Our view on cDNA chip analysis from engineering informatics standpoint

Chonghun Han, Sungwoo Kwon
Intelligent Process System Lab
Department of Chemical Engineering
Pohang University of Science and Technology
Outline

- Introduction to Bioinformatics
- Introduction to cDNA chip
- Classification of Tumor Classes
- Identification of Marker Genes
- Conclusions and Future Works
The Information Revolution

I need tools to extract important information from mountains of data

POSTECH IPSE Lab.
Data Mining

Data mining = ‘exploration and analysis by automatic and semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules’

Applications of data mining
- Search database
- Structural pattern recognition
- Medline abstract analysis
- DNA chip data analysis

Who has information and uses it wins
(Watterson, K)
Data Mining + Domain Knowledge = Engineering Informatics

**Informatics (정보 과학):** *the study of the structure, behavior, and interactions of natural and artificial computational systems*

**Informatics = Information + Mathematics**

**Application Areas**
- Bioinformatics (or Biomolecular Informatics)
- Cheminformatics
- Environmental Informatics
- Medical Informatics
- Neuro Informatics
- Process Informatics
- Many More …
Bioinformatics = Biology + Informatics

- Artificial Intelligence
- Combinatorial Optimization
- Data Mining
- Digital Signal Processing
- Machine Learning
- Mathematical Modeling
- Multivariate Statistics
- Pattern Recognition
- System identification
- …
DNA chip
The biological meaning of DNA chip

- Genome map is completed
  - Need to study functional genomic

- Know who, when, where, why, how much gene expressed
  - To classify different types of diseases (ex. Cancer types)
  - To understand the behavior of a biological system
  - To understand cell dynamics

- Can systematically disturb cell

- DNA chip experiment and data analysis are different matter
  - Methods of data analysis variant result of DNA chip experiment
  - Require suitable method of data analysis for DNA chip experiment objective
Application of DNA chip

- Analysis of gene expression and regulation
  - Genetic network, pathway analysis, metabolic engineering

- Disease diagnosis
  - Molecular cancer classification, the discovery of disease subtype,
    The marker gene discovery

- And many more…

- Cancer diagnosis

- Because significantly different groups of genes are expressed by many type of cell, we can fingerprint characteristic cell
cDNA chip: Lab Experiment
cDNA chip: Data Analysis

**Statistical analysis (data mining)**
- Hierarchical clustering
- K-means clustering
- Self-organizing map
- Neural network
- Bayesian decision theory
- Principal component analysis
- And many more…

**Molecular cancer classification**

**Identification of the potential marker gene**
cDNA chip: Procedure

DNA chip experiment

For Functional study using data mining techniques

Informatics validation

Biological validation

1. Expression profile data warehousing
2. Other database integration

Data mining techniques

Inference new hypothesis for biological experiment
Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

Moving from morphological to molecular classification

Acute lymphoblastic leukemia (ALL)

Acute myelogenous leukemia (AML)

Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

There are new cancer class discovery, two molecularly distinct forms of B-cell lymphoma (DLBCL) that are composed of GC B-like and Activated B-like DLBCL.

Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profile

- **Limitation,**
  - Because weighting method based on univariative or bivariative statistical analysis, we cannot capture correlated structure in the data.
  - When multi-class cancer classify, it is hard to know whether highly express or not.

Classification and Diagnostic Prediction of Cancers using Gene Expression Profiling and Artificial Neural Network

**Limitation,**

- Because relevant gene extraction method based on univariate statistical analysis, we can not capture correlated structure in the data.

Gene-expression profiles in hereditary breast cancer

- **Limitation**, 
  - Because relevant gene extraction method based on univariate statistical analysis, we can not capture correlated structure in the data.
  - When multi-class cancer classify, it is hard to know whether highly express or not

Gene Selection for Sample Classification based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method

Limitation,

- It is difficult to determine parameter value
- When multi-class cancer classify, it is hard to know whether highly express or not
- Computing time take a long time

Limitations and Improvements

Limitations of Previous Approaches
- Small number of samples vs many variables
- Strong variable interaction
- Lack of interpretation based on biological meanings
- Limitation in the identification of marker genes due to the black box model
- Limitations due to univariate approaches
- Procedure of analysis are very complex and take a long time

Improvements
- Overcome interaction of many variables
- Develop to a method to select potential marker genes
- Develop multivariate approach
- Develop simple and ease procedure of data analysis
Proposed Procedure

1. **Data preprocessing**
   - 1. Dimension reduction
   - 2. Modeling of correlation structure

2. **Feature selection**
   - Select the potential marker genes

3. **Classification of tumor classes**

**DNA Chip Data**
- High dimensional data
- Highly correlated variables

**PCA Analysis**
- Data preprocessing

**Stepwise Discriminant Analysis**
- Feature selection
  - (Select highly discriminant PC)

**Bayesian Decision Theory (Classifier)**

**Contribution Analysis**
Major Steps

**Stepwise Discriminant Analysis**

- Select subset where Wilks’ lambda value is minimum
- Maximize the discriminant power

$$\Lambda = \frac{SS_w}{SS_t}$$

- $SS_t$: Class heterogeneity
- $SS_w$: Class homogeneity

**Contribution Analysis**

- Discover potential marker genes to discriminate cancer classes

$$C_j = \sum_{n=1}^{k} w_n \times p_{n,j} \times (t_{i,n} - t_{r,n})$$

- $C_j$: the contribution of gene $j$
- $p_{n,j}$: the loading of the $j$-th gene on the $n$-th PC
- $t_{i,n}$: the average score of cancer class $i$
- $t_{r,n}$: the average score of reference cancer class
- $w_n$: the weight factor (eigenvalue of $n$-th PC)
Case Study: Classification of Small Round Blue Cell Tumor

- **Cancer DNA Chip data**
  - Total samples: 88-by-2308 (samples-by-variables)
  - Training samples (63), Testing samples (20), Noise samples (5)

- **Small round blue cell tumor**
  - EWS (Ewing family tumor)
  - BL (non-Hodgkin lymphoma)
  - RMS (rhabdomyosarcoma)
  - NB (neuroblastoma)
## Classification Results with SDA

### Classification power 100%

<table>
<thead>
<tr>
<th>Bayesian decision theory</th>
<th>Parametric method</th>
<th>Nonparametric method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear discriminant function</td>
<td>Quadratic discriminant function</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without SDA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-validation of training set</td>
<td>0.8359</td>
<td>0.75</td>
</tr>
<tr>
<td>Classification of the test set</td>
<td>0.7833</td>
<td>0.75</td>
</tr>
<tr>
<td>Using SDA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-validation of training set</td>
<td>0.0238</td>
<td>0.5575</td>
</tr>
<tr>
<td>Classification of the test set</td>
<td>0</td>
<td>0.5417</td>
</tr>
</tbody>
</table>

Training sample: 63, test sample: 25
## The number of identified potential marker genes

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of genes identified using the proposed method</th>
<th>Number of genes identified (Khan et al., 2001)</th>
<th>Number of matched genes</th>
<th>Number of mismatched genes</th>
<th>Image ID number</th>
</tr>
</thead>
<tbody>
<tr>
<td>EWS</td>
<td>54</td>
<td>16</td>
<td>16</td>
<td>0</td>
<td>200814</td>
</tr>
<tr>
<td>BL</td>
<td>45</td>
<td>10</td>
<td>9</td>
<td>1</td>
<td>82225, 813266</td>
</tr>
<tr>
<td>NB</td>
<td>95</td>
<td>15</td>
<td>13</td>
<td>2</td>
<td>788107, 809901, 122159, 245330, 246377, 1409509, 204545, 233721, 563673</td>
</tr>
<tr>
<td>RMS</td>
<td>68</td>
<td>20</td>
<td>14</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Not BL</td>
<td>61</td>
<td>12</td>
<td>8</td>
<td>4</td>
<td>45291, 204545</td>
</tr>
<tr>
<td>Not EWS</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>233721, 563673</td>
</tr>
<tr>
<td>Overlap</td>
<td>24</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>311</td>
<td>74</td>
<td>61</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>

Khan et al. misjudgment 5 : image ID 82225, 813266, 233721, 245330, 122159
Redefine 2: image ID 45291, 563673
Overall trend agree 6 : image ID 204545, 788107, 1409509, 809901, 246377, 200814
The expression profile of potential marker gene (1)

Results are consistent with that of Khan et al., (2001)

Genes expressed in NB class
The expression profile of potential marker gene (2)

Not matched results

NB samples concurrently expressed in the BL, EWS, RMS

Expression level

Tumor class

POSTECH IPSE Lab.
The expression profile of potential marker gene (3)

New discovered potential marker gene

Genes expressed in RMS class
Interpretation of the analysis results

- Hierarchical clustering based on 311 selected potential marker genes
- Correct classification for each class

- Derived from same cell line (GICAN)
- Derived from same cell line (ST486)
- Noise sample
- Noise sample of same cell (sk-muscle)

Highly expressed gene group of each tumor class
## Interpretation of the analysis results:
### Biological validation

### Marker genes for cancer classes

<table>
<thead>
<tr>
<th>Gene Image ID</th>
<th>Cancer class</th>
<th>Biological gene function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1435862</td>
<td>EWS</td>
<td>antigen identified by monoclonal antibodies 12E7, F21 and O13</td>
</tr>
<tr>
<td>291756</td>
<td>EWS</td>
<td>tubulin, beta, 5</td>
</tr>
<tr>
<td>43733</td>
<td>EWS</td>
<td>glycogenin 2</td>
</tr>
<tr>
<td>52076</td>
<td>EWS</td>
<td>olfactomedinrelated ER localized protein</td>
</tr>
<tr>
<td>377731</td>
<td>EWS</td>
<td>glutathione S-transferase M5</td>
</tr>
<tr>
<td>784224</td>
<td>RMS</td>
<td>fibroblast growth factor receptor 4</td>
</tr>
<tr>
<td>470128</td>
<td>RMS</td>
<td>Myosin IC</td>
</tr>
<tr>
<td>296448</td>
<td>RMS</td>
<td>insulin-like growth factor 2 (somatomedin A)</td>
</tr>
<tr>
<td>207274</td>
<td>RMS</td>
<td>Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF</td>
</tr>
<tr>
<td>461425</td>
<td>RMS</td>
<td>Myogenesis</td>
</tr>
<tr>
<td>377671</td>
<td>RMS</td>
<td>integrin, alpha 7</td>
</tr>
<tr>
<td>823886</td>
<td>RMS</td>
<td>Smooth muscle myosin heavy chain isoform SMemb [human, umbilical cord, fetal</td>
</tr>
</tbody>
</table>

### Marker genes not matched for cancer classes

<table>
<thead>
<tr>
<th>Gene Image ID</th>
<th>Chip data Cancer class</th>
<th>Normal Cancer class</th>
<th>Gene Image ID</th>
<th>Chip data Cancer class</th>
<th>Normal Cancer class</th>
</tr>
</thead>
<tbody>
<tr>
<td>823886</td>
<td>Not BL</td>
<td>RMS</td>
<td>782488</td>
<td>All class</td>
<td>Not NB</td>
</tr>
<tr>
<td>897667</td>
<td>EWS</td>
<td>RMS</td>
<td>814773</td>
<td>EWS</td>
<td>NB</td>
</tr>
<tr>
<td>162208</td>
<td>BL</td>
<td>RMS</td>
<td>29054</td>
<td>NB</td>
<td>RMS</td>
</tr>
<tr>
<td>626502</td>
<td>BL</td>
<td>RMS</td>
<td>308231</td>
<td>NB</td>
<td>RMS</td>
</tr>
<tr>
<td>785793</td>
<td>BL</td>
<td>RMS</td>
<td>823886</td>
<td>NB</td>
<td>RMS</td>
</tr>
<tr>
<td>868304</td>
<td>BL</td>
<td>RMS</td>
<td>377048</td>
<td>NB</td>
<td>RMS</td>
</tr>
<tr>
<td>781018</td>
<td>BL</td>
<td>RMS</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Contributions

- Accurate **multivariate** classification method based on Bayesian method
- Potential **marker gene selection** method
- Simple and easy **procedure** for data analysis
- 250 **new candidate marker genes** discovered
- **new hypothesis testing** based on the candidate marker genes for drug discovery or cancer research
Biotechnology meets data mining

- Time to dance!!!
- **Contacts** between the established ‘data mining community’ and ‘bio/medical scientists’ seem to be **rare**
- There will be more dances, and new biotechnology will be forthcoming as we learn the steps

**Coming dance !!!!**
Questions?

Contacts and full paper request: sw74@postech.ac.kr

POSTECH IPSE Lab.