CSE 527: Computational Biology
Fall 2004, Professor Larry Ruzzo
Notes by Angela Collie, 10/11

Lecture 4: Clustering Expression Data

I. Overview: Clustering Expression Data

Definition: Clustering is a data exploratory tool. It is an "unsupervised" approach to organizing expression data that allows distinct groupings (clusters) of similar data points. Expression data can be clustered according to samples or genes.

Why cluster?

- Tissue classification
- Finding biologically relevant genes
- First step in inferring regulatory networks
    - For example: the expression of genes in a pathway may be coordinated, so these genes would have similar expression data.
- Identifying common promoter elements
- Hypothesis generation
    - Clustering is not hypothesis testing!!! It is possible to say that the clusters are well-separated, coherent groups, but there is *no* p value or statistical measure to say that these groups fit a preconceived or null hypothesis.
- Clustering is one of the tools of expression analysis. There are others.

What clustering methods have been used?

Clustering is not a new concept and was often used in the 1950s for "numerical taxonomy" with the increase of digital data and digital computing techniques. Clustering experienced an increase in use in the 1990s with the generation of large amounts of data, such as the web and gene expression data. There are several clustering methods:

- Hierarchical clustering (Eisen '98)
  Eisen MB et al. Proc Natl Acad Sci. 1998; 95(25):14863-8.
  http://www.pnas.org/cgi/content/full/95/25/14863

- Self-Organizing Maps (SOM) (Tamayo '99)
  Tamayo P et al. Proc Natl Acad Sci. 1999; 96(6):2907-12.
  .http://www.pnas.org/cgi/content/full/96/6/2907

- CAST (Ben-Dor '99)
  Ben-Dor A. J Comput Biol. 1999;6(3-4):281-97.

- Support Vector Machines (SVM) (Grundy '00)

Brown MPS. Proc Natl Acad Sci. 2000;97(1):262-7.
http://www.pnas.org/cgi/content/full/97/1/262

- Etc.

Why are there so many clustering methods?

- Expression data is essentially an NP-hard problem, meaning that an exact solution is probably impossible since the problem is so complex.
- There is high dimensionality and high noise in array data.
- There is no best method. The choice of method also depends on the questions the experimenter wants to answer.

II. Clustering Algorithms Overview

- Partitional – Data is initially organized into one big cluster and is then separated into smaller clusters.
  - k-means (Hartigan '75)

    Tavazoie S Nat Genet. 1999; 22(3):281-5.
    http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v22/n3/full/ng0799_281.html

- Hierarchical Agglomerative – This is a bottom-up merge algorithm in which each data point is initially a cluster. Clusters are then joined together, until a specified endpoint is reached. That endpoint can be based on the similarity of members in a cluster or on the desired final number of clusters. Variations on the joining of clusters include:
  - Single-link
  - Average-link
  - Complete-link
- Random – The random assignment of data points to clusters can be used as a control to determine the reasonableness of a clustered data set.

III. Clustering 101

Ref: http://faculty.washington.edu/kayee/research.html

a. Defining Similarity

Similarity metric: a measure of pairwise similarity or dissimilarity

1. *Correlation coefficient*:

$$\frac{\sum_{j=1}^{p}(X[j]-\overline{X})(Y[j]-\overline{Y})}{\sqrt{\sum_{j=1}^{p}(X[j]-\overline{X})^2\sum_{j=1}^{p}(Y[j]-\overline{Y})^2}}, \quad where\ \overline{X}=\frac{\sum_{j=1}^{p}X[j]}{p}$$

The solution is a value between -1 and 1. If the correlation coefficient is -1, the pair is anti-correlated. 0 means the pair has no correlation. 1 means the pair is perfectly

correlated. This measure examines movement or parallels or patterns in the direction of the data. The minimum number of attributes necessary to compute this measurement is 3.

2. *Euclidian distance*

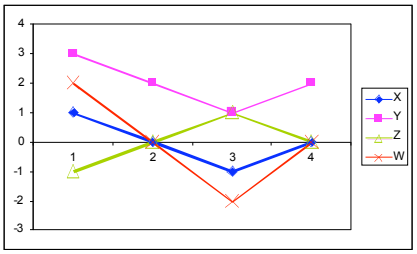
$$\sqrt{\sum_{j=1}^{p}(X[j]-Y[j])^2}$$

This measurement examines magnitude (absolute expression) and direction. The minimum number of attributes necessary to compute this measurement is 2.

b. Similarity Example

Example

X	1	0	-1	0
Y	3	2	1	2
Z	-1	0	1	0
W	2	0	-2	0



Correlation (X,Y) = 1 Distance (X,Y) = 4

Correlation (X,Z) = -1 Distance (X,Z) = 2.83

Correlation (X,W) = 1 Distance (X,W) = 1.41

c. Clustering Algorithms Overview

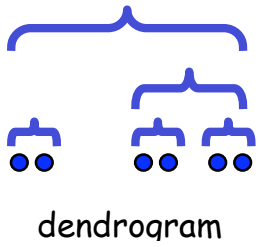
Inputs for these algorithms:

- Raw data matrix or similarity matrix
- Number of clusters or other endpoint parameter

Different classifications of algorithms

- Hierarchical vs. Partitional
- Heuristic-based vs. model-based
- Soft (fuzzy, partial membership) vs. hard (must define membership for every point)

d. Hierarchical Agglomerative Clustering



This is an *agglomerative, bottom-up* approach.

Algorithm:

Initialize: Each item is a cluster

Iterate: Merge two most similar clusters

Halt: When required number of clusters is reached

Types/Variations (See slide deck for examples)

i. ***Single-link****:* Cluster similarity is the ***similarity of the two most similar members***. This method may result in long and skinny clusters, so it is able to pull out long shapes of dense data, such as a spiral galaxy or written letters. It is computationally fast.

ii. ***Complete-link***: Cluster similarity is the ***similarity of the two least similar members***. This method may result in tight, round clusters. It is computationally slow.

iii. ***Average-link****:* Cluster similarity is the ***average similarity of all pairs***. This method may result in tight clusters. It is computationally slow.

iv. ***Centroid Link*** (also sometimes called average link): The cluster ***centroid is the average of all the points in the cluster***. ***Cluster similarity is the distance between the centroids***. This method discards the shape and orientation of a cluster. It is computationally faster since there are less distances to compute.

Software

Treeview

http://rana.lbl.gov/EisenSoftware.htm

Early Eisen paper on hierarchical clustering:
Eisen MB et al. Proc Natl Acad Sci. 1998; 95(25):14863-8.
http://www.pnas.org/cgi/content/full/95/25/14863?ijkey=ead7baa5e57dabccf438ae94e920c58b10fa7773

Serum stimulation of fibroblasts:
Iyer VR et al. Science. 1999; 283: 83-87.
http://www.sciencemag.org/cgi/content/full/283/5398/83

Note: Ordering of genes within the dendrogram

TMeV: TIGR Multiexperiment Viewer

http://www.tigr.org/software/tm4

e. Hierarchical Divisive Clustering Algorithms

These top down algorithms start with all the objects in one cluster and then split the cluster(s) into smaller clusters. This is less efficient than agglomerative clustering. An example of software

that uses this algorithm (and others) is Rosetta Resolver (http://www.rosettabio.com/products/resolver/default.htm), which used a deterministic annealing approach, meaning that some decisions could be reconsidered. (Alon, '99)

Alon U et al. Proc Natl Acad Sci. 1999;96(12):6745-50.
http://www.pnas.org/cgi/content/full/96/12/6745

f. Partitional – K-Means

Ref: (MacQueen 1965)
Initialize: Select location of k number of centroids
Iterate: Assign each datapoint to the closest centroid
Compute new centroid
Minimize:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \left(x - Centroid(C_i)\right)^2$$

Halt: Convergence

This method is fast and tends to result in spherical, equal-sized clusters. It will converge to a *local* minimum, not necessarily the global minimum. It could be run multiple times with different starting locations for the centroids. It is related to a model-based approach (see next lecture).

g. Self-organizing Maps (SOMs)

Ref: (Kohonen 1995)

The basic idea is to map high dimensional data onto a 2-dimensional map, so that neighboring nodes are more similar than those farther away. A grid of nodes is created, and input vectors that are close to each other are mapped to the same or neighboring nodes. A feature of SOMs are that they update incrementally, reassigning nodes with the addition of each point. This nudges the centroid belonging to the point as well as the neighboring nodes.

Properties of SOMs:

- Specify partial structure (grid)
- Easy visualization
- Many tunable parameters (such as nudging amount)
- Sensitive to parameters