CSE 527 Lecture 4, 10/08/03

Notes by Ana Kristine Torgerson" atorgers@u

- Case study, continued
  - Sporulation summary
    - What they did
      - Measured mRNA expression levels of all 6200 yeast genes at 7 times points in a (loosely synchronized) sporulating yeast culture
      - Plus some more standard tests and controls
    - What they learned
      - 3-10X increase in number of genes implicated in various subprocesses
      - Several subsequently verified by direct knockouts
    - Where computation fits in
      - Automated sample handling
      - Image analysis
      - Data storage, retrieval, integration
      - Visualization
      - Clustering
      - Sequence analysis
    - More on computation
      - Similarity search – given a loosely defined sequence "motif", e.g. a transcription factor binding site, scan genome for matches
        - Which genes have MSE element?
        - e. g. weight matrix models, Markov models
      - Motif discovery – given a collection of sequences presumed to contain a common pattern, e.g. a transcription factor binding site, find and characterize it
        - What motifs are common to early-middle genes?
        - e. g. MEME, Gibbs sampler, footprinter, …
      - Finding groups of sequences that plausibly contain common sequence motifs
        - e. g. clustering (co-varying because co-regulated?)
  - Chu's "supervised" clustering
    - Hand picked ~40 prototype genes
      - Significant variation in data set
      - Known function
    - Hand segregated into 7 groups ("early", …)
    - Assigned all others to "nearest" group
      - Based on Pearson correlation to average of prototypes
    - Ordered within groups by correlating to neighboring groups
  - Pearson correlation

$$\text{Pearson} := \left| \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \cdot \sigma_y} \right|$$

Larger → more similar

Euclidean distance

$$\text{Euclidean} := \left| \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \right|$$

Smaller → more similar

Critique

+	-
Informed clustering	Subjective, maybe counter to data
Knockout verification	
	"Peak time" simplistic
	Reproducibility? Replicatability?
	"Normalize" data?

Other analysis possible

- Lagged correlation
- Correlation other then Pearson's
- Other approaches to clustering

Principle component axis is line with greatest variance in data

- Used in statistical analysis

Turns out that two axis (variables) define most variation in data

- No clear division in data (not easily divided into 7 groups)
- Projection onto 3 axes gives more distinct groupings

Homework

Choose paper on microarrays, read it, try to think about issues (i.e. analysis, other research, etc). What surprised you about this paper? Send a paragraph of interesting observations via email by Wednesday or Thursday.

Clustering

Traced back to Aristotle, big push in 1950's

Ways to use

- Cluster genes
  - Those clustered together may be linked
- Cluster experiments
  - Drug/tumor function may be separated into groups
- Both

Exploratory technique

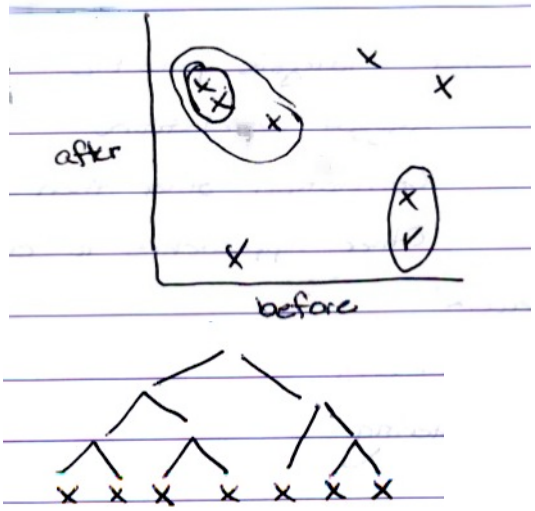
- Used to generate, rather then test, hypothesis

Many methods

- All find clusters, some more applicable then others
- How to compare methods?

Hierarchical clustering

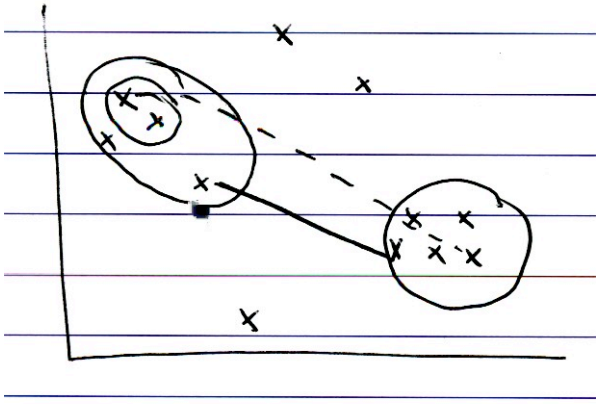
- (Dis)similarity measure individuals and groups



Each point a gene

At every stage of algorithm, merge two most similar "pairs"
Gives a tree structure

Height indicates similarity between "pairs"



___ Most similar pair comparison (single link)
- - - Least similar pair comparison (complete link)

Average link can also be used