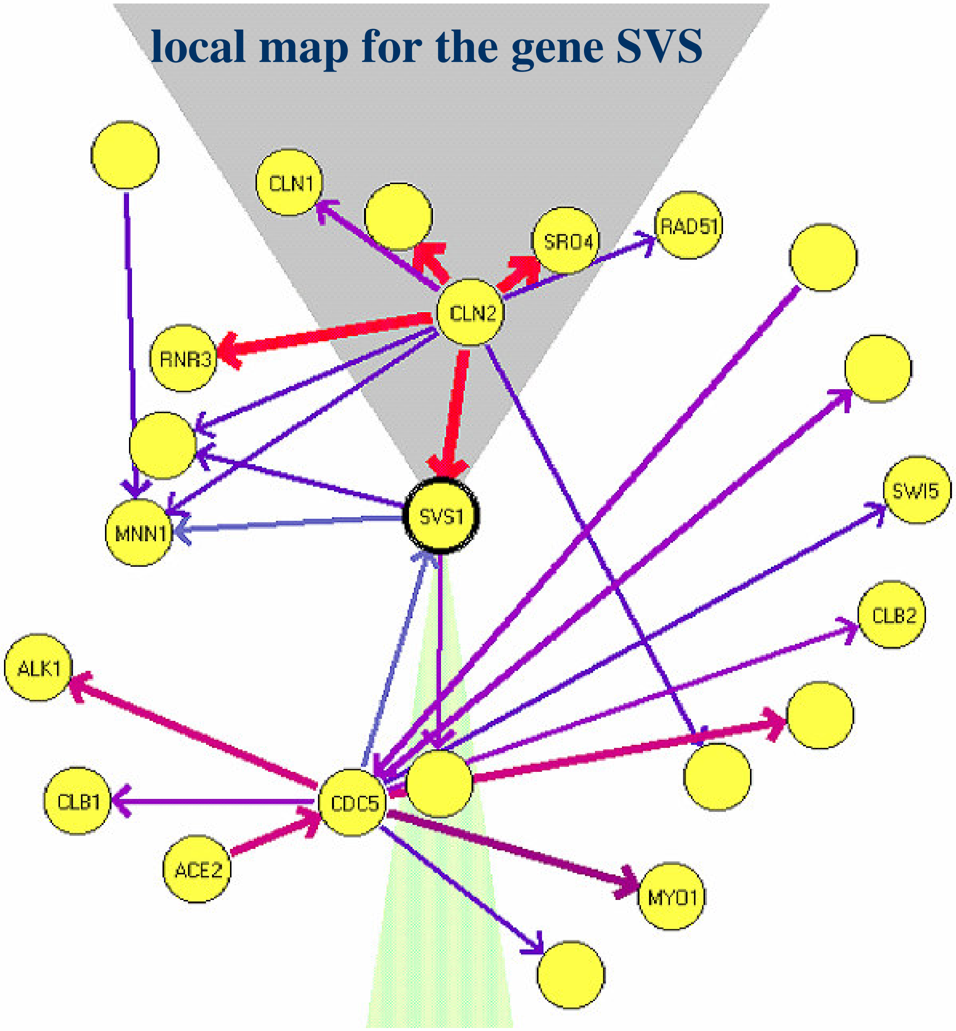
- 6177 genes, 76 samples of all the yeast genome, six time series.
- Identified 800 cell cycle regulated genes, and clustered them 250 genes in 8 clusters.
- Friedman et al analyzed these 250 genes by a Bayesian network.
- Multinomial model: treat each variable as discrete and learn a multinomial distribution that describes the probability of each possible state of the child variable given the state of its parents.
- Discretize the gene expression values:
  - under-expressed (-1),
  - normal (0), and
  - over-expressed (1), depending on whether the expression rate is significantly lower than, similar to, or greater than control.



Spellman et. al. (1998).
http://cellcycle-www.stanford.edu/

Analyzing Expression Data

- Width (color) of edges: the computed confidence level.
- CLN2 separates SVS1 from several other genes
- There is a strong connection between CLN2 to all these genes, there are no other edges connecting them
- These genes are conditionally independent given the expression level of CLN2.



Friedman N, Linial M, Nachman I and Pe'er D (2000), Using Bayesian networks to analyze expression data. Journal of Computational Biology, 7:601-620.

http://www.cs.huji.ac.il/labs/compbio/expression/

Small datasets with many variables:
many different networks are reasonable explanations of the data

⇨ Focus on features that are common to most of these networks.

Biological Analysis

Order relations (global property)

- Is gene X an ancestor of gene Y in all network of a given equivalence class?
- **Dominant Genes**: out of 800 genes, only a few seem to dominate the order appear before many genes.
- These gene are indicative of potential causal sources of the cell-cycle process directly involved in initiation of the cell cycle and its control:

CLN1, CLN2, CDC15 and RAD53 (functional relation has been established)

Markov relations (local property)

- Is gene X an direct relative of gene Y?
- Top scoring Karkov relations between genes were found to indicate a relation in biological function.

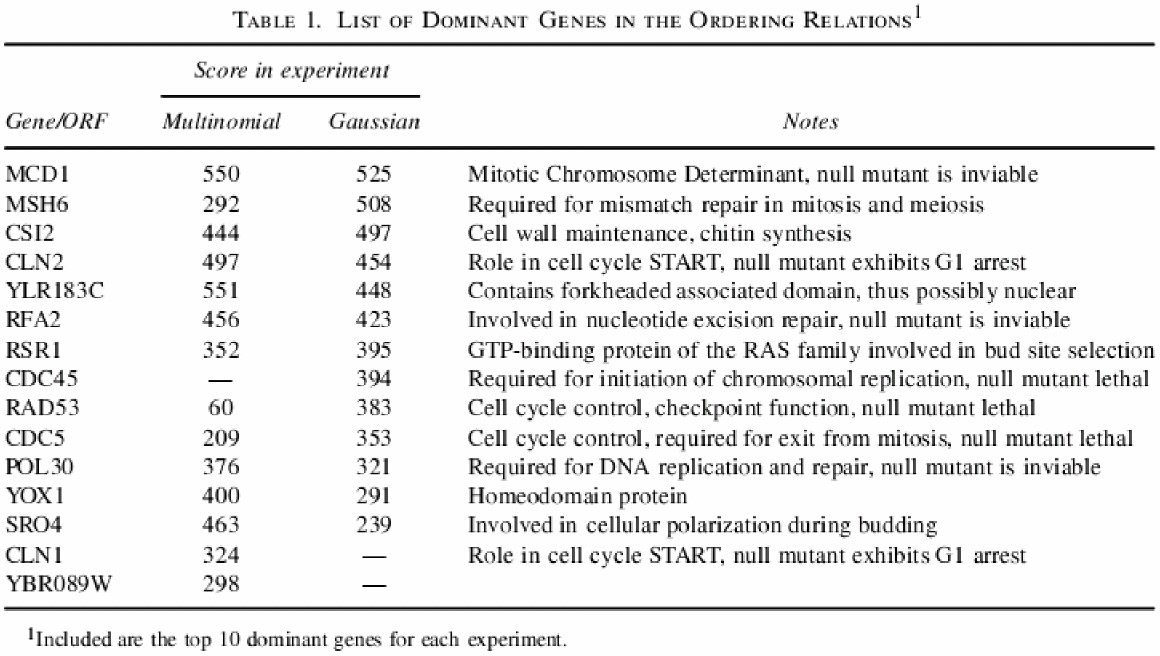


TABLE 1. LIST OF DOMINANT GENES IN THE ORDERING RELATIONS[1]

	Score in experiment		
Gene/ORF	*Multinomial*	*Gaussian*	*Notes*
MCD1	550	525	Mitotic Chromosome Determinant, null mutant is inviable
MSH6	292	508	Required for mismatch repair in mitosis and meiosis
CSI2	444	497	Cell wall maintenance, chitin synthesis
CLN2	497	454	Role in cell cycle START, null mutant exhibits G1 arrest
YLR183C	551	448	Contains forkheaded associated domain, thus possibly nuclear
RFA2	456	423	Involved in nucleotide excision repair, null mutant is inviable
RSR1	352	395	GTP-binding protein of the RAS family involved in bud site selection
CDC45	—	394	Required for initiation of chromosomal replication, null mutant lethal
RAD53	60	383	Cell cycle control, checkpoint function, null mutant lethal
CDC5	209	353	Cell cycle control, required for exit from mitosis, null mutant lethal
POL30	376	321	Required for DNA replication and repair, null mutant is inviable
YOX1	400	291	Homeodomain protein
SRO4	463	239	Involved in cellular polarization during budding
CLN1	324	—	Role in cell cycle START, null mutant exhibits G1 arrest
YBR089W	298	—	

[1] Included are the top 10 dominant genes for each experiment.

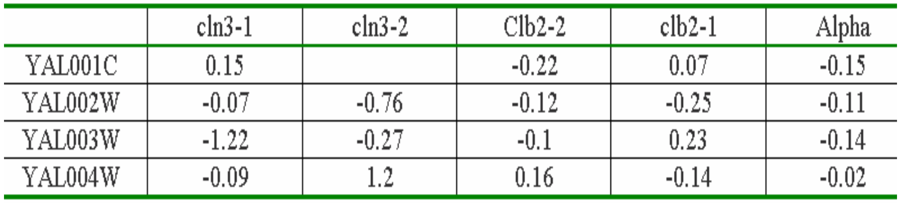
In general

BN provide us with a tool that allows biologically plausible conclusion from the data

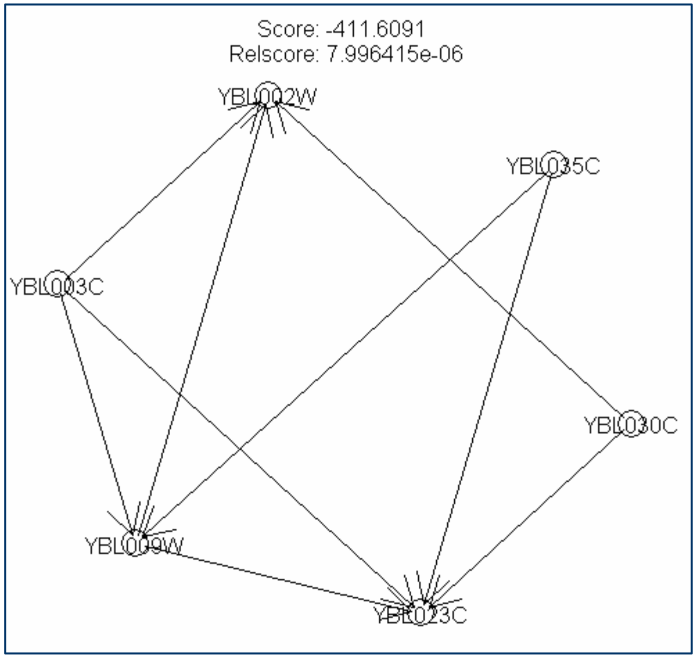
Software: BNArray

BNArray: http://www.cls.zju.edu.cn/binfo/BNArray/

- Impute missing values (LLSimpute)
- Construct Bayesian network and Bootstrap Bayesian networks
- Reconstruct significant coherent regulatory modules



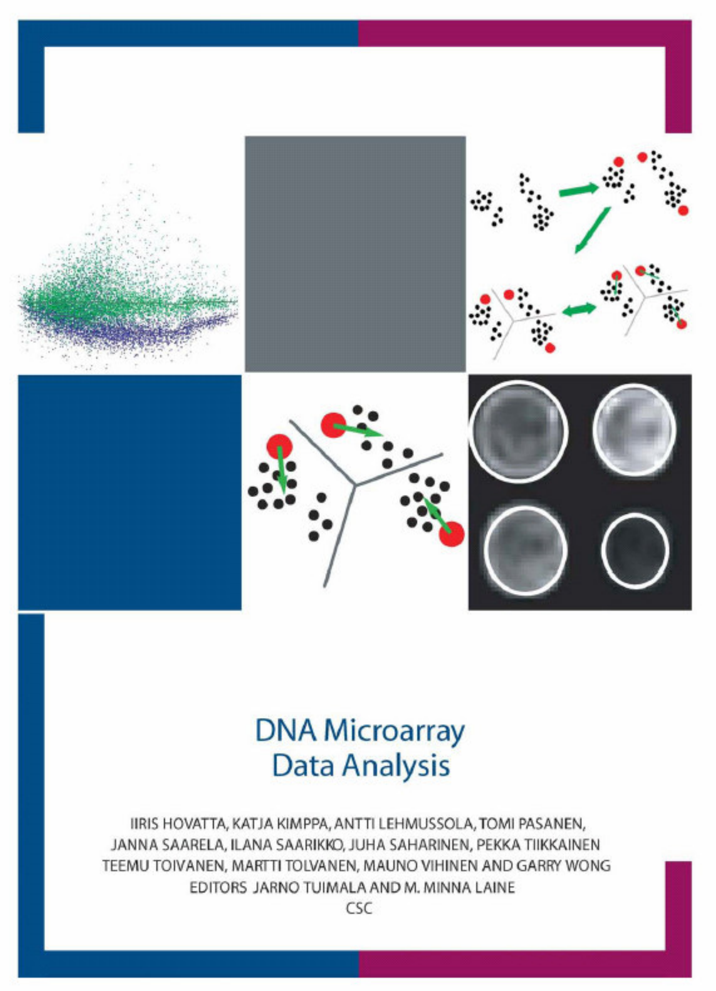
	cln3-1	cln3-2	Clb2-2	clb2-1	Alpha
YAL001C	0.15		-0.22	0.07	-0.15
YAL002W	-0.07	-0.76	-0.12	-0.25	-0.11
YAL003W	-1.22	-0.27	-0.1	0.23	-0.14
YAL004W	-0.09	1.2	0.16	-0.14	-0.02



- Gene YBL009W (unknown ORF) co-regulates H2A (YBL003C) and H2B (YBL002W).
- H2A and H2B form a compound during DNA replication process.
- YBL009W is a haspin which is involved in the meiosis process annotated in GO Biological Process database (check in SGD).

Chen X, Chen M, Ning K. BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. Bioinformatics. 2006 Sep 27

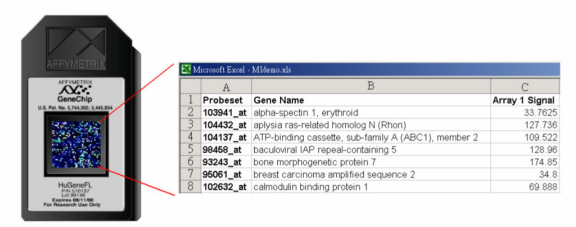
E-Book



DNA Microarray
Data Analysis

IIRIS HOVATTA, KATJA KIMPPA, ANTTI LEHMUSSOLA, TOMI PASANEN,
JANNA SAARELA, ILANA SAARIKKO, JUHA SAHARINEN, PEKKA TIIKKAINEN
TEEMU TOIVANEN, MARTTI TOLVANEN, MAUNO VIHINEN AND GARRY WONG
EDITORS JARNO TUIMALA AND M. MINNA LAINE
CSC

http://www.csc.fi/molbio/arraybook/



	A	B	C
1	**Probeset**	**Gene Name**	**Array 1 Signal**
2	**103941_at**	alpha-spectin 1, erythroid	33.7625
3	**104432_at**	aplysia ras-related homolog N (Rhon)	127.736
4	**104137_at**	ATP-binding cassette, sub-family A (ABC1), member 2	109.522
5	**98458_at**	baculoviral IAP repeal-containing 5	128.96
6	**93243_at**	bone morphogenetic protein 7	174.85
7	**95061_at**	breast carcinoma amplified sequence 2	34.8
8	**102632_at**	calmodulin binding protein 1	69.888

吳漢銘
E-mail: hmwu@stat.sinica.edu.tw
http://www.sinica.edu.tw/~hmwu

中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica

Thank You!