Finding Co-clusters of Genes and Clinical Parameters

Sungroh Yoon, *Stanford University*
Luca Benini, *University of Bologna*
Giovanni De Micheli, *EPF Lausanne*
Linking Genes with Clinical Traits

- Given specific clinical traits of interest, to determine which genes are responsible
- Can have a major impact on clinical diagnosis and prognosis
- Typically done by correlating gene expression with trait measurement
  - DNA microarray technology
    - Enables us to monitor expression levels of thousands of genes simultaneously
    - Sparked development of new methods
Correlation Analysis

- Compute the Pearson correlation coefficient between trait vector and gene expression vector

<table>
<thead>
<tr>
<th>Trait 1</th>
<th>Trait 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>0.7</td>
<td>0.5</td>
</tr>
<tr>
<td>0.8</td>
<td>0.6</td>
</tr>
<tr>
<td>0.9</td>
<td>0.5</td>
</tr>
<tr>
<td>1.0</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>0.2</td>
</tr>
<tr>
<td>0.3</td>
</tr>
<tr>
<td>0.4</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>0.6</td>
</tr>
<tr>
<td>0.7</td>
</tr>
<tr>
<td>0.8</td>
</tr>
<tr>
<td>0.5</td>
</tr>
</tbody>
</table>

Patients

(Nardini et al., 2004)
Challenges

• As the number of clinical traits increases, the inspection method breaks down
  – We need a systematic approach

• The values of clinical traits are not necessarily continuous or numeric
  – We need a more generalized statistic than the Pearson correlation coefficient

Our Approach

- Construct a “correlation matrix”
  - Use the SAM statistic (Tusher et al., 2001)
- Find local structures in that matrix
  - Use the technique of co-clustering
Co-clustering

- Simultaneous clustering of rows and columns in a data matrix
  - Co-clusters are represented by a submatrix
  - Co-clusters can overlap with each other

- Various applications in data mining
  - Text mining: word-document
  - Gene expression analysis: gene-patient

- Computationally challenging
  - Can be reduced to the problem of finding maximum edge bicliques
### Examples of Co-clusters

- **All constant**
  - Hartigan, 1972
  - Tibshirani et al., 1999

- **Constant rows/columns**
  - Getz et al., 2000
  - Califano et al., 2000
  - Segal et al., 2001

- **Common trend**
  - Cheng and Church, 2000
  - Wang et al., 2002
  - Kluger et al., 2003
  - Ben-Dor et al., 2002
  - Tanay et al., 2002
Definition

• **Co-cluster** of genes and clinical traits
  – A submatrix of the correlation matrix
  – For any pair of column vectors, the inter-column distance is less than a threshold
Measuring Inter-column Distance

- For a vector $V$
  \[ \text{range}(V) = \max(V) - \min(V) \]
- For two vectors $V$ and $W$
  \[ \text{distance}(V, W) = \text{range}(V - W) \leq \tau \]
- Rationale
  - Computational efficiency
Our Co-clustering Method

• Input
  – A matrix of clinical traits
  – A gene expression matrix

• Output
  – Co-clusters of genes and traits

• Two-step process
  – Step 1: Find “seed” co-clusters
  – Step 2: Merge seeds to find co-clusters

• Can find all the co-clusters
  – That satisfy specific input conditions
  – That are statistically significant
Experimental Results

• AML data set
  – Bullinger et al., 2004
  – 6283 genes
  – 119 patients
  – 15 clinical traits
• 43 GT co-clusters
Biological Validation

• “Survival” with TGFB1/2 and CD1a
  – Riedl et al., 1997, Jcobsen et al., 1996, Weisberg et al., 2002

• Genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Item</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALT1</td>
<td>GO term</td>
<td>Defense response</td>
</tr>
<tr>
<td>NFIL3</td>
<td>Corrected p-value</td>
<td>0.0359</td>
</tr>
<tr>
<td>APOH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FCGRT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SERPINA1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1QA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OAS2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD3G</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Description

CD3G antigen, gamma polypeptide

Mal tissue lymphoma translocation gene 1

Nuclear factor, interleukin 3 regulated

CD3G, 1996, Weisberg
Summary

- Linking genes and clinical traits
  - Can lead to a major impact on diagnosis
  - DNA microarray opened a door for new methods

- Our approach
  - Construct a correlation matrix
  - Find co-clusters of genes and traits appearing on the correlation matrix

- Experimental results
  - Tested with AML data set
  - Successfully identify 43 co-clusters
Thank You!