Analysis of Temporal Gene Expression Data of Astrocyte Differentiation

In order of appearance:
Meghan Dierks, Gene Yeo,
Alex Rakhlin, Melissa Kosinski,
Katie Steece
Biological System

- Multi-potent neuronal cells treated with CNTF differentiate into astrocytes.
- A number of different pathways are activated upon stimulation with CNTF.
- However, very limited prior knowledge about the early transcription and translation events underlying astrocyte development.
Goal

- Identify genes or gene classes involved in early differentiation of neuronal stem cells
- Characterize the temporal relationships in gene expression during this critical period
The data (source: John Park, MD, PhD)

- Rat Genome on 3 parts: A, B and C
- Four time points: 0, 45, 90, 180min post tx with CNTF
- Two sets of experiments
A, P, M calls

- Absent or unreliable??? Affy calculates decision boundaries empirically...
- Use ‘em or not?
- Use raw data?
- Negative – floor them. Also have large negative for P-calls...
- Idea #1: PPPP
- Idea #2: AP**
Normalizing the data: ML

Working on log of data
GAPDH normalization

- Didn’t seem to help
ALL P’s in both experiments
2-fold in both experiments

Reference: Butte, et al
Same trend over two experiments.

Final working set: 123 genes
Clustering Time Series Data

- Clustering Problems
- Current Clustering Solutions
- Feature Vectors
- Similarity measures
- Gene Relationships
- What we did...
Clustering Problems

- Hierarchical clustering (Eisen, 1998; Alon, 1999)
  - Problems with robustness, uniqueness and optimality of linear ordering
- Cost function optimization (Tamayo, 1999)
  - No guarantee that solution converges to global optimum
- Optimum number of clusters?
  - Hierarchical clustering – observer dictates number of clusters from dendrogram
  - Cost function – num of clusters is an external parameter
Current Clustering Solutions

- Clustering methods
  - Clustering by simulated annealing (Lukashin, 2001)
    - Guarantees global optimality
  - Gene shaving (Hastie, Brown & Botstein, 2000)
    - Genes may belong to more than one cluster, can be unsupervised or supervised
    - Not set up for time series, but is not a problem

- Optimum number of clusters – depends primarily on the variation between profiles within given datasets
  - Expected distribution of profiles over clusters (Lukashin, 2001)

- Optimum num genes per cluster:
  - Gap statistic (Brown & Botstein, 2000)
Feature Vectors

- Normalized gene expression values (e.g., values 0 to 1)
- Augmented vectors: normalized time series augmented with different values between time points – emphasized similarity between closely parallel but offset expression pattern
Similarity Measures

- Geometric Distances
- Standard correlation coefficients (dot product of two normalized vectors)
  Eisen et al PNAS [95] 14863-14868, 1998
Gene Relationships

- Linear Correlation, Rank Correlation & Information Theory to determine significant relationships (Somogyi, 1998)
What we Did...

- Feature Vector
  - Sign($X_{t+1} - X_t$) -> -1, 1 binary values
- Similarity Measure
  - Hamming Distance
- Gene Relationship
  - Ranked Correlation coefficients
Number of Clusters
Cluster I

Clustered by Binarized Trend Similarity: Hamming Distance
Cluster II

Clustered by Binarized Trend Similarity: Hamming Distance
Cluster III

Clustered by Binarized Trend Similarity: Hamming Distance
Cluster IV

Clustered by Binarized Trend Similarity: Hamming Distance

[Graphs showing trend similarity over four categories]
Cluster V

Clustered by Binarized Trend Similarity: Hamming Distance
Correlation Coeff Exp 1
Open Questions

- How do we incorporate biological prior knowledge into choice of:
  - Similarity measure
  - Representation of feature vectors
  - Number of functional clusters
  - Clustering algorithm
Histogram of Corr Coeff Exp2
Differences between Corr Coeffs
Functional Analysis

- Find GenBank UniGene identity
  - If true gene, keep data
  - If EST keep only if >50% homology to known
- Determine conserved domains
- Assign functional relevance to domains
- Compare to random 121 gene sample
- Group genes by probable biological function
## Functional Classes

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Cluster 5</th>
<th>Random Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample Size</strong></td>
<td>49</td>
<td>44</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>121</td>
</tr>
<tr>
<td><strong>Transcription/Translation Machinery</strong></td>
<td>2%</td>
<td>9%</td>
<td>11%</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td><strong>Membrane Transport</strong></td>
<td>0%</td>
<td>5%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td><strong>Immunological Response Protein</strong></td>
<td>0%</td>
<td>5%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td><strong>Protein Kinase</strong></td>
<td>4%</td>
<td>0%</td>
<td>11%</td>
<td>0%</td>
<td>0%</td>
<td>3%</td>
</tr>
<tr>
<td><strong>Transcription Factors</strong></td>
<td>4%</td>
<td>11%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Cell Adhesion/Structure</strong></td>
<td>2%</td>
<td>7%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>3%</td>
</tr>
<tr>
<td><strong>Membrane Ion Channels</strong></td>
<td>0%</td>
<td>2%</td>
<td>0%</td>
<td>33%</td>
<td>0%</td>
<td>&lt;1%</td>
</tr>
<tr>
<td><strong>Cell Cycle Proteins</strong></td>
<td>2%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>&lt;1%</td>
</tr>
<tr>
<td><strong>Signal Transduction</strong></td>
<td>4%</td>
<td>5%</td>
<td>11%</td>
<td>33%</td>
<td>0%</td>
<td>7%</td>
</tr>
<tr>
<td><strong>Metabolism</strong></td>
<td>4%</td>
<td>7%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>9%</td>
</tr>
<tr>
<td><strong>Lipid Binding Protein</strong></td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>&lt;1%</td>
</tr>
<tr>
<td><strong>Protein Folding/Degradation</strong></td>
<td>6%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>4%</td>
</tr>
<tr>
<td><strong>DNA Replication/Repair</strong></td>
<td>4%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td><strong>Protein Kinase Inhibitors</strong></td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td><strong>Protein Trafficking/PostTranslational Modification</strong></td>
<td>0%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>3%</td>
</tr>
<tr>
<td><strong>Other Unidentified Protein Domains</strong></td>
<td>2%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>EST</strong></td>
<td>65%</td>
<td>41%</td>
<td>67%</td>
<td>33%</td>
<td>50%</td>
<td>54%</td>
</tr>
</tbody>
</table>
Cluster 1 Significance

- More in Folding/degradation Proteins
  - Not needed early in differentiation
  - Reactivated after differentiation to regulate protein activity
- Fewer “housekeeping genes”
- More ESTs
  - Mature astrocytes have not been well characterized
  - More unknown genes being activated
Cluster 2 Significance

- More “housekeeping genes”
  - Inactivated early, reactivated after differentiation
  - Diverting resources to differentiation
- More Transcription factors
  - Proteins regulating “housekeeping genes” may be inactivated and then reactivated after differentiation is established
- More Transcriptional/Translational Machinery proteins
Correlation Relationships?

Epidermal Growth Factor Receptor (oncogene)
Polypyrimidine Tract Binding Protein 1
Correlation Relationships

Predicted Zinc Finger Motif Transcription Factors
Future Directions

- Retrieve more time points
- Sort genes by location, function, and pathway
- Perform true “before and after” experiment with more than 2 day time lapse
  - Determine the overall difference in gene expression levels
  - Determine which genes are needed in each stage
- Need a larger sample set to observe genes that are turned on early in differentiation (i.e. Cluster 4 and Cluster 5)
Conclusions

- Essential to establish quality of data –
  - Internal consistency measures
- Data are *very* sensitive to normalization technique...choose cautiously
- When limited with respect to number of trials, use combination of quantitative and qualitative (sequence analysis, domain class, etc.) techniques to characterize and classify
References (to pick a few)

- GeneChip Expression Analysis Algorithm Tutorial.