

## Gene Expression Data Analysis Using Bayesian Networks

Kyu-Baek Hwang Byoung-Tak Zhang Yung Taek Kim

School of Computer Science and Engineering, Seoul National University

DNA

### 1.

(gene) 가 (cell) (protein) (gene expression) 가 (microarray) [7], [3] (leukemia) [6] (Bayesian network) [5] (acute leukemia) AML(acute myeloid leukemia) ALL(acute lymphoblastic leukemia) AML ALL 가

### 2. (Bayesian networks)

( ) DAG(directed acyclic graph) 가 (joint probability distribution)  $\mathbf{X} = \{X_1, \dots, X_n\}$  2가 (conditional independence  $S$  (local probability distribution)  $P$   $S$  DAG  $S$   $\mathbf{X}$   $X_i$  가  $\mathbf{pa}_i$   $S$   $X_i$  ( ) 가  $\mathbf{X}$   $p(\mathbf{x}) = \prod_{i=1}^n p(x_i | \mathbf{pa}_i)$   $P$   $(S, P)$ 가  $\Pi$

### 3.

가 [5] 38 7129 (Bayesian network classifier)

leukemia class leukemia class leukemia class

AML, ALL class 가 leukemia 가

### 3.1 (Discretization of Gene Expression Levels)

Interval Binning [2] Equal Width Binning (discretization) 가 가 [2] Equal Width Binning (discretization) 0 1 Equal Width Binning 0 1 [5] 가 , 가 0 1

### 3.2 (Selection of Genes Related to the Classification of Leukemia)

[5] 7129 (training example) 38 가 [6] P-metric leukemia class (mutual information) AML ALL

3.3 leukemia class 가 가

가 BDe (Bayesian Dirichlet metric)

$$p(D, B) = p(B) \cdot \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

Dirichlet prior  $\alpha_{ij} \equiv \sum_{k=1}^{r_i} \alpha_{ijk}$   $N_{ij} \equiv \sum_{k=1}^{r_i} N_{ijk}$   $\alpha_{ijk}$  1.0 [4]

probability)  $p(B)$   $\Gamma(\cdot)$  log

NP-hard [1]. greedy search algorithm 5

### 4. (Selected Genes)

3.1 가 (discretization) 3.2 가 leukemia class (mutual information value)

1.

		Equal Width Binning	
		1	2
		FAH (0.737361) LTC4S (0.497781) Liver mRNA for IGIF (0.471081) HoxA9 (0.442018)	Zyxin (0.704252) ADM (0.591874) LTC4S (0.501848) NADPH (0.497781)
AML	ALL	3	4
		FAH (0.737361) LTC4S (0.497781) Platelet IB (0.311731) N-Methyl (0.265755)	Zyxin (0.704252) ADM (0.591874) C-myb (0.37707) MB-1 (0.316313)

4.2 가 1 4 가

가

5.

(mutual information)

34 leukemia class

0.5가

2

2.

AML, ALL  
(Bayesian network classifier)

3, 4

[6]

4

1

가

(gene

[3].

가

network)

2

1, 2

0

1, 2

3, 4

AML

ALL

[3].

bootstrap

BK21

4

가 가

0 1

[6]

6817

50

2/38,

5/34

1

가

4

[1] Chickering, D. M., Learning Bayesian networks is NP-complete, *Lecture Notes in Statistics*, 1995.

[2] Dougherty, J., Kohavi, R., and Sahami, M., Supervised and Unsupervised Discretization of Continuous Features, In *Proceedings of ICML '95*, pp. 194-202, 1995.

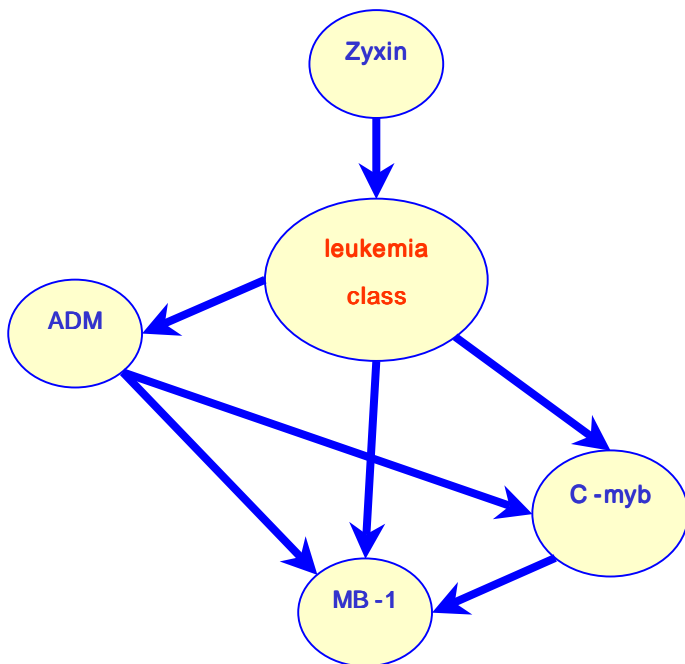
[3] Friedman, N., Linial, M., Nachman, I., and Pe'er, D., Using Bayesian Networks to Analyze Expression Data, In *Proceedings of RECOMB '00*, pp. 127-135, 2000.

[4] Heckerman, D., Geiger, D., and Chickering, D. M., Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning*, vol. 20, pp. 197-244, 1995.

[5] <http://waldo.wi.mit.edu/MPR>

[6] Slonim, D. K. et al., Class Prediction and Discovery Using Gene Expression Data, In *Proceedings of RECOMB '00*, pp. 263-272, 2000.

[7] Spellman, P. T. et al., Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization, *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297, 1998.



1.

4