For written notes on this lecture, please read chapter 14 of *The Practical Bioinformatician*.

**CS2220: Introduction to Computational Biology**
**Lecture 4: Gene Expression Analysis**

Liasong Wong

---

**Plan**

- Microarray background
- Gene expression profile classification
- Gene expression profile clustering
- Normalization
- Extreme sample selection
- Intersection Analysis

---

**Background on Microarrays**

- What is a Microarray?
  - Contain large number of DNA molecules spotted on glass slides, nylon membranes, or silicon wafers
  - Detect what genes are being expressed or found in a cell of a tissue sample
  - Measure expression of thousands of genes simultaneously

---

**Affymetrix GeneChip Array**

**Making Affymetrix GeneChip Array**

Exercise: What is the other commonly used type of microarray? How is that one different from Affymetrix's?
Some Advice on Affymetrix Gene Chip Data

- Ignore AFFX genes
  - These genes are control genes
- Ignore genes with “Abs Call” equal to “A” or “M”
  - Measurement quality is suspect
- Upperbound 40000, lowerbound 100
  - Accuracy of laser scanner
- Deal with missing values

Exercise: Suggest 2 ways to deal with missing value
Generally available only in major advanced hospitals

- Conventional risk assignment procedure requires difficult expensive tests and collective judgement of multiple specialists
- Generally available only in major advanced hospitals

⇒ Can we have a single-test easy-to-use platform instead?

Childhood ALL

- Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid≥50
- Diff subtypes respond differently to same Tx
- Over-intensive Tx
  - Development of secondary cancers
  - Reduction of IQ
- Under-intensive Tx
  - Relapse
- The subtypes look similar

Mission

- Conventional diagnosis
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
- Unavailable in most ASEAN countries

Application: Treatment Prognosis

Type of Gene Expression Datasets

Gene Expression Profile Classification

Diagnosis of Childhood Acute Lymphoblastic Leukemia and Optimization of Risk-Benefit Ratio of Therapy

Application: Drug Action Detection

Which group of genes are the drug affecting on?
Overall Strategy

- For each subtype, select genes to develop classification model for diagnosing that subtype
- For each subtype, select genes to develop prediction model for prognosis of that subtype

Subtype Diagnosis by PCL

- Gene expression data collection
- Gene selection by $\chi^2$
- Classifier training by emerging pattern
- Classifier tuning (optional for some machine learning methods)
- Apply classifier for diagnosis of future cases by PCL

Childhood ALL Subtype Diagnosis Workflow

A tree-structured diagnostic workflow was recommended by our doctor collaborator

Training and Testing Sets

<table>
<thead>
<tr>
<th>Pseudoclass</th>
<th>Sample Sizes</th>
<th>Test Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-ALL vs OTHERS</td>
<td>15 vs 82</td>
<td>42 vs 27</td>
<td>0 vs 82</td>
</tr>
<tr>
<td>BCR-ABL vs OTHERS</td>
<td>9 vs 106</td>
<td>27 vs 81</td>
<td>0 vs 82</td>
</tr>
<tr>
<td>CALL vs OTHERS</td>
<td>14 vs 91</td>
<td>42 vs 27</td>
<td>0 vs 82</td>
</tr>
<tr>
<td>CALL vs ALL</td>
<td>14 vs 91</td>
<td>42 vs 27</td>
<td>0 vs 82</td>
</tr>
<tr>
<td>CALL vs BCR-ABL</td>
<td>9 vs 106</td>
<td>27 vs 81</td>
<td>0 vs 82</td>
</tr>
<tr>
<td>CALL vs CALL</td>
<td>9 vs 106</td>
<td>27 vs 81</td>
<td>0 vs 82</td>
</tr>
<tr>
<td>CALL vs OTHERS</td>
<td>14 vs 91</td>
<td>42 vs 27</td>
<td>0 vs 82</td>
</tr>
<tr>
<td>OTHERS vs OTHERS</td>
<td>42 vs 27</td>
<td>0 vs 82</td>
<td>0 vs 82</td>
</tr>
</tbody>
</table>

Signal Selection Basic Idea

- Choose a signal with low intra-class distance
- Choose a signal with high inter-class distance
Signal Selection by $\chi^2$

The $\chi^2$ value of a signal is defined as:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$

where $m$ is the number of intervals, $k$ the number of classes, $A_{ij}$ the number of samples in the $i$th interval, $j$th class, $E_{ij}$ the number of samples in the $i$th interval, $j$th class, $N$ the total number of samples, and $E_{ij}$ the expected frequency of $A_{ij}$ ($E_{ij} = R_i \times C_j/N$).

Emerging Patterns

- An emerging pattern is a set of conditions
  - usually involving several features
  - that most members of a class satisfy
  - but none or few of the other class satisfy

- A jumping emerging pattern is an emerging pattern that
  - some members of a class satisfy
  - but no members of the other class satisfy

- We use only jumping emerging patterns

Examples

<table>
<thead>
<tr>
<th>Patterns</th>
<th>Frequency (P)</th>
<th>Frequency (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[9, 36]</td>
<td>38 instances</td>
<td>0</td>
</tr>
<tr>
<td>[9, 23]</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>[4, 9]</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>[9, 14]</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>[6, 9]</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>[7, 21]</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>[7, 11]</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>[7, 43]</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>[7, 39]</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>[24, 29]</td>
<td>0</td>
<td>34</td>
</tr>
</tbody>
</table>

Easy interpretation

Reference number 9: the expression of gene 37720_at > 215
Reference number 36: the expression of gene 38028_at \leq 12

PCL: Prediction by Collective Likelihood

Let $EP^P_1, \ldots, EP^P_k$ be the most general EPs of $D^P$ in descending order of support.

Suppose the test sample $T$ contains one or more general EPs of $D^P$ in descending order of support:

$$KP^P_1, KP^P_2, \ldots, KP^P_k$$

Define the score of $T$ in the $D^P$ class as:

$$score(T, D^P) = \sum_{i=1}^{k} \frac{\text{frequency}(EP^P_i)}{\text{frequency}(EP^P)}$$

If $score(T, D^P) > score(T, D^N)$, then $T$ is in class $P$.

Otherwise, it is in class $N$.

PCL Learning

Top-Ranked EPs in Positive class

- $EP^P_1$ (90%)  
- $EP^P_2$ (86%)  
- $EP^P_3$ (68%)

Top-Ranked EPs in Negative class

- $EP^N_1$ (100%)  
- $EP^N_2$ (95%)  
- $EP^N_3$ (80%)

The idea of summarizing multiple top-ranked EPs is intended to avoid some rare tie cases

PCL Testing

Most freq EP of pos class in the test sample

$$Score^P = EP^P_1 / EP^P_1 + \ldots + EP^P_k / EP^P_k$$

Most freq EP of pos class

Similarly,$$Score^N = EP^N_1 / EP^N_1 + \ldots + EP^N_k / EP^N_k$$

If $Score^P > Score^N$, then positive class.

Otherwise negative class
Accuracy of PCL (vs. other classifiers)

<table>
<thead>
<tr>
<th>Testing Data</th>
<th>Error rate of different models</th>
<th>C0S</th>
<th>SVM</th>
<th>NB</th>
<th>PCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-ALL vs OTHERS</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>KIA-vs OTHERS</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>TEL-AML vs OTHERS</td>
<td>1.0</td>
<td>0.1</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>ALL vs OTHERS</td>
<td>2.0</td>
<td>3.0</td>
<td>1.0</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Hyperdiploid-84 vs OTHERS</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Total Error</td>
<td>14</td>
<td>6</td>
<td>8</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

The classifiers are all applied to the 20 genes selected by χ² at each level of the tree

Understandability of PCL

- E.g., for T-ALL vs. OTHERS, one ideally discriminatory gene 38319_at was found, inducing these 2 EPs
  \[ \text{gene}._{38319\_at} @ (-\infty, 159.756) \] and
  \[ \text{gene}._{38319\_at} @ [159.756, +\infty) \].

- These give us the diagnostic rule
  If the expression of 38319.at is less than 159.756, then this ALL sample must be a T-ALL. Otherwise it must be a subtype in OTHERS1.

Multidimensional Scaling Plot for Subtype Diagnosis

Obtained by performing PCA on the 20 genes chosen for each level

Childhood ALL Cure Rates

- Conventional risk assignment procedure requires difficult expensive tests and collective judgement of multiple specialists
  - Not available in less advanced ASEAN countries

Childhood ALL Treatment Cost

- Treatment for childhood ALL over 2 yrs
  - Intermediate intensity: US$60k
  - Low intensity: US$36k
  - High intensity: US$72k

- Treatment for relapse: US$150k

- Cost for side-effects: Unquantified

Current Situation (2000 new cases/yr in ASEAN)

- Over intensive for 50% of patients, thus more side effects
- Under intensive for 10% of patients, thus more relapse

- US$120m (US$60k * 2000) for intermediate intensity tx
- US$30m (US$150k * 2000 * 10%) for relapse tx
- Total US$150m/yr plus un-quantified costs for dealing with side effects

Intermediate intensity conventionally applied in less advanced ASEAN countries
Using Our Platform

- Low intensity applied to 50% of patients
- Intermediate intensity to 40% of patients
- High intensity to 10% of patients

⇒ Reduced side effects
⇒ Reduced relapse
⇒ 75-80% cure rates

- US$36m (US$36k * 2000 * 50%) for low intensity
- US$48m (US$60k * 2000 * 40%) for intermediate intensity
- US$14.4m (US$72k * 2000 * 10%) for high intensity

⇒ Total US$98.4m/yr
⇒ Save US$51.6m/yr

A Nice Ending…

- Asian Innovation
Gold Award 2003

Is there a new subtype?

- Hierarchical clustering of gene expression profiles reveals a novel subtype of childhood ALL

Gene Expression Profile Clustering

Novel Disease Subtype Discovery

Hierarchical Clustering

- Assign each item to its own cluster
  – If there are N items initially, we get N clusters, each containing just one item
- Find the “most similar” pair of clusters, merge them into a single cluster, so we now have one less cluster
  – “Similarity” is often defined using
    - Single linkage
    - Complete linkage
    - Average linkage
- Repeat previous step until all items are clustered into a single cluster of size N

Single, Complete, & Average Linkage

- Single linkage defines distance between two clusters as min distance between them
- Complete linkage defines distance between two clusters as max distance between them

Exercise: Name and describe one bi-clustering method
Some Patient Samples

- Does Mr. A have cancer?

Let's rearrange the rows...

- Does Mr. A have cancer?

and the columns too...

Normalization

Sometimes, a gene expression study may involve batches of data collected over a long period of time...

Time Span of Gene Expression Profiles

In such a case, batch effect may be severe... to the extent that you can predict the batch that each sample comes!

⇒ Need normalization to correct for batch effect
Approaches to Normalization

- **Aim of normalization:**
  - Reduce variance w/o increasing bias

- **Scaling method**
  - Intensities are scaled so that each array has the same average value
  - E.g., Affymetrix’s

- **Quantile normalization**
  - Xform data so that distribution of probe intensities is the same on all arrays
  - E.g., \( \frac{x - \mu}{\sigma} \)

Quantile Normalization

- **Given** \( n \) arrays of length \( p \), form \( X \) of size \( p \times n \) where each array is a column
- **Sort** each column of \( X \) to give \( X_{\text{sort}} \)
- **Take** means across rows of \( X_{\text{sort}} \) and assign this mean to each element in the row to get \( X'_{\text{sort}} \)
- **Get** \( X_{\text{normalized}} \) by arranging each column of \( X'_{\text{sort}} \) to have the same ordering as \( X \)

- Implemented in some microarray s/w, e.g., EXPANDER

Selection of Patient Samples and Genes for Disease Prognosis

- **Gene Expression Profile + Clinical Data** \( \rightarrow \) **Outcome Prediction**
  - Univariate & multivariate Cox survival analysis (Beer et al. 2002, Rosenwald et al. 2002)
  - Fuzzy neural network (Ando et al. 2002)
  - Partial least squares regression (Park et al. 2002)
  - Weighted voting algorithm (Shipp et al. 2002)
  - Gene index and “reference gene” (LeBlanc et al. 2003)
  - ……

Our Approach

- “extreme” sample selection
- ERCOF
**Extreme Sample Selection**

Short-term Survivors vs. Long-term Survivors

<table>
<thead>
<tr>
<th>Short-term survivors</th>
<th>Long-term survivors</th>
</tr>
</thead>
<tbody>
<tr>
<td>who died within a short period</td>
<td></td>
</tr>
<tr>
<td>$F(T) &lt; c_1$ and $E(T) = 1$</td>
<td></td>
</tr>
<tr>
<td>who were alive after a long follow-up time</td>
<td></td>
</tr>
<tr>
<td>$F(T) &gt; c_2$</td>
<td></td>
</tr>
</tbody>
</table>

- $T$: sample
- $F(T)$: follow-up time
- $E(T)$: status (1: unfavorable; 0: favorable)
- $c_1$ and $c_2$: thresholds of survival time

**Risk Score Construction**

Linear Kernel SVM regression function

$$G(T) = \sum a_jy_jK(T,x(i)) + b$$

- $T$: test sample, $s(i)$: support vector,
- $y_j$: class label (1: short-term survivors; -1: long-term survivors)

Transformation function (posterior probability)

$$S(T) = \frac{1}{1 + e^{-cT}}$$

$S(T)$: risk score of sample $T$

**Diffuse Large B-Cell Lymphoma**

- DLBC lymphoma is the most common type of lymphoma in adults
- Can be cured by anthracycline-based chemotherapy in 35 to 40 percent of patients
- DLBC lymphoma comprises several diseases that differ in responsiveness to chemotherapy
- Not very good for stratifying DLBC lymphoma patients for therapeutic trials
- Use gene-expression profiles to predict outcome of chemotherapy?

Rosenwald et al., NEJM 2002

- 240 data samples
  - 160 in preliminary group
  - 80 in validation group
  - each sample described by 7399 microarray features
- Rosenwald et al.’s approach
  - identify gene: Cox proportional-hazards model
  - cluster identified genes into four gene signatures
  - calculate for each sample an outcome-predictor score
  - divide patients into quartiles according to score

Knowledge Discovery from Gene Expression of “Extreme” Samples
### Discussions: Sample Selection

<table>
<thead>
<tr>
<th>Application</th>
<th>Data set</th>
<th>Status</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLBCL</td>
<td>Original</td>
<td>88</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>Informative</td>
<td>47+1(*)</td>
<td>72</td>
</tr>
</tbody>
</table>

Number of samples in original data and selected informative training set. (*): Number of samples whose corresponding patient was dead at the end of follow-up time, but selected as a long-term survivor.

### Discussions: Gene Identification

<table>
<thead>
<tr>
<th>Gene selection</th>
<th>DLBCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>493(*)</td>
</tr>
<tr>
<td>Phase I</td>
<td>132(2.7%)</td>
</tr>
<tr>
<td>Phase II</td>
<td>84(1.7%)</td>
</tr>
</tbody>
</table>

Number of genes left after feature filtering for each phase. (*): number of genes after removing those genes who were absent in more than 10% of the experiments.

### Kaplan-Meier Plot for 80 Test Cases

- p-value of log-rank test: < 0.0001
- Risk score thresholds: 0.7, 0.3

### Improvement Over IPI

(A) IPI low, p-value = 0.0063
(B) IPI intermediate, p-value = 0.0003

### Merit of “Extreme” Samples

(A) W/o sample selection (p = 0.38)
(B) With sample selection (p=0.009)

No clear difference on the overall survival of the 80 samples in the validation group of DLBCL study, if no training sample selection conducted.

### About the Inventor: Huiqing Liu

- **Huiqing Liu**
  - PhD, NUS, 2004
  - Currently Senior Scientist at Centocor
  - Asian Innovation Gold Award 2003
  - New Jersey Cancer Research Award for Scientific Excellence 2008
  - Gallo Prize 2008
Beyond Disease Diagnosis & Prognosis

Beyond Classification of Gene Expression Profiles

• After identifying the candidate genes by feature selection, do we know which ones are causal genes, which ones are surrogates, and which are noise?

Genes for class distinction (n=271)

TEL-AML1
BCR-ABL
Hyperdiploid >50
E2A-PBX1
MLL
T-ALL

Copyright 2011 © Limsoon Wong

Gene Regulatory Circuits

• Genes are “connected” in “circuit” or network

• Expr of a gene in a network depends on expr of some other genes in the network

• Can we “reconstruct” the gene network from gene expression and other data?

Copyright 2011 © Limsoon Wong

Hints to extend reach of prediction

• Each disease subtype has underlying cause
  ⇒ There is a unifying biological theme for genes that are truly associated with a disease subtype.

• Uncertainty in reliability of selected genes can be reduced by considering molecular functions and biological processes associated with the genes

• The unifying biological theme is basis for inferring the underlying cause of disease subtype

Intersection Analysis

• Intersect the list of differentially expressed genes with a list of genes on a pathway

• If intersection is significant, the pathway is postulated as basis of disease subtype or treatment response

Exercise: What is a good test statistics to determine if the intersection is significant?

Gene Interaction Prediction
Beyond Classification of Gene Expression Profiles

- After identifying the candidate genes by feature selection, do we know which ones are causal genes and which ones are surrogates?

Key Questions

- For each gene in the network:
  - Which genes affect it?
  - How they affect it?
    - Positively?
    - Negatively?
    - More complicated ways?

Some Techniques

- Bayesian Networks
  - Friedman et al., JCB 7:601–620, 2000
- Boolean Networks
  - Akutsu et al., PSB 2000, pages 293–304
- Differential equations
  - Chen et al., PSB 1999, pages 29–40
- Classification-based method
  - Soinov et al., "Towards reconstruction of gene network from expression data by supervised learning", Genome Biology 4:R6.1–9, 2003

A Classification-Based Technique

- Given a gene expression matrix X
  - each row is a gene
  - each column is a sample
  - each element $x_{ij}$ is expression of gene $i$ in sample $j$
- Find the average value $a_i$ of each gene $i$
- Denote $s_i$ as state of gene $i$ in sample $j$,
  - $s_{ij} = \text{up}$ if $x_{ij} > a_i$
  - $s_{ij} = \text{down}$ if $x_{ij} \leq a_i$

- To see whether the state of gene $g$ is determined by the state of other genes
  - see whether $s_g$ can predict $s_i$ from $x_{ij}$
  - if can predict with high accuracy, then "yes"
  - Any classifier can be used, such as C4.5, PCL, SVM, etc.
- To see how the state of gene $g$ is determined by the state of other genes
  - apply C4.5 (or PCL or other "rule-based" classifiers) to predict $s_i$ from $x_{ij}$
  - and extract the decision tree or rules used
Advantages of this method

- Can identify genes affecting a target gene
- Don’t need discretization thresholds
- Each data sample is treated as an example
- Explicit rules can be extracted from the classifier (assuming C4.5 or PCL)
- Generalizable to time series

Concluding Remarks

Bcr-Abl

- Targeted drug development
  - Know what molecular effect you want to achieve
  - E.g., inhibit a mutated form of a protein
  - Engineer a compound that directly binds and causes the desired effect
- Gleevec (imatinib)
  - 1st success for real drug
  - Targets Bcr-Abl fusion protein (i.e., Philadelphia chromosome, Ph)
  - NCI summary of clinical trial of imatinib for ALL at http://www.cancer.gov/clinicaltrials/results/ALLimatinib1109/print

What have we learned?

- Technologies
  - Microarray
  - PCL, ERCOF
- Microarray applications
  - Disease diagnosis by supervised learning
  - Subtype discovery by unsupervised learning
- Important tactics
  - Extreme sample selection
  - Intersection analysis, Gene network reconstruction

References