Clustering

Intro to clustering

Overview

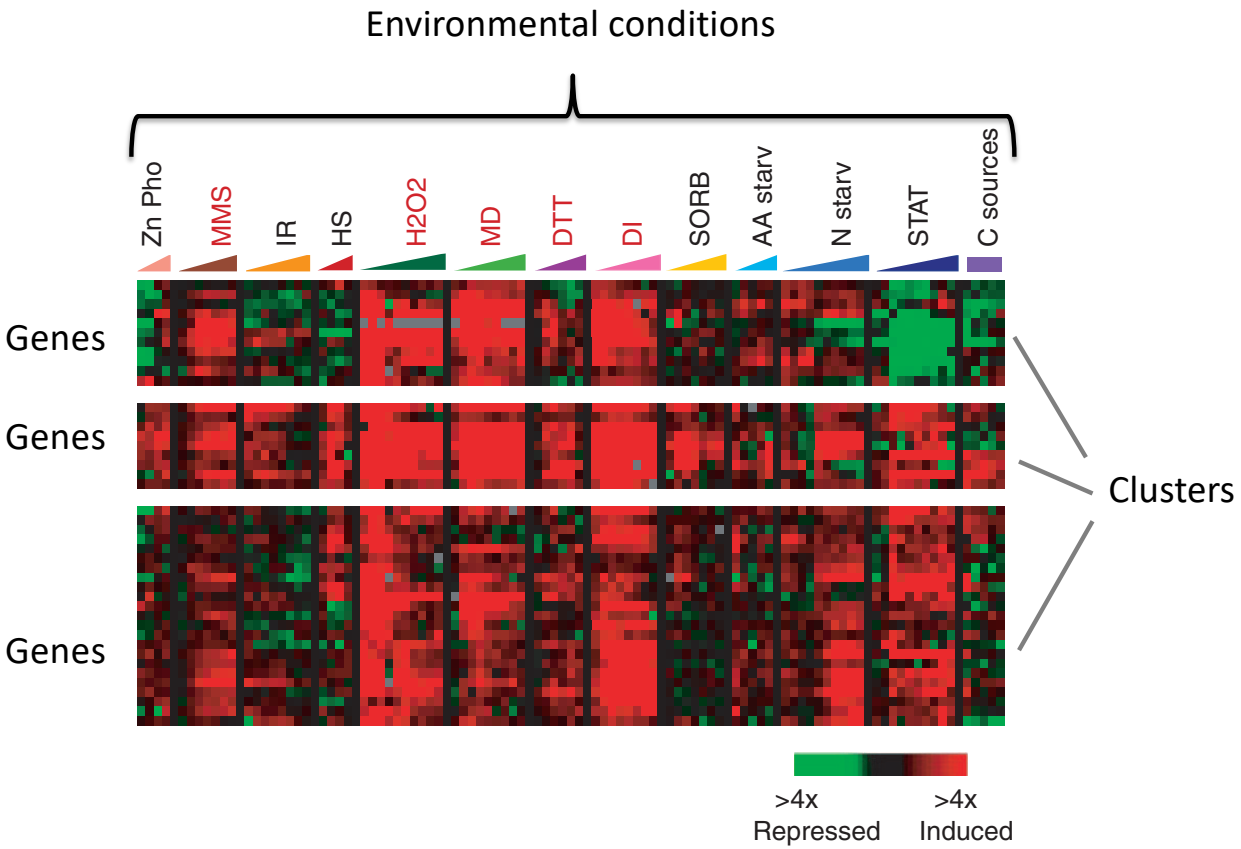
- Clustering task definition
- Classes of clustering methods
  - Flat
  - Hierarchical
  - etc.
- Distance metrics

Task definition: clustering gene expression profiles

- *Given*: expression profiles for a set of genes or experiments/individuals/time points (whatever columns represent)

- *Do*: organize profiles into clusters such that
  - profiles in the same cluster are highly similar to each other
  - profiles from different clusters have low similarity to each other

Example output of clustering



Gasch & Eisen, 2002

Motivation for clustering

- Exploratory data analysis
  - understanding general characteristics of data
  - visualizing data

- Generalization
  - infer something about an object (e.g. a gene) based on how it relates to other objects in the cluster

The clustering landscape

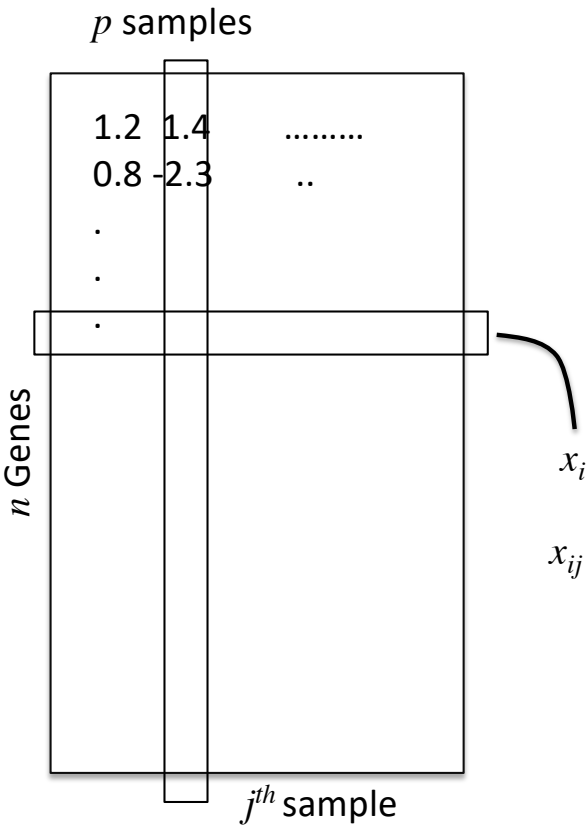
- There are many different clustering algorithms

- They differ with respect to several properties
  - hierarchical vs. flat
  - hard (no uncertainty about which profiles belong to a cluster) vs. soft clusters
  - overlapping (a profile can belong to multiple clusters) vs. non-overlapping
  - deterministic (same clusters produced every time for a given data set) vs. stochastic
  - distance (similarity) measure used

Distance measures

- Central to all clustering algorithms is a measure of distance between objects being clustered
- Clustering algorithms aim to group “similar” things together
  - Optimize some measure of within cluster similarity
- Defining the right similarity or distance is an important factor in getting good clusters
- Most algorithms will work with symmetric dissimilarities
- Dissimilarities may not be distances

Notation for gene expression data



Data matrix:

Is a tall matrix of numbers with rows corresponding to genes and columns corresponding to conditions

x_i Expression profile of i^{th} gene

x_{ij} Expression value of i^{th} gene in j^{th} sample

Different dissimilarity measures

- Euclidean distance between two vectors x_i and x_k

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2}$$

- Manhattan distance

$$d(x_i, x_k) = \sum_{j=1}^{p} \left| x_{ij} - x_{kj} \right|$$

Distance Metrics

Properties of a **metric** or **distance function**

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

$\text{dist}(x_i, x_j) = 0$ if and only if $x_i = x_j$ (identity)

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

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

Summary

- Clustering involves grouping similar entities into sets
- A common exploratory data analysis technique
- There are many variations of clustering methods
- Core to clustering is the definition of “similarity” or “distance”