Bioinformatics and Biomarker Discovery

Part 3: Examples

Limsoon Wong
3 September 2014
Outline

• **ALL**
  – Gene expression profile classification
  – Beyond diagnosis and prognosis

• **WEKA**
  – Breast cancer
  – Dermatology
  – Pima Indians
  – Echocardiogram
  – Mammography
Gene Expression Profile Classification

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

- Conventional diagnosis
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics

- Unavailable in most ASEAN countries
Subtype Diagnosis by Machine Learning

- Gene expression data collection
- Gene selection by e.g. $\chi^2$
- Classifier training by e.g. emerging pattern
- Classifier tuning (optional for some machine learning methods)
- Apply classifier for diagnosis of future cases by e.g. PCL
A tree-structured diagnostic workflow was recommended by our doctor collaborator.
# Training and Testing Sets

<table>
<thead>
<tr>
<th>Paired datasets</th>
<th>Ingredients</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-ALL vs OTHERS1</td>
<td>OTHERS1 = {E2A-PBX1, TEL-AML1, BCR-ABL, Hyperdip&gt;50, MLL, OTHERS}</td>
<td>28 vs 187</td>
<td>15 vs 97</td>
</tr>
<tr>
<td>E2A-PBX1 vs OTHERS2</td>
<td>OTHERS2 = {TEL-AML1, BCR-ABL, Hyperdip&gt;50, MLL, OTHERS}</td>
<td>18 vs 169</td>
<td>9 vs 88</td>
</tr>
<tr>
<td>TEL-AML1 vs OTHERS3</td>
<td>OTHERS3 = {BCR-ABL, Hyperdip&gt;50, MLL, OTHERS}</td>
<td>52 vs 117</td>
<td>27 vs 61</td>
</tr>
<tr>
<td>BCR-ABL vs OTHERS4</td>
<td>OTHERS4 = {Hyperdip&gt;50, MLL, OTHERS}</td>
<td>9 vs 108</td>
<td>6 vs 55</td>
</tr>
<tr>
<td>MLL vs OTHERS5</td>
<td>OTHERS5 = {Hyperdip&gt;50, OTHERS}</td>
<td>14 vs 94</td>
<td>6 vs 49</td>
</tr>
<tr>
<td>Hyperdip&gt;50 vs OTHERS</td>
<td>OTHERS = {Hyperdip47-50, Pseudodip, Hypodip, Normo}</td>
<td>42 vs 52</td>
<td>22 vs 27</td>
</tr>
</tbody>
</table>
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, $R_i$ the number of samples in the $i$th interval, $C_j$ the number of samples in the $j$th class, $N$ the total number of samples, and $E_{ij}$ the expected frequency of $A_{ij}$ ($E_{ij} = R_i * C_j / N$).
# Accuracy of Various Classifiers

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

<table>
<thead>
<tr>
<th>Testing Data</th>
<th>Error rate of different models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C4.5</td>
</tr>
<tr>
<td>T-ALL vs OTHERS1</td>
<td>0:1</td>
</tr>
<tr>
<td>E2A-PBX1 vs OTHERS2</td>
<td>0:0</td>
</tr>
<tr>
<td>TEL-AML1 vs OTHERS3</td>
<td>1:1</td>
</tr>
<tr>
<td>BCR-ABL vs OTHERS4</td>
<td>2:0</td>
</tr>
<tr>
<td>MLL vs OTHERS5</td>
<td>0:1</td>
</tr>
<tr>
<td>Hyperdiploid&gt;50 vs OTHERS</td>
<td>2:6</td>
</tr>
<tr>
<td>Total Errors</td>
<td>14</td>
</tr>
</tbody>
</table>
Normalization
Sometimes, a gene expression study may involve batches of data collected over a long period of time…

Time Span of Gene Expression Profiles

Image credit: Dong Difeng
In such a case, batch effect may be severe… to the extent that you can predict the batch that each sample comes!

\[ \Rightarrow \text{Need normalization to correct for batch effect} \]
Approaches to Normalization

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

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

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

- **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 \( X \)

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

Figure 3.6: GEPs after the batch effects removing.
Percentage of Overlapping Genes

- Low % of overlapping genes from diff expt in general
  - Prostate cancer
    • Lapointe et al, 2004
    • Singh et al, 2002
  - Lung cancer
    • Garber et al, 2001
    • Bhattacharjee et al, 2001
  - DMD
    • Haslett et al, 2002
    • Pescatori et al, 2007

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DEG</th>
<th>POG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Lung Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>DMD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>

Zhang et al, Bioinformatics, 2009
Individual Genes

• Suppose
  – Each gene has 50% chance to be high
  – You have 3 disease and 3 normal samples

• How many genes on a microarray are expected to perfectly correlate to these samples?

• Prob(a gene is correlated) = $1/2^6$

• # of genes on array = 30,000

⇒ $E(\# \text{ of correlated genes}) = 468$

⇒ Many false positives
• These cannot be eliminated based on pure statistics!
Group of Genes

• Suppose
  – Each gene has 50% chance to be high
  – You have 3 disease and 3 normal samples

• What is the chance of a group of 5 genes being perfectly correlated to these samples?

• Prob(group of genes correlated) = \( \left( \frac{1}{2} \right)^6 \)^5
  – Good, \( \ll 1/2^6 \)

• # of groups = \( \binom{30000}{5} \)

\[ E(\text{# of groups of genes correlated}) = \binom{30000}{5} \times \left( \frac{1}{2^6} \right)^5 = 2 \times 10^{11} \]

⇒ Even more false positives?

• Perhaps no need to consider every group
Gene Regulatory Circuits

- Each disease phenotype has some underlying cause
- There is some unifying biological theme for genes that are truly associated with a disease subtype
- Uncertainty in selected genes can be reduced by considering biological processes of the genes
- The unifying biological theme is basis for inferring the underlying cause of disease subtype
Taming false positives by considering pathways instead of all possible groups

Group of Genes

- Suppose
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples

- What is the chance of a group of 5 genes being perfectly correlated to these samples?

\[ \text{Prob(group of genes correlated)} = (1/2^6)^5 \]
- Good, \(< 1/2^6\)

\[ \# \text{ of groups} = \binom{30000}{5} \]

\[ E(\# \text{ of groups of genes correlated}) = \binom{30000}{5} \cdot (1/2^6)^5 = 2 \times 10^{11} \]

\[ \Rightarrow \text{Even more false positives?} \]
- Perhaps no need to consider every group

\# of pathways = 1000

\[ E(\# \text{ of pathways correlated}) = 1000 \cdot (1/2^6)^5 = 9.3 \times 10^{-7} \]
Towards More Meaningful Genes

- ORA
  - Khatri et al
  - *Genomics, 2002*

- FCS
  - Pavlidis & Noble
  - PSB 2002

- GSEA
  - Subramanian et al
  - *PNAS, 2005*

- SNet
  - Soh et al
  - *BMC Genomics, 2011*

Overlap Analysis

Direct-Group Analysis

Network-Based Analysis
Intersection Analysis (ORA)

- 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?
Disappointing Performance

DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data
- PathwayAPI, Soh et al., 2010
Issue #1 with ORA

- Its null hypothesis basically says “Genes in the given pathway behaves no differently from randomly chosen gene sets of the same size”

- This null hypothesis is obviously false

⇒ Lots of false positives

- A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Thus necessarily the behaviour of genes in a pathway is more coordinated than random ones
Issue #2 with ORA

- It relies on a pre-determined list of DE genes
- This list is sensitive to the test statistic used and to the significance threshold used
- This list is unstable regardless of the threshold used when sample size is small
Issue #3 with ORA

- It tests whether the entire pathway is significantly differentially expressed.

- If only a branch of the pathway is relevant to the phenotypes, the noise from the large irrelevant part of the pathways can dilute the signal from that branch.
ORA-Paired: Paired Test and New Null Hypothesis

- Let $g_i$ be genes in a given pathway $P$
- Let $p_j$ be patients
- Let $q_k$ be normals

- Let $\Delta_{i,j,k} = \text{Expr}(g_i, p_j) - \text{Expr}(g_i, q_k)$

- Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

- **Issue #1 is solved**
  - The null hypothesis is now “If a pathway $P$ is irrelevant to the difference between patients and normals, then the genes in $P$ are expected to behave similarly in patients and normals”

- **Issue #2 is solved**
  - No longer need a pre-determined list of DE genes

- **Issue #3 is unsolved**

- **Is sample size now larger?**
  - $|\text{patients}| \times |\text{normals}| \times |\text{genes in } P|$
NEA-Paired: Paired Test on Subnetworks

• Given a pathway P

• Let each node and its immediate neighbourhood in P be a subnetwork

• Apply ORA-Paired on each subnetwork individually

• Issues #1 & #2 are solved as per ORA-Paired

• Issue #3 is partly solved
  – Testing subnetworks instead of whole pathways
  – But subnetworks derived in fragmented way
ESSNet: Larger Subnetworks

- Compute the average rank of a gene based on its expression level in patients
- Use the top $\alpha\%$ to extract large connected components in pathways
- Test each component using ORA-Paired

- Gene rank is very stable
- Issues #1 - #3 solved
Fantastic Performance

upregulated in DMD

sample size (N)

subnetwork agreement

ESSNet
NEA-Paired
ORA-Paired
PFSNet
GSEA
ORA
Concluding Remarks

• Consistent successful gene expression profile analysis needs deep integration of background knowledge

• Most gene expression profile analysis methods fail to give reproducible results when sample size is small (and some even fail when sample size is quite large)

• Logical analysis to identify key issues and simple logical solution to the issues can give fantastic results
References

A Popular Software Package: WEKA
http://www.cs.waikato.ac.nz/ml/weka

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Exercise: Download a copy of WEKA. What are the names of classifiers in WEKA that correspond to C4.5 and SVM?
Let’s try WEKA on …

• Breast cancer

• Dermatology

• Pima Indians

• Echocardiogram

• Mammography