Tumor Classification Using Gene Expression Data

Three main types of statistical problems associated with tumor classification:

- Identification of new/unknown tumor classes using gene expression profiles (unsupervised learning – clustering)
- Classification of malignancies into known classes (supervised learning – discrimination)
- Identification of “marker” genes that characterize the different tumor classes (feature or variable selection).

Clustering

- “Clustering” is an exploratory tool for looking at associations within gene expression data.
- These methods allow us to hypothesize about relationships between genes and classes.
- We should use these methods for visualization, hypothesis generation, selection of genes for further consideration.
- We should not use these methods inferentially.
- Generally, there is no convincing measure of “strength of evidence” or “strength of clustering structure” provided.
- Hierarchical clustering specifically: we are provided with a picture from which we can make many/any conclusions.
More specifically….

- Cluster analysis arranges samples and genes into groups based on their expression levels.
- Arrangements are sensitive to choices made with regards to cluster components.
- In hierarchical clustering, the VIZUALIZATION of the arrangement (the dendrogram) is not unique!

  Just because two samples are situated next to each other does not mean that they are similar.

Generic Clustering Tasks

- Assigning objects to the groups
- Estimating number of clusters
- Assessing strength/confidence of cluster assignments for individual objects

Basic principles of clustering

**Aim:** to group observations that are “similar” based on predefined criteria.

Clustering can be applied to rows (genes) and / or columns (arrays) of an expression data matrix.

Clustering allows for reordering of the rows/columns of an expression data matrix which is appropriate for visualization.

Basic principles of clustering

**Issues:**
- Which genes / arrays to use?
- Which similarity or dissimilarity measure?
- Which method to use to join clusters/observations?
- Which clustering algorithm?
- How to validate the resulting clusters?

It is advisable to **reduce** the number of genes from the full set to some more manageable number, before clustering. The basis for this reduction is usually quite context specific and varies depending on what is being clustered, genes or arrays.
Clustering microarray data

- Clustering leads to readily interpretable figures and can be helpful for identifying patterns in time or space.

Examples:
- We can cluster cell samples (cols), e.g. 1) for identification (profiles). Here, we might want to estimate the number of different neuron cell types in a set of samples, based on gene expression.
  2) the identification of new / unknown tumor classes using gene expression profiles.
- We can cluster genes (rows), e.g. using large numbers of yeast experiments, to identify groups of co-regulated genes.
- We can cluster genes (rows) to reduce redundancy (cf. variable selection) in predictive models.

Clustering samples

Expression data – set of samples to cluster

Determine the set of genes to be used in clustering (DO NOT use class labels in the set determination).

Descriptive interpretation of genes separating novel subgroups of the samples

Commonly used measure?

- A metric is a measure of the similarity or dissimilarity between two data objects and it’s used to form data points into clusters

Two main classes of distance:
- 1- Correlation coefficients (scale-invariant)
- Distance metric (scale-dependent)
Some correlations to choose from

- **Pearson Correlation:**
  
  \[
  s(x_i, x_j) = \frac{\sum_{k=1}^{K} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{K} (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^{K} (x_{jk} - \bar{x}_j)^2}}
  \]

- **Uncentered Correlation:**

  \[
  s(x_i, x_j) = \frac{\sum_{k=1}^{K} x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{K} x_{ik}^2} \sqrt{\sum_{k=1}^{K} x_{jk}^2}}
  \]

- **Absolute Value of Correlation:**

  \[
  s(x_i, x_j) = \frac{\sum_{k=1}^{K} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{K} (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^{K} (x_{jk} - \bar{x}_j)^2}}
  \]

The difference is that, if you have two vectors X and Y with identical shape, but which are offset relative to each other by a fixed value, they will have a standard Pearson correlation (centered correlation) of 1 but will not have an uncentered correlation of 1.

Correlation (a measure between -1 and 1)

- Others include Spearman’s \( \rho \) and Kendall’s \( \tau \)
- You can use *absolute correlation* to capture both positive and negative correlation

Distance metrics

- **City Block (Manhattan) distance:**
  - Sum of differences across dimensions
  - Less sensitive to outliers
  - Diamond shaped clusters

- **Euclidean distance:**
  - Most commonly used distance
  - Sphere shaped cluster
  - Corresponds to the geometric distance into the multidimensional space

\[
 d(X,Y) = \sum_{i} |x_i - y_i|
\]

\[
 d(X,Y) = \sqrt{\sum_{i} (x_i - y_i)^2}
\]

where gene \( X = (x_1, ..., x_n) \) and gene \( Y = (y_1, ..., y_n) \)
How to Compute Group Similarity?

Four Popular Methods:

Given two groups $g_1$ and $g_2$,

- Single-link algorithm: $s(g_1, g_2) = \text{similarity of the closest pair}$
- Complete-link algorithm: $s(g_1, g_2) = \text{similarity of the furtherest pair}$
- Average-link algorithm: $s(g_1, g_2) = \text{average of similarity of all pairs}$
- Centroid algorithm: $s(g_1, g_2) = \text{distance between centroids of the two clusters}$

Comparison of the Three Methods

- Single-link
  - Elongated clusters
  - Individual decision, sensitive to outliers
- Complete-link
  - Compact clusters
  - Individual decision, sensitive to outliers
- Average-link or centroid
  - “In between”
  - Group decision, insensitive to outliers

Which one is the best? Depends on what you need!
**Clustering algorithms**
- Clustering algorithm comes in 2 basic flavors:
  - **Partitioning**
  - **Hierarchical**

**Hierarchical methods**
- Hierarchical clustering methods produce a tree or dendrogram.
- They avoid specifying how many clusters are appropriate by providing a partition for each k obtained from cutting the tree at some level.
- The tree can be built in two distinct ways:
  - bottom-up: agglomerative clustering.
  - top-down: divisive clustering.

**Hierarchical Clustering**
- The most overused statistical method in gene expression analysis
- Gives us pretty red-green picture with patterns
- But, pretty picture tends to be pretty unstable.
- Many different ways to perform hierarchical clustering
- Tend to be sensitive to small changes in the data
- Provided with clusters of every size: where to "cut" the dendrogram is user-determined

**Agglomerative Methods**
- Start with n mRNA sample (or g gene) clusters
- At each step, merge the two closest clusters using a measure of between-cluster dissimilarity which reflects the shape of the clusters
- The distance between clusters is defined by the method used (e.g., if complete linkage, the distance is defined as the distance between furthest pair of points in the two clusters)
Partitioning methods

- Partition the data into a pre-specified number $k$ of mutually exclusive and exhaustive groups.

- Iteratively reallocate the observations to clusters until some criterion is met, e.g., minimize within cluster sums of squares. Ideally, dissimilarity between clusters will be maximized while it is minimized within clusters.

- Examples:
  - $k$-means, self-organizing maps (SOM), PAM, etc.;
  - Fuzzy (each object is assigned probability of being in a cluster): needs stochastic model, e.g., Gaussian mixtures.
**Partitioning methods**

K = 2

- Partitioning Method
- Don’t get pretty picture
- MUST choose number of clusters K a priori
- More of a “black box” because output is most commonly looked at purely as assignments
- Each object (gene or sample) gets assigned to a cluster
- Begin with initial partition
- Iterate so that objects within clusters are most similar

K = 4

**K-means and K-medoids**

- Partitioning Method
- Don’t get pretty picture
- MUST choose number of clusters K a priori
- More of a “black box” because output is most commonly looked at purely as assignments
- Each object (gene or sample) gets assigned to a cluster
- Begin with initial partition
- Iterate so that objects within clusters are most similar

**How to make a K-means clustering**

1. Choose samples and genes to include in cluster analysis
2. Choose similarity/distance metric (generally Euclidean or correlation)
3. Choose number of clusters K.
4. Perform cluster analysis.
5. Assess cluster fit and stability
6. Interpret resulting cluster structure

Adapted from Elizabeth Garrett-Mayer
**K-means Algorithm**

1. Choose K centroids at random
2. Make initial partition of objects into k clusters by assigning objects to closest centroid
3. Calculate the centroid (mean) of each of the k clusters.
   a. For object i, calculate its distance to each of the centroids.
   b. Allocate object i to cluster with closest centroid.
   c. If object was reallocated, recalculate centroids based on new clusters.
4. Repeat 3 for object i = 1,...,N.
5. Repeat 3 and 4 until no reallocations occur.
6. Assess cluster structure for fit and stability

*Adapted from Elizabeth Garrett-Mayer*
PAM: Partitioning Around Medoids
or K-medoids

- A little different
- Centroid: The average of the samples within a cluster
- Medoid: The “representative object” within a cluster.
- Initializing requires choosing medoids at random.

Mixture Model for Clustering

\[ P(X|\text{Cluster}_i) \]

\[ P(X) = \sum_{i=1}^{k} \lambda_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \]

- Likelihood function (generally Gaussian)
- Parameters: e.g., \( \lambda_i, \sigma_i, \mu_i \)
- Using EM algorithm
  - Similar to “soft” K-mean
- Number of clusters can be determined using a model-selection criterion, e.g., AIC or BIC (Raftery and Fraley, 1998)
Some digression into model selection

- Principle of Parsimony: use the smallest number of parameters necessary to represent the data adequately
  - with increasing K (number of parameters), trade-off
  - low K: underfit, miss important effects
  - high K: overfit, include spurious effects and "noise"
  - parsimony – "proper" balance between these 2 effects so that you can repeat results across replications

AIC/BIC approach – seek a balance between overfit and underfit

\[ AIC = -2 \ln (\text{likelihood}) + 2K; \text{K} = \text{number of parameters}. \]

Partitioning vs. hierarchical

Partitioning:
- Advantages
  - Optimal for certain criteria.
  - Genes automatically assigned to clusters
- Disadvantages
  - Need initial k;
  - Often require long computation times.
  - All genes are forced into a cluster.

Hierarchical
- Advantages
  - Faster computation.
  - Visual
- Disadvantages
  - Unrelated genes are eventually joined
  - Rigid, cannot correct later for erroneous decisions made earlier.
  - Hard to define clusters.

How many clusters?

Global Criteria:
3. Graph theory (e.g.: cliques in CAST) (Ben-Dor et al., 1999).

Resampling methods:
2. WADP (Bittner et al., 2000).

Estimating number of clusters using silhouette (see PAM)

Define silhouette width of the observation is:

\[ S = (b-a)/\max(a,b) \]

Where \( a \) is the average dissimilarity to all the points in the cluster and \( b \) is the minimum distance to any of the objects in the other clusters.

Intuitively, objects with large \( S \) are well-clustered while the ones with small \( S \) tend to lie between clusters.

How many clusters: Perform clustering for a sequence of the number of clusters \( k \) and choose the number of components corresponding to the largest average silhouette.

*Issue of the number of clusters in the data is most relevant for novel class discovery, i.e. for clustering samples.*
Estimating number of clusters

There are other resampling (e.g. Dudoit and Fridlyand, 2002) and non-resampling based rules for estimating the number of clusters (for review see Milligan and Cooper (1978) and Dudoit and Fridlyand (2002)).

The bottom line is that none work very well in complicated situations and, to a large extent, clustering lies outside a usual statistical framework.

It is always reassuring when you are able to characterize a newly discovered clusters using information that was not used for clustering.

Estimating number of clusters using reference distribution

Idea: Define a goodness of clustering score to minimize, e.g. pooled Within clusters Sum of Squares (WSS) around the cluster means, reflecting compactness of clusters.

$$W_k = \frac{1}{n} \sum_{i=1}^{k} \frac{1}{2n_i} D_i$$

where n and D are the number of points in the cluster and sum of all pairwise distances, respectively.

Then gap statistic for k clusters is defined as:

$$\text{Gap}_n(k) = E_k^*(\log(W_k)) - \log(W_k)$$

Where $E^*$ is the average under a sample of the same size from the reference distribution. Reference distribution can be generated either parametrically (e.g. from a multivariate) or non-parametrically (e.g. by sampling from marginal distributions of the variables. The first local maximum is chosen to be the number of clusters (slightly more complicated rule) (Tibshirani et al, 2001).
Clest

Combines supervised and unsupervised approaches:

For each K in 2 … Kmax
  - Repeatedly split the observations into training and test set
  - Cluster training and test sets into K clusters
  - Use training set to build a predictor using the resulting cluster labels
  - Assess how well predicted labels match the cluster results on the training set

Assessment is done by considering null distribution of the “agreement” statistic.

WADP: Weighted Average Discrepancy Pairs

- Add perturbations to original data
- Calculate the number of paired samples that cluster together in the original cluster that didn’t in the perturbed
- Repeat for every cutoff (i.e. for each k)
- Do iteratively
- Estimate for each k the proportion of discrepant pairs.

Confidence in of the individual cluster assignments

Want to assign confidence to individual observations of being in their assigned clusters.

- Model-based clustering: natural probability interpretation
- Partitioning methods: silhouette

Dudoit and Fridlyand (2003) have presented a resampling-based approach that assigns confidence by computing the proportion of resampling times that an observation ends up in the assigned cluster.
Using aggregation to assign confidence to the observations’ labels

• Number of clusters K needs to be fixed a-priori
• Has been shown on simulated data to improve quality of cluster assignment
• Interesting alternative by-product:
  - For each pair of samples, compute proportion of bootstrap iterations where they were co-clustered
  - Use 1-proportion as a new distance metric
  - Re-cluster observations using this new distance metric

Tight clustering (genes)

Identifies small stable gene clusters by not attempting to cluster all the genes. Thus, it does not necessitate estimation of the number of clusters and assignment of all points into the clusters. Aids interpretability and validity of the results. (Tseng et al, 2003)

Algorithm:

For sequence of $k > k_0$:

1. Identify the set of genes that are consistently grouped together when genes are repeatedly sub-sampled. Order those sets by size. Consider the top largest $q$ sets for each $k$.
2. Stop when for $(k, (k+1))$, the two sets are nearly identical. Take the set corresponding to $(k+1)$. Remove that set from the dataset.
3. Set $k_0 = k_0 - 1$ and repeat the procedure.
Hybrid methods: HOPACH

- Hierarchical Ordered Partitioning and Collapsing Hybrid.

- Apply a partitioning algorithm iteratively to produce a hierarchical tree of clusters.
- At each node, a cluster is partitioned into two or more smaller clusters. Splits are not restricted to be binary. E.g., choose $K$ based on average silhouette.

Two-way clustering of genes and samples.

Refer to the methods that use samples and genes simultaneously to extract information. These methods are not yet well developed.

Some examples of the approaches include Block Clustering (Hartigan, 1972) which repeatedly rearranges rows and columns to obtain the largest reduction of total within block variance.

Another method is based on Plaid Models (Lazzeroni and Owen, 2002)

Friedman and Meulmann (2002) present an algorithm allowing to cluster samples based on the subsets of attributes, i.e. each group of samples could have been characterized by different gene sets.

Applications of clustering to the microarray data


- Three subtypes of lymphoma (FL, CLL and DLBCL) have different genetic signatures. (81 cases total)
- DLBCL group can be partitioned into two subgroups with significantly different survival. (39 DLBCL cases)
Clustering cell samples
Discovering sub-groups

Taken from Alizadeh et al (Nature, 2000)

Fall 2004, BMI 209 - Statistical Analysis of Microarray Data

Attempt at validation of DLBCL subgroups

Taken from Alizadeh et al (Nature, 2000)

Summary

Which clustering method should I use?
- What is the biological question?
- Do I have a preconceived notion of how many clusters there should be?
- Hard or soft boundaries between clusters

Keep in mind:
- Clustering cannot NOT work. That is, every clustering methods will return clusters.
- Clustering helps to group / order information and is a visualization tool for learning about the data. However, clustering results do not provide biological "proof".
- Clustering is generally used as an exploratory and hypotheses generation tool.

Yeast Cell Cycle (Cho et al, 1998)
6 \times 5 SOM with 828 genes

Taken from Tamayo et al, (PNAS, 1999)

Fall 2004, BMI 209 - Statistical Analysis of Microarray Data
Some clustering pitfalls

The procedure should not bias results towards desired conclusions.

Question: Do expression data cluster according to the survival status.

Design: Identify genes with high t-statistic for comparison short and long survivors. Use these genes to cluster samples. Get excited that samples cluster according to survival status.

Issues: The genes were already selected based on the survival status. Therefore, it would rather be surprising if samples did "not" cluster according to their survival.

Conclusion: None are possible with respect to clustering as variable selection was driven by class distinction.

P-values for differential expression are only valid when the class labels are independent of the current dataset.

Question: Identify genes distinguishing among “interesting” subgroups.

Design: Cluster samples into K groups. For each gene, compute F-statistic and its associated p-value to test for differential expression among two subgroups.

Issues: Same data was used to create groups as to test for DEs – p-values are invalid.

Conclusion: None with respect to DEs p-values. Nevertheless, it is possible to select genes with high value of the statistic and test hypotheses about functional enrichment with, e.g., Gene Ontology. Also, can cluster these genes and use the results to generate new hypotheses.

Acknowledgements

SFGH
- Agnes Paquet
- David Erle
- Andrea Barczac
- UCSF Sandler Genomics Core Facility.

UCSF /CBMB
- Ajay Jain
- Mark Segal
- UCSF Cancer Center Array Core
- Jain Lab

UCB
- Terry Speed
- Jean Yang
- Sandrine Dudoit
Some references

6. Petricoin et al, "Use of proteomics patterns in serum to identify ovarian cancer", Lancet, 2002 (and relevant correspondence)

13. Dudoit and Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset ", Genome Biology, 2002
15. Kaufmann and Rousseeuw, "Clustering by means of medoids.", Elsevier/North Holland 1987
16. See many articles by Leo Breiman on aggregation