Microarray Data Analysis

CENG 734

Fall 2006

Quiz #2

1. Describe the contents (i.e., the format) of the gene expression data used by the algorithm described in the paper that uses GO graph structure to find significant GO terms. (E.g., Is the data a matrix of expression values, a graph of genes, etc.?)

2. Which technique is used to find the best parameters of the patient-gene probabilistic model for temporal expression profiles?

Microarrays

- Brief introduction
- Paper #1: Enrichment of top genes with GO terms using the underlying GO graph structure
- Paper #2: How to generate a probabilistic model to combine multiple temporal expression profiles by account for patient differences.

Gene expression

- Cells are different because of differential gene expression.
- About 40% of human genes are expressed at any one time.
- Gene is expressed by transcribing DNA into single-stranded mRNA
- mRNA is later translated into a protein
- Microarrays measure the level of mRNA expression
- Differential gene expression is also observed in different conditions such as disease, heat, starvation, etc.

A demonstration of the underlying technology

- DNA microarray animation by A. Malcolm Campbell.
  - chip.swf (Shockwave flash animation from Ceng 465)

Experimental conditions

- Different tissues
- Different developmental stages
- Different disease states
- Different treatments
Data mining challenges

• Too few experiments (samples), usually < 100
• Too many genes, usually > 1,000
• Too many genes lead to false positives
• For exploration, a large set of all relevant genes is desired
• For diagnostics or identification of therapeutic targets, the smallest set of genes is needed

Paper #1

• Improved scoring of functional groups from gene expression data by decorrelating GO graph structure
• Authors from Max-Planck Institute for Informatics
• Published in Bioinformatics journal July 2006 issue (Advance access April 2006).

What is the problem they try to solve?

• Biologists want to interpret the results of a microarray experiment.
• What is the result of a microarray experiment?
  – A set of differentially expressed genes if we are comparing genes of two different conditions. Usually ranked.
  – We need to characterize the genes, say something about them to reason about the disease. Example: All the differentially expressed genes are involved in cell-cycle. So, this disease affect cell division?

The Gene Ontology

• It is a controlled vocabulary, a hierarchical dictionary.
• If you know something about a gene, you can state it as a GO term, i.e., annotate the gene with a GO term.
• Example:
  – response to stress is_a response to stimuli is_a physiological process is_a biological process
• There are three GO hierarchies:
  – Biological Process, Cellular Component, Molecular Function
Gene Expression and GO ontology

- Analyze the top differentially expressed genes and identify their GO terms. In other words: which GO terms are most highly represented among the top genes.
- This is also called GO term enrichment of differentially expressed genes.
- Usually GO terms are ordered by statistical significance
  - i.e., is it by chance that I saw that GO term annotated to some of the top genes?

Classic methods

- Classic methods evaluate the statistical significance of each GO term separately using the set of genes annotated with that GO term and the top genes annotated with that term.

Computing the significance

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>sigGenes</td>
<td>sigGenes</td>
<td>allGenes</td>
</tr>
<tr>
<td>genes[u]</td>
<td>genes[u]</td>
<td>genes[u]</td>
</tr>
<tr>
<td>genes[u]</td>
<td>genes[u]</td>
<td>genes[u]</td>
</tr>
<tr>
<td>genes[u]</td>
<td>genes[u]</td>
<td>genes[u]</td>
</tr>
<tr>
<td>genes[u]</td>
<td>genes[u]</td>
<td>genes[u]</td>
</tr>
</tbody>
</table>

- Fisher’s exact test is applied to test the associations of $A$ with $B$.
- $p$-value gives the probability of observing at least the same amount of GO terms when the top genes were selected randomly among all genes.

Contribution of this paper

- When computing the significance make use of the neighborhood and GO hierarchy level information.
- In the classic approach, if a GO term contains the same genes as one of its children, they will get the same significance score. However, intuitively the child gives a biologist more information, e.g., response to external stimuli vs. response to stimuli.
- Therefore, it could have been better if we gave a higher score to the child node.

The $elim$ algorithm

- A very simple heuristic that gives priority to significant children nodes:
  - Proceed bottom-up in the GO hierarchy
  - If a node $u$ is found significant, remove all the genes of $u$ from the gene list of its parents (i.e., the upperInducedGraph($u$))
- This way, a parent node will not be able to use the genes of $u$ when computing its significance

```
Algorithm 1 elim

markedGenes ← Ø; nodeSig ← Ø
get the DAG levels list dagLevels
for i from max(dagLevels) to 1
  for u in nodes(dagLevels, i)
    genes[u] ← genes[u] \ markedGenes[u]
    nodeSig[u] ← FisherTest(genes[u], sigGenes)
    if nodeSig[u] ≤ threshold then
      for x in upperInducedGraph(u)
        markedGenes[x] ← markedGenes[x] \ genes[u]
  end
end
return nodeSig
```
The weight algorithm

- The *elim* algorithm can be considered as giving 0 or 1 weights to the genes when computing the significance. A more general strategy would be to allow for arbitrary weights from \([0..1]\).
- The weight is accomplished by a significance ratio, \(\text{sigRatio}\):

\[
\text{w} = \frac{\text{sigRatio}(\text{score}(\mathbf{e}), \text{score}(\mathbf{u}))}{\text{f}(\mathbf{a})}
\]

Computing the significance

- The same Fisher’s exact test can be applied. This time the number of genes in a set is replaced by the summation of weights.

\[
x = \sum_{i \in \text{sigGenet}(\text{set}(\mathbf{a}))} \text{weight}[i]
\]

The weight algorithm

Algorithm 2: weight

```pseudo
for \(a\) in \text{node}(\text{dag})
    \text{node}\[a\] = \text{null}
end

get the DAG levels list \(\text{algLevels}\)
for \(i\) from \(\text{max}(\text{algLevels})\) to 1
    for \(a\) in \text{node}(\text{algLevels}[i])
        \text{computeTermSign}(\text{node}(\text{algLevels}[i]), \text{children}(\text{node}(\text{algLevels}[i])))
    end
end

return \text{node}\[\text{dagg}]
```

The combined algorithm

- Mean of \(p\)-values obtained with *classic*, *elim* and *weight* algorithms, computed on a log scale

\[
\hat{p} = \exp\left(\frac{1}{3} \sum_{i=1}^{3} \log(p\text{-value}_i)\right)
\]

Experiments

- Two real expression data sets and on simulated experiment.
- Real expression data set was about Leukemia patients. The first experiment had 515 differentially expressed genes and the second experiment had 682 differentially expressed genes.
- Some general GO terms that are found significant by the *classic* method are eliminated by *elim* and *weight* methods.

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Term</th>
<th>Observed</th>
<th>Expected</th>
<th>Adjusted</th>
<th>classic</th>
<th>elim</th>
<th>weight</th>
<th>reranking</th>
<th>allRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>signal transduction</td>
<td>22</td>
<td>3.9e-17</td>
<td>6.26E-3</td>
<td>1.0e-17</td>
<td>8.2e-17</td>
<td>7.8e-17</td>
<td>4.5e-17</td>
<td>3.6e-17</td>
</tr>
<tr>
<td>2</td>
<td>immune response</td>
<td>105</td>
<td>1.2e-1</td>
<td>1.9e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
</tr>
<tr>
<td>3</td>
<td>antigen processing</td>
<td>104</td>
<td>1.2e-1</td>
<td>1.9e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
</tr>
<tr>
<td>4</td>
<td>immune response</td>
<td>104</td>
<td>1.2e-1</td>
<td>1.9e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
</tr>
<tr>
<td>5</td>
<td>immune response</td>
<td>104</td>
<td>1.2e-1</td>
<td>1.9e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
</tr>
<tr>
<td>6</td>
<td>immune response</td>
<td>104</td>
<td>1.2e-1</td>
<td>1.9e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
</tr>
<tr>
<td>7</td>
<td>immune response</td>
<td>104</td>
<td>1.2e-1</td>
<td>1.9e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
</tr>
<tr>
<td>8</td>
<td>immune response</td>
<td>104</td>
<td>1.2e-1</td>
<td>1.9e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
</tr>
<tr>
<td>9</td>
<td>immune response</td>
<td>104</td>
<td>1.2e-1</td>
<td>1.9e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
</tr>
<tr>
<td>10</td>
<td>immune response</td>
<td>104</td>
<td>1.2e-1</td>
<td>1.9e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
<td>1.5e-1</td>
</tr>
</tbody>
</table>

Note: The GO ID 0000000 represents the expected number of interesting genes assigned to the GO term, the interesting genes were manually thresholded to control FDR.
Simulated Gene Expression Sets

- Selected 50 GO terms that contain a certain number of genes (in total). Select the genes of all these 50 GO terms as top-ranked genes. Add 10% noise. The score of a method is given by the amount of enriched genes it can place in top-k of most significant GO terms.

\[
\text{score}_k^M = |\text{top}_k(M) \cap \text{enriched}| 
\]

Results

<table>
<thead>
<tr>
<th>k</th>
<th>class</th>
<th>weight_log</th>
<th>weight_ratio</th>
<th>elm</th>
<th>all.M</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>5.5</td>
<td>13</td>
<td>14.2</td>
<td>16.9</td>
<td>15.4</td>
</tr>
<tr>
<td>50</td>
<td>14.5</td>
<td>25.5</td>
<td>28</td>
<td>27</td>
<td>28.5</td>
</tr>
<tr>
<td>75</td>
<td>22.5</td>
<td>36</td>
<td>38</td>
<td>31</td>
<td>38</td>
</tr>
<tr>
<td>100</td>
<td>31</td>
<td>42</td>
<td>39.5</td>
<td>33.5</td>
<td>43</td>
</tr>
</tbody>
</table>

Discussion

- Pros and Cons?
- Possible improvements:
  - Fisher's exact test does not consider the “order” of differentially expressed genes. How can we incorporate the “order” information into the scoring (i.e., the statistical significance) function?
  - When we are determining the weight of a term \( u \), what happens when two of its more significant child nodes contain the same gene? What will be the weight?
Paper #2

• Combining multiple temporal expression profiles considering the differences in patient response (amount and rate) using a probabilistic model
• Authors from University of Pittsburgh and Carnegie Mellon University
• Paper appeared in RECOMB 2006 conference.

The problem

• Given a number of different temporal expression profiles collected from different patients. Can we find the genes that are common or specific to a patient? And what will be the behavior of these genes under the studied condition? Can we also find the time at which a patient starts responding along with the response rate?

A Gene-Patient Model

• A gene-patient model is introduced with several parameters

The Goal

• Given the amount of expression for several genes and several patients and several time intervals. How can we find the parameters of the gene-patient model that will fit the data best?
• Solution:
  - Use Expectation-Maximization algorithm
  - Starting from an initial set of parameters, and assuming part of the model is known (\( \gamma_g \) and \( z_q \)) and the rest is unknown, predict the unknown components (maximization), recompute \( \gamma_g \) and \( z_q \) (expectation). Iterate until convergence
  - Does not guarantee optimum parameter set, may stuck at a local optima.

Results

Fig. 4. Top row: Three genes expressed in a similar way in all six patients according to the posterior computed by our algorithm. Bottom row: Three genes that were expressed similarly in five of the patients, but different in the sixth. Time is on a log scale due to the sampling rate. Note that the average computed in the bottom row (dotted line) is affected by the outlier, while the consensus computed by our algorithm (blue line) is not.

Discussion

• Questions?
• Pros and Cons?
• Improvements?
Projects

• Project proposals due by November 1.
  – One short paragraph description of your project to keep track of who is doing what.
• Schedule and appointment if you have not decided on your project yet.