CMSC423: Bioinformatic Algorithms, Databases and Tools
Lecture 20

Motif finding

Microarray data analysis

forward-backward why backward

Motif finding

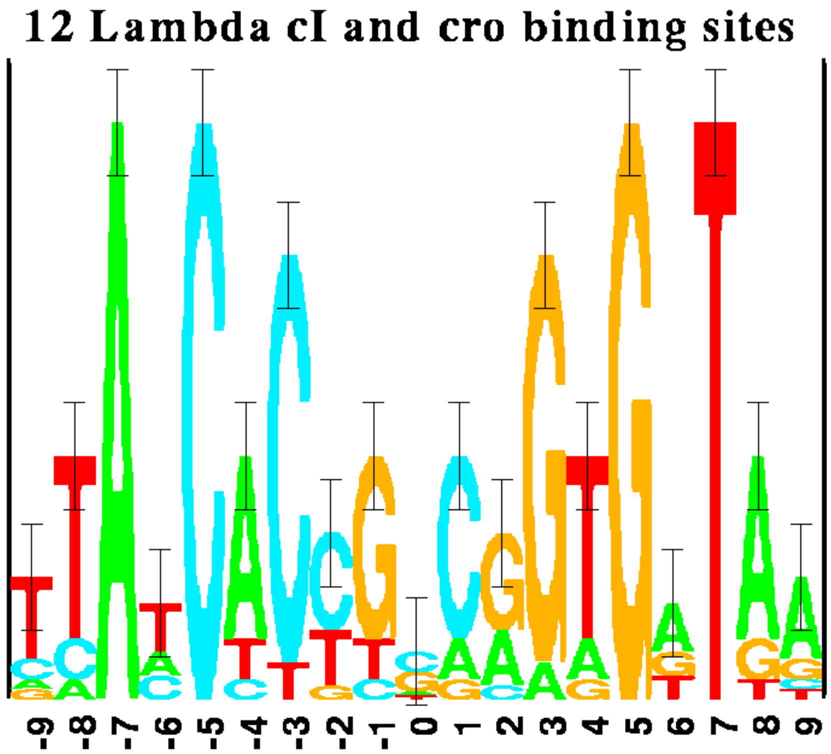
- Problem: given a set of genes, are there any "motifs" common in the upstream region?
- Motifs could be transcription factor binding sites or other regulatory elements
- Parameters:
  - length of upstream region (e.g. 5kbp)
  - length of motif (10 bp)
- Complexity: HIGH
  - look through all possible combinations of k-mers for N genes
- Solution: probabilistic local search (Gibbs sampling, expectation-maximization, etc.)

```
atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
acccctattttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa
tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga
gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
tcccttttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag
gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat
aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag
ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa
```

N - # of genes, L - # length of upstream region, K-motif length
$(L-K+1)^N$ possible choices

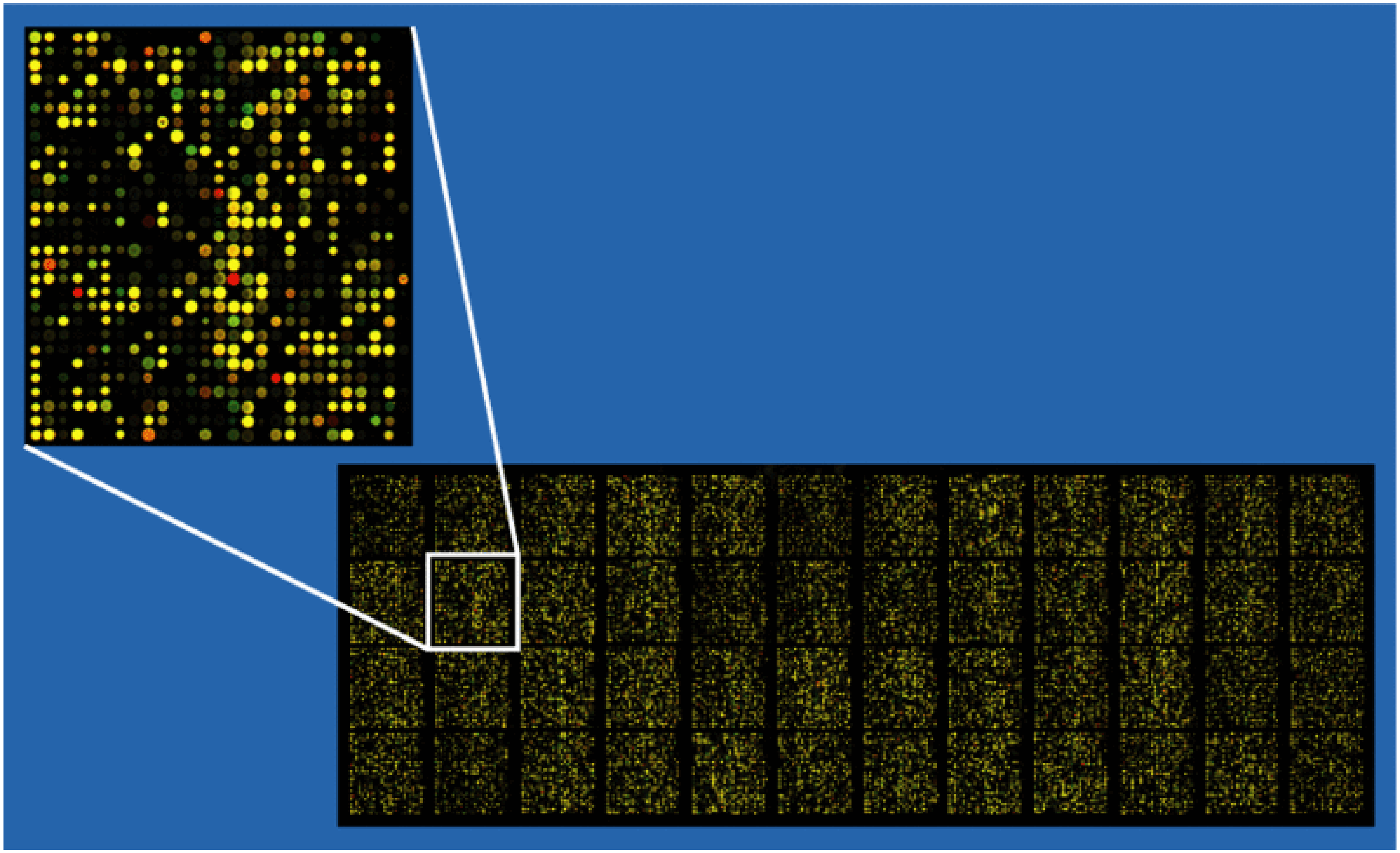
Probabilistic search

- Outline:
  - Pick a set of random k-mers (one from each sequence)
  - Build a multiple-alignment profile – frequency of each nucleotide at each of the k positions
  - Remove one sequence at random and find the k-mer within it that best matches the profile (p(k-mer|profile)= product of frequencies for k-mer nucleotides in profile table)
  - Recompute profile and repeat



	0	1	2	3
A	0.2	0.05	0.3	0.0
C	0.7	0.1	0.0	0.3
G	0.03	0.5	0.7	0.3
T	0.07	0.35	0.0	0.4

Microarray data analysis

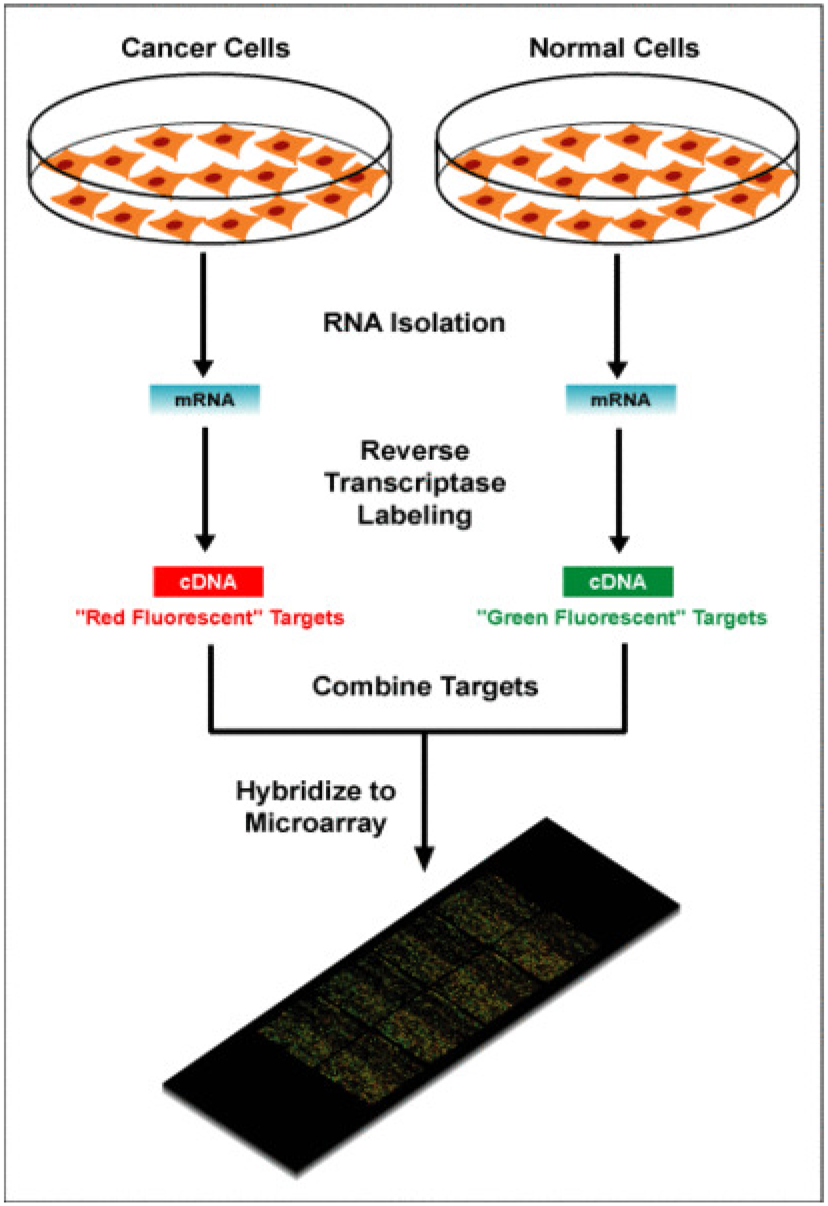


Types of microarrays

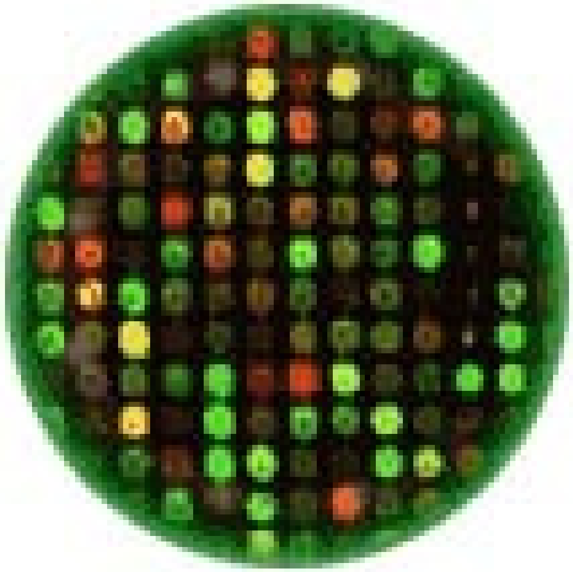
- By technology
  - Spotted
  - Affymetrix
  - Nimblegen
  - Illumina

- By information
  - cDNA (genes or parts of genes)
  - DNA (e.g. sequencing by hybridization)
  - Tiling arrays (whole genome)
  - Protein

Typical microarray experiment



- Difference in color intensity indicate differences in gene expression levels
- Red – expressed in sample
- Green – expressed in control
- Yellow – expressed in both
- Black – expressed in neither



Typical data analysis process

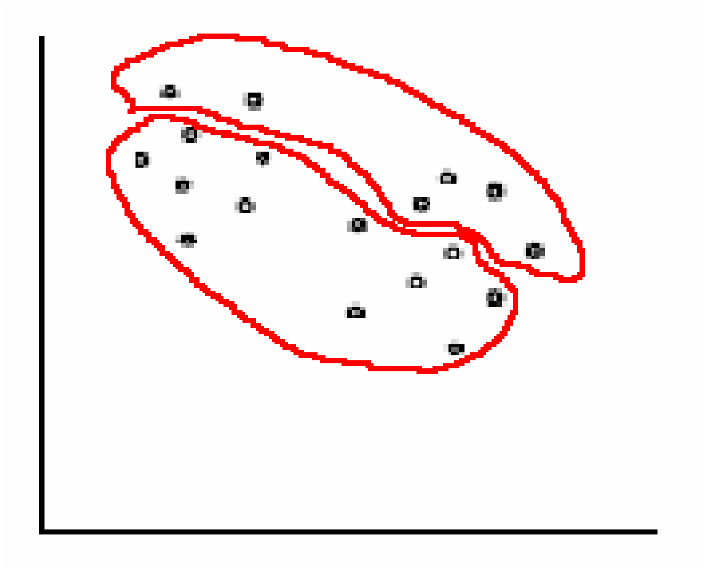
- Image analysis
  - find spots
  - find errors (air bubbles, fingerprints, smears, etc.)
- Normalization
  - make sure total intensity for green and red is the same (otherwise cannot compare intensities)
- Clustering
  - which genes have similar expression?
  - which genes are expressed similarly during a disease?
  - which genes have similar expression patterns over time (time-course experiments)?

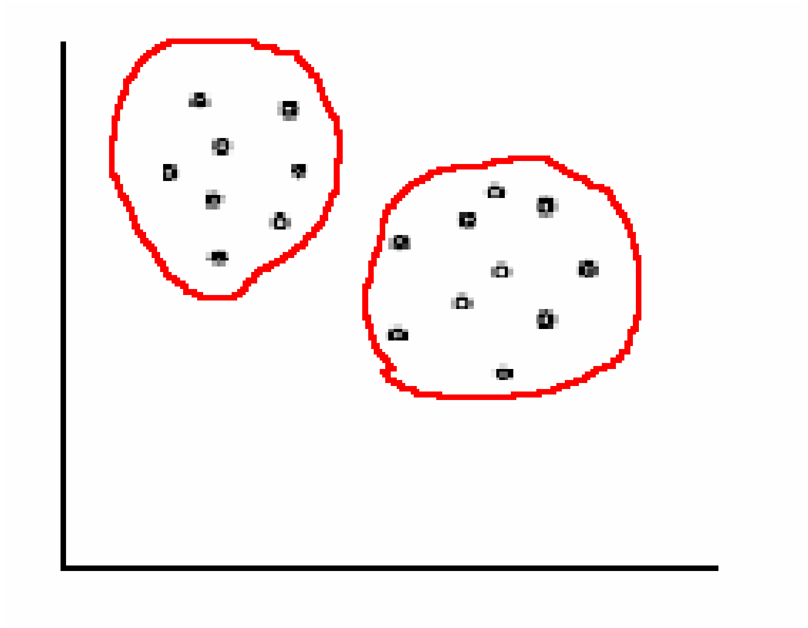
Data clustering

- Agglomerative
  - Start with single observations
  - Group similar observations into the same cluster
- Divisive
  - All datapoints start in the same cluster
  - Iteratively divide cluster until you find good clustering
- Hierarchical
  - Build a tree – leaves are datapoints, internal nodes represent clusters

Measures of goodness of clustering

- Homogeneity
  - All points in a cluster must be similar
- Separation
  - Points in different clusters are disimilar



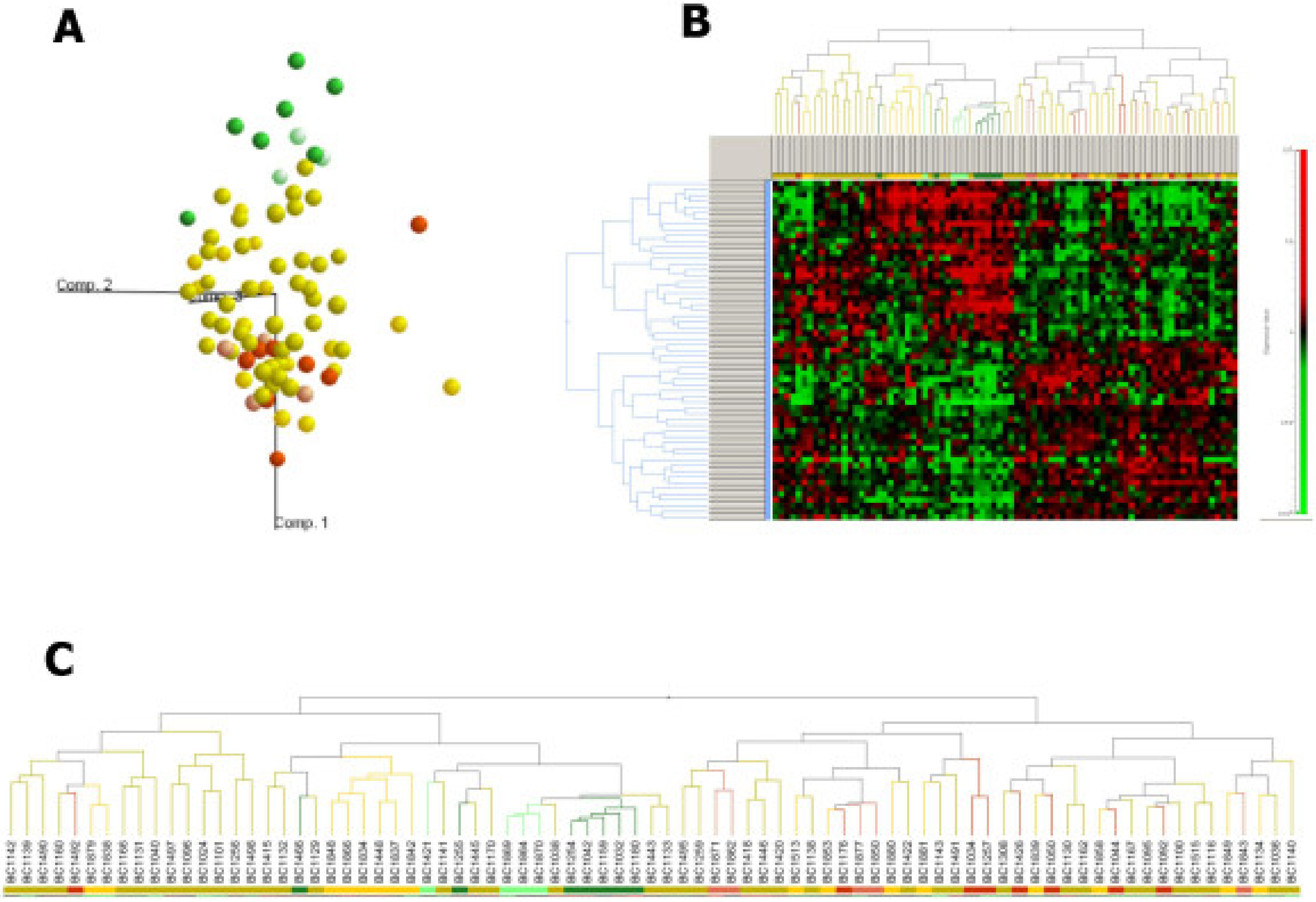


Microarray clustering

- For each gene can be viewed as an array of numbers
  - expression of gene at different time-points
  - expression of gene in different conditions (normal, variants of a disease, etc.)
- Each time-point or tissue sample can also be viewed as an array of numbers
  - expression levels for all genes
- Basic idea: cluster genes and/or samples to highlight genes involved in disease

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


A
B
C
Comp. 2
Comp. 3
Comp. 1

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!