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

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

Limsoon 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

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 values
Application: Treatment Prognosis

- 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

Application: Drug Action Detection

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

Type of Gene Expression Datasets

- Gene-Conditions or Gene-Sample (numeric or discretized)

Gene Expression Profile Classification

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

Mission

- 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?
Single-Test Platform of Microarray & Machine Learning

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>Dataset</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-ALL vs. ALL</td>
<td>28 vs 117</td>
<td>13 vs 27</td>
</tr>
<tr>
<td>B-ALL vs. ALL</td>
<td>18 vs 109</td>
<td>9 vs 29</td>
</tr>
<tr>
<td>TEL-AML vs. ALL</td>
<td>52 vs 111</td>
<td>27 vs 42</td>
</tr>
<tr>
<td>BCR-ABL vs. ALL</td>
<td>9 vs 50</td>
<td>6 vs 35</td>
</tr>
<tr>
<td>MLL vs. ALL</td>
<td>14 vs 94</td>
<td>6 vs 49</td>
</tr>
<tr>
<td>HyperDip-56 vs.</td>
<td>32 vs 62</td>
<td>27 vs 27</td>
</tr>
<tr>
<td>OTHERS unknown</td>
<td>14 vs 94</td>
<td>6 vs 49</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:

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

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, $R_j$ the number of samples in the $j$th interval, $C_i$ the number of samples in the $i$th class, $N$ the total number of samples, and $E_{ij}$ the expected frequency of $A_{ij}$ ($E_{ij} = R_i \cdot C_j/N$).

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

PCL Learning

Top-Ranked EPs in Positive class

EP_1^P (90%)  
EP_2^P (86%)  
EP_3^P (68%)  

Top-Ranked EPs in Negative class

EP_1^N (100%)  
EP_2^N (95%)  
EP_3^N (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

$$\text{Score}^P = \frac{\text{EP}_1^P}{\text{EP}_1^P + \ldots + \text{EP}_k^P}$$

Most freq EP of pos class

Similarly,

$$\text{Score}^N = \frac{\text{EP}_1^N}{\text{EP}_1^N + \ldots + \text{EP}_k^N}$$

If $\text{Score}^P > \text{Score}^N$, then positive class, Otherwise negative class
Accuracy of PCL (vs. other classifiers)

<table>
<thead>
<tr>
<th>Testing Data</th>
<th>Error rates of different models</th>
<th>CLS</th>
<th>SVM</th>
<th>NB</th>
<th>PCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-ALL vs OTHERS1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>ESA-PERK vs OTHERS2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>TEL-AML1 vs OTHERS3</td>
<td>1.1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>BCL-ALL vs OTHERS4</td>
<td>2.0</td>
<td>3.0</td>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>MLL vs OTHERS5</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Hyperdiploid&gt;50 vs OTHERS</td>
<td>2.6</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

The classifiers are all applied to the 20 genes selected by $\chi^2$ at each level of the tree.

Understandability of PCL

- E.g., for T-ALL vs. OTHERS, one ideally discriminative gene 38319_at was found, inducing these 2 EPs:

  \[ \text{gene}_{38319\_at} \begin{cases} \leq (15975.6) \text{ if} & \text{gene}_{38319\_at} \geq (15975.6) \end{cases} \text{ and} \]

-These give us the diagnostic rule:

  If the expression of 38319_at is less than 15975.6, 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

Exercise: Give definition of “average linkage”
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

- Xform data so that distribution of probe intensities is same on all arrays
  - E.g., \((x - \mu) / \sigma\)

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

- Quantile normalization

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 elem in the row to get \(X'_{\text{sort}}\)
- Get \(X'_{\text{normalized}}\) by arranging each column of \(X'_{\text{sort}}\) to have same ordering as \(X\)

Selection of Patient Samples and Genes for Disease Prognosis
Gene Expression Profile + Clinical Data ⇒ 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 (Strang et al 2002)
- Gene index and “reference gene” (LeBlanc et al 2003)
- ……

Our Approach

“extreme” sample selection

ERCOF

Extreme Sample Selection

Short-term Survivors v.s. Long-term Survivors

Short-term survivors who died within a short period ↓

\[ F(T) < c_2 \text{ and } E(T) = 1 \]

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

Long-term survivors who were alive after a long follow-up time ↓

\[ F(T) > c_2 \]

Risk Score Construction

Linear Kernel SVM regression function

\[ G(T) = \sum_{i} a_i y_i K(T, x(i)) + b \]

\( T \): test sample, \( y_i \): support vector, \( K(T, x(i)) \): kernel function

Transformation function (posterior probability)

\[ S(T) = \frac{1}{1 + e^{-a(T)}} \quad (S(T) \in (0,1)) \]

\( S(T) \): 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
- Intl Prognostic Index (IPI)
  - age, “Eastern Cooperative Oncology Group” Performance status, tumor stage, lactate dehydrogenase level, sites of extranodal disease, ...
- 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**

- “extreme” sample selection: < 1 yr vs > 8 yrs
- knowledge discovery from gene expression
  - T is long-term if S(T) < 0.3
  - T is short-term if S(T) > 0.7

**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>Dead</td>
<td>88</td>
</tr>
<tr>
<td>DLBCL</td>
<td>Informative</td>
<td>Alive</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>160</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>4937(*)</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

- 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?

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?

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.

**Caution:**
- Initial list of differentially expressed genes is defined using test statistics with arbitrary thresholds.
- Diff test statistics and diff thresholds result in a diff list of differentially expressed genes.
- Outcome may be unstable.

**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?

---

**Gene Regulatory Circuits**

- Genes are “connected” in “circuit” or network.
- Expression of a gene in a network depends on expression of some other genes in the network.
- Can we reconstruct the gene network from gene expression data?

---

**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**
A Classification-Based Technique
Soinov et al., Genome Biology 4:R6.1-9, Jan 2003

• 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 $s_{ij}$ 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 $(s_{ij} | i \neq g)$ can predict $s_g$
  – if can predict with high accuracy, then “yes”
  – Any classifier can be used, such as C4.5, PCL, SVM, etc.

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

Bcr-Abl

• Targeted drug dev
  – 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/citeli
    c results/ALL Imatinib1109/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

Concluding Remarks
Any Question?

References

• E.-J. Yeoh et al., “Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling”, Cancer Cell, 1:133–143, 2002
• J. Li, L. Wong, “Techniques for Analysis of Gene Expression”, The Practical Bioinformatician, Chapter 14, pages 319—346, WSPC, 2004