

Clustering Gene Expression Data

BMI/CS 776

www.biostat.wisc.edu/~craven/776.html

Mark Craven

craven@biostat.wisc.edu

April 2002

Announcements

- milestone #2 for project due next Monday: description of your experiments
 - how you will test your hypotheses
 - data to be used
 - what will be varied (algorithm, parameter of alg, etc.)
 - methodology
- reading for next week
 - Brazma et al., Predicting Gene Regulatory Elements in Silico on a Genomic Scale, *Genome Research* 1998

Clustering Gene Expression Profiles

- given: expression profiles for a set of genes or experiments/patients (whatever columns represent)
- do: organize profiles into clusters such that
 - instances in the same cluster are highly similar to each other
 - instances from different clusters have low similarity to each other

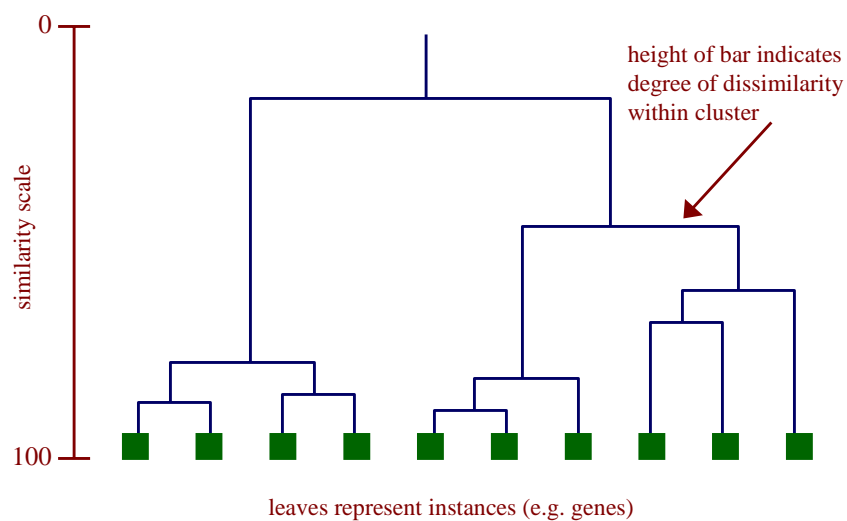
Motivation for Clustering

- *exploratory data analysis*
 - understanding general characteristics of data
 - visualizing data
- generalization
 - infer something about an instance (e.g. a gene) based on how it relates to other instances

The Clustering Landscape

- there are many different clustering algorithms
- they differ along several dimensions
 - hierarchical vs. partitional
 - hard vs. soft clusters
 - disjunctive (an instance can belong to multiple clusters) vs. non-disjunctive
 - deterministic (same clusters produced every time for a given data set) vs. stochastic
 - distance (similarity) measure used

Hierarchical Clustering: A Dendrogram



Distance (Similarity) Matrix

- based on the distance/similarity measure we can construct a symmetric matrix of pairwise distances
- (i, j) entry in the matrix is the distance (similarity) between instances i and j

| | I_1 | I_2 | \dots | I_n |
|-------|----------|----------|---------|----------|
| I_1 | • | d_{12} | \dots | d_{1n} |
| I_2 | d_{21} | • | \dots | d_{2n} |
| I_n | d_{n1} | d_{n2} | \dots | • |

Note that $d_{ij} = d_{ji}$ (i.e., the matrix is symmetric). So, we only need the lower triangle part of the matrix.

The diagonal is all 1's (similarity) or all 0's (distance)

Bottom-Up Hierarchical Clustering

given : a set $X = \{x_1 \dots x_n\}$ of instances

for $i := 1$ to n do

$c_i := \{x_i\}$ /* each object is initially its own cluster */

$C := \{c_1 \dots c_n\}$

$j := n + 1$

while $|C| > 1$

$(c_a, c_b) := \underset{(c_u, c_v)}{\operatorname{argmax}} \operatorname{sim}(c_u, c_v)$ /* find most similar pair */

$c_j = c_a \cup c_b$ /* create a new cluster for pair */

$C := C - \{c_a, c_b\} \cup \{c_j\}$

$j := j + 1$

Bottom-Up Hierarchical Clustering

- keep track of history of merges and distances in order to reconstruct the tree

Similarity of Two Clusters

- the similarity of two clusters can be determined in several ways
 - *single link*: similarity of two most similar instances
 - *complete link*: similarity of two least similar instances
 - *average link*: average similarity between instances

Similarity/Distance Metrics

- *distance* = inverse of *similarity*
- properties of metrics

$$\text{dist}(x_i, x_j) \geq 0$$

$$\text{dist}(x_i, x_j) = \text{dist}(x_j, x_i)$$

$$\text{dist}(x_i, x_i) = 0$$

$$\text{dist}(x_i, x_j) \leq \text{dist}(x_i, x_k) + \text{dist}(x_k, x_j)$$

Genome-Wide Cluster Analysis

- Eisen et al., *PNAS* 1998
- *S. cerevisiae* (baker's yeast)
 - all genes (~ 6200) on a single array
 - measured during several processes
- human fibroblasts
 - 8600 human transcripts on array
 - measured at 12 time points during serum stimulation

The Data

- 79 measurements for yeast data
- collected at various time points during
 - diauxic shift (shutting down genes for metabolizing sugars, activating those for metabolizing ethanol)
 - mitotic cell division cycle
 - sporulation
 - temperature shock
 - reducing shock

The Data

- each measurement G_i represents

$$\log \frac{\text{red}_i}{\text{green}_i}$$

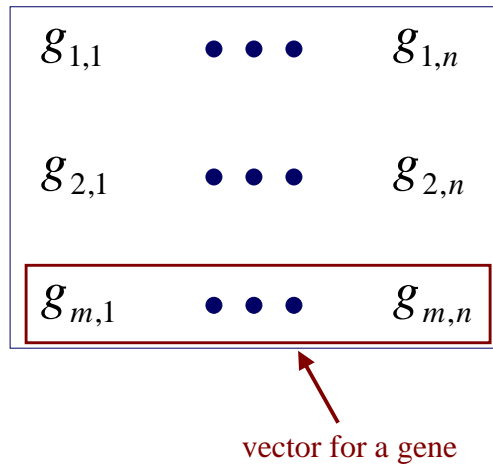
where red is the test expression level, and green is the reference level for gene G in the i th experiment

- the expression profile of a gene is the vector of measurements across all experiments

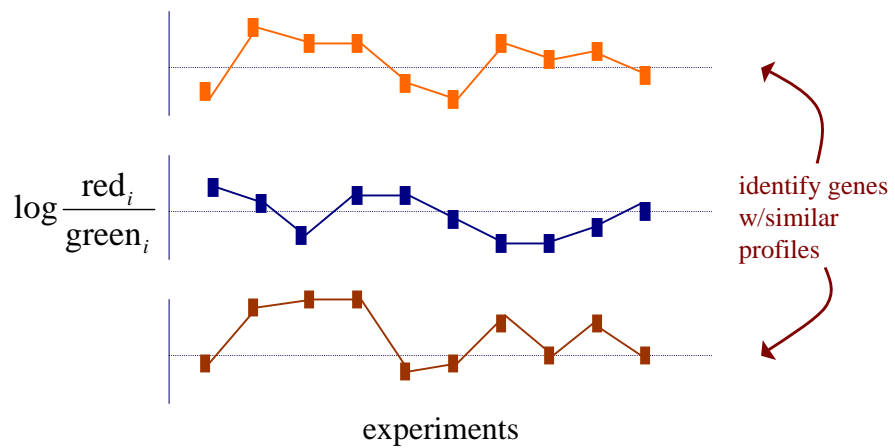
$$\langle G_1 \dots G_n \rangle$$

The Data

- m genes measured in n experiments



The Task



Gene Similarity Metric

- to determine the similarity of two genes X, Y

$$S(X, Y) = \frac{1}{n} \sum_{i=1}^n \left| \frac{X_i - X_{\text{offset}}}{\Phi_X} \right| \left| \frac{Y_i - Y_{\text{offset}}}{\Phi_Y} \right|$$

measurements
for each gene

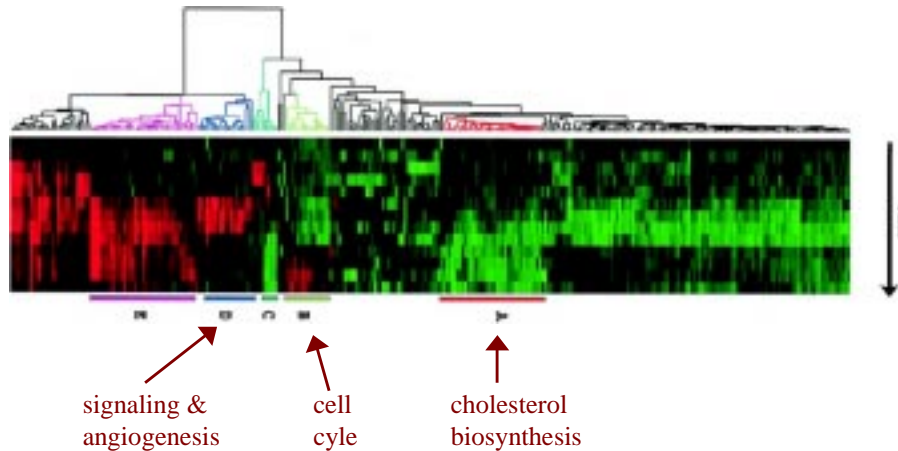
$$\Phi_G = \sqrt{\frac{\sum_{i=1}^n (G_i - G_{\text{offset}})^2}{n}}$$

Gene Similarity Metric

- since there is an assumed reference state (the gene's expression level didn't change), G_{offset} is set to 0 for all genes

$$S(X, Y) = \frac{1}{n} \sum_{i=1}^n \left| \frac{X_i}{\sqrt{\sum_{i=1}^n \frac{X_i^2}{n}}} \right| \left| \frac{Y_i}{\sqrt{\sum_{i=1}^n \frac{Y_i^2}{n}}} \right|$$

Dendrogram for Serum Stimulation of Fibroblasts



Eisen et al. Results

- redundant representations of genes cluster together
 - but individual genes can be distinguished from related genes by subtle differences in expression
- genes of similar function cluster together
 - e.g. 126 genes strongly down-regulated in response to stress

Eisen et al. Results

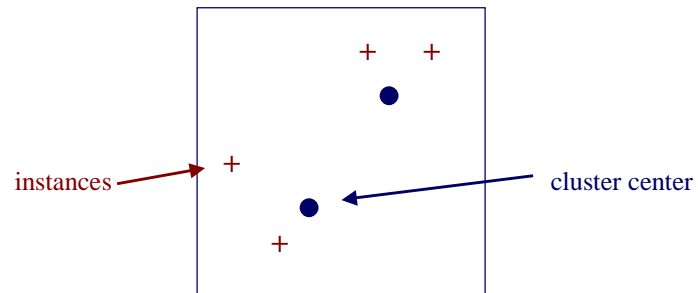
- 126 genes down-regulated in response to stress
 - 112 of the genes encode ribosomal and other proteins related to translation
 - agrees with previously known result that yeast responds to favorable growth conditions by increasing the production of ribosomes

Partitional Clustering

- divide instances into disjoint clusters
 - flat vs. tree structure
- key issues
 - how many clusters should there be?
 - how should clusters be represented?

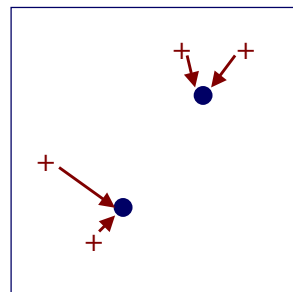
K-Means Clustering

- assume our instances are represented by vectors of real values
- put k cluster centers in same space as instances
- now iteratively move cluster centers

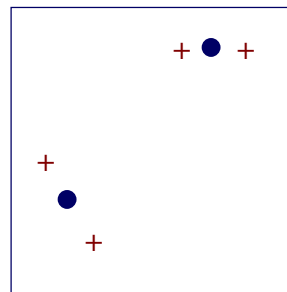


K-Means Clustering

- each iteration involves two steps
 - assignment of instances to clusters
 - re-computation of the means



assignment



re-computation of means

K-Means Clustering

given : a set $X = \{\vec{x}_1 \dots \vec{x}_n\}$ of instances

select k initial cluster centers $\vec{f}_1 \dots \vec{f}_k$

while stopping criterion not true do

 for all clusters c_j do

$$c_j = \{ \vec{x}_i \mid \forall f_l \text{ sim}(\vec{x}_i, \vec{f}_j) \geq \text{sim}(\vec{x}_i, \vec{f}_l) \}$$

 for all means \vec{f}_j do

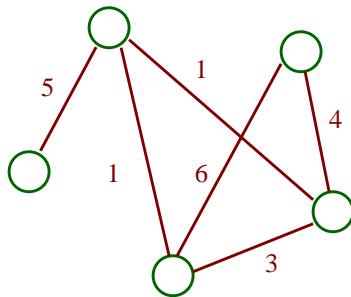
$$\vec{f}_j = \mu(c_j)$$

K-Means Clustering

- in k -means as just described, instances are assigned to one and only one cluster
- can do “soft” k -means clustering via EM
 - each cluster represented by a normal distribution
 - E step: determine how likely is it that each cluster “generated” each instance
 - M step: move cluster centers to maximize likelihood of instances

The CLICK Algorithm

- Sharan & Shamir, ISMB 2000
- instances to be clustered (e.g. genes) represented as vertices in a graph
- weighted, undirected edges represent similarity of instances



CLICK: How Do We Get Graph?

- assume pairwise similarity values are normally distributed

$$N(\mu_T, \sigma_T^2) \quad \text{for } \textit{mates} \text{ (instances in same "true" cluster)}$$

$$N(\mu_F, \sigma_F^2) \quad \text{for } \textit{non-mates}$$

- estimate the parameters of these distributions and $\Pr(\textit{mates})$ (the prob that two randomly chosen instances are mates) from the data

CLICK: How Do We Get Graph?

- let $f(S_{ij} | i, j \text{ are mates})$ be the probability density function for similarity values when i and j are mates
- then set the weight of an edge by:

$$w_{ij} = \log \frac{\Pr(\text{mates}) f(S_{ij} | i, j \text{ are mates})}{(1 - \Pr(\text{mates})) f(S_{ij} | i, j \text{ are non - mates})}$$

- prune edges with weights $<$ specified non-negative threshold t

The Basic CLICK Algorithm

BasicCLICK (G) :

if $V(G) = \{v\}$ then /* does graph have just one vertex? */

 move v to singleton set R

else if G is a kernel /* does graph satisfy stopping criterion? */

 return $V(G)$

else /* partition graph, call recursively */

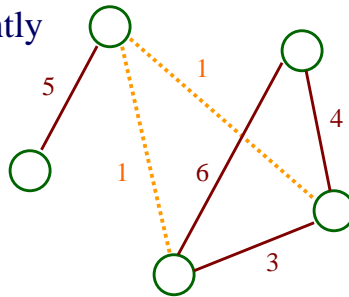
$(H, \bar{H}) \leftarrow \text{MinWeightC ut}(G)$

 BasicCLICK (H)

 BasicCLICK (\bar{H})

Minimum Weight Cuts

- a *cut* of a graph is a subset of edges whose removal disconnects the graph
- a *minimum weight cut* is the cut with the smallest sum of edge weights
- can be found efficiently



Deciding When a Subgraph Represents a Kernel

- we can test a cut C against two hypotheses
 - H_0^C : C contains only edges between non - mates
 - H_1^C : C contains only edges between mates
- we can then score C by

$$\log \frac{\Pr(H_1^C | C)}{\Pr(H_0^C | C)}$$

Deciding When a Subgraph Represents a Kernel

- if we assume a complete graph, the minimum weight cut algorithm finds a cut that minimizes this ratio, i.e.

$$\text{weight}(C) = \log \frac{\Pr(H_1^C | C)}{\Pr(H_0^C | C)}$$

- thus, we accept H_1^C and call G a kernel iff $\text{weight}(C) > 0$

Deciding When a Subgraph Represents a Kernel

- but we don't have a complete graph
- we call G a kernel iff $\text{weight}(C) + \text{weight}'(C) > 0$ where $\text{weight}'(C)$ approximates the contribution of missing edges

The Full CLICK Algorithm

- the basic CLICK algorithm produces kernels of clusters
- add two more operations
 - adoption*: find singletons that are similar, and hence can be adopted by kernels
 - merge*: merge similar clusters

The Full CLICK Algorithm

CLICK(G_N):

$R \leftarrow N$

while some change occurs do

 BasicCLICK(G_R)

 let L be the set of kernels produced

 let R be the set of singletons produced

 Adoption(L, R)

Merge(L)

Adoption(R)

CLICK Experiment: Fibroblast Serum Response Data

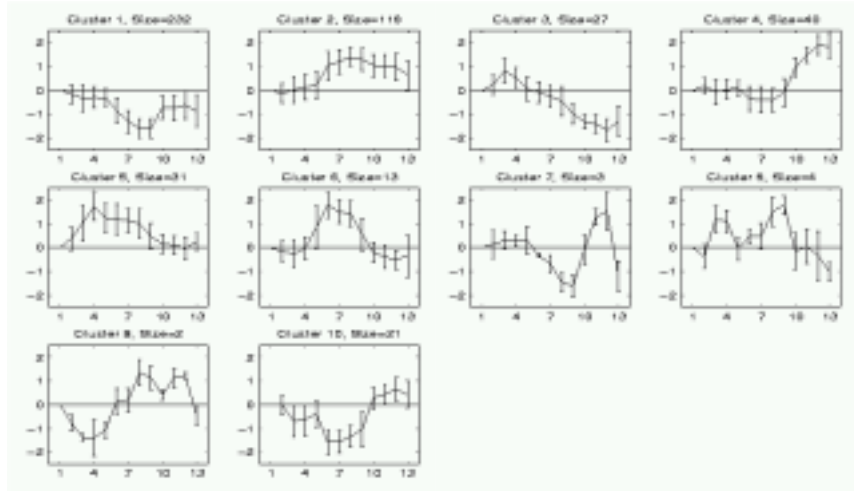


figure from: Sharan & Shamir, *ISMB 2000*

Measuring Homogeneity

- average similarity of instances to their clusters

$$H_{ave} = \frac{1}{|N|} \sum_{u \in N} \text{sim}(F(u), F(\text{cluster}(u)))$$

- minimum similarity of an instances to its cluster

$$H_{\min} = \min_{u \in N} \text{sim}(F(u), F(\text{cluster}(u)))$$

Measuring Separation

- average separation of pairs of clusters

$$S_{ave} = \frac{1}{\sum_{i \neq j} |X_i \parallel X_j|} \sum_{i \neq j} |X_i \parallel X_j| \text{sim}(F(X_i), F(X_j))$$

- maximum separation of a pair of clusters

$$S_{max} = \max_{i \neq j} \text{sim}(F(X_i), F(X_j))$$

- note that under these definitions, low separation is good!

CLICK Experiment: Fibroblast Serum Response Data

| Program | #Clusters | Homogeneity | | Separation | |
|--------------|-----------|-------------|-----------|------------|-----------|
| | | H_{Ave} | H_{Min} | S_{Ave} | S_{Max} |
| CLICK | 10 | 0.88 | 0.13 | -0.34 | 0.65 |
| Hierarchical | 10 | 0.87 | -0.75 | -0.13 | 0.9 |

table from: Sharan & Shamir, ISMB 2000

Evaluating Clustering Results

- given random data without any “structure”, clustering algorithms will still return clusters
- the gold standard: do clusters correspond to natural categories?
- do clusters correspond to categories we care about? (lots of ways to partition the world)
- how probable does held aside data look
- how well does clustering algorithm optimize intra-cluster similarity and inter-cluster dissimilarity