## raphical Models for DNA Microarray Data Mining

양 진산 박사 (서울대 바이오정보기술연구센터 책임연구원)
(02-880-7302, jsyang@scai.snu.ac.kr)

**Abstract:**
현대적 실험방법 및 유전공학의 발전으로 최근 생물학적 자료는 비약적으로 늘어나고 있다. 이러한 자료의 기계학습 을 이용한 분석방법은 많은 비용과 시간을 요구하는 전통적인 생물학적 실험에 있어서 실험 시간을 단축시켜주고 실험비용을 줄여주게 된다.
본 논문에서는 특별히 micro array data 의 분석에 있어서 graphical model 에 기반한 기계학습 방법들을 소개한다. 이중 GTM 은 특히 시각화 효과가 뛰어난 방법으로 Graphical model 에 기반한 GTM 의 제반 특성을 소개하고 이를 yeast data 의 분석에 적용시킨 결과를 자세히 알아보고자 한다.(**Presentation file을 수신 보관 중)

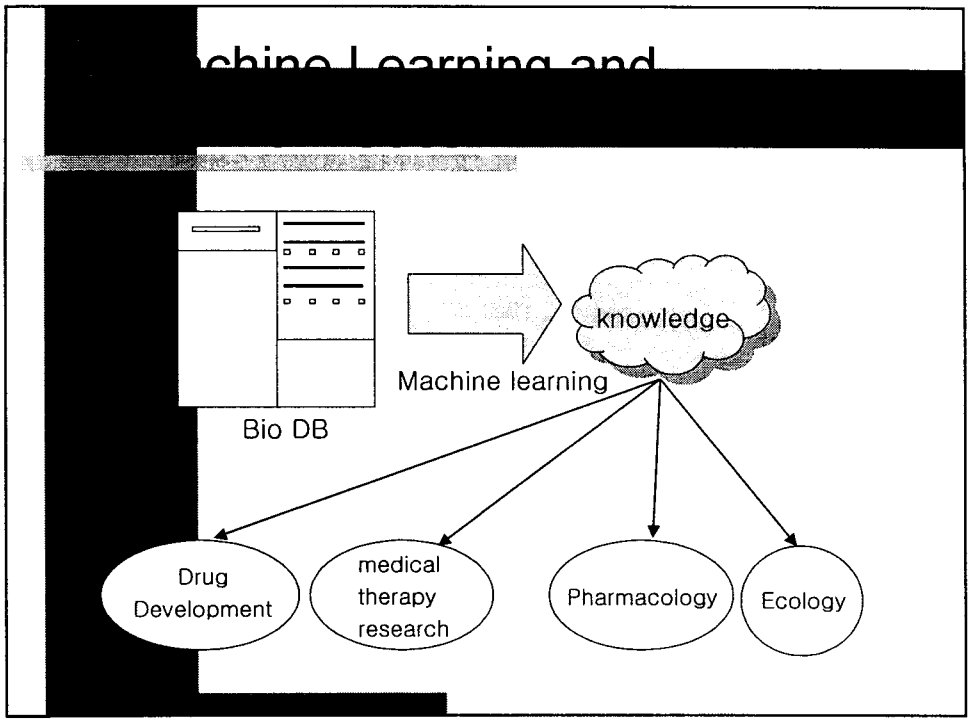
**약력**
1986 년 한양대학교 물리학사
1995 년 미시시피 주립대학교 통계학 박사
1997 년 ~ 1998 년 학술진흥원 후원 postdoc 연구원
1999 년 ~ 현재 서울대 바이오정보기술연구센터 연구원

관심분야:
바이오 데이터 마이닝, Latent Variable model, 확률 그래프 모형

# Graphical Models for DNA Microarray Data Mining

장병탁 양진산
Biointelligence Lab

Byoung-Tak Zhang and Jinsan Yang
Biointelligence Laboratory
Seoul National University
E-mail: {btzhang, jsyang}@bi.snu.ac.kr
Web: http://bi.snu.ac.kr/




vised Learning

imate an unknown mapping from known input- output pairs

arn $f_w$ from training set $D=\{(\mathbf{x},y)\}$ s.t. $f_w(\mathbf{x}) = y = f(\mathbf{x})$

ssification: $y$ is discrete, categorical

gression: $y$ is continuous

pervised Learning

ly input values are provided

arn $f_w$ from $D=\{(\mathbf{x})\}$ $f_w(\mathbf{x}) = y$

mpression

stering

## Problems

- Sequence analysis
  - quence alignment
  - ucture and function prediction
  - ne finding
- ure analysis
  - tein structure comparison
  - tein 3D structure prediction
  - A 3D structure modeling
- ork analysis
  - nome analysis
  - **Expression analysis**
  - **Gene clustering**
  - hway analysis

## Sequence Analysis

| Problems in Biological Science | | | Machine Learning Methods |
|---|---|---|---|
| Sequence alignment (homology search) | | Pairwise sequence alignment<br>Database search for similar sequences<br>Multiple sequence alignment<br>Phylogenetic tree reconstruction<br>Protein 3D structure alignment | Optimization algorithms<br>- Dynamic programming<br>- Simulated annealing<br>- Genetic algorithms<br>- Neural networks<br>- Hidden Markov models |
| Structure/function prediction | | RNA secondary structure prediction<br>RNA 3D structure prediction<br>Protein 3D structure prediction | |
| | | Motif extraction<br>Functional site prediction<br>Cellular localization prediction<br>Coding region prediction<br>Transmembrane segment prediction<br>Protein secondary structure prediction<br>Protein 3D structure prediction | Pattern recognition and learning algorithms<br>- Discriminant analysis<br>- Hierarchical neural networks<br>- Hidden Markov models<br>- Formal grammar |
| Molecular Clustering/Classification | | Superfamily classification<br>Ortholog/paralog grouping of genes<br>3D fold classification | Clustering algorithms<br>- Hierarchical cluster analysis<br>- Kohonen neural networks<br>Classification algorithms<br>- Bayesian Networks<br>- Neural Networks<br>- Support Vector Machines<br>- Decision Trees |

## Machine Learn

- Probabilistic Models
  - Hidden Markov Models
  - Bayesian Networks
  - Generative Topographic Mapping (GTM)
- Neural Networks
  - Multilayer Perceptrons (MLPs)
  - Self-Organizing Maps (SOM)
- Genetic Algorithms
- Other Machine Learning Algorithms
  - Support Vector Machines
  - Nearest Neighbor Algorithms
  - Decision Trees

## Applications of ML Methods for Bio Data Mining

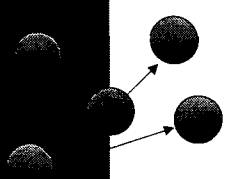
- Sequence Alignment
  - Simulated Annealing
  - Genetic Algorithms
- Structure and Function Prediction
  - Hidden Markov Models
  - Multilayer Perceptrons
  - Decision Trees
- Molecular Clustering and Classification
  - Support Vector Machines
  - Nearest Neighbor Algorithms
- Expression (DNA Chip Data) Analysis:
  - Self-Organizing Maps
  - Bayesian Networks
  - Generative Topographic Mapping

sian Networks

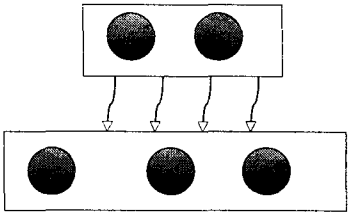
hical model for probabilistic relationships among a set of les

rative Topographic Mapping

hical model through a nonlinear relationship between the latent les and observed features.

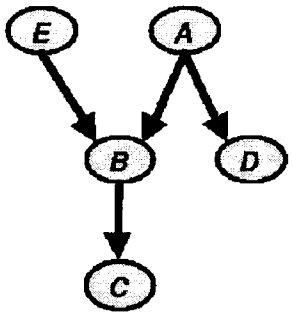


Network)



(GTM)

ian networks represent statistical relationships g random variables (e.g. genes).




- $B$ and $D$ are independent given $A$.
- $B$ asserts dependency between $A$ and $E$.
- $A$ and $C$ are independent given $B$.

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

## ...sian Networks (2)

(Directed Acyclic Graph)

...ian Network: Network Structure (S) + Local ...obability (P).

...ss dependence relations between variables

...se prior knowledge on the data (parameter)

...ichlet for multinomial data

...rmal-Wishart for normal data

...ods of searching:

Greedy, Reverse, Exhaustive

## ...

...missing values:

...ibbs sampling

...aussian Approximation

...M

...ound and Collaps etc.

...pretations:

...epends on the prior order of nodes or prior structure.

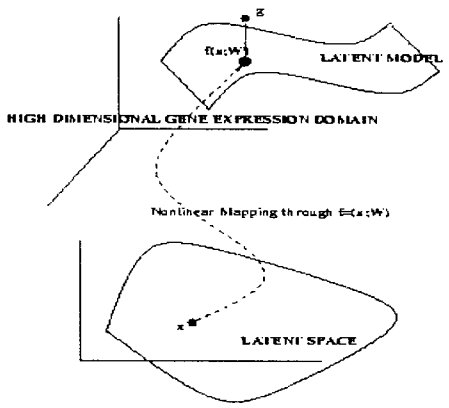
...ocal conditional probability

...hoice of nodes

...verall nature of data

## Gen

is a non-linear mapping model between latent space
ata space.




## Gene

mplex data structure is modeled from an intrinsic
space through a nonlinear mapping.

$$t = \Phi(x)W + E$$

: data point
: latent point
: matrix of basis functions
: constant matrix
: Gaussian noise

## Generative model

- A distribution of x induces a probability distribution in the data space for non-linear y(x,w).

$$p(\mathbf{t}|\mathbf{x},\mathbf{W},\beta) = \mathcal{N}(\mathbf{y}(\mathbf{x},\mathbf{W}),\beta) = \left(\frac{\beta}{2\pi}\right)^{-D/2} \exp\left\{-\frac{\beta}{2}\sum_{d}^{D}(t_d - y_d(\mathbf{x},\mathbf{W}))^2\right\}$$

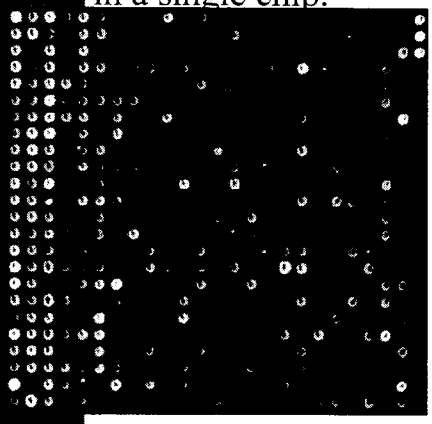
- Likelihood for the grid of K points

$$p(\mathbf{x}) = \frac{1}{K}\sum_{k}^{K}\delta(\mathbf{x}-\mathbf{x}_k), \quad p(\mathbf{t}|\mathbf{W},\beta) = \frac{1}{K}\sum_{k}^{K}p(\mathbf{t}|\mathbf{x}_k,\mathbf{W},\beta).$$

## Generative model

- lly the latent distribution is assumed to be uniform ).
- data point is assigned to a grid point probabilistically.
- can be visualized by projecting each data point onto the space to reveal interesting features
- lgorithm for training.
  - **itialize** parameter $W$ for a given grid and basis function set.
  - -Step) Assign each data point's probability of belonging to each id point.
  - l-Step) Estimate the parameter $W$ by maximizing the corresponding likelihood of data.
  - **ıtil** some convergence criterion is met.

microarray data provides the whole genomic
in a single chip.



ttp://www.gene-

- The intensity and color of each spot encode information on a specific gene from the tested sample.

- The microarray technology is having a significant impact on genomics study, especially on drug discovery and toxicological research.

lect cell cycle-regulated genes out of 6179 yeast
nes. (cell cycle-regulated : transcript levels vary
iodically within a cell cycle )

ere are 104 known cell cycle-regulated genes of 6
sters

S/G2 phase : 9 (train:5 / test:2)

S phase : 8 (Histones) (train:5 / test:3)

M/G1 boundary (SWI5 or ECB (MCM1) or STE12/MCM1 dependent) : 19 (train:13 / test:6)

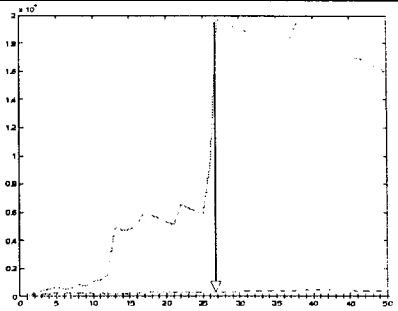
G2/M phase: 15 (train: 10 / test:5)

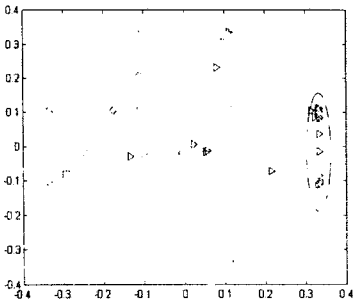
Late G1, SCB regulated : 14 (train: 9 / test:5)

Late G1, MCB regulated : 39 (train: 25 / test:12)

1-G1-S-G2-M)

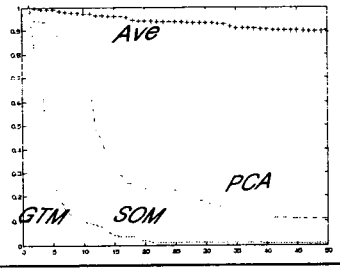
| cluster | size | mean response to Cln3p | mean response to Clb2p |
|---|---|---|---|
| 1 | 2758 | -0.125 | -0.130 |
| 2 | 28 | -0.095 | -0.094 |
| 3 | 27 | -0.085 | -0.070 |
| 4 | 37 | 1.142 | -0.510 |
| 5 | 42 | -0.406 | 1.316 |
| 6 | 110 | -0.180 | -0.152 |
| 7 | 132 | 0.795 | -0.552 |
| 8 | 26 | -0.629 | -0.180 |
| 9 | 25 | -0.447 | 0.066 |
| 10 | 47 | -0.194 | 0.102 |
| 11 | 32 | 0.025 | -0.104 |
| 12 | 225 | -0.086 | -0.120 |
| 13 | 48 | -0.284 | 0.050 |
| 14 | 25 | 0.079 | -0.010 |
| 15 | 53 | -0.437 | -0.138 |
| 16 | 53 | -0.058 | -0.088 |
| 17 | 23 | -0.178 | -0.050 |
| 18 | 45 | -0.122 | -0.204 |
| 19 | 86 | -0.230 | -0.138 |
| 20 | 76 | -0.002 | -0.084 |
| 21 | 28 | -0.140 | -0.104 |
| 22 | 35 | -0.046 | -0.058 |
| 23 | 34 | -0.167 | -0.168 |
| 24 | 117 | -0.213 | -0.120 |
| 25 | 52 | -0.214 | -0.198 |
| *unclassified* | 2014 | -0.189 | -0.088 |





| cluster | GTM1 | GTM2 | SOM1 | SOM2 | PCA | Aveage Linkage |
|---|---|---|---|---|---|---|
| size (S) | 71 | 37 | 21 | - | - | - |
| # known genes | 8 | 8 | | | | |
| Cln3p | 0.856 | 1.142 | 1.34 | | | |
| Clb2p | -0.415 | -0.510 | -0.554 | - | | |
| size (G2/M) | 123 | 42 | - | 63 | - | - |
| # known genes | 11 | 12 | | 12 | | |
| Cln3p | -0.446 | -0.406 | - | -0.452 | | |
| Clb2p | 0.563 | 1.316 | | 0.962 | | |
| size (G1) | 202 | 132 | 62 | 47 | - | 147 |
| # known genes | 29 | 25 | 21 | 19 | | 9 |
| Cln3p | 0.684 | 0.795 | -0.890 | 1.283 | | 0.439 |
| Clb2p | -0.481 | -0.552 | -0.103 | -0.993 | | -0.361 |

[Clusters identified by various methods]




→ [The comparison of entropies for each method]

## Summary

- Challenges of Artificial Intelligence and Machine Learning Applied to Biosciences
  - Large data size
  - Noise and data sparseness
  - Unlabeled and imbalanced data
  - Dynamic Nature of DNA Microarray Data
- Our study for DNA Microarray Data by GTM
  - Modeling of dynamic nature
  - Effective data selections
  - Proper measure of clustering ability

## References


- [Bishop C.M., Svensen M. and Wiliams C.K.I. (1988)]. GTM: The Generative Topographic Mapping, *Neural Computation*, 10(1).
- [Kohonen T. (1990)]. The Self-organizing Map. *Proceedings of the IEEE*. 78(9):
- [Eisen Michael, Gavin Sherlock, M.Q. Zhang, V.R. Iyer, Kirk Anders, M.B. Eisen, P.O. Brown, David Botstein, and Bruce Futcher. (1998)]. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae. *Molecular Biology of the Cell*, Vol. 9. 3273-3297.
- [Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric S. Lander, and Todd R. Golub (1999)] Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA Vol. 96, Issue 6, 2907-2912*
- [Cho *et al.* (1998)]. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell 2, 65-7[3.*
- [Pearl (1994)]. Operations for learning with graphical models. *Journal of Artificial Intelligence Research* .2. pp. 159-225.