Gene regulation 1: Expression Clustering

At a glance: Entering module III

The building blocks of genomes and gene regulation

Goal: understand the DNA elements responsible for gene regulation …

- The regulators: Transcription Factors, microRNAs, and their specificities
- The regions: the components of enhancers, promoters, silencers, other
- The targets: reliable identification of individual regulatory motif instances
- Tissue-specific activity: in vivo binding, chromatin state, active genes … and gain insights into their function.

Our tools: Comparative genomics & large-scale experimental datasets.

- Evolutionary signatures: distinct function ◯ distinct selective constraints
- Chromatin signatures: specific functions ◯ specific epigenomic codes
- Activity signatures: functional links ◯ correlated activity patterns
- Towards predictive models of gene expression, development, disease

Enter ENCODE/modENCODE

- Coordinated cell types/developmental stages
- Expression, replication, chromatin, TF binding

Challenges in Computational Biology

Module III: Gene regulation and networks

- Foundations
  - Clustering: Parameteric (K-means, EM), Non-param (Hierarchical)
  - Classification: Generative (Naive Bayes)/Feature selection
  - Regulatory Motif Discovery: Gibbs Sampling, EM, Information
- Frontiers:
  - Regulation: de novo motifs, target prediction, TFs/microRNAs
  - Networks: network inference, function prediction, predict expression
  - Epigenomics: chromatin states, activity signatures, disease SNPs
Predicting stage-specific regulators

- Their targets associated with expression changes
- TF expression changes confirm predicted regulators

Binding, motifs reveal physical regulatory network

- Hierarchical network: master regulators, 94% down
- Feed-forward, cooperation, feedback through miRs

Inferring a functional regulatory network

- Combine TF binding, motifs, correlated TF/TG activity
- Reveal ‘functional’ edges, response determinants
- Functional net shows increased predictive value

Functional network leads to new gene annotations

- Shared activity and regulation \(\Rightarrow\) shared function
- Tissue expression confirms functional predictions

Learning predictive models of gene expression

- Linear regression model: \(\text{Target}\_\text{expr}=F([\text{TF}\_\text{expr}, \ldots])\)
- Learn coefficients in 27 time-points, predict in other 3
- ‘Unpredictable’ genes are also less reproducible
- ‘Predictable’ genes: learned weights work in cell lines

Putting it all together: regulatory logic

- Underline regulatory logic specifying development
Goals for today: Gene Expression Clustering

1. Probing Gene Regulation
   - Microarray technology, data matrix
   - Questions / applications

2. The clustering problem
   - Finding structure in unlabeled data
   - Two types of algorithms

3. K-means clustering
   - The basic algorithm
   - Fuzzy k-means

4. Machine learning formulation
   - Generative models and learning
   - K-means vs. Expectation Maximization

5. Hierarchical Clustering
   - Basic algorithm
   - Distance measures

6. Evaluating Clustering Results
   - Hypergeometric distribution

Central Dogma

DNA $\rightarrow$ mRNA $\rightarrow$ protein $\rightarrow$ phenotype

We can measure amounts of mRNA for every gene in a cell

Expression Microarrays

- A way to measure the levels of mRNA in every gene

- Two basic types
  - Affymetrix gene chips
  - Spotted oligonucleotides

- Both work on same principle
  - Put DNA probe on slide
  - Complementary hybridization

Expression Microarray Data Matrix

- Measure 6,000 genes in 100s of conditions
- Study resulting matrix

Natural first step: Bi-clustering

- Cluster 1000s of ‘data points’ and 100s of ‘dimensions’

- Learn what distinguishes different ‘conditions’
- Study common patterns of behavior in regulatory ‘modules’ of many genes

Expression Microarrays

- Measure the level of mRNA messages in a cell

- Begin with probes for each predicted gene

- Variations on the theme:
  - Several probes per gene: tiled k-mers, averaged
  - Whole-genome tiling microarrays forego gene prediction
  - RNAseq: every base in genome becomes a bin
Gene regulation questions now possible:

- groups of genes that share similar function have similar expression patterns – identify regulons
- group related proteins to infer function, localization, etc
- classify cell state (i.e. AML vs ALL) using expression data
- predict cell function
- ...

Applications of clustering genes

- Data Exploration
  - Summarize data
  - Explore without getting lost in each data point
  - Enhance visualization

- Co-regulated Genes
  - Common expression may imply common regulation
  - Predict cis-regulatory promoter sequences

- Functional Annotation
  - Similar function from similar expression

Gene regulation questions now possible:

- groups of genes that share similar function have similar expression patterns – identify regulons
- group related proteins to infer function, localization, etc
- classify cell state (i.e. AML vs ALL) using expression data
- predict cell function
- ...

Applications of clustering genes

- Data Exploration
  - Summarize data
  - Explore without getting lost in each data point
  - Enhance visualization

- Co-regulated Genes
  - Common expression may imply common regulation
  - Predict cis-regulatory promoter sequences

- Functional Annotation
  - Similar function from similar expression

Goals for today: Gene Expression Clustering

1. Probing Gene Regulation
   - Microarray technology, data matrix
   - Questions / applications

2. The clustering problem
   - Finding structure in unlabelled data
   - Two types of algorithms

3. K-means clustering
   - The basic algorithm
   - Fuzzy k-means

4. Machine learning formulation
   - Generative models and learning
   - K-means vs. Expectation Maximization

5. Hierarchical Clustering
   - Basic algorithm
   - Distance measures

6. Evaluating Clustering Results
   - Hypergeometric distribution

Today: general clustering techniques

- Cluster Experiments
  - Group by similar expression profiles

- Cluster Genes
  - Group by similar expression in different conditions

Clustering questions pervasive in genomics

- Clustering
  - Microarray data: groups of genes that share similar function have similar expression patterns – identify regulons
  - Protein sequence: group related proteins to infer function
  - EST data: collapse redundant sequences

- Classification
  - Microarray data: classify cell state (i.e. AML vs ALL) using expression data
  - Protein/gene sequences: predict function, localization, etc.

Clustering vs Classification

- Objects characterized by one or more features

- Classification
  - Have labels for some points
  - Want a “rule” that will accurately assign labels to new points
  - Supervised learning

- Clustering
  - No labels
  - Group points into clusters based on how “near” they are to one another
  - Identify structure in data
  - Unsupervised learning
Structure in High-Dimensional Data

- Structure can be used to reduce dimensionality of data
- Structure can tell us something useful about the underlying phenomena
- Structure can be used to make inferences about new data

Clustering Algorithms

- Partitioning
  - Divides objects into non-overlapping clusters such that each data object is in exactly one subset
- Agglomerative
  - A set of nested clusters organized as a hierarchy

Goals for today: Gene Expression Clustering

1. Probing Gene Regulation
   - Microarray technology, data matrix
   - Questions / applications
2. The clustering problem
   - Finding structure in unlabelled data
   - Two types of algorithms
3. K-means clustering
   - The basic algorithm
   - Fuzzy k-means
4. Machine learning formulation
   - Generative models and learning
   - K-means vs. Expectation Maximization
5. Hierarchical Clustering
   - Basic algorithm
   - Distance measures
6. Evaluating Clustering Results
   - Hypergeometric distribution

K-Means Clustering

The Basic Idea

- Assume a fixed number of clusters, K
- Goal: create “compact” clusters

K-Means Algorithm

- Randomly Initialize Clusters
K-Means Algorithm

- Randomly Initialize Clusters
- Assign data points to nearest clusters
- Recalculate Clusters
- Repeat… until convergence
More Formally

1. Initialize K centers $u_k$

For each iteration $n$ until convergence

2. Assign each $x_i$ the label of the nearest center, where the distance between $x_i$ and $u_k$ is

$$d_{ij} = (x_i - u_k)^2$$

3. Move the position of each $u_k$ to the centroid of the points with that label

$$u_k(n+1) = \frac{\sum_{x_i \text{ with label } j} x_i P(x_i | u_k)^2}{\sum_{x_i \text{ with label } j} P(x_i | u_k)}$$

Iterate

Regular K-Means is a special case of fuzzy k-means where:

$$P(\text{label } K | x, u_k) = \begin{cases} 1 & \text{if } x_i \text{ is closest to } u_k \\ 0 & \text{otherwise} \end{cases}$$

Cost Criterion

We can think of K-means as trying to create clusters that minimize a cost criterion associated with the size of the cluster

$$\text{COST}(x_1, x_2, \ldots, x_n) = \sum_{x_i \text{ with label } k} (x_i - u_k)^2$$

Minimizing this means minimizing each cluster term separately:

$$\sum_{x_i \text{ with label } k} (x_i - u_k)^2$$

Fuzzy K-Means

- Initialize K centers $u_k$

- For each point calculate the probability of membership for each category

$$P(\text{label } K | x, u_k)$$

- Move the position of each $u_k$ to the weighted centroid:

$$u_k(n+1) = \frac{\sum_{x_i \text{ with label } j} x_i P(x_i | u_k)^2}{\sum_{x_i \text{ with label } j} P(x_i | u_k)}$$

- Iterate

Regular K-Means is a special case of fuzzy k-means where:

$$P(\text{label } K | x, u_k) = \begin{cases} 1 & \text{if } x_i \text{ is closest to } u_k \\ 0 & \text{otherwise} \end{cases}$$

Goals for today: Gene Expression Clustering

1. Probing Gene Regulation
   - Microarray technology, data matrix
   - Questions / applications

2. The clustering problem
   - Finding structure in unlabelled data
   - Two types of algorithms

3. K-means clustering
   - The basic algorithm
   - Fuzzy k-means

4. Machine learning formulation
   - Generative models and learning
   - K-means vs. Expectation Maximization

5. Hierarchical Clustering
   - Basic algorithm
   - Distance measures

6. Evaluating Clustering Results
   - Hypergeometric distribution

K-Means as a Generative Model

Model of $P(X, \text{Labels})$

Generate

Observations

Samples drawn from two equally normal distributions with unit variance - a Gaussian Mixture Model

$$P(x_i | u_k) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x_i - u_k)^2}{2} \right\}$$

Unsupervised Learning

Model of $P(X, \text{Labels})$

Learn?

Observations

Samples drawn from two equally normal distributions with unit variance - a Gaussian Mixture Model

$$P(x_i | u_k) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x_i - u_k)^2}{2} \right\}$$
If labeled points \( \rightarrow \) derive centers

Need to estimate best unknown gaussian centers \((u_k)\) from data

Choose \( u_k \) to maximize probability of model

Given a set of \( x_i \) all of label \( k \), find max likelihood \( u_k \):

\[
\arg \max \left\{ \log \prod_x P \left( x \mid \mu \right) \right\} = \arg \max \sum \left\{ -\frac{1}{2} (x_i - \mu)^2 \right\}
\]

Find best center \( \mu \) Solution is the centroid of the \( x_i \)

If centers known \( \rightarrow \) assign points

Need to estimate labels for the data

That maximizes data likelihood

\[
\arg \max \left\{ \log P \left( x \mid \mu \right) \right\} = \arg \max \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x_i - \mu)^2}{2} \right\} = \arg \min \left\{ \mu - x_i \right\}
\]

Find best label \( k \)

Solution is given by the nearest center

If we have neither: iterate

The key idea:
1. Imagine we start with some \( u_k^0 \)
2. We could calculate the most likely labels for \( x_i \) given these \( u_k^0 \)
3. We could then use these labels to choose \( u_k^1 \)
4. And iterate (to convergence)

Expectation Maximization (EM)

1. Initialize parameters

2. E Step Estimate probability of hidden labels \( Q \), given parameters and sequence

   \( Q = P(hidden \mid data, parameters^{t-1}) \)

3. M Step Choose new parameters to maximize expected likelihood of parameters given the data and \( Q \)

   \( parameter^{t}_i = \arg \max \sum_{parameter} P(x \mid parameter^{t-1}) \)

4. Iterate

\( P(x \mid Model) \) guaranteed to increase each iteration

Goals for today: Gene Expression Clustering

1. Probing Gene Regulation
   - Microarray technology, data matrix
   - Questions / applications

2. The clustering problem
   - Finding structure in unlabeled data
   - Two types of algorithms

3. K-means clustering
   - The basic algorithm
   - Fuzzy k-means

4. Machine learning formulation
   - Generative models and learning
   - K-means vs. Expectation Maximization

5. Hierarchical Clustering
   - Basic algorithm
   - Distance measures

6. Evaluating Clustering Results
   - Hypergeometric distribution
Revisiting K-Means

1. Initialize K centers $u_k$
2. Assign each $x_i$ the label of the nearest center, where the distance between $x_i$ and $u_k$ is
   \[ d_{ik} = (x_i - u_k)^2 \]
3. Move the position of each $u_k$ to the centroid of the points with that label
4. Iterate

The most likely label $k$ for a point $x_i$

Maximum likelihood parameter $\mu_k$ given most likely label

Generative Model Perspective

Revisiting Fuzzy K-Means

1. Initialize K centers $u_k$
2. For each point calculate the probability of membership for each category
   \[ P(\text{label } K | x_i, \mu_k) \]
3. Move the position of each $u_k$ to the weighted centroid:
   \[
   \mu_k^{(n+1)} = \frac{\sum_{x_i \text{ with label } k} P(\mu_k | x_i) x_i}{\sum_{x_i \text{ with label } k} P(\mu_k | x_i)}
   \]
4. Iterate

This is analogous to Baum Welch from HMMs

Revisiting Fuzzy K-Means

1. Initialize K centers $u_k$
2. For each point calculate the probability of membership for each category
   \[ P(\text{label } K | x_i, \mu_k) \]
3. Move the position of each $u_k$ to the weighted centroid:
   \[
   \mu_k^{(n+1)} = \frac{\sum_{x_i \text{ with label } k} P(\mu_k | x_i) x_i}{\sum_{x_i \text{ with label } k} P(\mu_k | x_i)}
   \]
4. Iterate

This is analogous to Baum Welch from HMMs

Revisiting K-Means

1. Initialize K centers $u_k$
2. Assign each $x_i$ the label of the nearest center, where the distance between $x_i$ and $u_k$ is
   \[ d_{ik} = (x_i - u_k)^2 \]
3. Move the position of each $u_k$ to the centroid of the points with that label
4. Iterate

The most likely label $k$ for a point $x_i$

Maximum likelihood parameter $\mu_k$ given most likely label

Generative Model Perspective

Revisiting K-Means

1. Initialize K centers $u_k$
2. Assign each $x_i$ the label of the nearest center, where the distance between $x_i$ and $u_k$ is
   \[ d_{ik} = (x_i - u_k)^2 \]
3. Move the position of each $u_k$ to the centroid of the points with that label
4. Iterate

The most likely label $k$ for a point $x_i$

Maximum likelihood parameter $\mu_k$ given most likely label

Generative Model Perspective

Revisiting Fuzzy K-Means

1. Initialize K centers $u_k$
2. For each point calculate the probability of membership for each category
   \[ P(\text{label } K | x_i, \mu_k) \]
3. Move the position of each $u_k$ to the weighted centroid:
   \[
   \mu_k^{(n+1)} = \frac{\sum_{x_i \text{ with label } k} P(\mu_k | x_i) x_i}{\sum_{x_i \text{ with label } k} P(\mu_k | x_i)}
   \]
4. Iterate

This is analogous to Baum Welch from HMMs

Revisiting Fuzzy K-Means

1. Initialize K centers $u_k$
2. For each point calculate the probability of membership for each category
   \[ P(\text{label } K | x_i, \mu_k) \]
3. Move the position of each $u_k$ to the weighted centroid:
   \[
   \mu_k^{(n+1)} = \frac{\sum_{x_i \text{ with label } k} P(\mu_k | x_i) x_i}{\sum_{x_i \text{ with label } k} P(\mu_k | x_i)}
   \]
4. Iterate

This is analogous to Baum Welch from HMMs

K-Means, Viterbi learning & EM

K-Means and Fuzzy K-means are two related methods that can be seen performing unsupervised learning on a gaussian mixture model

Revealing assumptions about underlying data model

Can relax assumptions by relaxing constraints on model
- Including explicit covariance matrix
- Relaxing assumption that all gaussians are equally likely
Implications: Non-globular Clusters

Actual Clustering  K-means (K = 2)

Goals for today: Gene Expression Clustering
1. Probing Gene Regulation
   - Microarray technology, data matrix
   - Questions / applications
2. The clustering problem
   - Finding structure in unlabelled data
   - Two types of algorithms
3. K-means clustering
   - The basic algorithm
   - Fuzzy k-means
4. Machine learning formulation
   - Generative models and learning
   - K-means vs. Expectation Maximization
5. Hierarchical Clustering
   - Basic algorithm
   - Distance measures
6. Evaluating Clustering Results
   - Hypergeometric distribution

But How Many clusters?

• How do we select K?
  – We can always make clusters “more compact” by increasing K
  – e.g. What happens if K=number of data points?
  – What is a meaningful improvement?
• Hierarchical clustering side-steps this issue

Hierarchical clustering

Most widely used algorithm for expression data

• Start with each point in a separate cluster
• At each step:
  – Choose the pair of closest clusters
  – Merge

Phylogeny (UPGMA)

Visualization of results

Optimal leaf-ordering algorithms

But how do we define distances between clusters?
Distance between clusters

- CD(X,Y) = \min_{x \in X, y \in Y} D(x,y)  
  Single-link method
- CD(X,Y) = \max_{x \in X, y \in Y} D(x,y)  
  Complete-link method
- CD(X,Y) = \frac{1}{n} \sum_{x \in X, y \in Y} D(x,y)  
  Average-link method
- CD(X,Y) = D(\text{avg}(X), \text{avg}(Y))  
  Centroid method

(Dis)Similarity Measures

Evaluation Cluster Performance

In general, it depends on your goals in clustering

- Robustness
  - Select random samples from data set and cluster
  - Repeat
  - Robust clusters show up in all clusters

- Category Enrichment
  - Look for categories of genes “over-represented” in particular clusters
  - Also used in Motif Discovery

Evaluating clusters – Hypergeometric Distribution

N experiments, p labeled +, (N-p) –
Cluster: k elements, m labeled +
P-value of uniformity in computed cluster
Prob that a randomly chosen set of k experiments would result in m positive and k-m negative

Similar Genes Can Cluster

Clustered 8600 human genes using expression time course in fibroblasts

- (A) Cholesterol biosynthesis
- (B) Cell cycle
- (C) Immediate early response
- (D) Signalling and angiogenesis
- (E) Wound healing
Clusters and Motif Discovery

Expression from 15 time points during yeast cell cycle

Tavazoie & Church (1999)

Goals for today: Gene Expression Clustering

1. Probing Gene Regulation
   - Microarray technology, data matrix
   - Questions / applications

2. The clustering problem
   - Finding structure in unlabelled data
   - Two types of algorithms

3. K-means clustering
   - The basic algorithm
   - Fuzzy k-means

4. Machine learning formulation
   - Generative models and learning
   - K-means vs. Expectation Maximization

5. Hierarchical Clustering
   - Basic algorithm
   - Distance measures

6. Evaluating Clustering Results
   - Hypergeometric distribution

Module III: Gene regulation and networks

- Foundations
  - Clustering: Parametric (K-means, EM), Non-param (Hierarchical)
  - Classification: Generative (Naïve Bayes) / Feature selection
  - Regulatory Motif Discovery: Gibbs Sampling, EM, Information

- Frontiers:
  - Regulation: de novo motifs, target prediction, TFs/microRNAs
  - Networks: network inference, function prediction, predict expression
  - Epigenomics: chromatin states, activity signatures, disease SNPs