Cluster Validation for Gene Expression Data

Ka Yee Yeung [1]

David R. Haynor [2]

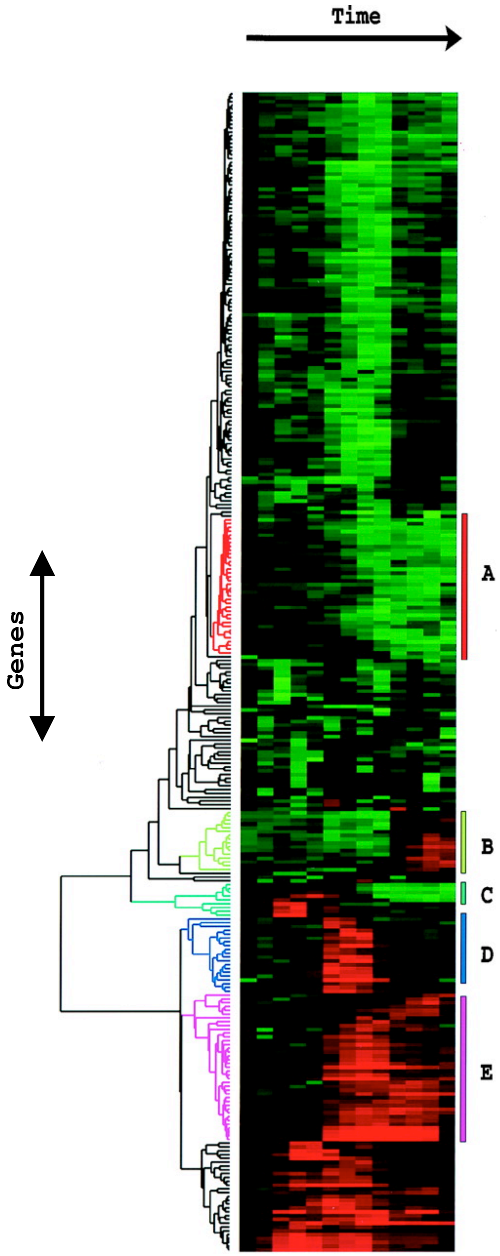
Walter L. Ruzzo [1]

[1] Department of Computer Science & Engineering

[2] Department of Radiology

University of Washington

Bioinformatics v17 #4, (2001) pp 309-318

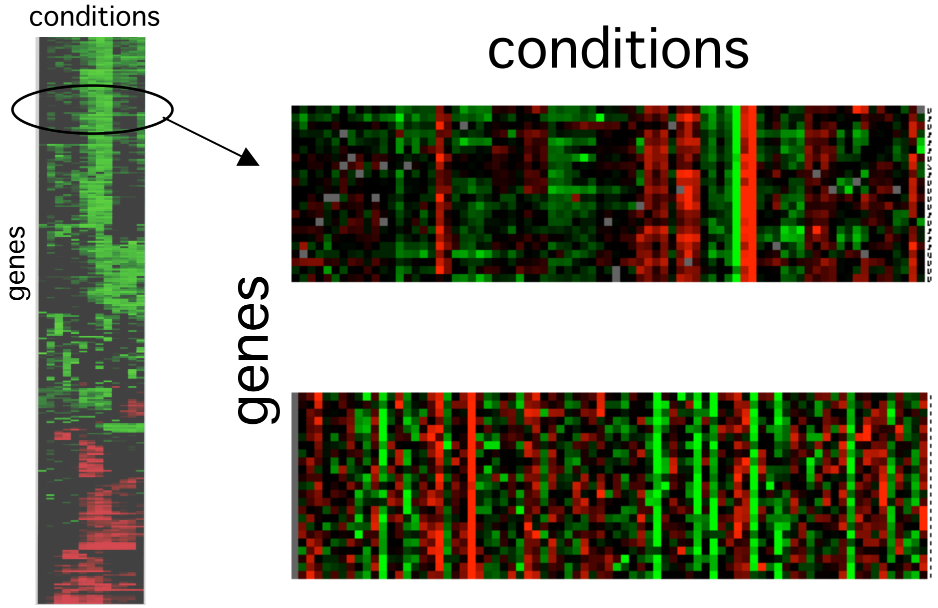


Eisen’s Cluster Software (PNAS 1998)

- Centroid-link hierarchical clustering algorithm
- Reorder for display
- Decide on your own cluster!

Why Validate clusters?

- All clustering algorithms find “clusters”:
  - Are they real?
  - Are they good?



A cluster from Eisen et al. (1998) on a yeast data set

A simulated data set with no intrinsic clusters.

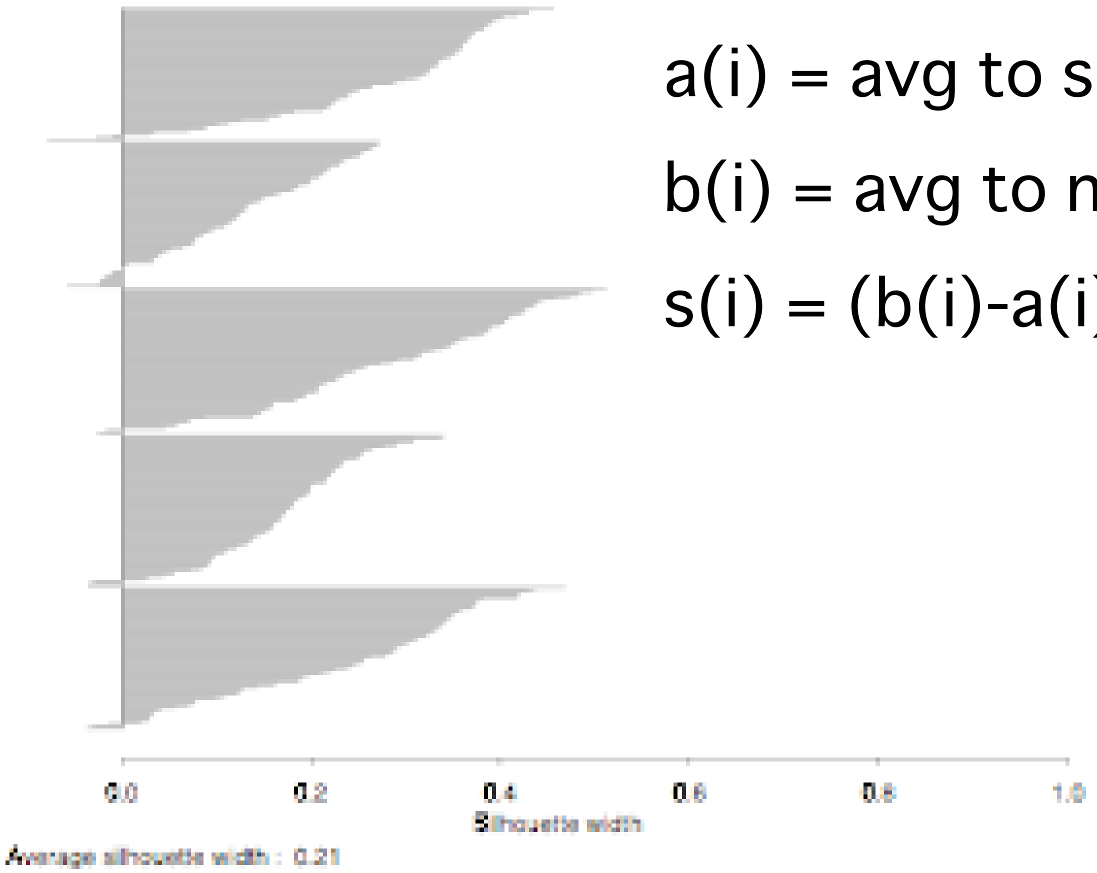
Approaches to Cluster Validation: External Criteria

- Agreement with an external “gold standard” answer (rarely available)
- Uniformity of clusters w.r.t. related external information, e.g. Gene Ontology or MIPS categories
- Either is quantifiable in various ways -- Jaccard, Hubert, adjusted Rand indices, relative entropy, hypergeometric, …

Approaches to Cluster Validation: Internal Criteria

- “Compactness” & “separation”
- E.g. residual sum of squares to cluster centers vs sums of squares between centers
- E.g. Silhouette - average distance to points in same cluster vs nearest other cluster

Silhouette



a(i) = avg to same

b(i) = avg to neighbor

s(i) = (b(i)-a(i))/ max(b(i),a(i))

Figure 15.4. A silhouette plot of 5 clusters from PAM on the cell cycle data.

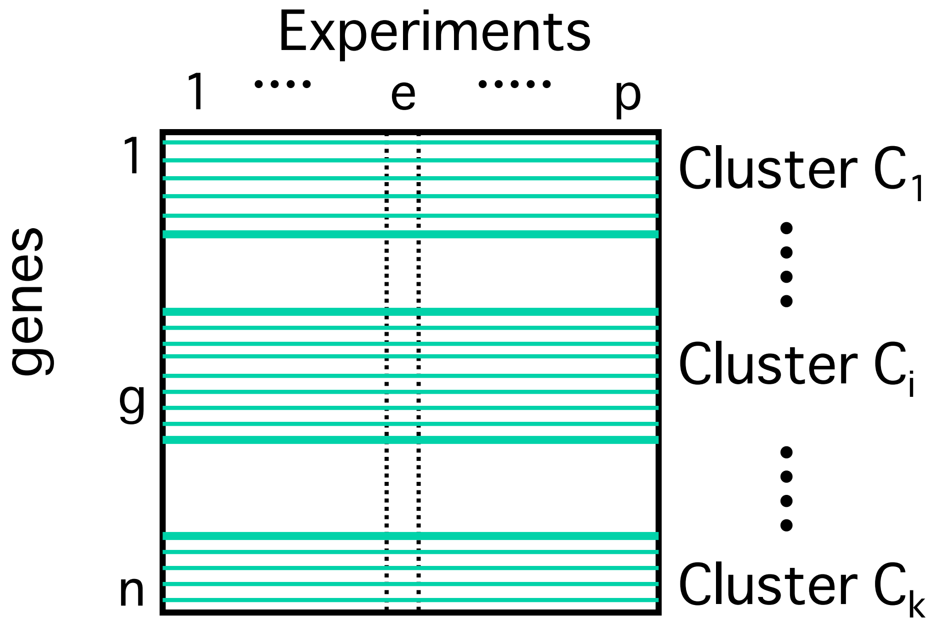
Approaches to Cluster Validation: Model-based

- Given (statistical) model of data, how well does model fit
- E.g. look at likelihood ratio that data could have been generated by one model vs alternative
- More on this topic later in the quarter

Our Methodology for Algorithm Comparison

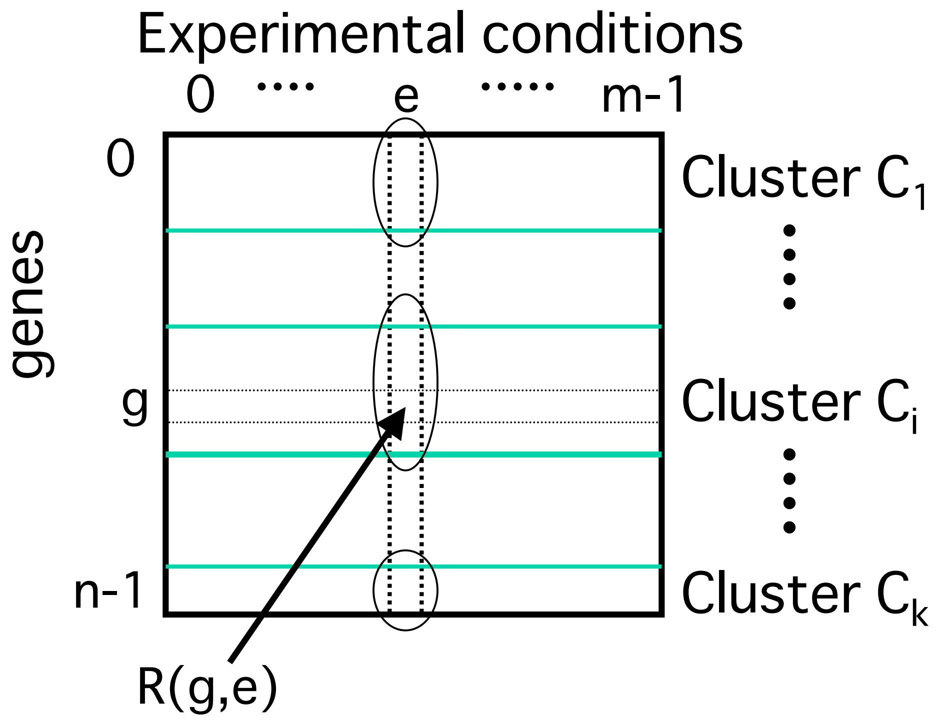
- A form of “Leave Out One Cross Validation”
  - Cluster genes based on all but one condition.
  - Use left-out condition to check cohesiveness of clusters.
    - I.e., within each cluster, how uniform are expression levels in the left-out condition?
    - Meaningful clusters should be more uniform that chance aggregations
  - Repeat for each condition
- Compare algorithms based on performance.

Figure of Merit (FOM)



- FOM measures uniformity of gene expression levels in each cluster in the left-out experiment (basically mean squared error)
- Low FOM => High predictive power
- Leave out each experiment in turn

“Figure of Merit”



FOM(e,k) = mean squared deviation of expression level from cluster mean:

$$\mathrm{FOM(e,k)} = \frac{1}{n}\sum_{i=1}^{k}\sum_{g\in C_i}\left(R(g,e)-\mu_{\mathrm{C_i}}(e)\right)^2$$

$$FOM(k) = \sum_{e=0}^{m-1} FOM(e,k)$$

$$\mathrm{adjFOM(k) = FOM(k)\cdot n/(n - k)}$$

In clusters formed, how uniform are expression levels in the left-out condition?

Other approaches

- S. Datta & S. Datta ‘03 -- look at agreement between clusterings with all data & leaving out different conditions

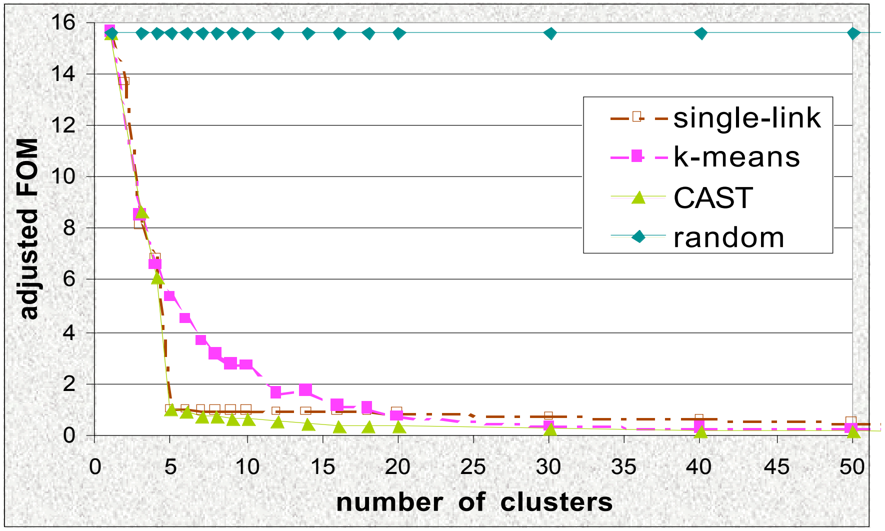
Three Successes

- We can distinguish clustered from non-clustered data

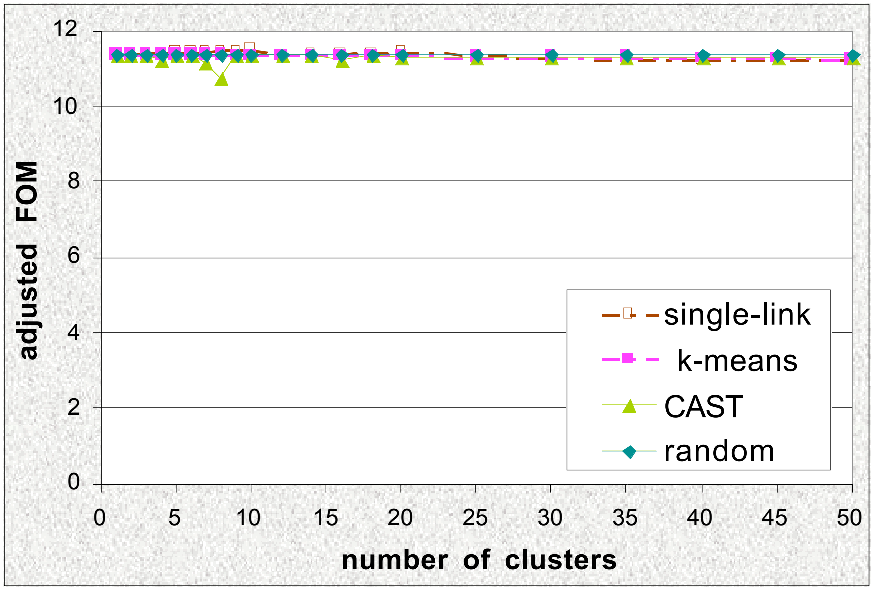
- We can tell algorithms apart

- Better FOM generally signals better clusters

Are there clusters?



A simulated data set with 5 clusters



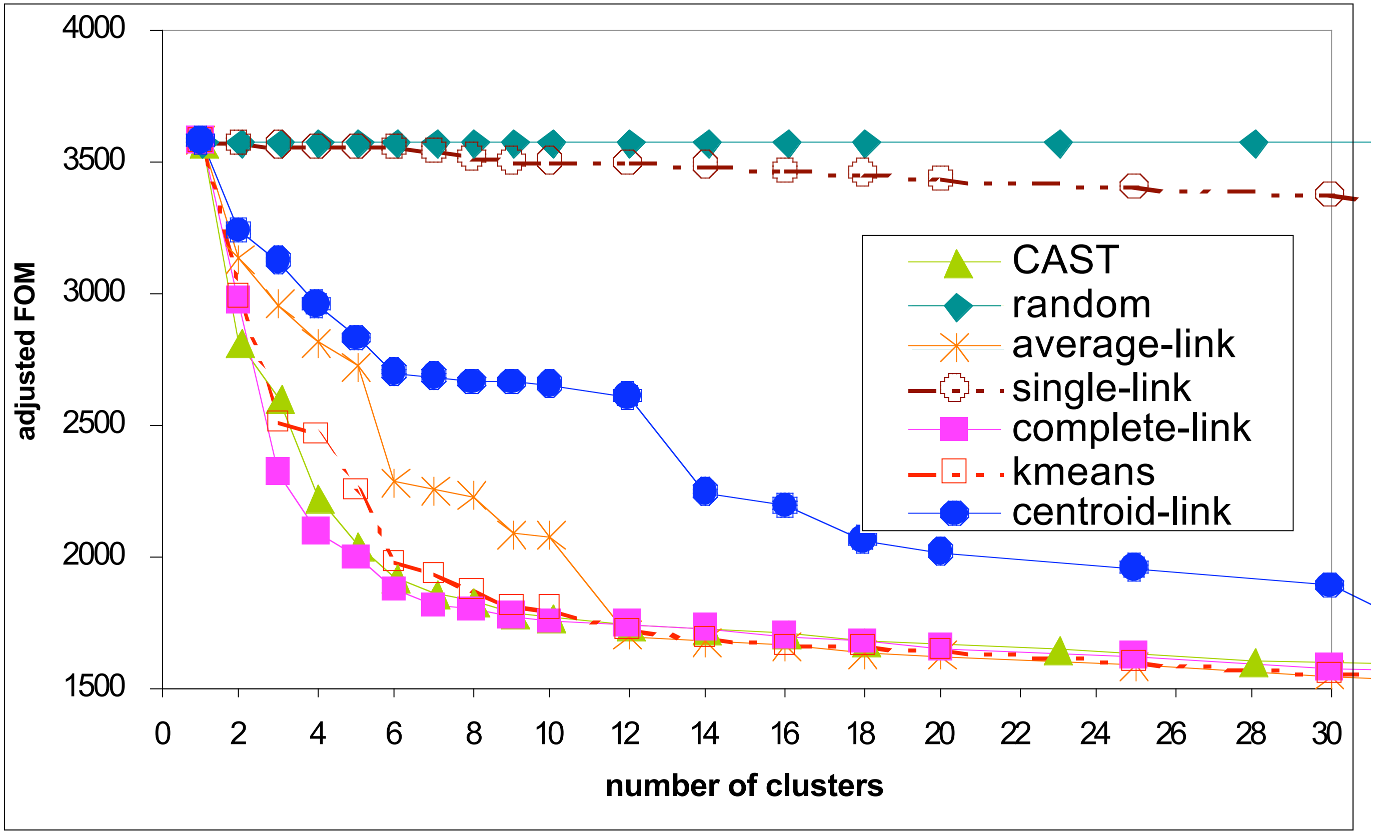
A simulated data set with no clusters

Gene expression data sets

- Ovarian cancer data set
  (Michel Schummer, Institute of Systems Biology)
  - Subset of data: 235 clones

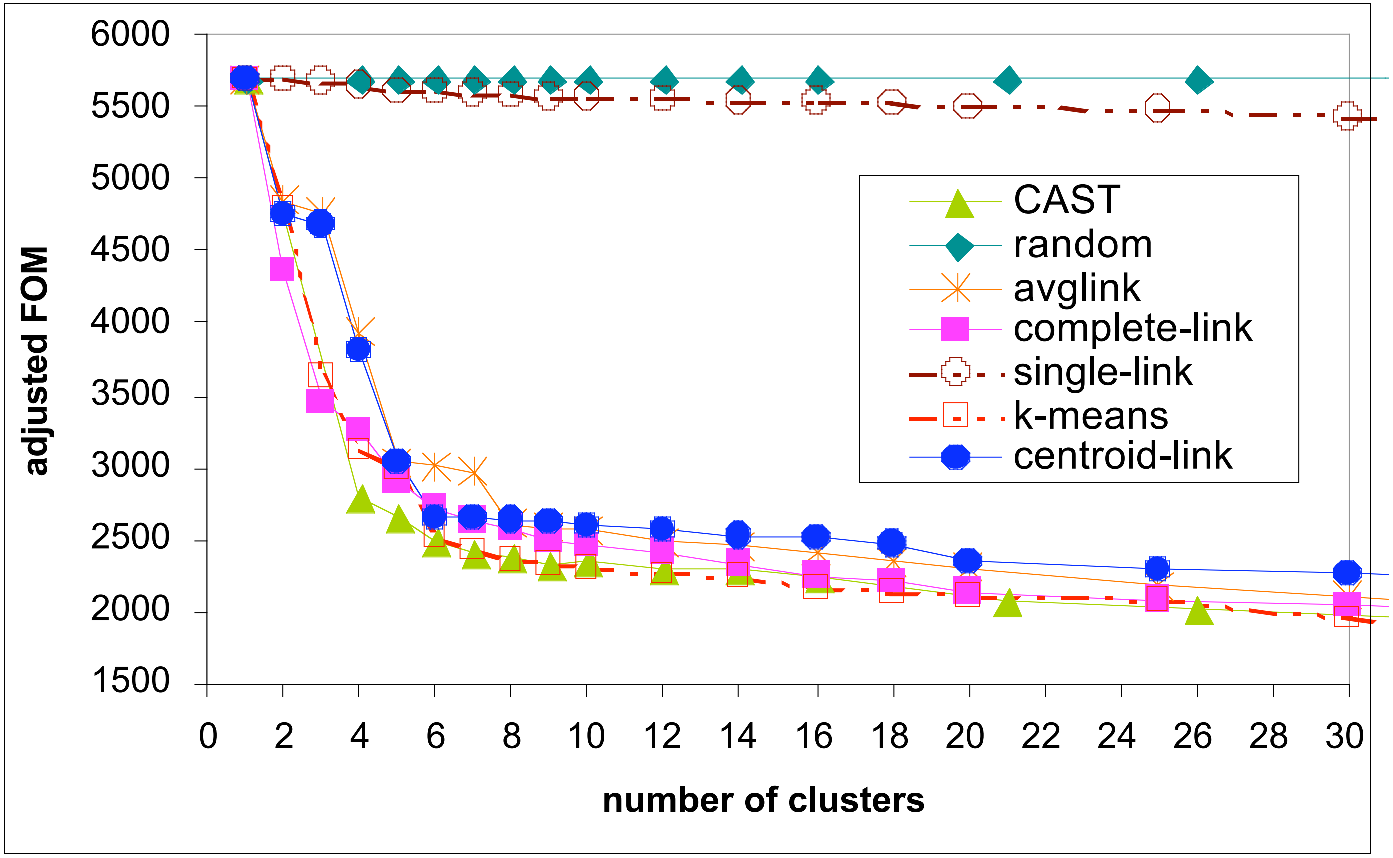
    24 experiments (cancer/normal tissue samples)
  - 235 clones correspond to 4 genes
- Yeast cell cycle data (Cho *et al* 1998)
  - 17 time points
  - Subset of 384 genes associated with 5 phases of cell cycle

Results: ovary data



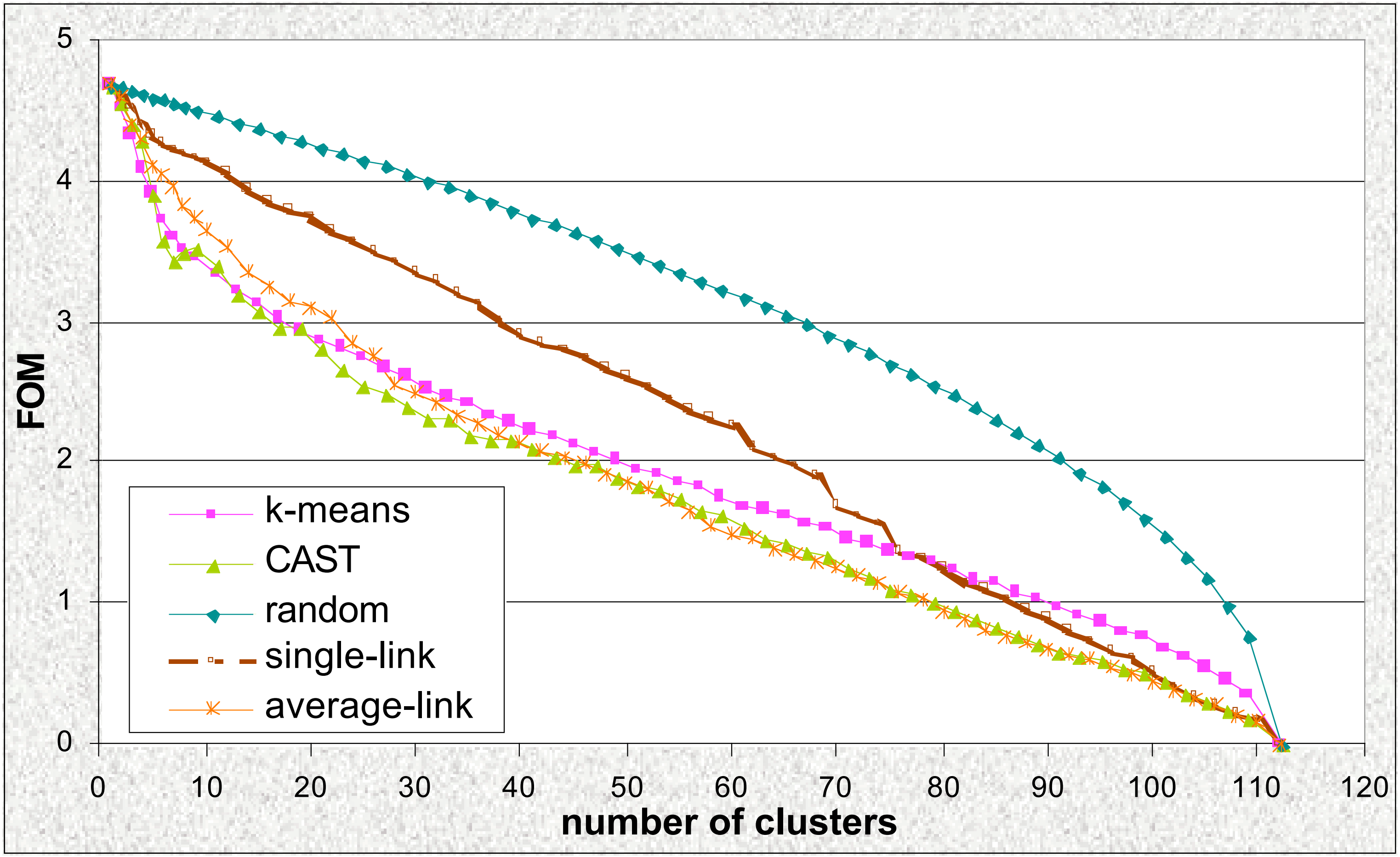
- CAST, k-means and complete-link : best performance

Results: yeast cell cycle data



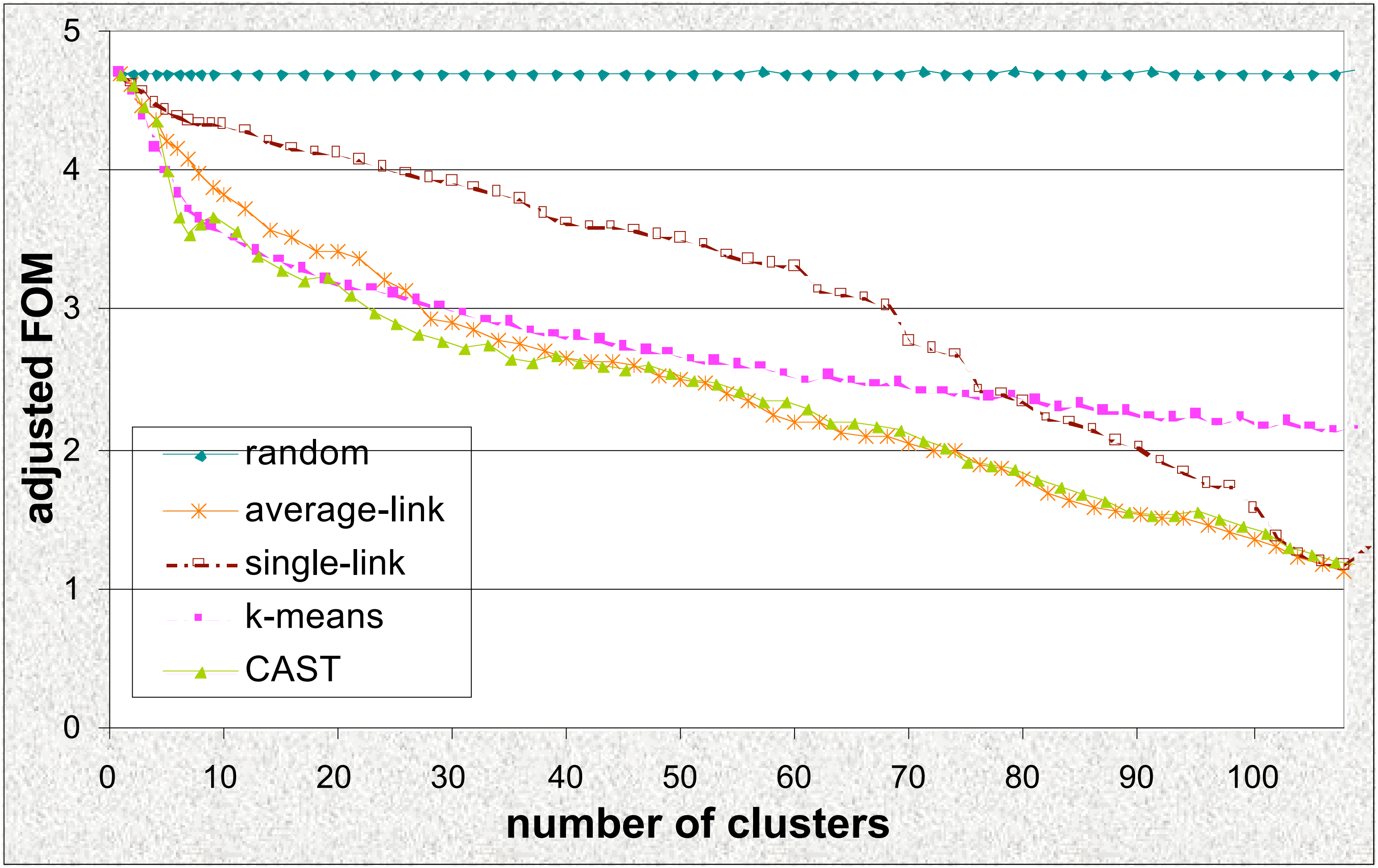
CAST, k-means: best performance

Rat CNS data



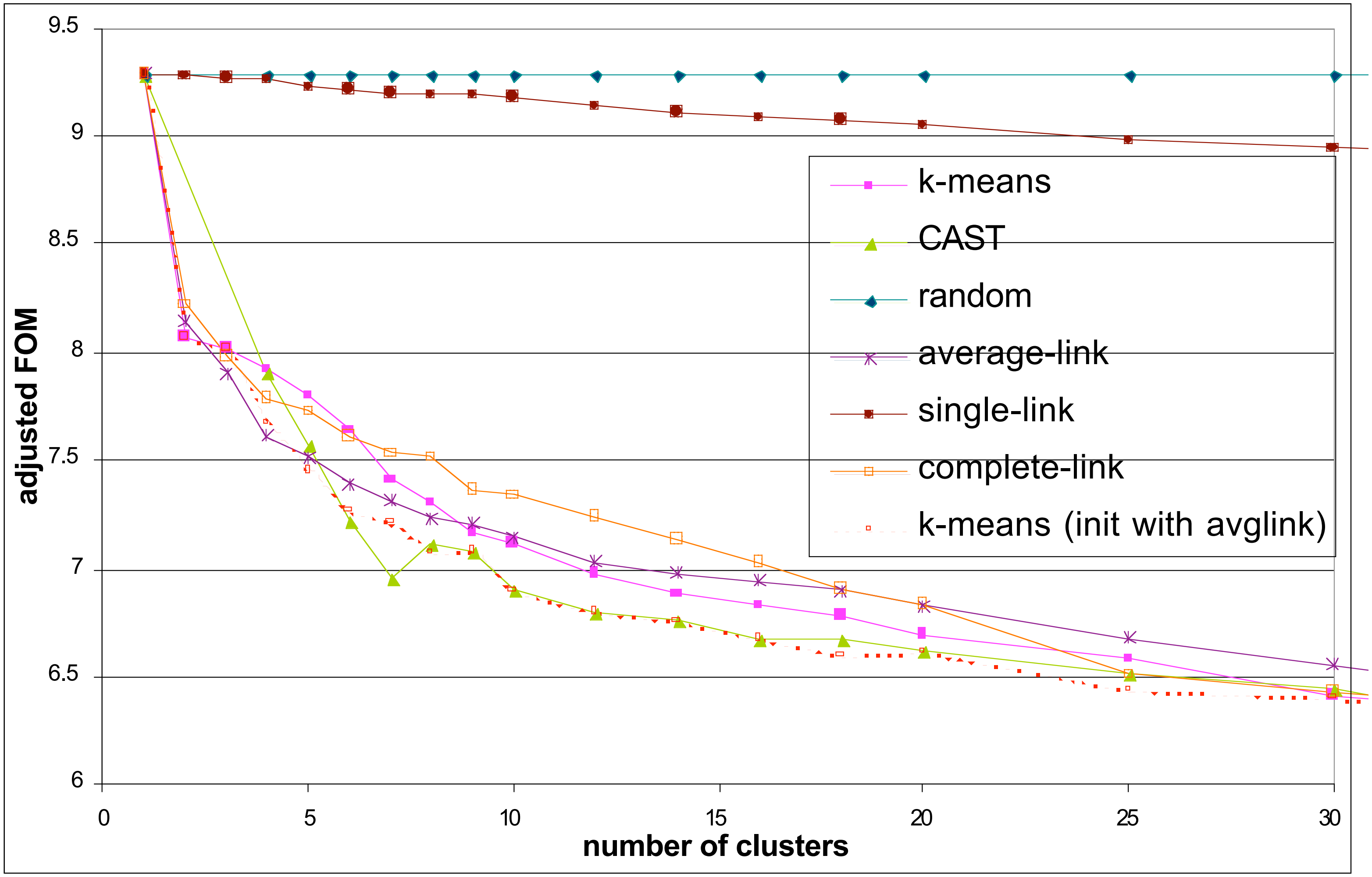
Full range, non-adjusted FOM

Rat CNS data



Full range, adjusted FOM

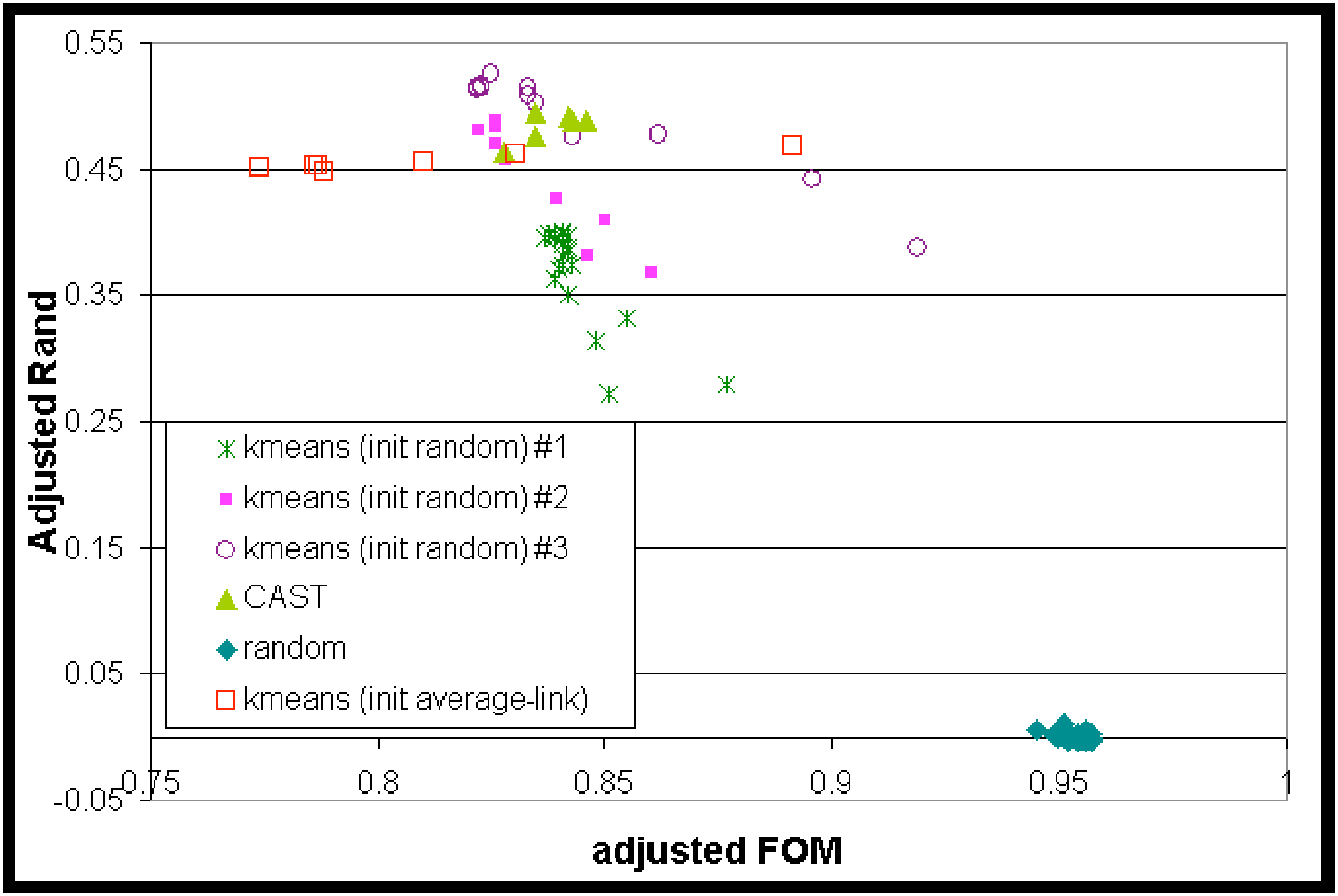
FOM on the Barrett’s data



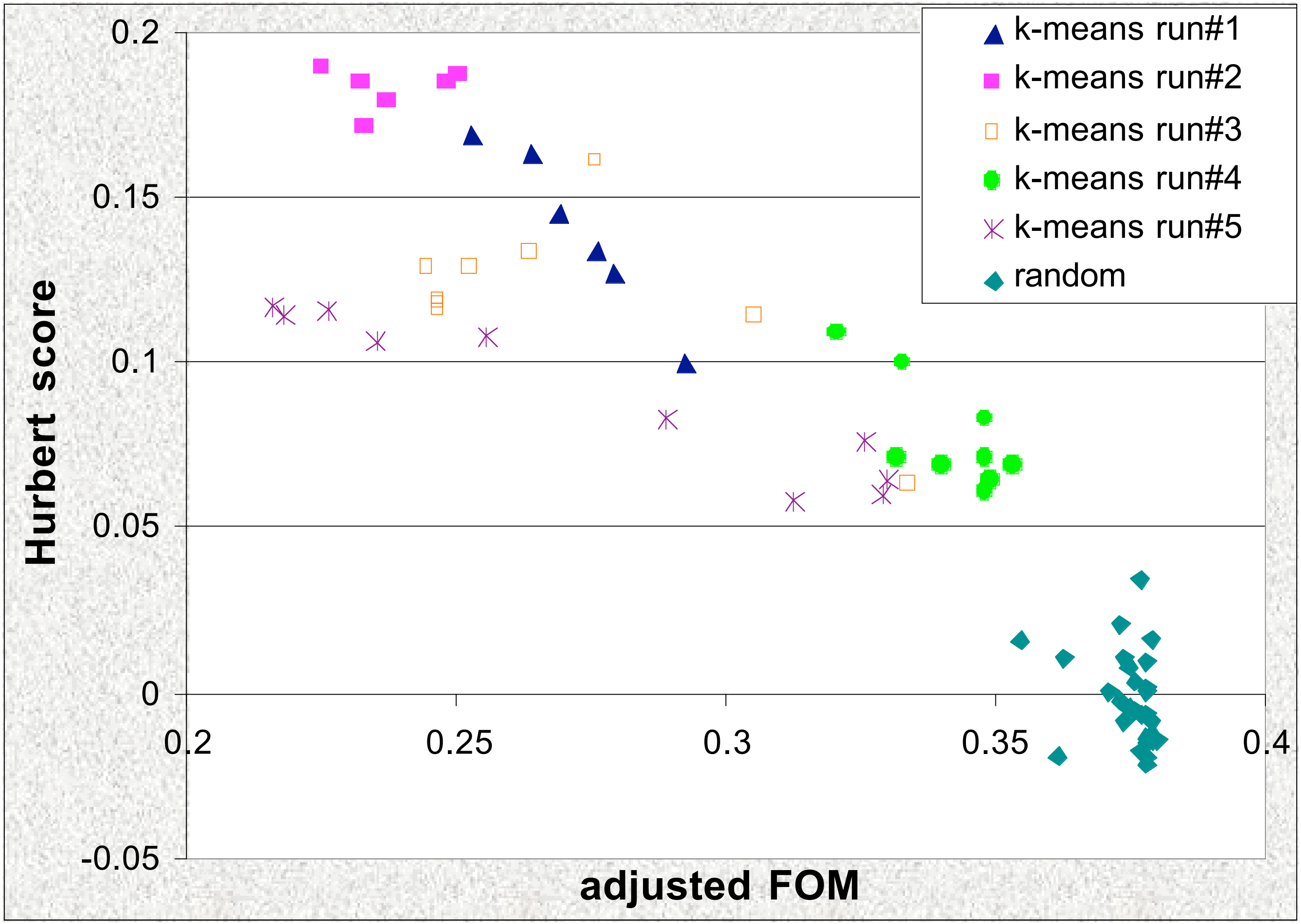
FOM □ Cluster Quality

- On ovary data:
  - – Lowest FOM clusters in good agreement with the right answer
  - – Next lowest incorrectly split/merged true classes
- On Barrett’s data, 10 clusters:
  - – the lowest FOM clusters (CAST & k-means initialized with average-link) correctly grouped the 20 cytokeratins that passed the variation filter
  - – the next lowest FOM (average-link) did NOT

FOM □ Cluster Quality



FOM □ Cluster Quality



FOM Summary

- Simple quantitative methodology to compare different clustering algorithms on any data set without using any external knowledge
- Reduced FOM generally signals improved clusters
- Omitting one condition doesn't destroy cluster quality

FOM Summary, cont.

- All clustering algorithms not created equal
- Some algorithm comparisons (on this data):
  - CAST and k-means produce higher quality clusters than the hierarchical algorithms
  - Single-link has the worst performance among the hierarchical algorithms

Acknowledgements

- Ka Yee Yeung
- David Haynor
- Michael Barrett
- Michèl Schummer

More Info

http://www.cs.washington.edu/homes/{kayee,ruzzo}

Adjusted Rand Example

	c#1(4)	c#2(5)	c#3(7)	c#4(4)
class#1(2)	2	0	0	0
class#2(3)	0	0	0	3
class#3(5)	1	4	0	0
class#4(10)	1	1	7	1

$$a = \binom{2}{2} + \binom{3}{2} + \binom{4}{2} + \binom{7}{2} = 31$$

$$b = \binom{4}{2} + \binom{5}{2} + \binom{7}{2} + \binom{4}{2} - a = 43 - 31 = 12$$

$$c = \binom{2}{2} + \binom{3}{2} + \binom{5}{2} + \binom{10}{2} - a = 59 - 31 = 28$$

$$d = \binom{20}{2} - a - b - c = 119$$

$$Rand, R = \frac{a + d}{a + d + c + d} = 0.789$$

$$\text{Adjusted Rand} = \frac{R - E(R)}{1 - E(R)} = 0.469$$