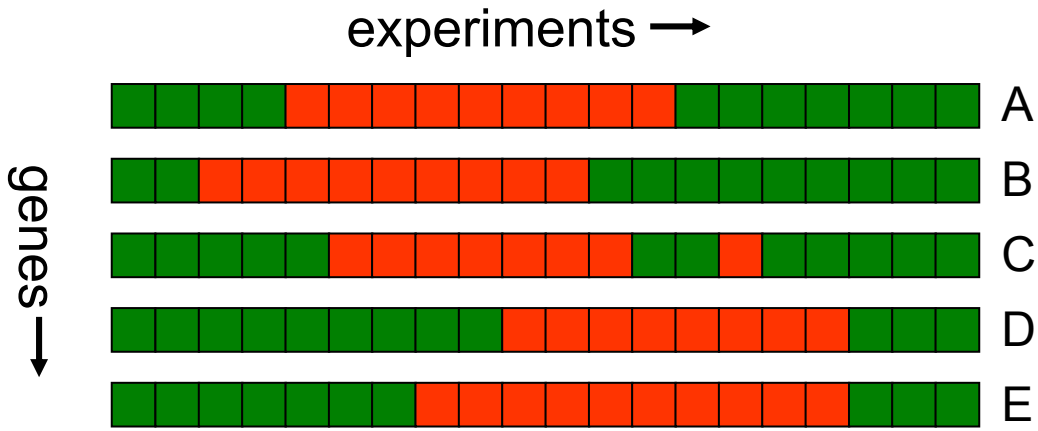
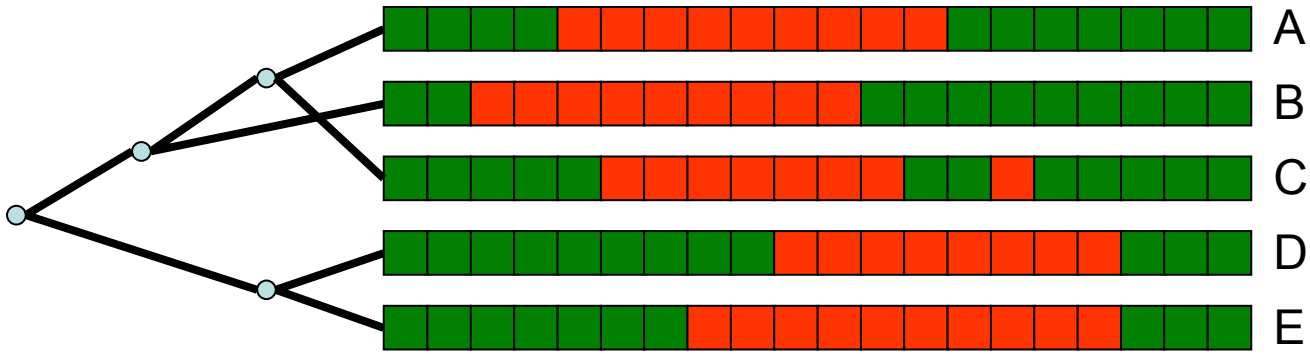
K-ary Clustering With Optimal Leaf Ordering for Gene Expression Data

Ziv Bar-Joseph, Erik Demaine, David Gifford, Angèle Hamel, Tommi Jaakkola, Nathan Srebro



Goal: Quickly and easily arrange the data for further inspection

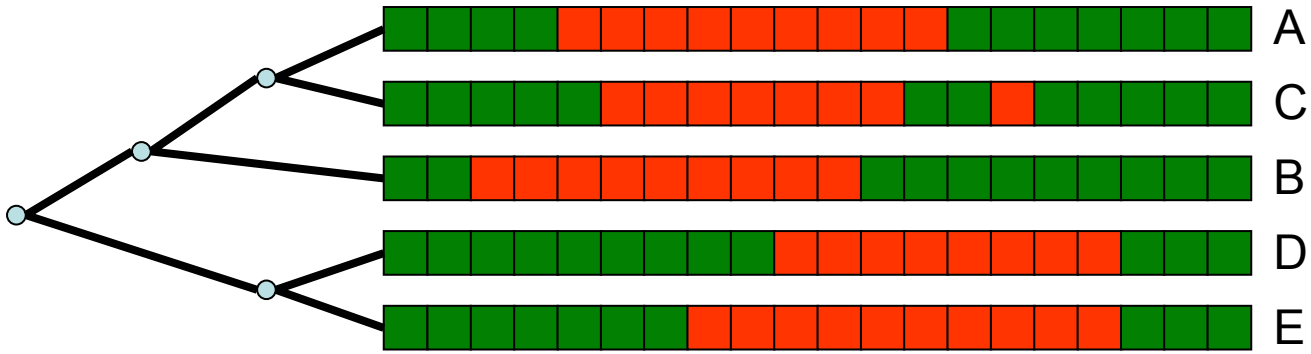
Hierarchical Clustering



- Greedily join nearest cluster pair [Eisen 1998]

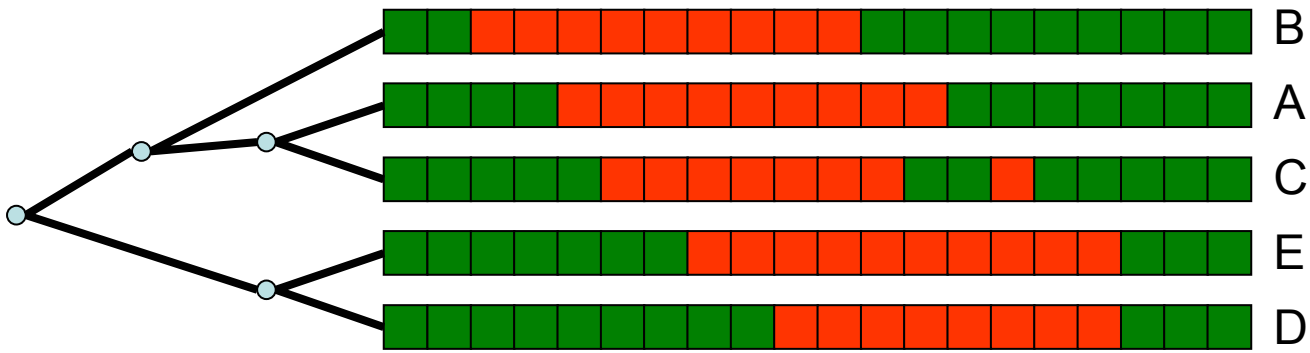
nearest: we use correlation coefficient (normalized dot product)
can use other measures as well

Hierarchical Clustering

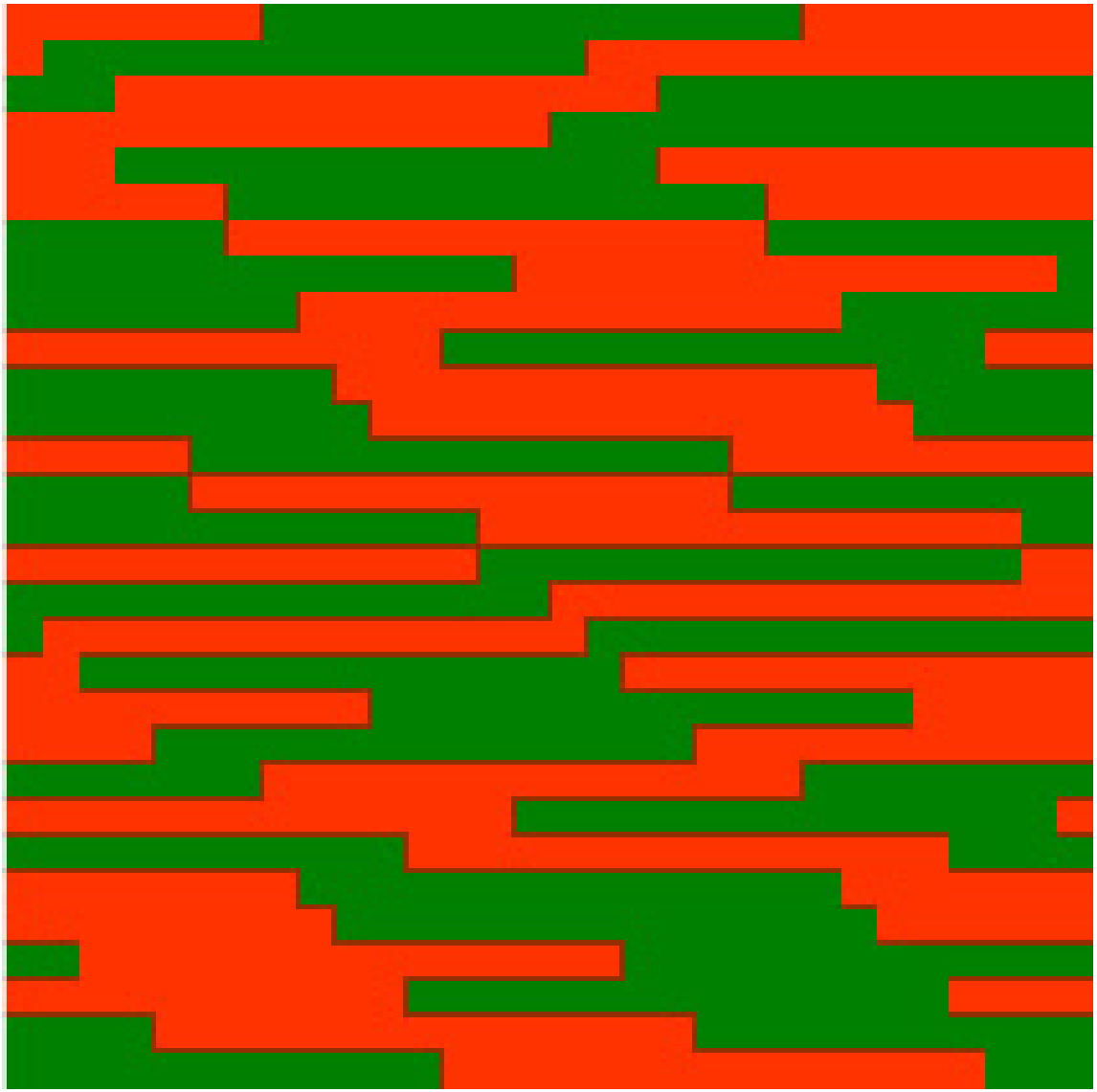


- Greedily join nearest cluster pair [Eisen 1998]
- Optimal ordering: minimize summed distance between consecutive genes
  - Criterion suggested by Eisen
  - n^3 algorithm [Bar-Joseph et al 2001 + improvements]

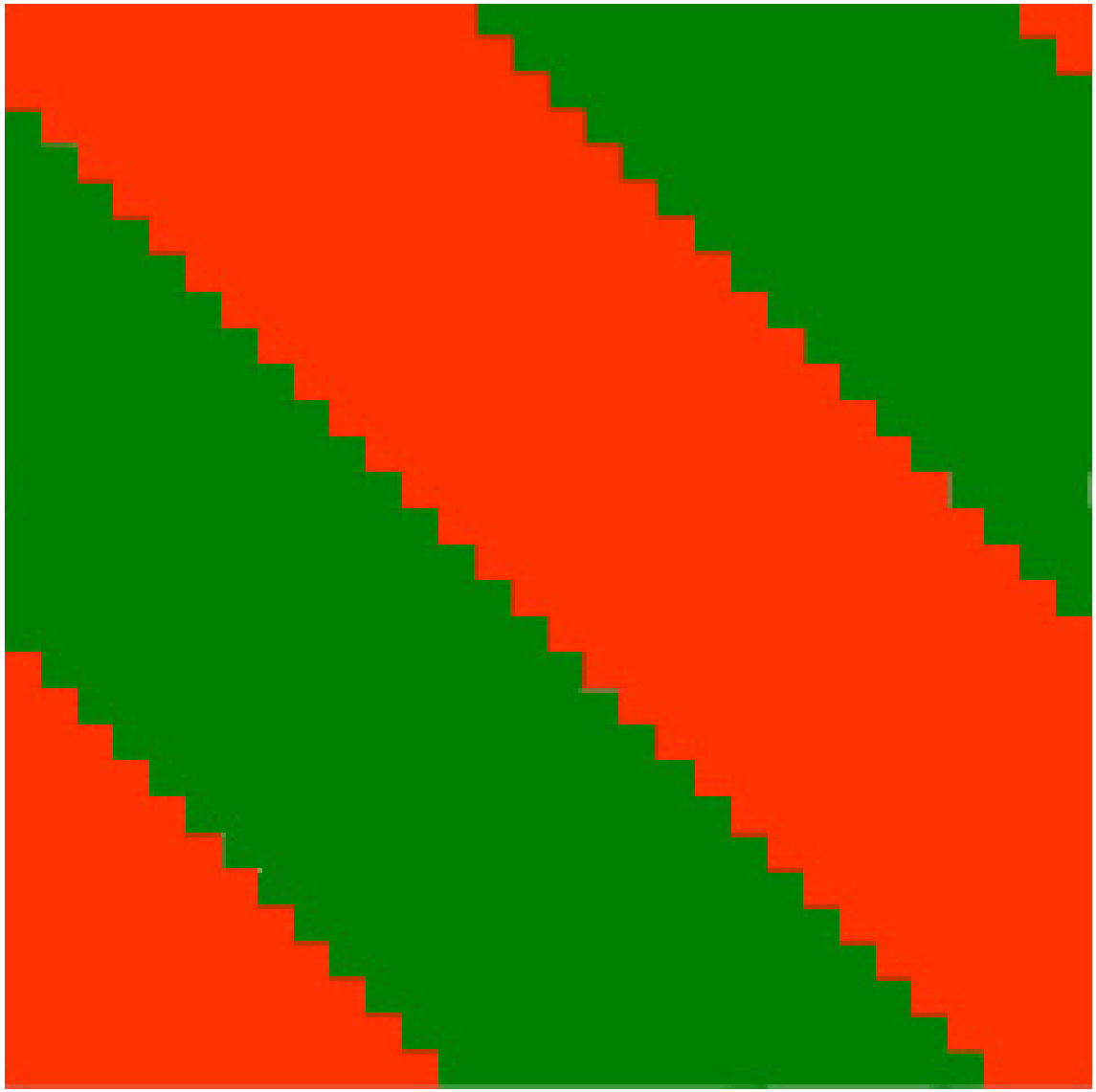
Hierarchical Clustering

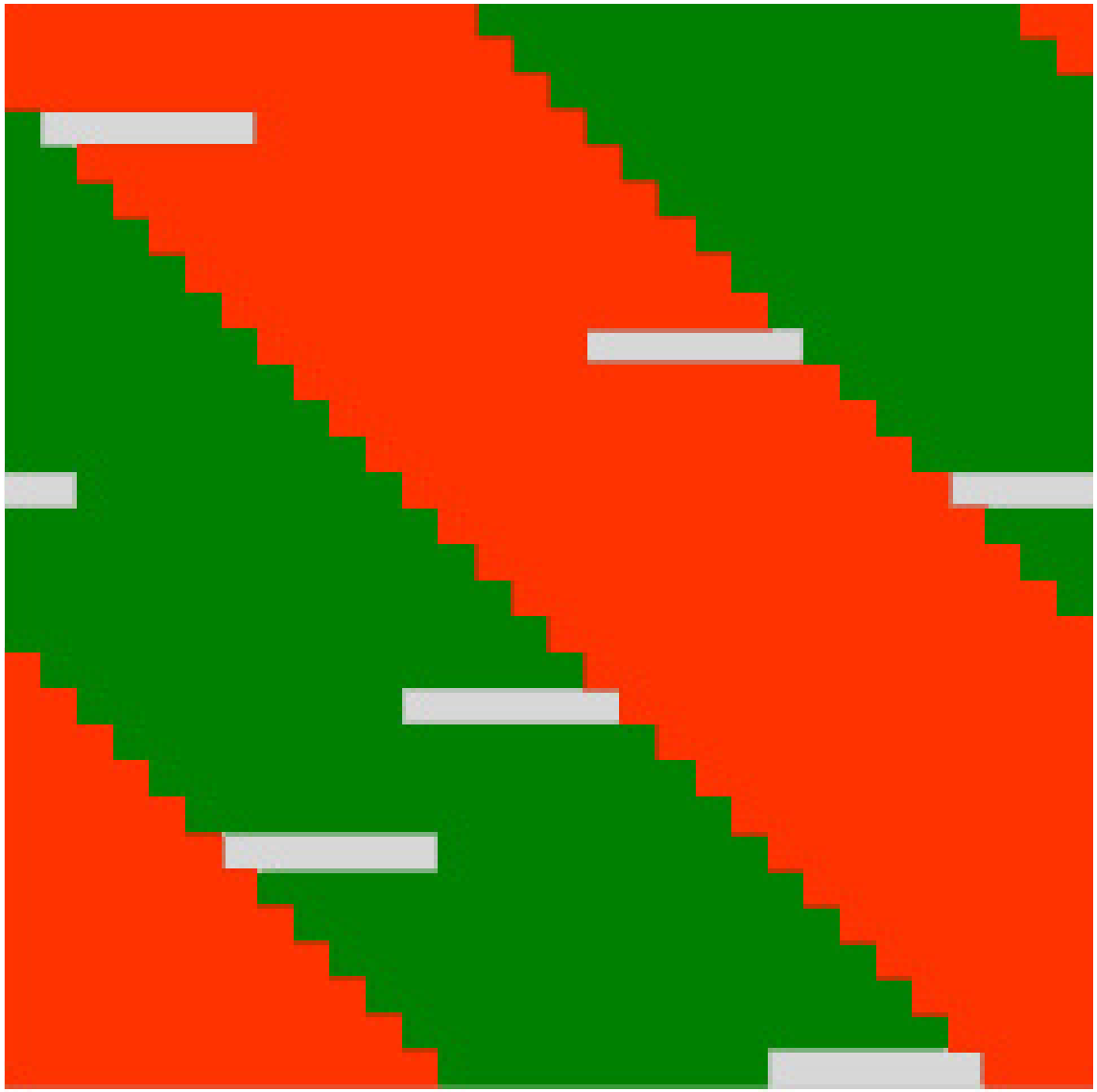


- Greedily join nearest cluster pair [Eisen 1998]
- Optimal ordering: minimize summed distance between consecutive genes
  - Criterion suggested by Eisen
  - n^3 algorithm [Bar-Joseph et al 2001 + improvements]

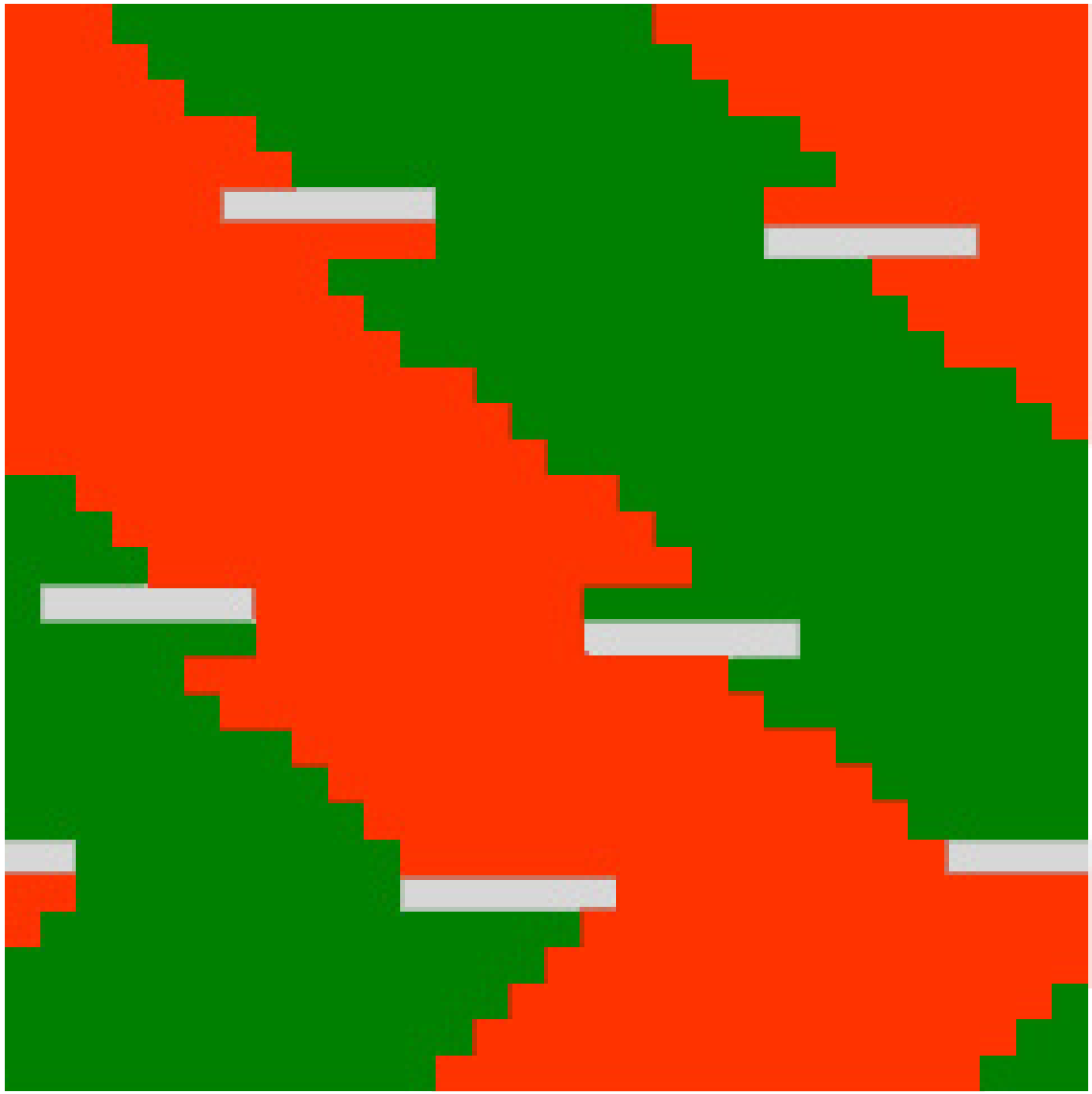


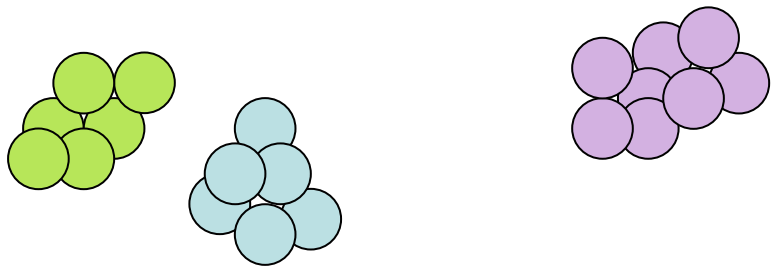
Optimally Ordered using Binary Clustering





Optimally Ordered using Binary Clustering



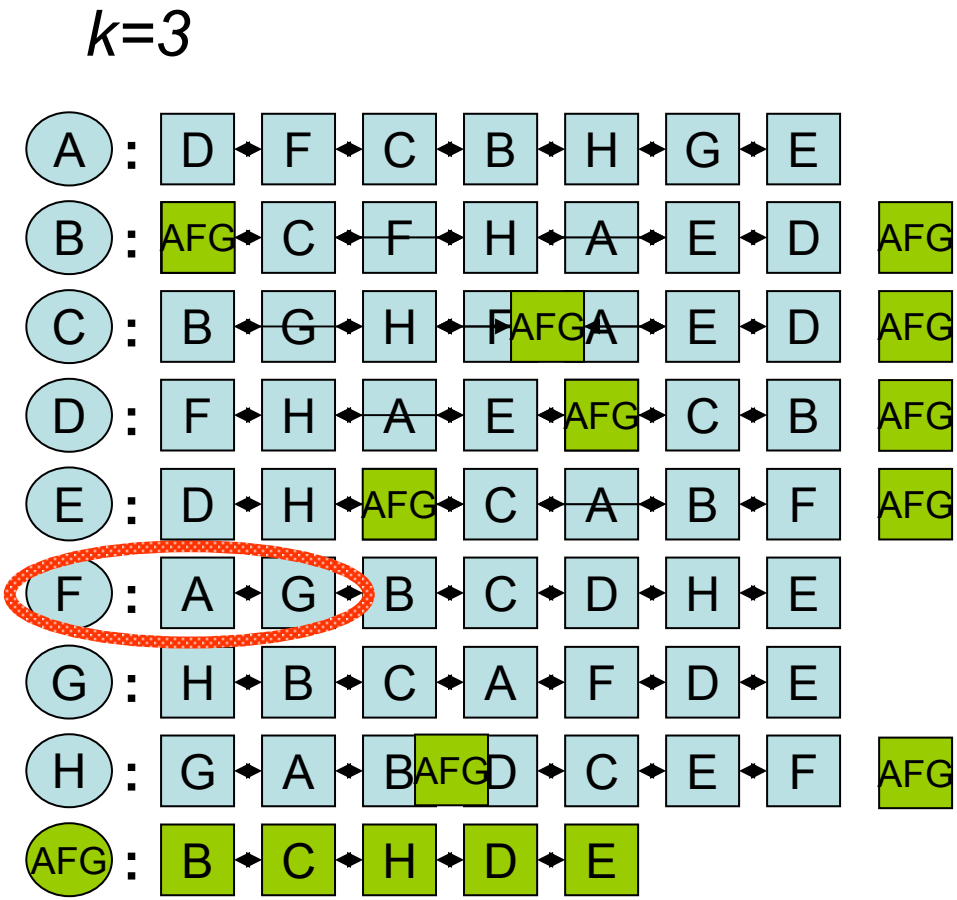


Main motivation for k-ary clustering: more information in k-wise distances leads to more robust clustering.

K-ary Hierarchical Clustering

- Greedily join tightest k clusters to form new cluster.
  - Tightest: maximal summed pairwise similarity
- Finding tightest k clusters: $O(n^k)$
- Running time: $O(n^{k+1})$
- Finding tightest k cluster is as hard as max-clique → W[1] hard, fixed parameter intractable: no *poly(n)·f(k)* algorithm

Sub-optimal Heuristic



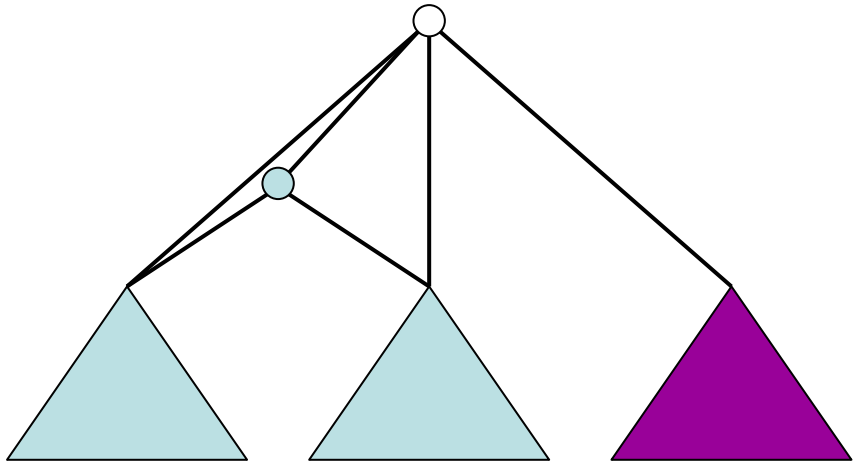
- In order to find tight group of *k* clusters:
  - For each cluster, consider the cluster at its *(k-1)* nearest neighbors
  - Choose tightest of these neighborhoods
- For each cluster, maintain ordered linked-list of neighbors

$$n^2 \log n + n\left(nk^2 + n \cdot \log n + n^2 + n^2\right) = O(n^3)$$

Heuristic vs. Optimal

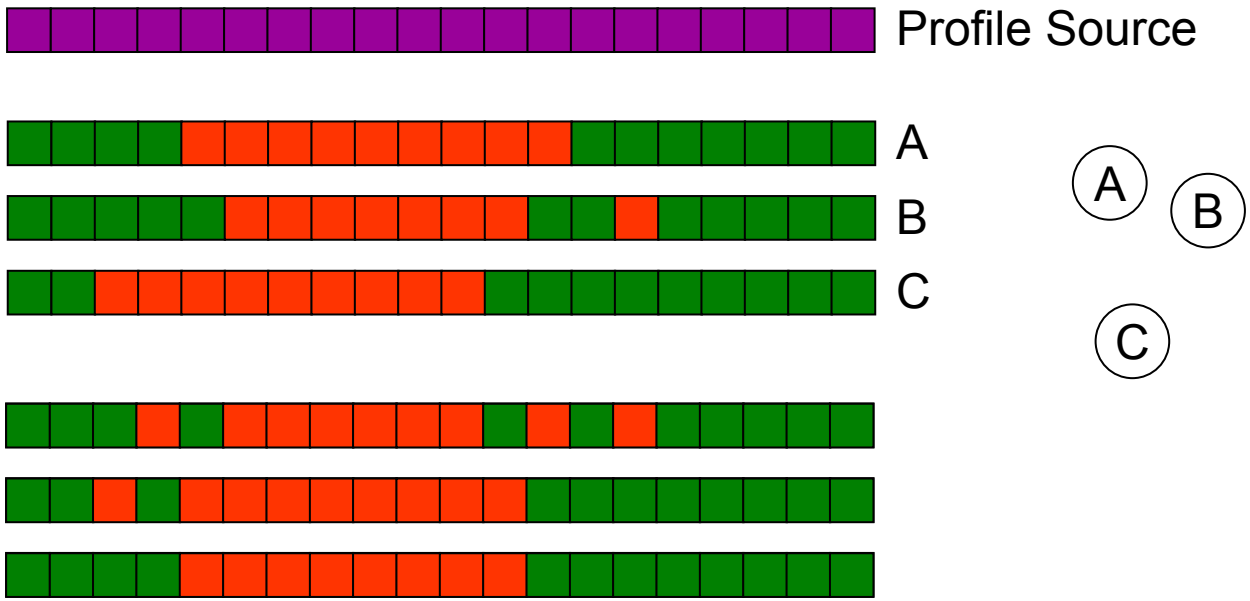
- Real data from 979 genes using *k*=*3*
- Heuristic 35 seconds vs Optimal 57 minutes
  (1.4 GHz Pentium III)
- Average node similarity: 0.6366 vs 0.6371
- For k=4, optimal is impractical

Allowing Variable Number of Children



If the data clearly indicates a subset of less than *k* clusters that are significantly distinct from the rest, cluster them.
(Otherwise, cluster the tightest *k*-neighborhood)

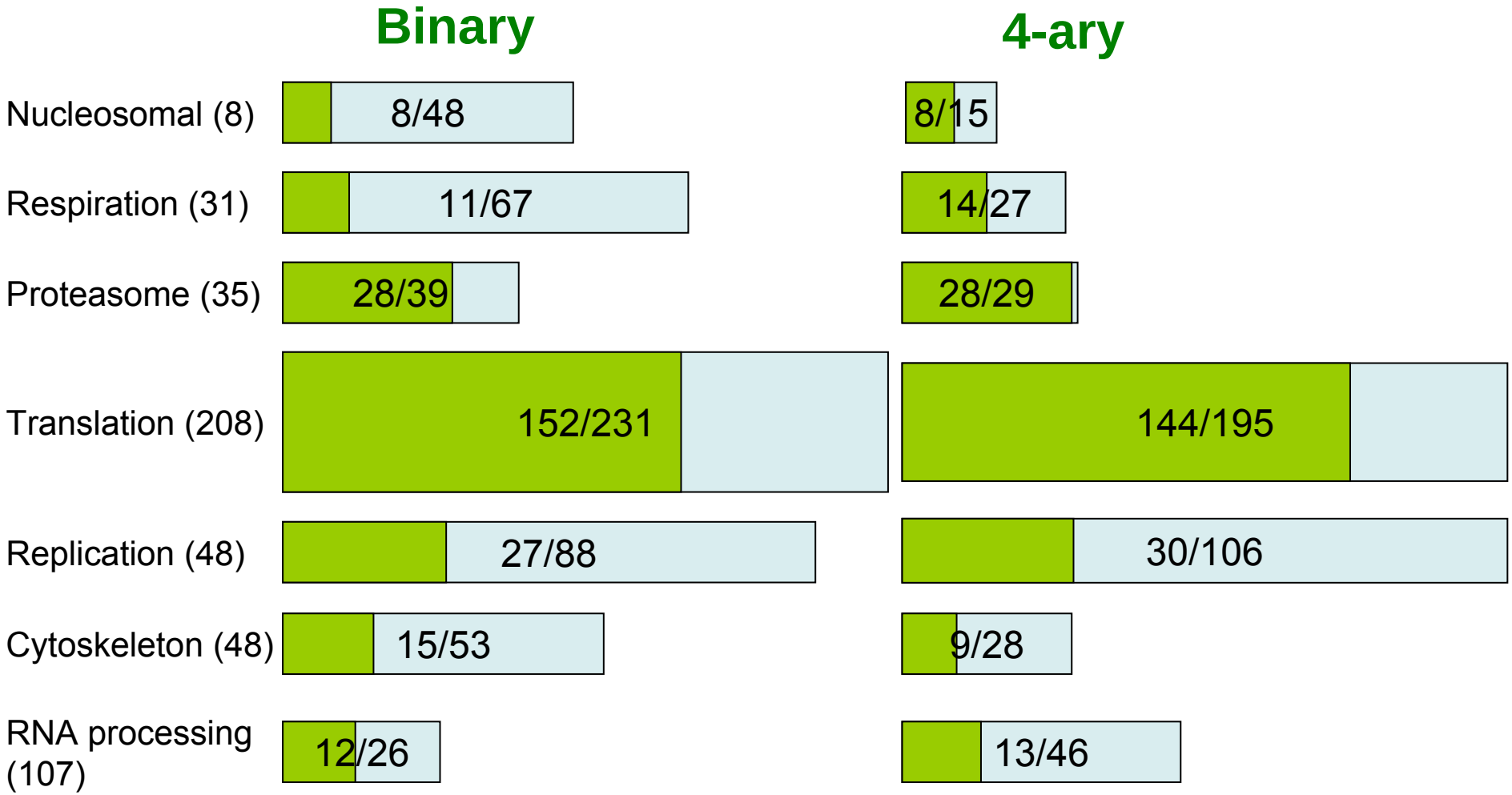
Permutation Test for Significance



Might A,B and C all result from the same profile source, with the variations in distances being random variations?

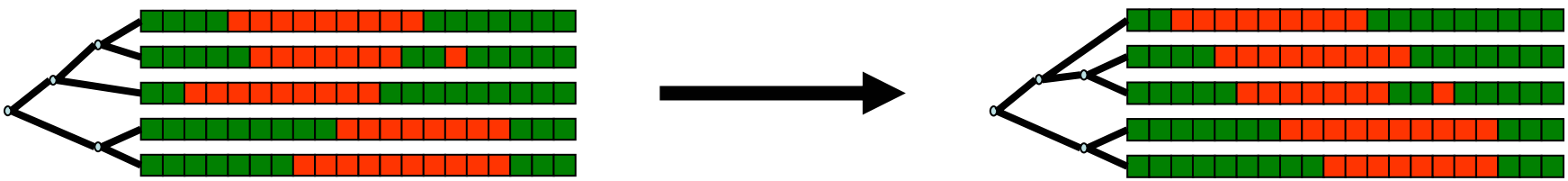
Pr(2nd closest pair in bootstrap < 2nd closest pair in data)

Binary vs. 4-ary Hierarchical Clustering



Clusters of yeast condition response expression [Eisen et al 1998] corresponding to MIPS categories: genes from category/cluster size

Optimal Ordering



Gene order consistent with hierarchical clustering tree that minimizes summed distance between consecutive genes

(constrained TSP)

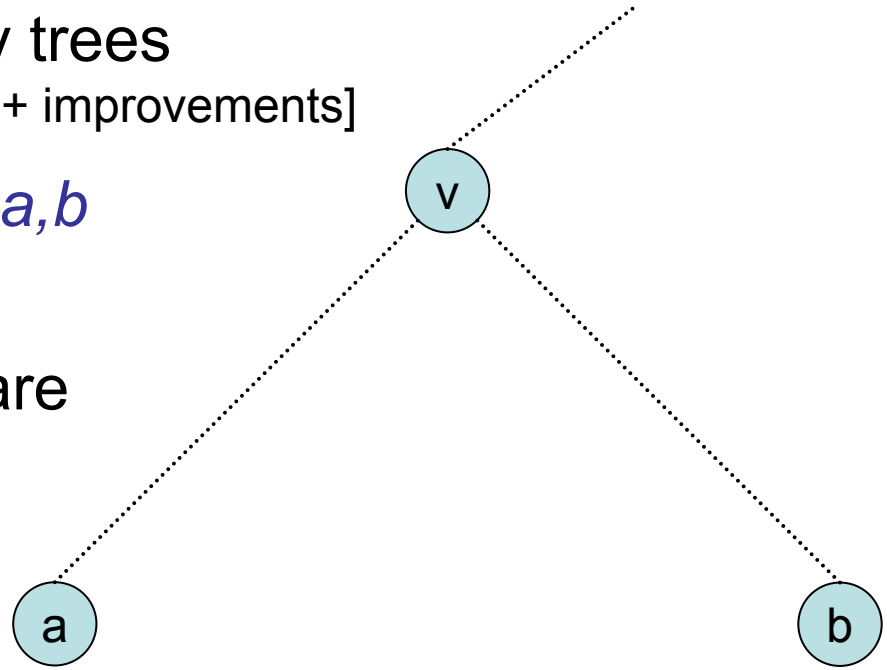
Optimal Ordering

for binary trees

[Bar-Joseph et al 01 + improvements]

v is least-common-ancestor of *a,b*

S(*a,b*)=Cost of opt ordering of descendents of *v*, such that *a,b* are at the two ends



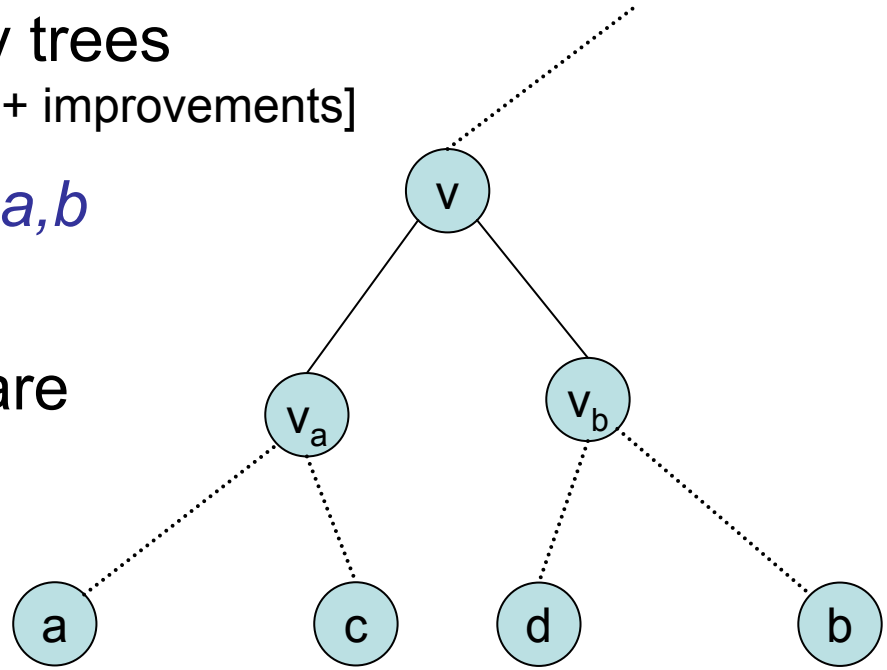
Optimal Ordering

for binary trees

[Bar-Joseph et al 01 + improvements]

v is least-common-ancestor of *a,b*

S(*a,b*)=Cost of opt ordering of descendents of *v*, such that *a,b* are at the two ends



$$S(a,b) = \min_{c,d} S(a,c) + \delta(c,d) + S(d,b) \quad \Rightarrow O(n^4)$$

$$T(a,d) = \min_{c} S(a,c) + \delta(c,d)$$

$$S(a,b) = \min_{d} T(a,d) + S(d,b) \quad \Rightarrow O(n^3)$$

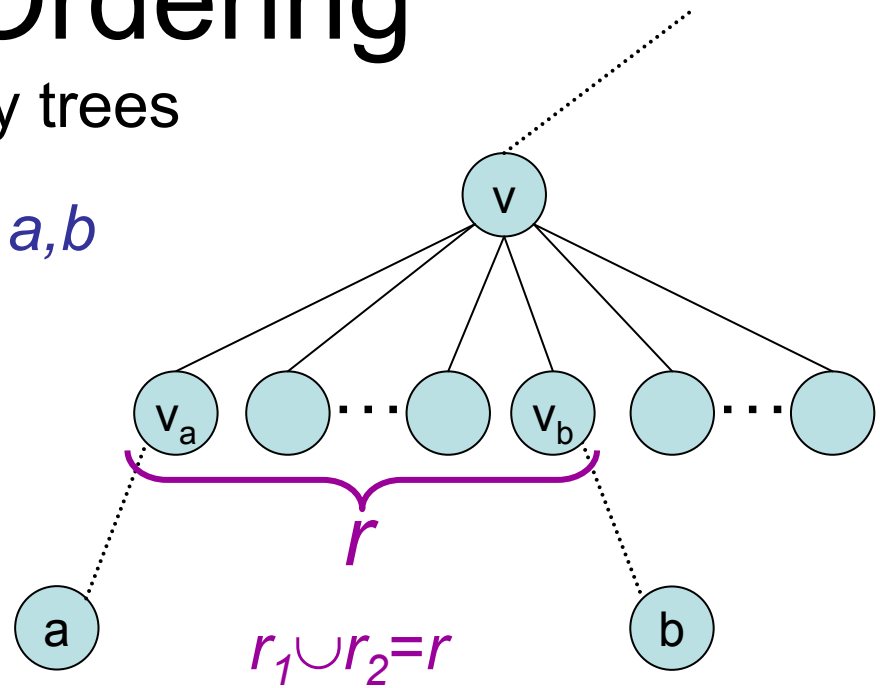
Optimal Ordering

for k-ary trees

v is least-common-ancestor of *a,b*

S(*a,b,r*)=Cost of opt ordering of descendents of *nodes in r*, such that *a,b* are at the two ends

r is a subset of children of *v*



$$S(a,b,r) = \min_{c \in r_1,\ d \in r_2} S(a,c,r_1) + \delta(c,d) + S(d,b,r_2)$$

Dynamic Programming

$r_2=\{v_b\}$

$$T(a,d,r_1) = \min_{c \in r_1} S(a,c,r_1)+\delta(c,d)$$

$$S(a,b,r) = \min_{d \in r_2} T(a,d,r_1)+S(d,b,r_2)$$

$O(n^3 2^k)$

$O(n^2 2^k)$ space

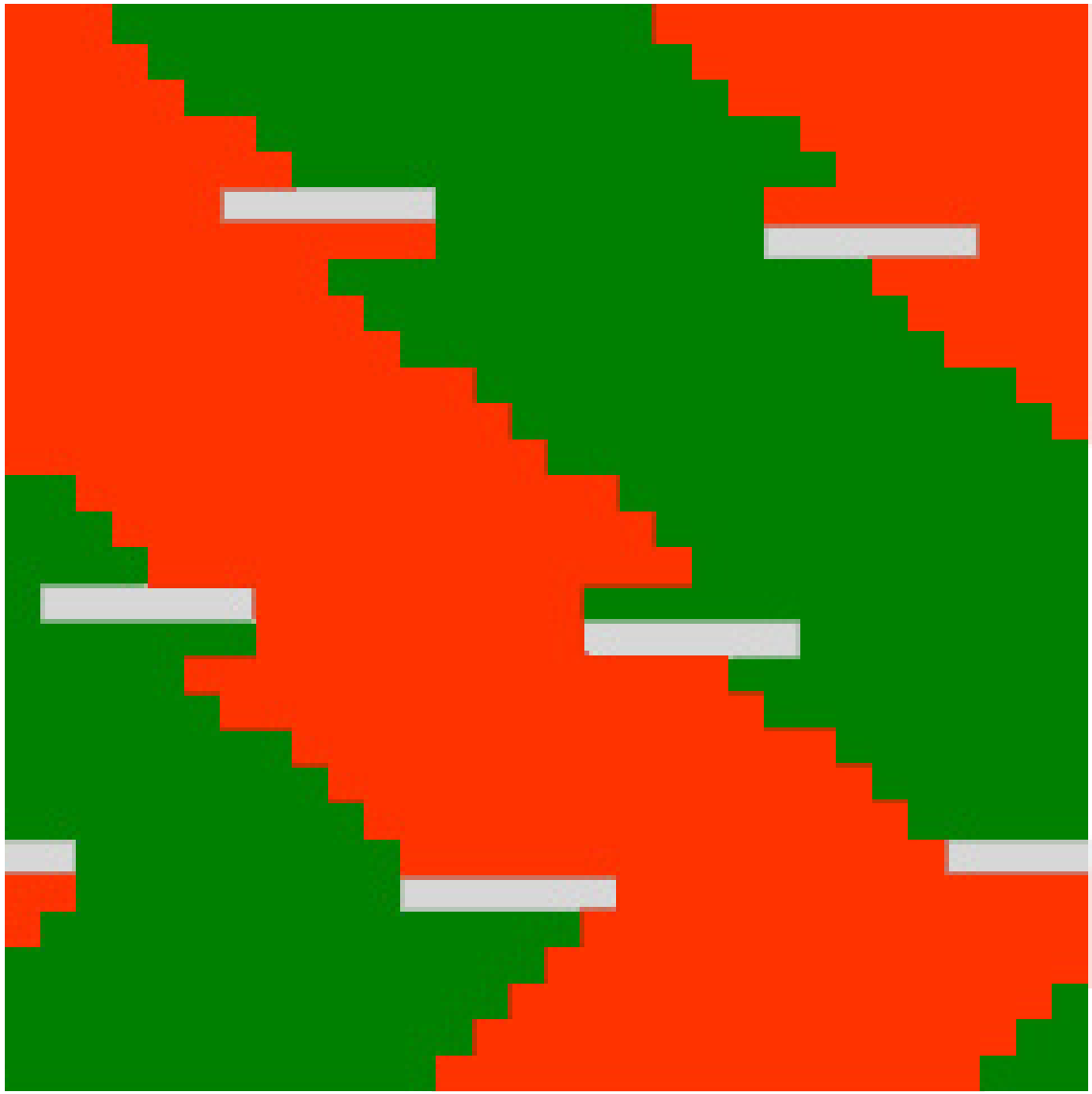
Divide and Conquer: no memoization

equipartitions: $|r_1|=|r_2|$

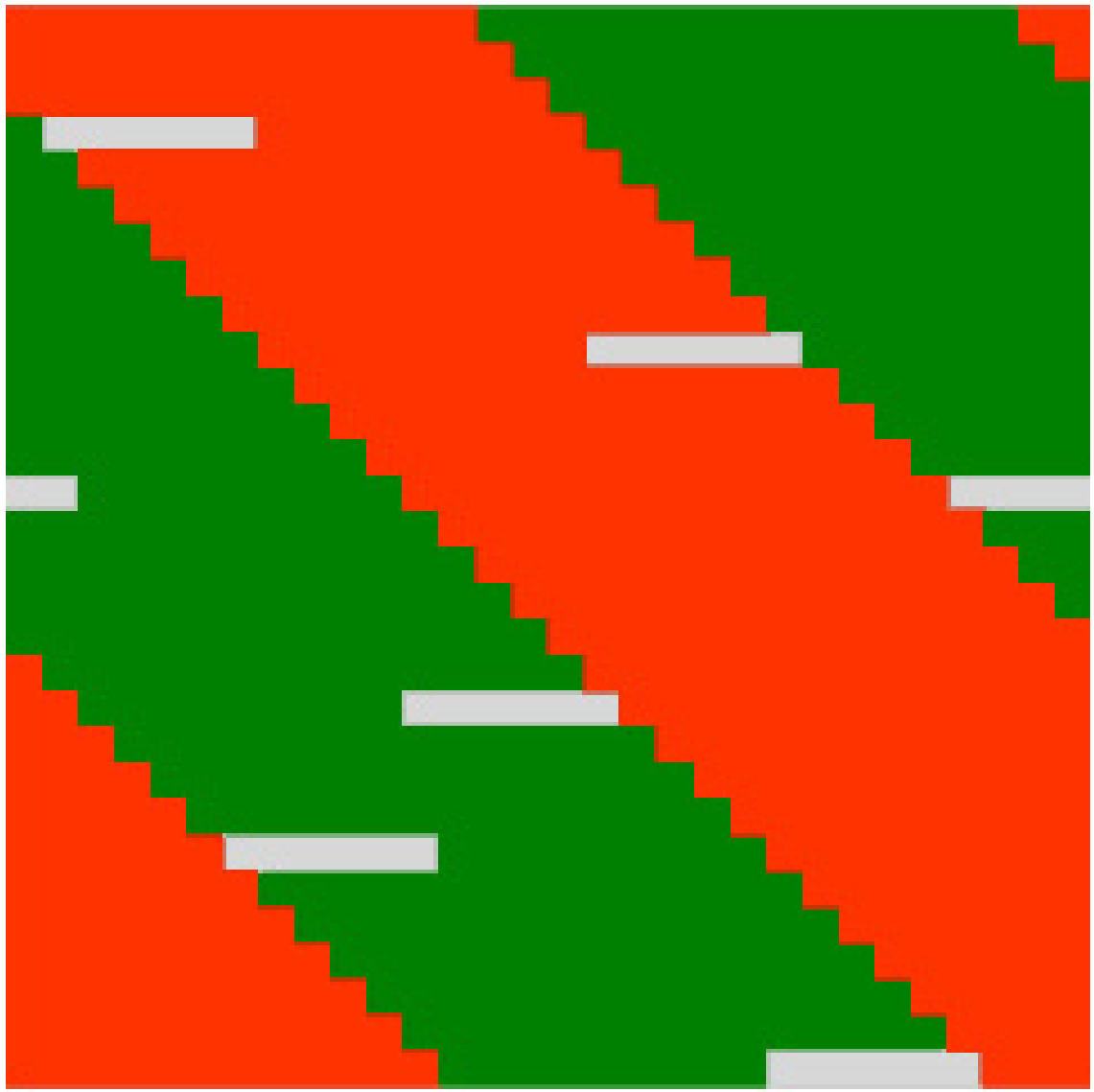
$O(n^3 4^k)$

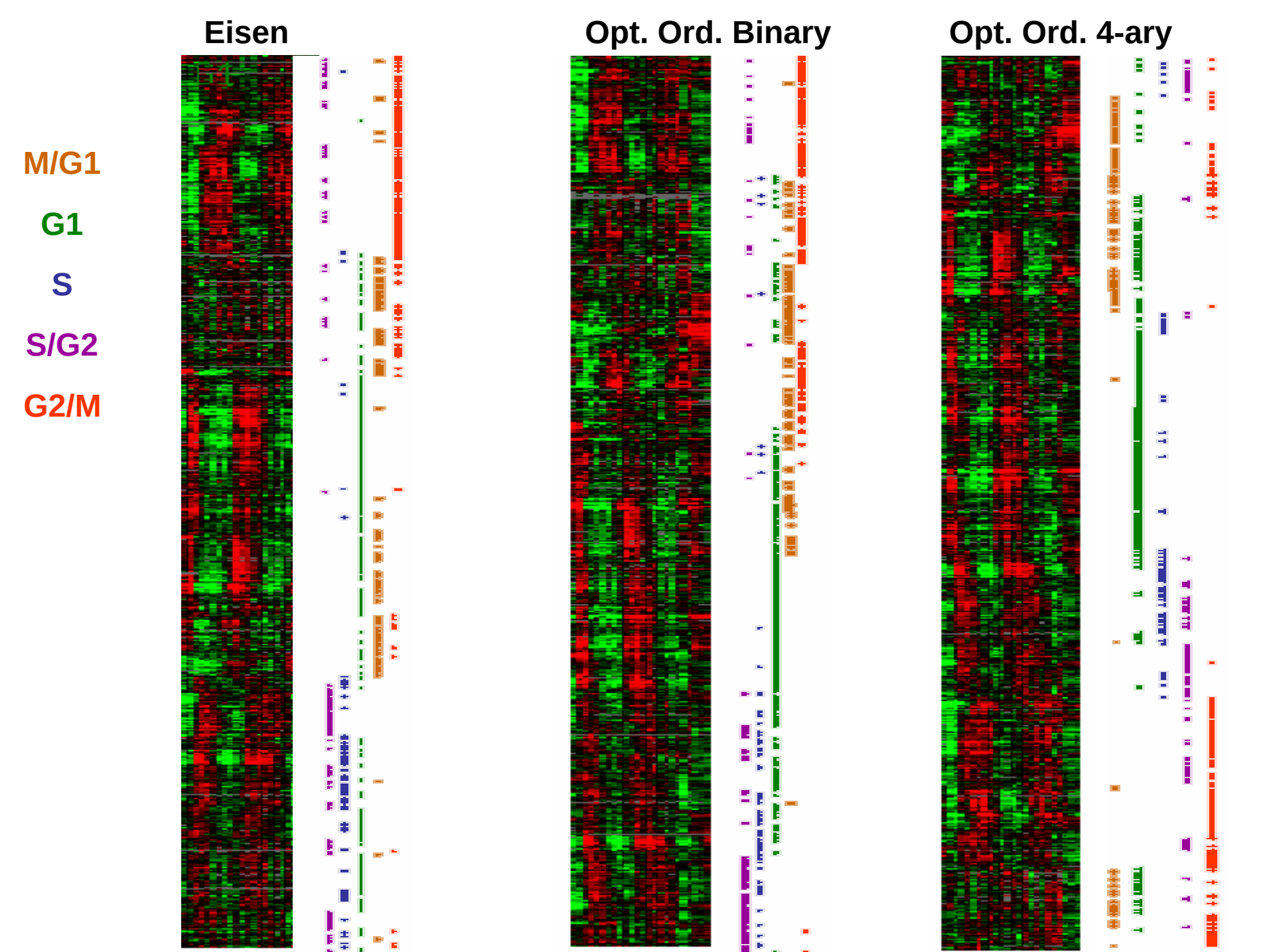
$O(n^2 k)$ space

Optimally Ordered using Binary Clustering



Optimally Ordered using 4-ary Clustering




Eisen
Opt. Ord. Binary
Opt. Ord. 4-ary
M/G1
G1
S
S/G2
G2/M

Summary

We suggest using optimally ordered K-ary hierarchical clustering, instead of binary hierarchical clustering, for initial arrangement of gene expression data

- Simplicity and efficiency comparable to binary hierarchical clustering
- Clusters and order better reflect biology

http://psrg.lcs.mit.edu/~zivbj/