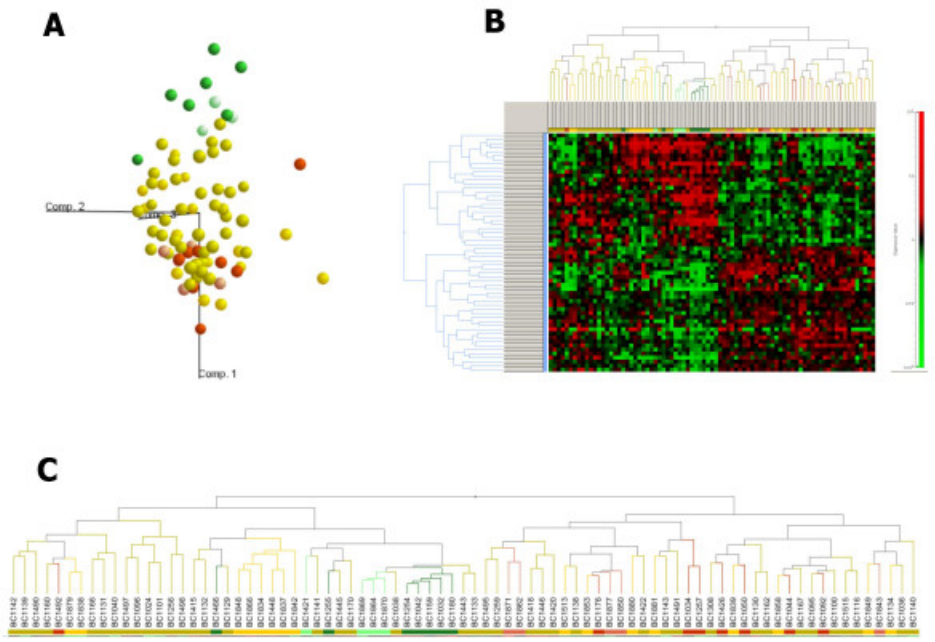
CMSC423: Bioinformatic Algorithms, Databases and Tools
Lecture 21

Microarray data analysis

RNA folding

Hierarchical clustering

- UPGMA (remember from phylogenetic trees?)
  - compute distance between genes (e.g. euclidean distance of expression vectors)
  - join most similar genes
  - repeat
  - Key element – compute distance between a gene and a cluster, or between two clusters – average distance between all genes in the two clusters



k-means clustering

- Split data into exactly k clusters
- Basic algorithm:
  - Create k arbitrary clusters - pick k points as cluster centers and assign each other point to the closest center
  - Re-compute the center of each cluster
  - Re-assign points to clusters
  - Repeat
- Another approach: pick a point at and see if moving it to a different cluster will improve the quality of the overall solution. Repeat!

k-means clustering

- Measure of cluster goodness: mean square distance of each point to its nearest cluster center.
- d(Points, Centers) = sum(d(point i, center)^2) / n

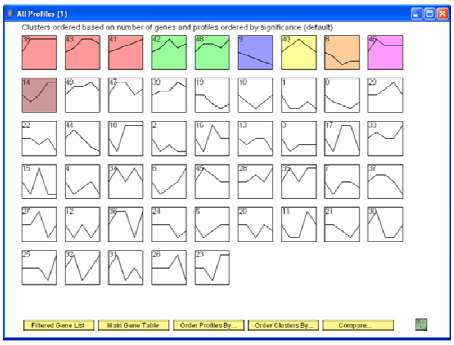
k-means clustering demo

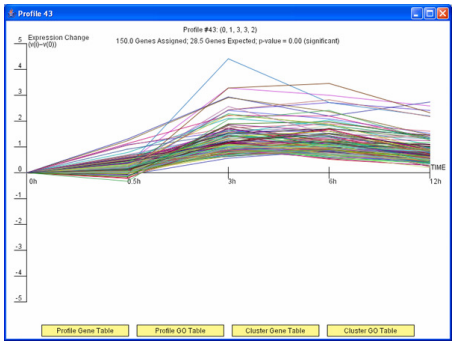
Other clustering methods

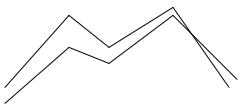
- Principal component analysis
  - "rotate" cloud of points until clusters become obvious
  - essentially projection onto the appropriate plane or line
- Self Organizing Maps
  - based on neural networks

- Clustering of time-series data

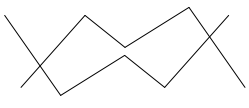
Clustering of time-series data



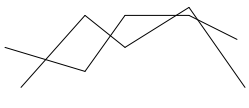




correlated



anti-correlated



un-correlated

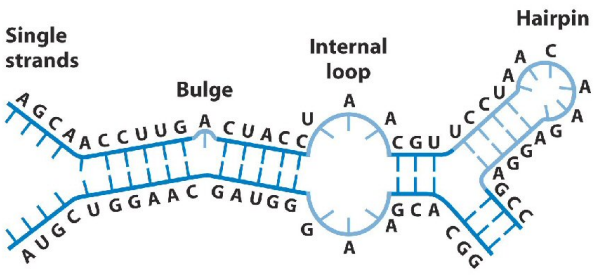
http://www.cs.cmu.edu/~jernst/stem/

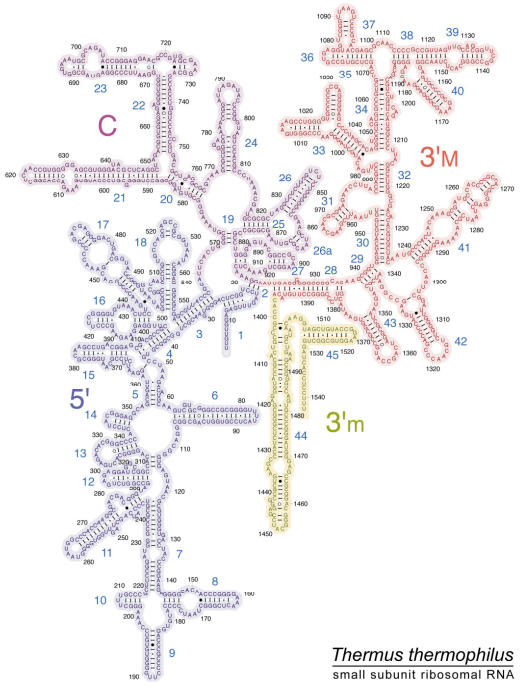
Assessing significance

- All clustering methods produce clusters EVEN IF NO CLUSTERS EXIST!!!
- Need to associate a confidence that the clusters are real
- Basic approach – bootstrapping
  - randomly shuffle data labels (e.g. disease/no disease, or time-point)
  - recompute clustering
  - count how often the initial clusters appear in random data

RNA folding

- Function of RNA molecules depends on how they fold, based on nucleotide base-pairing

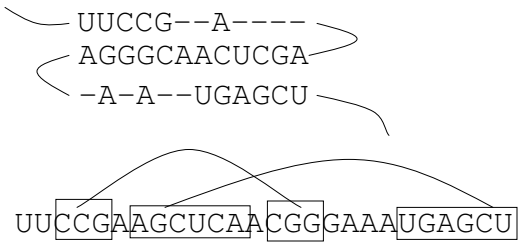




Types of structures

- Nested (hairpin)

- Pseudo-knots



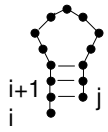
Nussinov's algorithm

- Assumes no pseudo-knots
- Dynamic programming approach – maximize # of pairings

- S – string of nucleotides representing the RNA molecule
- Sub-problem – F[i,j] – score of folding just S[i..j]
- Initial values: F[i-1,i] = F[i,i] = F[i, i+1] = 0

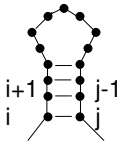
Nussinov's algorithm

F[i,j] is the maximum of:



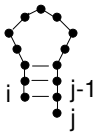
I. F[i+1,j]

S[i] unpaired



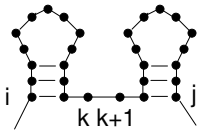
III. F[i+1,j-1] + 1
if S[i+1] complementary to S[j-1]

S[i] paired with S[j]



II. F[i,j-1]

S[j] unpaired



IV. $\max_k$ F[i,k]+F[k+1,j]

Branch

Questions

- In what order do we fill the dynamic programming table?
- How can we ensure that "loops" consist of at least k nucleotides?