StAM – Structured Analysis of Microarray Data

Claudio Lottaz and Rainer Spang

Computational Diagnostics Group
Computational Molecular Biology Department
Max Planck Institute for Molecular Genetics
Ihnestrasse 73, D-14195 Berlin (Germany)
Overview

• Introduction
• Class prediction using Gene Ontology annotations
• Performance evaluation
• Observations on weighs and nodewise predictions
• Discussion
Problem Statement

• **Goal**: medical diagnostics using gene expression patterns
• **Data**: many genes – few samples – no structure
• **Our approach**: add structure using functional annotations in addition to expression data
• **Implication**: relate prediction results to biological aspect ⇒ rationale for computational results
**Gene Ontology**

- Structure knowledge about genes
- Directed acyclic graph
- Represents knowledge on
  - Molecular function
  - Biological process
  - Cellular component
- Genes are annotated to nodes in the graph
Nearest Shrunken Centroids
[Tibshirani et al., 2002]

- **Centroids** represent classes
- **Shrinkage** weights influence of genes
- **Soft thresholding** leads to removal of undiscriminating genes
- **Classification** to nearest shrunken centroid.

![Diagram showing shrunken centroids and gene expression data points.](image-url)
One Diagnostic Predictor per GO-Node

- One predictor per GO-node $N$ contains 2 classifiers
- Only genes annotated with $N$ or its successors are used for classification.
- First classifier for directly annotated genes, based on nearest shrunken centroids
- Second classifier for children, a weighted sum with normalization

*Diagram showing the structure of GO annotations with $N_y$, direct genes, and children.*
Bottom-up Information Propagation through Weights

- Start with leaf-nodes (postorder traversal)
- Use results of CHILDREN to train their parents
- Edges carry weights for each class.
- Weights are chosen proportional to $p_{\text{correct}} - p_{\text{a-priori-correct}}$ in child (zero if negative)
- Scores computed as weighted sums are normalized to mimic probabilities.
Explaining Classification

- Weights on edges after supervised training as well as nodewise accuracy after cross validation:
  - Which biological aspects (nodes) are considered important in a classification task?
- Results in nodes after classification of a single case:
  - Which aspects favour the predicted class?
  - Which aspects are missing compared to a typical case of the predicted class?
Implementation

• Java-program (by Stefan Bentink)
  • Crawls through the Gene Ontology
  • Annotates probe-sets to GO nodes
  • Generates post-order list of GO-nodes
• Perl-script translates list of GO-nodes to R
• R-program implements training and classification
• Perl-scripts distribute cross validation on the Grid Engine
Annotating GO-Nodes

- 12625 probe-sets on Affymetrix HG-U95Av2
- 7115 probe-sets are annotated
- 6310 probe-sets are annotated several times, up to 23
- 2979 nodes have probe-set annotations below them
- 50 nodes have more than 100, up to 965 annotations
- 33 nodes have more than 10, up to 31 children
Expression Data from Leukemia Study

- Study on acute lymphoblastic leukemia (ALL) carried out at the St. Jude Children’s Research Hospital
  - 327 patients
  - 12625 genes (Affymetrix HG-U95Av2)
  - Various genetic subtypes of ALLs clinically confirmed
  - 269 patients with follow-up on relapse
- Gene expression values computed by average diff.
- Variance stabilisation and calibration
Performance Assessment - Recognizing Leukemia-Subtypes

- Leave-one-out cross validation for StAM, 10 fold cross validation for PAM
- Compare with St. Jude pretentions and plain PAM (nearest shrunken centroids)

![Graph showing Accuracy for different ALL translocations with bars for StAM, PAM, and St. Jude]
Nodewise Sensitivity and Specificity

Sensitivity vs. Specificity for E2A-PBX1

Sensitivity vs. Specificity for TEL-AML1
Nodewise Sensitivity and Specificity (continued)

Sensitivity vs. Specificity for hyper50

Sensitivity vs. Specificity for unspecific
Observation on Weights

- For prediction: few GO-nodes, sparse graphs:
- Overall 431 of 3835 edges connect 385 of 3180 GO-nodes

<table>
<thead>
<tr>
<th></th>
<th>BCR-ABL</th>
<th>E2A-PBX1</th>
<th>MLL</th>
<th>T-ALL</th>
<th>TEL-AML1yperdiploid</th>
<th>Unspecific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>35</td>
<td>182</td>
<td>117</td>
<td>321</td>
<td>244</td>
<td>156</td>
</tr>
<tr>
<td>Edges</td>
<td>36</td>
<td>201</td>
<td>128</td>
<td>360</td>
<td>270</td>
<td>170</td>
</tr>
<tr>
<td>Genes</td>
<td>3688</td>
<td>5543</td>
<td>4997</td>
<td>6109</td>
<td>5862</td>
<td>5470</td>
</tr>
</tbody>
</table>

Claudio Lottaz: StAM – Structured Analysis of Microarray Data
Thinned Graph for BCR-ABL
"Thinned" Graph for TEL-AML1

- Projects/StAM/R/
Observations on Nodewise Prediction

- When investigating class $C$:
- From cross validation results, select all samples predicted for class $C$
- Select all nodes used for classification
- Cluster samples hierarchically
- Can we find differences/groups among samples sharing the same prediction?
Predicted TEL-AML1-Probability
Discussion

• Summary
  • Competitive performance on simple problem
  • Small graphs are used for prediction
  • Alternative features leading to same prediction – not yet confirmed

• Future work
  • Fine-tuning on difficult problems (weighting)
  • Improve methods to find interesting nodes
  • Does all this mean something biologically?
  • Investigate other means to structure the data