Statistical Genomics and Bioinformatics Workshop: Genetic Association and RNA-Seq Studies

Genomic Clustering and Signature Development
Brooke L. Fridley, PhD
University of Kansas Medical Center

Genomic Clustering
Clustering Basics

- **Clustering** is the process of grouping a set of physical or abstract objects into classes of similar objects.
  - It is also called unsupervised learning.

- **Exploratory tool**: use these methods for visualization, hypothesis generation, selection of genes for further consideration
  - We should not use these methods inferentially.

- Hierarchical clustering specifically: we are provided with a picture from which we can make many/any conclusions.

Why cluster genes?

- Identify groups of possibly **co-regulated genes**
- Identify typical temporal or spatial gene expression patterns
- Arrange a set of genes in a linear order that is at least not totally meaningless (we hope).
- **Aids in the interpretation**
Why cluster samples?

- **Quality control:** Detect experimental artifacts/bad hybridizations
- Check whether samples are grouped according to known categories (though this might be better addressed using a *supervised* approach: statistical tests, classification)
- **Identify new “classes”** of biological samples
  - tumor subtypes
  - Disease heterogeneity

Human breast tumors cluster into 6 distinct molecular subtypes of breast cancer with differences in patient survival.

### Clustering vs Classification

- **Clustering** is *unsupervised*:
  - We don’t use any information about what class the samples belong to (e.g. disease status, cancer type) to determine cluster structure
  - Hierarchical, K-Means, PCA, SOM, model-based clustering
  - *Clustering finds groups in the data*

- Classification methods are *supervised*:
  - Identifying to which of a set of groups/categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.
  - Discriminant analysis, PAM (Shrunken centroids), Random Forests/CART, K-Nearest Neighbor,
  - *Classification methods finds ‘classifiers’*

### Cluster analysis

Generally, cluster analysis is based on two components:

1. **Distance measure**: Quantification of dissimilarity / similarity of objects.

2. **Cluster algorithm**: A procedure to group objects.
   - Aim: small within-cluster distances, large between-cluster distances.
Distance and Similarity

• Every clustering method is based **solely** on the measure of distance or similarity.
  – The clustering is only as good as the distance matrix

• Generally, not enough thought and time is spent on choosing and estimating the distance/similarity matrix.
  – Applying correlation to highly skewed data will provide misleading results.
  – Applying Euclidean distance to data measured on categorical scale will be invalid.

Hierarchical Clustering

• The most over used statistical method in gene expression analysis
• Tends to be pretty **unstable**.
  – Many different ways to perform hierarchical clustering
  – Sensitive to small changes in the data
• Provided with clusters of every size
  – where to “cut” the dendrogram is user-determined

Distances between clusters used for hierarchical clustering

- Distance between two clusters is based on the pairwise distances between members of the clusters.

- **Complete linkage**: largest distance

- **Average linkage**: average distance

- **Single linkage**: smallest distance

- Complete linkage gives preference to compact / spherical clusters.
- Single linkage can produce long stretched clusters.

Hierarchical Clustering Example: Visualizing DNA Methylation Data

Unsupervised hierarchical clustering based on Manhattan distance and average linkage.

Christensen et al. PLoS Genetics, 2009
**K-Means**

- Intuitive and very **easy to implement**
- **Pre-specification of the number of clusters** $K$
  - $K$ is typically unknown for most practical purposes
  - Misspecification may lead to poor results
- Choice of a distance measure may be difficult to justify
- Clusters are expected to be of similar size
- May not work well for irregular clusters
  - Clusters based on the first-moment (i.e., mean) only

**K-means clustering illustration**

![Gene A vs Gene B scatter plot](image)
• Need to first pre-specify the number of clusters
  • For simplicity, assume k = 2
• Step 1: Initialize the means for the two clusters

• Compute the distance of each of the points to each of the two means
• Assign points to the cluster with the closest mean
• Compute the distance of each of the points to each of the two means
• Assign points to the cluster with the closest mean

• Re-compute the means based on the observations within that cluster (i.e., m^{(2)}_1 and m^{(2)}_2)
• Compute the distance of each of the points to each of the two means
• Assign points to the cluster with the closest mean
• Re-compute the means based on the observations within that cluster (i.e., $m^{(2)}_1$ and $m^{(2)}_2$)
• Compute the distance of each of the points to each of the two means
• Assign points to the cluster with the closest mean

![Diagram of gene expression data with clusters and means](image1)

• After 5 iterations, we converge on our final solution!
• Consists of class labels for each of the $n$ observations

![Diagram of gene expression data with final solution](image2)
Cautions about Clustering

- Clustering can be a useful exploratory tool
- Cluster results are very sensitive to noise in the data
- Need to assess cluster structure and stability of results
- Different clustering approaches can give quite different results
  - Methods
  - Number Clusters
  - Distance measures
- For hierarchical clustering, interpretation is almost always subjective
- Doesn’t tell us anything about what features should be used for clustering

Genomic Classification
Bad prognosis recurrence < 5yrs
Good Prognosis recurrence > 5yrs
Good Prognosis Matesis > 5

Reference

Decision tree classifiers (CART)

Transparent rules and easy to interpret and implement
Ensemble classifiers

- Resample 1 → Classifier 1
- Resample 2 → Classifier 2
- Resample 499 → Classifier 499
- Resample 500 → Classifier 500

Aggregate classifier

Examples: Bagging, Boosting, Random Forest

Feature Selection

- Lead to **better classification performance** by removing variables that are noise with respect to the outcome
- May provide useful **biological insights**
- Can eventually lead to the **diagnostic tests**
Classifier Performance Assessment

- Any classification rule needs to be evaluated for its performance on the future samples.
- One needs to estimate future performance based on what is available
- Assessing performance of the classifier based on
  - Cross-validation
    - V-fold CV
    - Leave-one-out cross validation (LOOCV)
  - Training vs Testing set
  - Testing on independent dataset

Diagram of performance assessment
Pharmacogenomic (PGx) Classifiers

Benefits:
- Enables patients to be treated with drugs that actually work for them
- Avoids false negative trials for heterogeneous populations
- Avoids erroneous generalizations of conclusions from positive trials

Develop a PGx classifier to determine patients likely to benefit from a new TRT

Establish reproducibility of the PGx classifier

Use the PGx classifier to design and analyze a new clinical trial to evaluate effectiveness of TRT in the overall population or pre-defined subsets determined by the classifier.

Coming Full Circle: Integrating Many Methods & Data Types
Molecular Phenotype Based GWAS

- **Molecular Subtype GWAS**:  
  - For risk with existing controls  
  - For clinical outcome  
  - For quantitative trait

Example: Integrative Analysis  
Multiple Methods and Multiple Data Types

- Association analysis to determine epigenetic features associated with clinical outcome (TTR)
- Model-based clustering (semi-supervised) to determine clinically relevant methylation-based subtypes
- Nearest shrunken centroid (PAM) (supervised) analysis to determine genes in which mRNA levels differ between methylation subgroups
- Pathway Analysis of resulting differential expressed genes to determine enriched pathways.
Application to Ovarian Cancer

- Restricted to High Grade Serous (HGS) histology
- Pre-chemo tumor sample
- 450K Illumina Methylation Array
- Similar stage and recurrence status between testing and training data sets

337 HGS Ovarian Cancer Tumors

Training set
N=168
Recurrences = 110

Testing set
N=169
Recurrences = 116

337 high-grade serous tumors

Training set (168)

Random split

Testing set (169)

Analysis workflow of semi-supervised clustering used in this study.

SS-RPMM

1. Loci ranking
2. Cross-validation
3. Clustering and signature generation

optimal number of CpG loci = 60

SS: worse outcome
R: better outcome

All the HGS samples in testing set

169 samples / 104 events
L: worse outcome
R: better outcome

p-value=5.3E-4, HR=0.5

Kaplan Meier plot of association between groups (R or L) and recurrence time, for all the samples in testing set (n=169).
HGS testing samples with platinum and taxane treatment

Kaplan Meier plot of association between groups (R or L) and recurrence time, for the samples in testing set and received platinum and taxane treatment (n=130).

p-value=1.20E-5, HR=0.39

Nearest Shrunken Centroid Method (PAM Analysis)

- **Goal**: In a sample with K different classes and p variables, what variables contribute to the separation of these classes?

- **PAM**: Shrinks each class centroid towards the overall centroid. The shrinkage factor is determined by CV.

- The *shrinkage de-noises large effects* while setting small ones to zero (i.e., selection of key genes)
Ovarian Cancer Study

- 104 of the HGS cases have Agilent gene expression data

  rL (poor outcome): 48

  rR (better outcome): 56
Gene expression heatmap of PAM selected genes

Expression heatmaps of signature genes selected by PAM analysis using shrinkage factor 1.5, which was selected based on minimum cross-validation error.

What are the genes that distinguish between the molecular subtypes?

• 712 genes are over expressed in patients with poor outcome
  – moderately enriched in signaling pathways, such as
    Wnt/beta-catenin Signaling (p-value=8.71E-5).

• 958 genes are over expressed in patients with better outcome
  – extremely enriched in immune related pathways, such as
    Antigen Presentation Pathway (p-value=1.6E-32),
    Crosstalk between Dendritic Cells and Natural Killer Cells (p-value=2E-24), and Communication between Innate and Adaptive Immune Cells (p-value=5E-24).
  – Might explain why this group is associated with better outcome (blessed by protection of boosted immune mechanism)
Questions?

Thank You for Attending this Workshop.