Clustering Methods for Gene Expression Data

Katherine S. Pollard
Center for Biomolecular Science & Engineering
University of California, Santa Cruz

November 15, 2003

©Copyright 2003, all rights reserved
Outlines

- Background.

- Example Data Analyses.

- Clustering:
  - Distance Metrics,
  - Algorithms,
  - Output,
  - How Many Clusters?

- Inference & Variability.

- Visualization.

- Applications.
What is Clustering?

Exploratory data analysis methods for discovering patterns:

- Grouping,
- Dimension reduction,
- Ordering.

Clustering methods have the following components:

1. **Distance** measure.

2. **Parameters** of interest, *e.g.*
   - Number of clusters,
   - Cluster labels (partitioning),
   - Fuzzy clustering memberships,
   - Hierarchical tree.

3. **Algorithm** mapping distance into parameters.
Gene Expression Data

Each array produces a vector of $p$ (relative) gene expression measurements.

Typically, we observe data from $n \ll p$ arrays, which might represent different

- experimental conditions,
- subjects,
- time points.

Example: Tumor vs. healthy tissues of $n$ cancer patients.
Applications of Clustering to Gene Expression Studies

Clustering methods can be used to identify:

- Groups of genes similarly expressed across the arrays (co-regulation).

- Groups of arrays with similar expression profiles (sub-populations).

- Groups of genes and arrays simultaneously:
  - Cluster arrays within each gene cluster.
  - Find important genes for determining array clusters.

- Orderings of the genes or arrays based on expression.

- The number $k$ of distinct groups of genes.

- True dimension of the data, e.g. $k$ vs. $p$. 
Example Analysis: NCI60 Data

1. **Pre-processing**: image analysis, normalization, etc.

2. Focus on 3 types of cancer: colon, leukemia, melanoma.

3. Subset genes *(multtest)*.

4. Choose **algorithm** *(HOPACH)* and **distance** metric(s):
   - Genes = cosine-angle (uncentered correlation),
   - Arrays = Euclidean.

5. Cluster genes and select **main clusters** *(with MSS)*.

6. Cluster arrays using only the genes in each gene cluster.

7. Plot **heat maps** for each gene cluster with genes and arrays ordered (final level of hierarchical trees):
   - Red = over expressed vs. pooled control,
   - White = under expressed vs. pooled control.
Simultaneous Clustering of NCI60 Data

Hierarchical clustering (CLUSTER) of genes and arrays visualized in TREEVIEW.

Clustering of arrays reveals subpopulations of interest.

(a) $6 \times 5$ Self-Organizing Map, (b) cluster 29 detail, (c) centroids of clusters 29, 14, 1 and 5 correspond with G1, S, G2, and M phases, (d) centroids of all clusters with peaks in one of these phases.
Distance Metrics

Let $D_{ij}$ denote the dissimilarity between genes (arrays) $i$ and $j$, where each gene (array) is represented by an $n$ ($p$) dimensional vector. Possible dissimilarities are:

$$D_{ij} = 1 - \rho_{ij} \text{ correlation},$$

$$D_{ij} = 1 - \rho_{ij}^0 \text{ cosine-angle or uncentered correlation},$$

$$D_{ij} = \sum_{l=1}^{n} (Y_{il} - Y_{jl})^2 \text{ Euclidean},$$

where

$$\rho_{ij}^0 = \frac{\sum_{l=1}^{n} Y_{il} Y_{jl}}{\sqrt{\sum_{l=1}^{n} Y_{il}^2} \sqrt{\sum_{l=1}^{n} Y_{jl}^2}}.$$

NOTE: The absolute value of each can also be used.
Comparing Distance Metrics

Each distance metric quantifies a different notion of what it means for two vectors to be close to each other.

Two vectors are close if pairs of values in the two vectors have:

- **Correlation**: similar patterns of ups and downs,
- **Euclidean**: similar magnitudes,
- **Cosine-Angle**: similar patterns and magnitudes,
- **Absolute Correlation**: similar patterns of ups or downs,
- **Absolute Euclidean**: similar absolute magnitudes,
- **Absolute Cosine-Angle**: similar absolute patterns and magnitudes.

**NOTE**: $1 - \rho_{ij}^0$ equals 0.5 times the squared Euclidean distance of the two vectors standardized to have Euclidean norm 1.
Same Mean, Uncorrelated

Cor=1.1
Cos–Angle=1.1
Euclidean=1.5
Abs Cor=0.9
Perfectly Anti-Correlated

Cor=1.4
Cos–Angle=1.4
Euclidean=2.0
Abs Cor=0.0
Same Mean, No Variation

Cor=1.0
Cos–Angle=0.98
Euclidean=1.0
Abs Cor=1.0
How to Measure Distance Between Clusters?

- **Complete (minimum)**

- **Single (maximum)**

- **Average**
Algorithmic Approaches

- **Supervised** (COBWEB, SVMs, CART, gene-shaving, transformations) vs. **Unsupervised**.

- **Model-based** (AUTOCLASS, SNOB) vs. **Non-parametric**.

- **Partitioning** (SOMs, PAM, MASLOC, KMEANS) vs. **Hierarchical**.
  - Agglomerative (single, complete, and average linkage CLUSTER, AGNES),
    * bottom up
  - Divisive (SOTA, DIANA, TSVQ),
    * top down
  - Hybrid (HOPACH).

- **Graphical** approaches (CAST).
Nice Properties for Clustering Algorithms

- Use any distance metric.
- Perform well in the presence of noise.
- Identify patterns of biological interest.

Partitioning Methods.
- Small (or different sized) clusters,
- Overlapping clusters,
- Robust cluster profiles.

Hierarchical Methods.
- Sensible ordering.
**Clustering Output: Parameters of Interest**

- **Number of clusters.**
  - Data adaptive methods,
  - Global versus more detailed structure.

- **Partitioning Methods.**
  - Cluster labels,
  - Cluster profiles (means, medoids),
  - Fuzzy clustering memberships.

- **Hierarchical Methods.**
  - Tree structure,
  - Main clusters (pruning),
  - Final ordering.
Choosing the Number of Clusters

• **Global Criteria:**
  1. 30 methods reviewed by Milligan & Cooper, 1985.
  3. Graph theory (*e.g.*: cliques in CAST) (Ben-Dor *et al.*, 1999).

• **Resampling Methods:**
  2. WADP (Bittner *et al.*, 2000).
Inference: Reliability and Repeatability

We would like to know the variability of clustering results based on an observed data set.

- Clustering parameters can be viewed just like any other parameter (e.g.: a mean) in statistics.

- The output of a clustering algorithm applied to a data set is an estimate of the underlying (true) parameter.

- There will not, however, generally be a closed formula for the variance of the parameter estimate.

Resampling methods (e.g.: parametric or non-parametric bootstrap) can be used to estimate the reliability and repeatability of clustering results.
Inference with the Bootstrap

**Idea:** Use the computer to simulate the process of repeating the experiment many times in the lab.

1. The observed data is used to generate many similar resampled data sets (with variation),
2. The clustering algorithm is applied to each of these,
3. The variability of the clustering output over the resampled data sets estimates the variability of the original output.

**Applications:**

- Number of clusters: select most stable number,
- Variability of cluster memberships,
- Fuzzy clustering: overlapping clusters,
- Pairwise reappearance probabilities.
Visualization

• Value of criteria function vs. number of clusters.

• Heat maps:
  – reordered data matrices,
  – reordered distance matrices.

• Fuzzy Clustering:
  – membership of each element in each cluster,
  – bootstrap membership probabilities.

• Biochemical pathways.
MSS Plots

MSS versus number of clusters for eight values of sigma
n=360, mu=(1,2,5,6,14,15,18,19)

Plot the history of MSS as the tree unfolds.
Maple Tree (Lisa Simirenko, Eisen Lab)

1. Hierarchical view of the gene and array trees with heat map,

2. Fuzzy clustering view of bootstrap estimated membership probabilities for each gene (array) and each of the clusters.
Bootstrapping Clusplot

Reappearance proportions and cluster reproducibility

Subset contains 4373 genes.

Bar plot of bootstrap estimated membership probabilities.
Reordered distance matrices reveal the underlying cluster structure and the relationship between distance in gene expression and distance in the final ordered list (leaks of the tree).
GenMAPP Pathway (Conklin Lab)

Fatty Acid Degradation

- Triacylglycerol
- Lipoprotein Lipase
- Glycerol
- Dihydroxyacetone Phosphate
- Fatty Acid
- L-Glycerol-3-Phosphate
- Glycerol Kinase
- Glycerol-3-PO4 Dehydrogenase
- Dihydroxyacetone Phosphate
- Triosephosphate Isomerase
- Glyceraldehyde-3-Phosphate
- See Glycolysis MAPP
- Fatty Acid CoA Ligase 2
- Fatty Acid CoA Ligase 4
- Acyl-CoA
- Carnitine Acyltransferase
- Carnitine Palmitoyltransferase I
- Carnitine Palmitoyltransferase II
- Acyl-CoA Dehydrogenases:
  - Short Chain
  - Medium Chain
  - Long Chain
  - Very Long Chain
- 3-OH-acyl-CoA
- Enoyl-CoA Hydrolase
- 3-L-Hydroxyacyl-CoA
- 3-Ketoacyl-CoA
- 3-Hydroxyacyl-CoA

Gene Database
Mm-Std_20030922.gdb
Expression Dataset
Name: HOPACH Clustering 44 grps genes affy
Color Set: Main Clusters
Gene Value: [None]
Legend
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
No criteria met
Not found

Author: Charles H. Redfern, Nathan Salomonis
E-mail: nsalomonis@gladstone.ucsf.edu
Last modified: 2/12/02
Copyright © 2000, Gladstone Institutes
Other Applications in Computational Biology

Clustering algorithms can be applied to many problems:

- Association between gene expression and an outcome (supervised clustering),
- Other array data (e.g.: CGH, oligo tilings),
- Proteomics,
- Functional genomics (annotation),
- Protein families (based on sequence similarity),
- Haplotype block identification,
- Physical location of transcription factor binding sites in a genome,
- Relating gene expression to chromosome location,
- Phylogenetics and phylogenomics.
Summary: Things to Keep in Mind

• Clustering results can be sensitive to the subset of genes (or arrays) used in the analysis.

• What is the biological question?
  – Choose a distance metric that matches your idea of “close”,
  – Choose a clustering algorithm that returns the type of output you seek (e.g.: partitioning for hard labels vs. hierarchical for orderings and trees).

• Clustering algorithms will always produce some output:
  – Assess the variability of your results,
  – Assess the biological relevance of your results,
  – Confirm hypotheses with further experiments.