Clustering and Functional Analysis of Coordinately Regulated Genes

Gavin Sherlock

sherlock@genome.stanford.edu

February 11th 2010
Clustering Coordinately Regulated Genes

• What are the goals of typical expression experiments?
• How can we determine if two genes are ‘coexpressed’?
• What can we infer when we decide that two genes are coexpressed?
Visualizing Data
Extracting Data

<table>
<thead>
<tr>
<th>Cy3</th>
<th>Cy5</th>
<th>$\frac{\text{Cy5}}{\text{Cy3}}$</th>
<th>$\log_2 \left( \frac{\text{Cy5}}{\text{Cy3}} \right)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>10000</td>
<td>50.00</td>
<td>5.64</td>
</tr>
<tr>
<td>4800</td>
<td>4800</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>9000</td>
<td>300</td>
<td>0.03</td>
<td>-4.91</td>
</tr>
</tbody>
</table>
Visualizing Data (cont.)

Expression During Sporulation

Time (hours)

Log Ratio
In microarray studies, we often use clustering algorithms to help us identify patterns in complex data.

For example, we can randomize the data used to represent this painting and see if clustering will help us visualize the pattern.
Clustering algorithms

First, we represent the painting in black and white.
Clustering algorithms

The painting is “sliced” into rows which are then randomized.
Clustering algorithms

Rows ordered by hierarchical clustering with nodes flipped to optimize ordering
Clustering algorithms

Rows ordered by using a Self-Organizing Map (SOM)
Random vs. Biological Data

Goal of Clustering
Types of Clustering

- **Agglomerative**
  - Bottom up approach
  - Different variants of hierarchical clustering
  - This is the typical clustering you see

- **Partitioning / Divisive**
  - Top down approach
  - K-means Clustering
  - Self-Organizing Maps

- All require the ability to compare expression patterns to each other.
How do we compare expression profiles?

- Treat expression data for a gene as a multidimensional vector.

- Use a distance/correlation metric to compare the vectors.
Expression Vectors

- Crucial concept for understanding clustering

- Each gene is represented by a vector where coordinates are its values - log(ratio) - in each experiment

  - $x = \log(\text{ratio})_{\text{expt}1}$
  - $y = \log(\text{ratio})_{\text{expt}2}$
  - $z = \log(\text{ratio})_{\text{expt}3}$
  - etc.

  Similar expression
Distance metrics

- Distances or correlations are measured “between” expression vectors

- Many different ways to measure distance:
  - Euclidean distance
  - Pearson correlation coefficient(s)
  - Spearman’s Rank Correlation
  - Manhattan distance
  - Mutual information
  - Kendall’s Tau
  - etc.

- Each has different properties and can reveal different features of the data
Euclidean distance

- Euclidean distance metrics detect similar vectors by identifying those that are closest in space. In this example, Gene A and C are closest.
Pearson correlation

- The Pearson correlation disregards the magnitude of the vectors but instead compares their directions. In this example, Gene A and Gene B have the same slope, so would be most similar to each other.
Agglomerative Hierarchical Clustering

1. Compare all expression patterns to each other.
2. Join patterns that are the most similar out of all patterns.
3. Compare all joined and unjoined patterns.
4. Go to step 2, and repeat until all patterns are joined.

Need a rule to decide how to compare clusters to each other
Visualization of Hierarchical Clustering
Single linkage Clustering

Nearest Neighbor

This method produces long chains which form straggly clusters.
Complete Linkage Clustering

Uses the Furthest Neighbor

This method tends to produce very tight clusters of similar patterns.
Average Linkage Clustering

The red and blue ‘+’ signs mark the centroids of the two clusters.

Average (only shown for two cases)
The red and blue ‘+’ signs mark the centroids of the two clusters.
And we get a cluster:

- Single
- Complete
- Average
- Centroid
Two-way clustering

- Just as gene patterns are clustered, array patterns can be clustered.
- All the data points for an array can be used to construct a vector for that array and the vectors of multiple arrays can be compared.
Two-way Clustering

Two-way clustering can help show which samples are most similar, as well as which genes.

Proliferation Cluster
Agglomerative Hierarchical Clustering

Advantages:
• Simple
• Easy to implement
• Easy to visualize

Disadvantages:
• Can lead to artifacts
• Discarding of subtleties in 2-way clustering
Partitioning Methods

- Split data up into smaller, more homogenous sets
- Should avoid artifacts associated with incorrectly joining dissimilar vectors
- Can cluster each partition independently of others, by genes and arrays
- Self-Organizing Maps and k-means clustering are two possible partitioning methods
Self Organizing Maps

- Create a ‘Map’ of ‘n’ partitions, that is modeled on the expression data, where each partition in the map has an associated vector.

- Genes’ expression vectors are assigned to the partition with the most similar associated vector.

- Neighboring partitions are more similar to each other than they are to distant partitions.
The Map Is Disorganized

Repeat 100,000 times
K-means Clustering

- Split data into ‘n’ partitions, each with an associated vector.
- Assign genes to partitions, and recalculate the vector associated with each partition as the centroid of its associated genes.
- Repeat until solution converges, or for a fixed number of iterations.
Divisive Hierarchical Clustering

• Iteratively use k-means clustering, with k set to 2.

• Successively divide data into smaller and smaller subsets.

• Allows you to build a tree describing how the data were successively split, similarly to agglomerative hierarchical clustering.
Agglomerative vs. Divisive

Agglomerative:
Agglomerative vs. Divisive

Agglomerative
Bottom-Up

Divisive
Top-Down

Hybrid

Chipman and Tibshirani, 2006
Summary For Clustering

• Many different methods exist for finding groups and patterns in data (including some I haven’t mentioned).
• Many different parameters can be used in those methods.
• Caution should be exercised in interpreting the results.
Comparing Different Clustering Methods

Which technique is right?

• Hierarchical clustering?
  – Single, Average, Complete, Centroid linkage, etc.?

• Self Organizing Maps

• K-means clustering

• Other algorithms?
What is a ‘cluster’?

– And how do we know if it’s any good, or if one technique for producing clusters is better than another?

• Rather than think simply of clustering, think of all these methods as capable of producing groups of genes:
One cluster to two groups of genes
One cluster to three groups of genes
Now what?

• Try many methods, and demand they each produce the same number of groups of genes.
• Is there a metric that says which did best for a given number of groups?
• Can we come up with a metric for the best number of groups?
What do we think that co-expression means?

- Our general assumption is guilt by association:
  i.e. genes with similar expression patterns are more likely to participate in the same biological process.
- Therefore, we can exploit the Gene Ontology to assess our clusters:
How do we measure how ‘good’ the annotation is?

• Use a score that measures how coherent the level of annotation is compared to what would be expected from random clusters.
  – Developed system, such that the higher the score, the better the annotation fit the clustering.
Figure 2. Four data sets clustered using $k$-means, hierarchical, and self-organized map algorithms. The horizontal axis shows the number of clusters desired, and the vertical axis shows $z$-scores. Data sets are (a) Cho, (b) CJRR, (c) Gasch, and (d) Spellman.

Gibbons F. D., Roth F. P. Genome Res. 2002;12:1574-1581
Characterization of clusters

- Now we have groups of genes that best fit their annotation, find the best annotation(s) that fits those groups.
- Calculate P-values for each GO term’s association to a cluster, and choose those that are most significant.
Using the Gene Ontology to assess clusters

• Many microarray analyses result in a list of interesting genes
• Typically biologists can make up a story about any random list
• So, look at all GO annotations for the genes in a list, and see if the number of annotations for any GO node is significant
The Categories of GO
(The Gene Ontology)

- **Biological Process** = goal or objective \hspace{1cm} (Why)
  (e.g. DNA replication, Cell Cycle Control, Cell adhesion)

- **Molecular Function** = elemental activity/task \hspace{1cm} (What)
  (e.g. Transcription factor, polymerase, protein kinase)

- **Cellular Component** = location or complex \hspace{1cm} (Where)
  (e.g. pre-replication complex, kinetochore, membrane)

Each Category is a structured, controlled vocabulary
Parent-Child Relationships

A child is a subset of a parent’s elements

The cell component term Nucleus has 5 children
Determining P-values for GO annotation for a list of genes

We can calculate the probability of having \( x \) of \( n \) genes having an annotation to a GO node, given that in the genome, \( M \) of \( N \) genes have that annotation, using the hypergeometric distribution, as:

\[
p = \binom{M}{x} \binom{N - M}{n - x} \binom{N}{n}
\]
Determining GO significance

To calculate a P-value, we calculate the probability of having \textit{at least} $x$ of $n$ annotations:

$$ \text{P-value} = 1 - \sum_{i=0}^{x-1} \binom{M}{i} \binom{N-M}{n-i} \binom{N}{i} $$

Then do multiple hypothesis correction on the p-values
Methionine Cluster

YPL250C
MET11
YER042W
YLR302C
YPL274W
MET28
YGL184C
YLL061W
MET1
YNL276C
YIL074C
YLL062C
MET14
MET16
MET3
MET10
ECM17
YNL276C
MUP1
MET17
MET6
GO Annotations

- sulfur metabolic process: 2.43e-19 (12/18 vs 66/6608)
- methionine metabolic process: 1.40e-14 (10/18 vs 24/6608)
Recommended reading : Clustering


Recommended reading for Cluster Validation / Analysis