Nonstandard Machine Learning Algorithms for Microarray Data Mining

Byoung-Tak Zhang

Center for Bioinformation Technology (CBIT) & Biointelligence Laboratory
School of Computer Science and Engineering
Seoul National University

http://cbit.snu.ac.kr/
http://bi.snu.ac.kr/

Outline

- Microarray Bioinformatics
- Probabilistic Graphical Models for Gene Expression Analysis
- Gene-Drug Dependency Analysis with Bayesian Networks
- Gene and Tissue Clustering with Latent Variable Models
- Summary and Future Work
Microarray Bioinformatics

Molecular Biology: Central Dogma
Reductionistic and Synthetic Approaches in Biology

Biological System

(Organism)

Reductionistic Approach (Experiments)

Synthetic Approach (Bioinformatics)

Building Blocks

(Genes/Molecules)

Topics in Bioinformatics

Sequence analysis
- Sequence alignment
- Structure and function prediction
- Gene finding

Structure analysis
- Protein structure comparison
- Protein structure prediction
- RNA structure modeling

Expression analysis
- Gene expression analysis
- Gene clustering

Pathway analysis
- Metabolic pathway
- Regulatory networks
Applications of DNA Microarrays

- Gene discovery
- Analysis of gene regulation
- Disease diagnosis
- Drug discovery: pharmacogenomics
- Toxicological research: toxicogenomics

A Comparative Hybridization Experiment

Image analysis
Image Analysis

- Scanned images → probe intensities → numerical values for higher-level analysis

Array target segmentation

Background intensity extraction

Target detection

Ratio analysis

Target intensity extraction

Data Preparation for Data Mining

Sample 1 Sample 2 Sample i

Sample k Sample n

<Microarray image samples> <Numerical data for data mining>
Gene Expression Data Mining

Data preprocessing:
- Normalization
- Discretization
- Gene selection

Mining
- Classification
- Clustering
- Regulation analysis

Learning:
- Greedy search
- EM algorithm

Why Data Mining?

- Traditional analysis
  ♦ One gene in one experiment
  ♦ Small data sets
  ♦ Simple analytical methods

- High-throughput genomic analysis
  ♦ Simultaneous measurements of thousands of gene expression levels → massive data sets
  ♦ Statistical methods
  ♦ Machine learning approach
Analysis of DNA Microarray Data

Previous Work

- Characteristics of data
  - Analysis of expression ratio based on each sample
  - Analysis of time-variant data
- Clustering
  - Self-organizing maps [Golub et al., 1999]
  - Singular value decomposition [Orly Alter et al., 2000]
- Classification
  - Support vector machines [Brown et al., 2000]
- Gene identification
  - Information theory [Stefanie et al., 2000]
- Gene modeling
  - Bayesian networks [Friedman et al., 2000]
An Introductory Example

- A Bayesian network classifier for acute leukemias [Hwang et al. 2001]

\[
\prod_{i} P(X_i | \text{Pa}_i)
\]

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zyxin</td>
<td>P(Leukemia = ALL)</td>
</tr>
<tr>
<td>LTC4S</td>
<td>Leukotriene C4 synthase (LTC4S) gene</td>
</tr>
<tr>
<td>C-myb</td>
<td>C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds</td>
</tr>
<tr>
<td>MB-1</td>
<td>MB-1 gene</td>
</tr>
</tbody>
</table>

\[
P(X) = \prod_{i=1}^{n} P(X_i | \text{Pa}_i)
\]
Probabilistic Graphical Models

- The joint probability distribution over \( X = \{X_1, X_2, \ldots, X_n\} \)
  - Chain rule
    \[
    P(X) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
    \]

- Conditional independence \( X \perp Y | Z \)
  \[
  P(X | Y, Z) = P(X | Z)
  \]

- Efficient representation of the joint probability distribution using conditional independencies encoded by the graph (network) structure
  \[
  P(X) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
  = \prod_{i=1}^{n} P(X_i | V(X_i))
  \]

- \( X_i \perp \{X_1, \ldots, X_{i-1}\} - V(X_i) | V(X_i) \)

Analysis Procedure

- Gene Expression Data
- Discretization and Selection
  - Gene A
  - Gene B
  - Gene C
  - Gene D
  - Target
  - Selected genes and the target variable

<Learned Graphical Models>
- Classification
- Dependency analysis
- Clustering
Why Probabilistic Graphical Models?

- The joint probability distribution → all the knowledge about the (biological) system.
- Generative → modeling data sources
- Efficient probabilistic inference → prediction
- The network structure → an insight into the intricate relationships between components (i.e., genes) of the biological system → dependency analysis
- Robust to the noise and error in gene expression data

Classes of Graphical Models

- Boltzmann Machines
- Markov Random Fields
- Bayesian Networks
- Latent Variable Models
- Hidden Markov Models
- Generative Topographic Mapping
- Non-negative Matrix Factorization
Gene-Drug Dependency Analysis with Bayesian Networks

NCI Drug Discovery Program
NCI 60 Cell Lines Data Set

- 60 human cancer cell lines
  - Colorectal, renal, ovarian, breast, prostate, lung, and central nervous system origin cancers, as well as leukemias and melanomas.

- Drug activity patterns (Database A)
  - Sulphorhodamine B assay → changes in total cellular protein after 48 hours of drug treatment.

- Individual targets (matrix T_i)
  - Analysis of molecular characteristics other than mRNA expressions.

- This study focuses on
  - Sensitivity to therapy ↔ not molecular consequences of therapy

Gene-Drug Dependency Analysis Using Bayesian Networks

- To discover
  - Gene-gene expression dependency
  - Gene expression-drug activity dependency
  - Drug-drug activity dependency

![Bayesian Network Diagram]

Local probability distribution

<table>
<thead>
<tr>
<th>Drug A</th>
<th>Drug B</th>
<th>P(Gene A = 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.003</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.007</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.533</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.633</td>
</tr>
</tbody>
</table>
Bayesian Networks: the Qualitative Part

- The directed-acyclic graph (DAG) structure
  - Conditional independencies between variables → dependency analysis

```
<Conditional independencies>
- Gene B and Gene D are independent given Gene A.
- Gene B asserts dependency between Gene A and Gene E.
- Gene A and Gene C are independent given Gene B.
```

Bayesian Networks: the Quantitative Part

- The joint probability distribution over all the variables in a Bayesian network.
  \[ P(X) = \prod_{i=1}^{n} (P(X_i | \text{Pa}_i)) \]

\[ \Theta_i = (\theta_{i1}, ..., \theta_{iq_i}) \sim \text{parameter for } P(X_i | \text{Pa}_i) \]
\[ P(\theta_{ij}) = \text{Dir}(\theta_{ij} | \alpha_{j1}, ..., \alpha_{jq_j}) \]
\[ q_i : \# \text{ of configurations for } \text{Pa}_i \]
\[ r_i : \# \text{ of states for } X_i \]

Applications of Bayesian Networks

- Classification
  - Bayesian network classifiers
  - Probabilistic inference $\rightarrow$ prediction $\rightarrow$ cancer classification based on gene expression

- Dependency analysis
  - Bayesian network structure $\rightarrow$ dependency between components $\rightarrow$ putative causal relationships

Discretization of Gene Expression Levels

- The local probability distribution for each variable
  - Multinomial distribution with Dirichlet prior.
    \[
    p(\theta_{ij} | S^h) = \text{Dir}(\theta_{ij} | \alpha_{ij1}, \alpha_{ij2}, \ldots, \alpha_{ijr})
    \]
    \[
    = \frac{\Gamma(\alpha_{ij1} + \ldots + \alpha_{ijr})}{\prod_{k=1}^r \Gamma(\alpha_{ijk})} \cdot \prod_{k=1}^r \theta_{ijk}^{\alpha_{ijk} - 1}
    \]
    \[
    \Gamma(1) = 1, \quad \Gamma(x + 1) = x\Gamma(x)
    \]
    \[
    S^h: \text{ the network structure}
    \]

- Discretization of gene expression levels and drug activities
  - All values $\rightarrow$ 0 and 1 according to the mean values across the training samples.
Selection of Genes and Drugs for Analysis

Known drugs
- 566 human genes and ESTs
- 82 drugs

Missing values elimination

Learning Bayesian networks with 648 nodes

Learn the Bayesian Network from Data

- Two approaches to Bayesian network learning
  - Dependency analysis-based approach
  - Optimization-based approach
    - Network score: fitness of the network to training data $D$. 
    - Search the best-scoring network structure

- Scoring metric for the network
  - MDL (minimum description length) score
  - BD (Bayesian Dirichlet) score
**BD (Bayesian Dirichlet) Score**

- **Assumptions**
  - Multinomial sample
  - Parameter independence and modularity
  - Dirichlet prior
  
  $p(\theta_{ij} | S^h) = c \cdot \prod_{k=1}^{r_{ij}} \theta_{ijk}^{\alpha_{ijk} - 1}$

  - Complete data

- **BD score is calculated as follows:**
  $p(D, S) = p(S) \cdot p(D | S)$
  
  $= p(S) \cdot \prod_{i=1}^{n} \prod_{j=1}^{d_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_{ij}} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$

  - $S$: the network structure, $D$: the training data

---

**Search Strategy**

- **Finding the best network structure** $\rightarrow$ NP-hard problem.
  - Exponential time complexity.
  - In general, greedy search or simulated annealing is used.

- **General greedy search algorithm** is not applicable to the construction of large-scale Bayesian networks with hundreds of nodes.
  - Time and space complexity

- **Reduce the search space**
  - “Sparse candidate algorithm” by [Friedman et al. 1999]
  - “Local to global search algorithm”
    - Exploit the local search for the Markov blanket of each node to reduce the global search space.
Local to Global Search Algorithm

Input:
- A data set $D$.
- An initial Bayesian network structure $B_0$.
- A decomposable scoring metric,
  \[ Score(B, D) = \sum_i Score(X_i | Pa^B(X_i), D). \]

Output: A Bayesian network structure $B$.
Loop for $n = 1, 2, \ldots$, until convergence.
- Local Search Step:
  * Based on $D$ and $B_{n-1}$, select for each variable $X_i$, a set $CB_i^p| |CB_i^p| \leq k$ of candidate Markov blanket of $X_i$.
  * For each set $(X_i, CB_i^p)$, learn its local structure and determine the Markov blanket of $X_i$, $BL_i^p(X_i)$, from this local structure.
  * Merge all the local network structures $G(X_i, BL_i^p(X_i), E_i)$ into a global network structure $H_n$ (usually cyclic).
- Global Search Step:
  * Find the Bayesian network structure $B_n \subset H_n$ which maximizes $Score(B_n, D)$ and retains all non-cyclic edges in $H_n$.

Learning Curve

Learning takes about 8 minutes.
Drug-Drug Dependency

- The pyrimidine analogues
  - Aphidicolin-glycinate
  - Floxuridine
  - Cyclocytidine
  - Cytarabine
  - DNA synthesis inhibitor

- Shows similar activity pattern across 60 cancer cell lines
- Clustered together in [Scherf et al. 2000].

Experimental Results: Drug-Drug Dependency

- Drug-drug activity correlations
  - Three drugs “Aphidicolin-glycinate”, “Floxuridine”, and “Cytarabine” directly depend on each other.
  - “Cyclocytidine” directly depends on “Cytarabine” and vice versa.

Confirmation

<Part of the learned Bayesian network structure>
Gene-Drug Dependency

- Gene “ASNS”
  - Certain malignant cells, including many acute lymphoblastic leukemias (ALL) lack asparagine synthetase (ASNS) \( \rightarrow \) exogenous “L-asparagine” [Scherf et al. 2000].

Sensitivity of the drug “L-asparagine” \( \leftarrow \rightarrow \) Expression levels of the gene “ASNS”

High negative correlation

Experimental Results: Gene-Drug Dependency

- Gene expression-drug activity correlations

- The negative correlation between “ASNS” and “L-asparagine” is mediated by two other genes.
- The relationships revealed by the Bayesian network is putative and should be verified by biological experiments. \( \rightarrow \) exploratory analysis
CAMDA-2000 Data Set

- Gene expression data for cancer prediction
  - Training data: 38 leukemia samples (27 ALL, 11 AML)
  - Test data: 34 leukemia samples (20 ALL, 14 AML)
- Datasets contain measurements corresponding to ALL and AML samples from bone marrow and peripheral blood.
Cluster Analysis Using Non-negative Matrix Factorization

- Method
  - Using NMF for class clustering and prediction of gene expression data from acute leukemia patients
  - NMF (non-negative matrix factorization)

\[ G \approx WH \]

\[ (G)_{ij} = (WH)_{ij} = \sum_{a=1}^{r} W_{ia} H_{a\mu} \]

- G : gene expression data matrix
- W : basis matrix (prototypes)
- H : encoding matrix (in low dimension)
- \( G_{ij}, W_{ia}, H_{a\mu} \geq 0 \)

- NMF as a latent variable model

![Diagram](image)

Clustering Gene Expression Data

- Factors can capture the correlations between the genes using the values of expression level.
- Cluster training samples into 2 groups by NMF
  - Assign each sample to the factor (class) which has higher encoding value.
Learning Procedure

**Input**: Gene expression data matrix, $G \in \mathbb{R}^{n \times m}$

**Output**: base matrix $W \in \mathbb{R}^{n \times k}$, encoding matrix $H \in \mathbb{R}^{k \times m}$

$n$: data size, $m$: number of genes, $k$: number of latent variables

**Objective function**:

$$F = \sum_{j=1}^{n} \sum_{\mu=1}^{m} \left[ G_{j\mu} \log(H_{j\mu}) - (WH)_{j\mu} \right]$$

**Procedure**

1. Initialize $W$, $H$ with random numbers.

   $$W_{ij} \geq 0, \quad \sum_j W_{ij} = 1$$

   $$H_{ij} \geq 0$$

2. Update $W$, $H$ iteratively until max iteration or some criterion is met.

   $$W_{ia} \leftarrow W_{ia} \frac{G_{j\mu}}{(WH)_{j\mu}} H_{j\mu}$$

   $$H_{i\mu} = H_{i\mu} \frac{W_{ia} G_{j\mu}}{(WH)_{j\mu}}$$

   $$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$$

---

Learning Curve

![Learning Curve Diagram](https://via.placeholder.com/150)
Experimental Results: Cancer Clustering

Accuracy: 0 ~ 1 error for the training data set

![Graph showing cancer clustering results]

Diagnosis Using NMF

- For each test sample $g$, estimate the encoding vector $h$ that best approximates the sample.
  - $W$ is the basis matrix computed during training (fixed).
  - As in training, assign each sample to the factor (class) which has the highest encoding value.

Accuracy: 1 ~ 2 error(s) for the test data set
Clustering Using Generative Topographic Mapping

- GTM: a nonlinear, parametric mapping $y(x; W)$ from a latent space to a data space.

\begin{align*}
\text{Grid} & \quad \text{y}(x; W) \\
\begin{array}{c}
\begin{array}{c}
 x_2 \\
 x_1 \\
 \end{array}
\end{array} & \quad \begin{array}{c}
\begin{array}{c}
 t_3 \\
 t_2 \\
 t_1 \\
 \end{array}
\end{array}
\end{align*}

Learning Generative Topographic Mapping

- Learning algorithm
  - Generate the grid of latent points.
  - Generate the grid of latent function centers.
  - Compute the matrix of basis function activations $\Phi$.
  - Initialize weights $W$ in $Y = \Phi W$ and the noise variance $\beta$.
  - Compute $\Delta_{n,k} = \|t_n - \Phi_k W\|^2 = \|t_n - y_k(x, W)\|^2$ for each $n, k$.
  - Repeat
    - Compute the responsibility matrix $R$ using $\Delta$ and $\beta$. [E-Step]
    - Compute $G = R^T R$
    - Update $W$ by $\Phi^T G \Phi W = \Phi^T R T$ [M-step]
    - Compute $\Delta = \|t - \Phi W\|^2$
    - Update $\beta$
  - Until convergence
GTM: Visualization

- Posterior distribution in latent space given a data point $t$:
  \[ p(x_k | t) = \frac{p(t | x_k, W^*, \beta^*) p(x_k)}{\sum_{k'} p(t_n | x_{k'}, W^*, \beta^*) p(x_{k'})}. \]

- For a whole set of data: for each $t$, plot in the latent space.
  - Posterior mode:
    \[ x_{n, \text{mode}} = \arg\max_{x_k} p(x_k | t_n), \]
  - Posterior mean:
    \[ x_{n, \text{mean}} = \sum_{k} x_k p(x_k | t_n). \]

GTM: Clustering Experiment

- Gene Selection
  - Select about 50 genes out of 7,129 based on the three test scores of cancer diagnosis.
    - Correlation metric (similar as $t$-test)
    - Wilcoxon test scores (a nonparametric $t$-test)
    - Median test scores (a nonparametric $t$-test)

- Clustering & Visualization
  - After learning a model, genes are plotted in the latent space.
  - With the mapping in the latent space, clusters can be identified.
### List of Genes Selected

<table>
<thead>
<tr>
<th>Gene Description</th>
<th>Gene Accession Number</th>
<th>x(Normal)</th>
<th>x3(Cancer)</th>
<th>P2_WIL</th>
<th>P2_MED</th>
<th>P*Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHF Fynvase pseudogene</td>
<td>M50153_at</td>
<td>0.054</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>Leukotriene C4 synthase</td>
<td>U25956_at</td>
<td>0.056</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>Zynin</td>
<td>X80536_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>LYN Vars-K1 Yamauchi sarcoma viral related oncogene homolog</td>
<td>M16030_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>COX3 COX3 antigen (differentiation antigen)</td>
<td>M21917_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>DF 3 component of complement (adipon)</td>
<td>M50423_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>UEPR Leitin receptor</td>
<td>Y26070_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>GB DEP homolog domain protein</td>
<td>X70785_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>LHR mRNA for interferon-alpha inducible factor (KIF)</td>
<td>D49350_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>GSTC Cystatin C (Ceruloplasmin and central hexamethion)</td>
<td>M24091_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>PEG1 Proteasome 1 (secretory subunit)</td>
<td>X70785_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>Thrombospodin-1 gene extracted from Human thrombospodin-1</td>
<td>H12471_c01_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>Phosphatase dependent phosphatase for the L67-2-4 domain</td>
<td>U46878_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>INTERLEUKIN-6 PRECURSOR</td>
<td>Y09818_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>INDIUCED MYC/MYCID LEUKMA CELL DIFFERENTIATION PROTEIN</td>
<td>L08246_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>PEPTIDYL-PRENYL CO-TRANSPORTER, MITOCHONDRIAL PRECURSOR</td>
<td>M90254_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>ATP6C1Vacuum-like protein channel subunit</td>
<td>M28787_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>CDC25A Cell division cycle 25A</td>
<td>M19313_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>Acetylcholine receptor</td>
<td>M30826_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>Transcriptional activator HSP70</td>
<td>U2891_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>ACAD11 Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain</td>
<td>M29413_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>HEX1</td>
<td>S02233_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>CI12 mRNA, alternative splice variant alpha-1</td>
<td>AF003492_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>IRF2 interferon regulatory factor 2</td>
<td>X70785_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>CTPS CTP synthetase</td>
<td>X51225_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>TOP2B Topoisomerase (DNA) beta (18KD)</td>
<td>Z15151_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>COX19 COX19 antigen</td>
<td>M2913_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>Inducible protein M3 312 gene</td>
<td>L4773_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>Cytoplasmic dynactin-like chain 1 (hdc1) mRNA</td>
<td>U50344_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>TFE3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12E470)</td>
<td>M1052_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>Transcriptional activator HSP70</td>
<td>D26251_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>Thyroboxin beta mRNA</td>
<td>U60897_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>MYL1 Myosin light chain (myosin)</td>
<td>M11712_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>CBF1 Crystallo s (axone reductase)</td>
<td>L10278_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>RETINOBLASTOMA BINDING PROTEIN 48</td>
<td>X74812_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>CND3 Cystin D3</td>
<td>M52097_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>MB-1 gene</td>
<td>U50259_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>PROTEASOME CHAIN</td>
<td>M58411_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
<tr>
<td>C-kid gene extracted from Human (c-kid) gene, complete primary cDNA, and the cDNA:</td>
<td>X30237 cds2 s_at</td>
<td>0.052</td>
<td>0.0301</td>
<td>0.0001</td>
<td>1.4676</td>
<td></td>
</tr>
</tbody>
</table>

---

**GTM: Learning Curve**

![Learning Curve Diagram](image)
**GTM: Clustering Result**

Genes with high expression levels in case of ALL
(large P-metric value)

Genes with high expression levels in case of AML
(negative large P-metric value)

---

**Experimental Results: Clusters Found by GTM**

- Three cell cycle-regulated clusters found by GTM [Shin et al. 2000]

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cluster center</th>
<th>No. of train Data/ no. in cluster</th>
<th>Correct no. / test data</th>
<th>Overall mean expression levels (Cln/b) of known genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>S/G2</td>
<td></td>
<td>5 / 5</td>
<td>1 / 2</td>
<td>(.148 .184 -.367 -.044)</td>
</tr>
<tr>
<td>S</td>
<td>(0.111 -0.333)</td>
<td>5 / 5</td>
<td>5 / 5 (100%)</td>
<td>(1.075 1.482 -.233 -.375)</td>
</tr>
<tr>
<td>M/G1</td>
<td>c1 (0.111 0.333)</td>
<td>13 / 7</td>
<td>1 / 6</td>
<td>(-.171 -.573 .091 .311)</td>
</tr>
<tr>
<td></td>
<td>c2 (-0.111 -0.111)</td>
<td>/ 2</td>
<td>0 / 6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>c3 (0.323 0.1)</td>
<td>/ 2</td>
<td>0 / 6</td>
<td></td>
</tr>
<tr>
<td>G2/M</td>
<td>c1 (0.111 0.333)</td>
<td>10 / 5</td>
<td>0 / 5</td>
<td>(-.616 -1.01 1.832 1.596)</td>
</tr>
<tr>
<td></td>
<td>c2 (0.111 0.111)</td>
<td>/ 3</td>
<td>3 / 5 (80%)</td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>c1 (-0.111 0.333)</td>
<td>35 / 18</td>
<td>10 / 16 (62%)</td>
<td>(.894 .907 -.766 -.479)</td>
</tr>
<tr>
<td></td>
<td>c2 (-0.111 0.111)</td>
<td>/ 7</td>
<td>0 / 16</td>
<td></td>
</tr>
</tbody>
</table>
## Experimental Results:
### Comparison with Other Methods

#### Comparison of prototype expression levels

<table>
<thead>
<tr>
<th>Condition</th>
<th>No. of selected genes</th>
<th>Mean expression levels by GTM</th>
<th>No. of selected genes by Spellman</th>
<th>Mean expression levels by Spellman</th>
</tr>
</thead>
<tbody>
<tr>
<td>S/G2</td>
<td>92</td>
<td>(.13 -.06 -.1 .01)</td>
<td>121</td>
<td>(.13 .05 -.16 .03)</td>
</tr>
<tr>
<td>S</td>
<td>25</td>
<td>(.84 .81 -.42 -.33)</td>
<td>71</td>
<td>(.46 .47 -.43 -.18)</td>
</tr>
<tr>
<td>M/G1</td>
<td>120</td>
<td>(.82 .65 -.65 -.38)</td>
<td>113</td>
<td>(-.21 -.61 -.04 .07)</td>
</tr>
<tr>
<td>M/G1 c1</td>
<td>34</td>
<td>(-.04 -.37 -.01 -.11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M/G1 c2</td>
<td>10</td>
<td>(.32 .29 -.3 .05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M/G1 c3</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2/M</td>
<td>33</td>
<td>(.59 -.96 1.34 1.29)</td>
<td>195</td>
<td>(-.32 -.62 .49 .54)</td>
</tr>
<tr>
<td>G2/M c1</td>
<td>34</td>
<td>(.08 -.30 .51 .57)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2/M c2</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>122</td>
<td>(.92 .74 -.62 -.33)</td>
<td>300</td>
<td>(.66 .49 -.55 -.33)</td>
</tr>
<tr>
<td>G1 c1</td>
<td>74</td>
<td>(.79 .82 -.48 -.34)</td>
<td>(total = 800)</td>
<td></td>
</tr>
<tr>
<td>G1 c2</td>
<td>(total = 570)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary and Future Work

- A class of nonstandard learning algorithms, i.e. probabilistic graphical models, is presented for microarray data analysis:
  - Bayesian networks
  - Non-negative matrix factorization
  - Generative topographic mapping
- Probabilistic graphical models are useful for biological data mining:
  - Comprehensibility
  - Generative
  - Soft inference (clustering)
  - Dependency analysis
- Future work includes:
  - Construction of larger-scale networks
  - Acceleration of learning processes
  - Handling data sparseness and missing data problems

References

Acknowledgements

CAMDA-2000 & CAMDA-2001

- Sirk-June Augh (Molecular Biology)
- **Jeong-Ho Chang** (Helmholtz Machines, NMF, PLSA)
- Dong-Yeon Cho (Bayesian Evolution, Neural Trees)
- **Kyu-Baek Hwang** (Bayesian Networks)
- Sung-Dong Kim (Hierarchical Clustering)
- Sang-Wook Park (RBF Networks)
- Hyung-Joo Shin (Probabilistic LSA)
- **Jin-San Yang** (Statistics, GTM, SOM)
- Ho-Jean Chung (Statistical Learning)
- Seung-Woo Chung (ICA, PCA)
- Jae-Hong Eom (Hidden Markov Models)
- Tae-Jin Jeong (Reinforcement Learning)
- Je-Gun Joung (Evolution, Neural Trees)
- Chul-Joo Kang (Molecular Biology)
- Jun-Shik Kim (Statistical Physics)
- Sun Kim (Evolutionary Algorithms)
- Yoo-Hwan Kim (ICA, Naïve Bayes, Boosting)
- In-Hee Lee (Evolutionary Computation)
- Jae-Won Lee (Reinforcement Learning)
- Jong-Woo Lee (Latent Variable Models)
- Si-Eun Lee (Bayesian Evolution)
- Seung-Joon Lee (Reinforcement Learning)
- Jong-Yoon Lim (SOM, Graphical Models)
- Hyun-Koo Moon (HMM, Naïve Bayes)
- Jang-Min O (Support Vector Machines, Boosting)
- Seong-Bae Park (Decision Trees)
- Hyung-Joo Shin (Probabilistic LSA)
- Soo-Yong Shin (Evolution, Helmholtz Machines)

More information at

http://cbit.snu.ac.kr/
http://bi.snu.ac.kr/