Clustering

Katie Pollard

BMI 206

In this unit we will learn …

- Differences between supervised and unsupervised learning
- How commonly used distance measurements encode different notions of "close"
- Hierarchical versus partitioning algorithms for clustering bioinformatics data
- Strategies for statistical inference and assessing variability with clustering results

Clustering vs. Classification

Clustering = unsupervised learning

- Classes: unknown a priori
- Goal: discover groups from the data

Classification = supervised learning

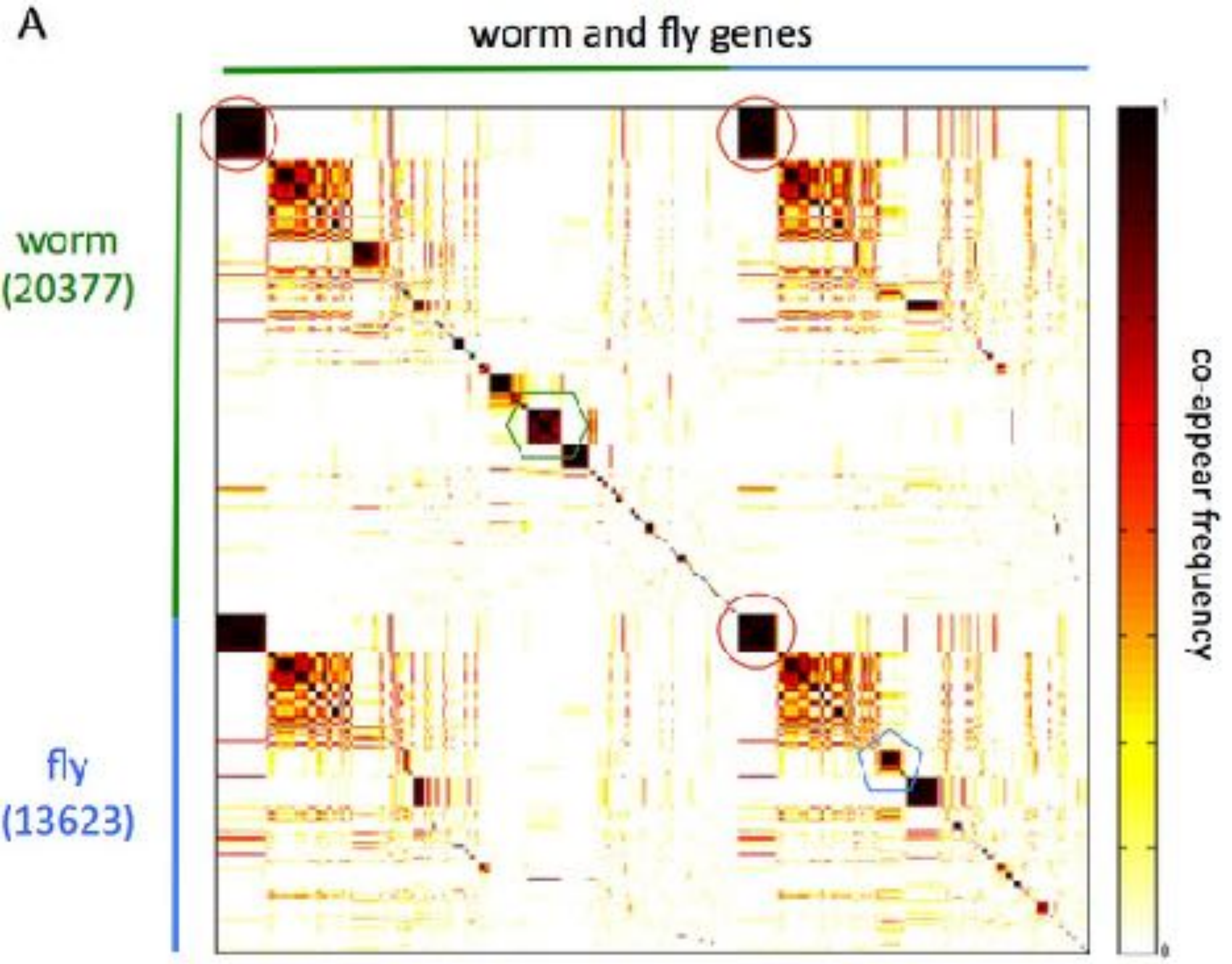
- Classes: known/predefined
- Goals: understand the basis for the classes and build a predictor (classify new data)
- Prediction with a categorical outcome variable

Cluster Analysis

Exploratory data analysis methods for:

- Discovering patterns
- Grouping
  - Variables
  - Samples
  - Both simultaneously
- Ordering and organizing
- Dimension reduction
  - How many distinct patterns?

Clustering in Bioinformatics



Worm orthologs of co-expressed genes in fly are some times also co-expressed.

Yan et al. (2014) Genome Biology

Clustering Methods

Two main components:

1. Distance measure
2. Algorithm

These produce a mapping from data to parameters of interest:

– Cluster labels, sizes, profiles

– Hierarchical tree structure, ordering

– Number of groups

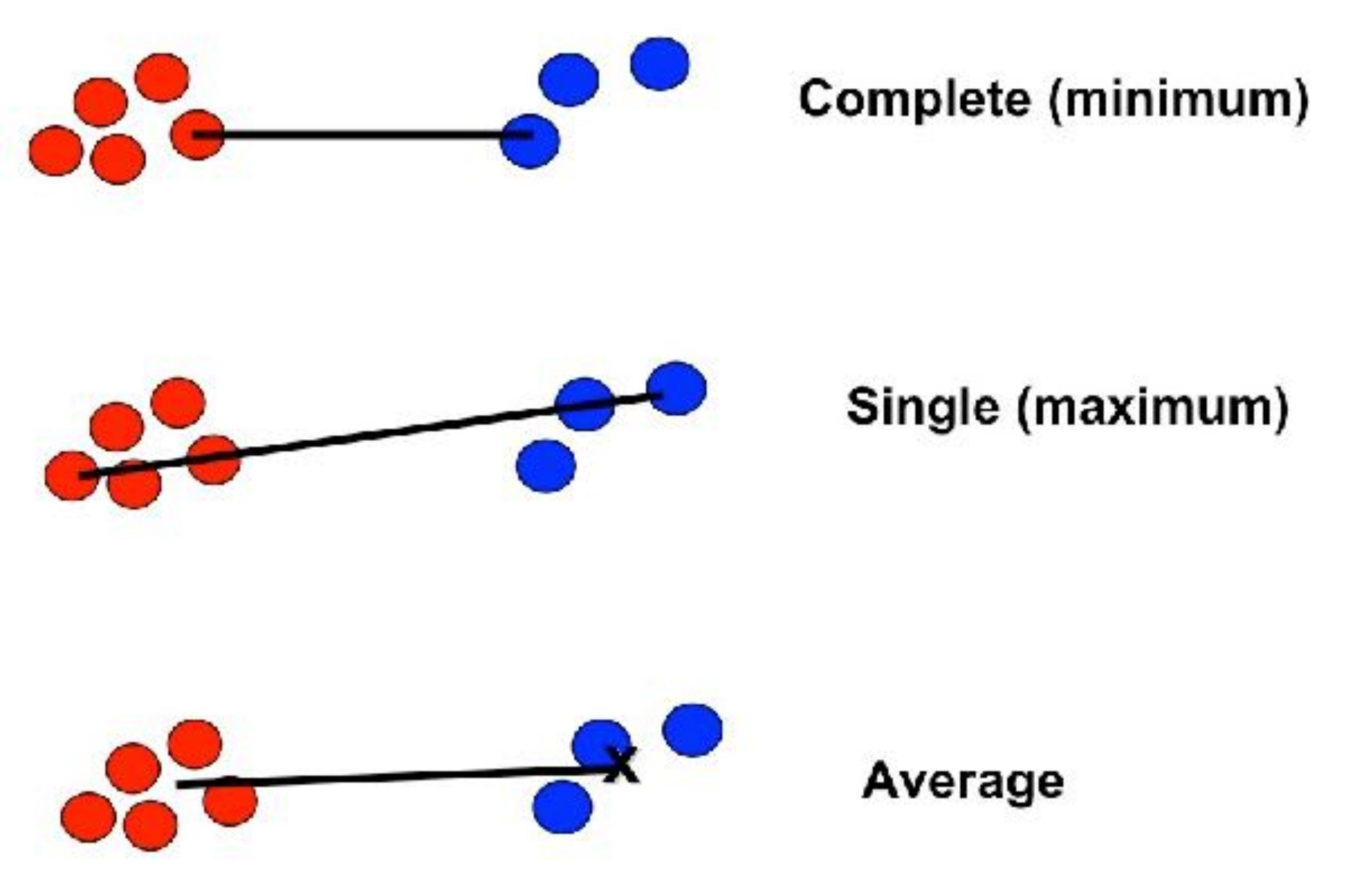
Distances

Clustering algorithms require a notion of pairwise distance between objects.

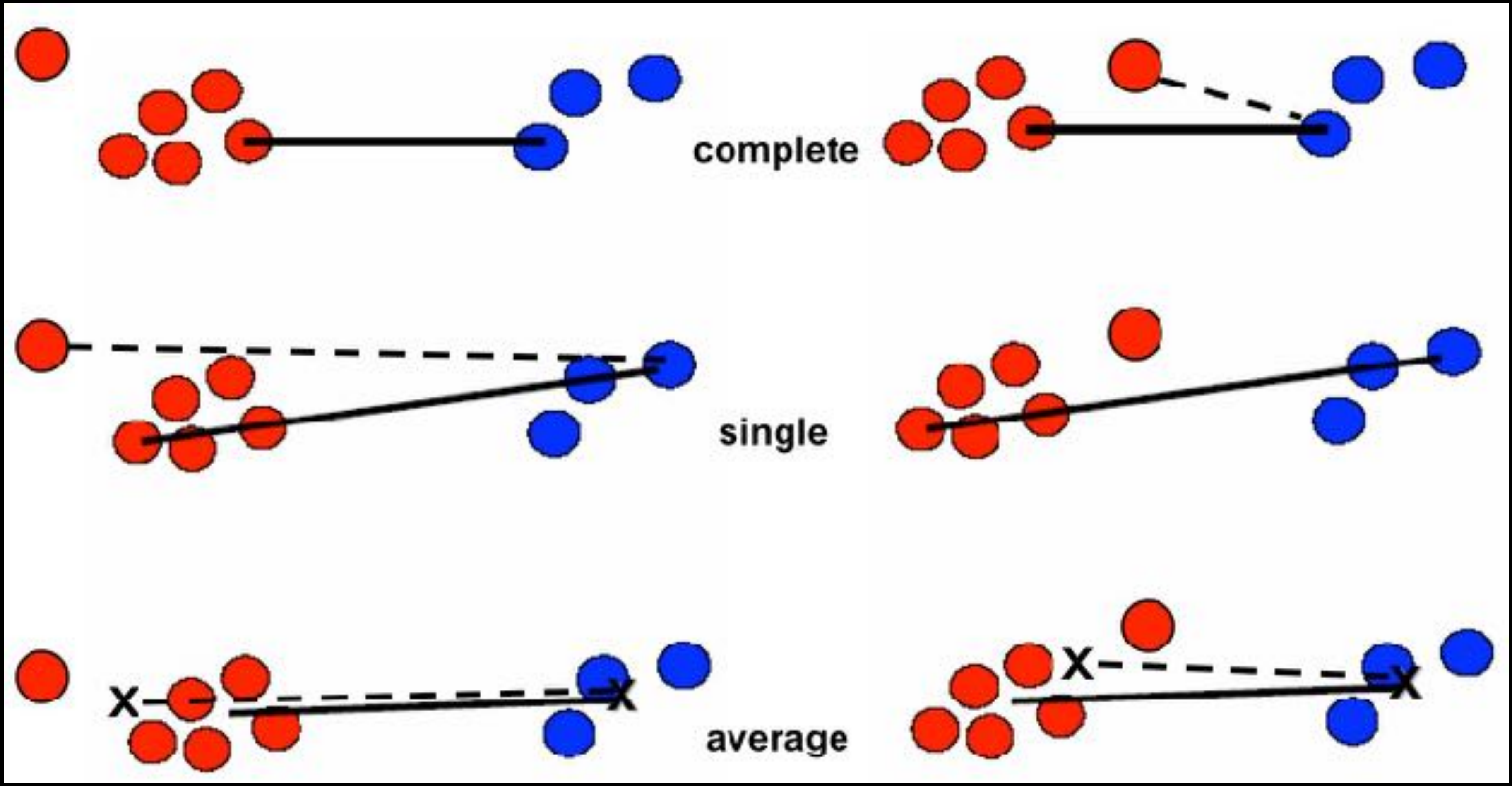
- Minkowski metrics: magnitude
- Correlation distances: pattern (or both)
- The absolute value of any distance can also be used, e.g.

$$d(x,y) = 1 - |r(x,y)| \in (0,1)$$

Distance Between Clusters



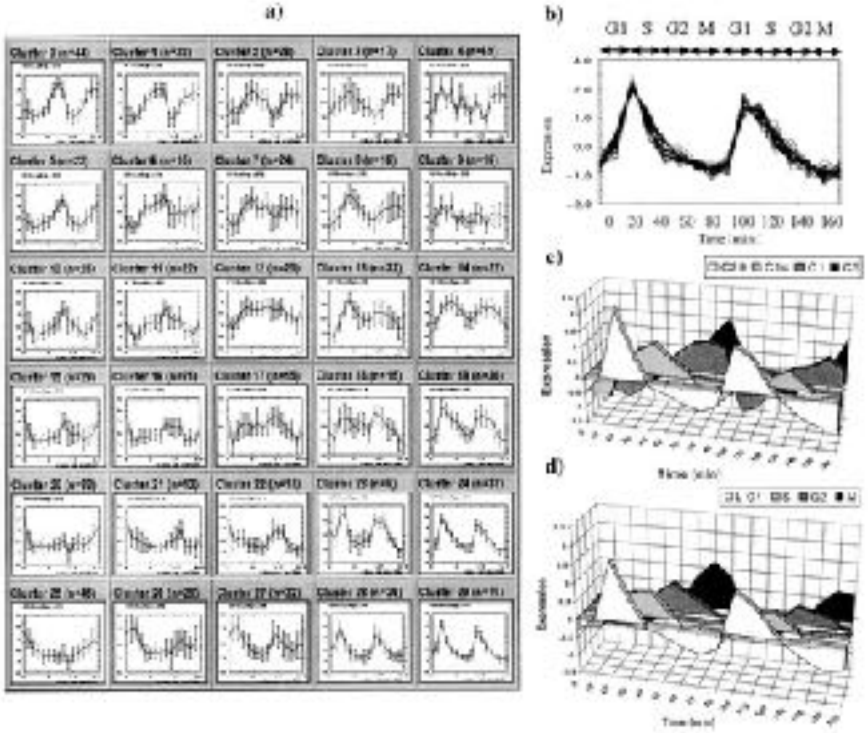
Effects of Outliers



Clustering Algorithms

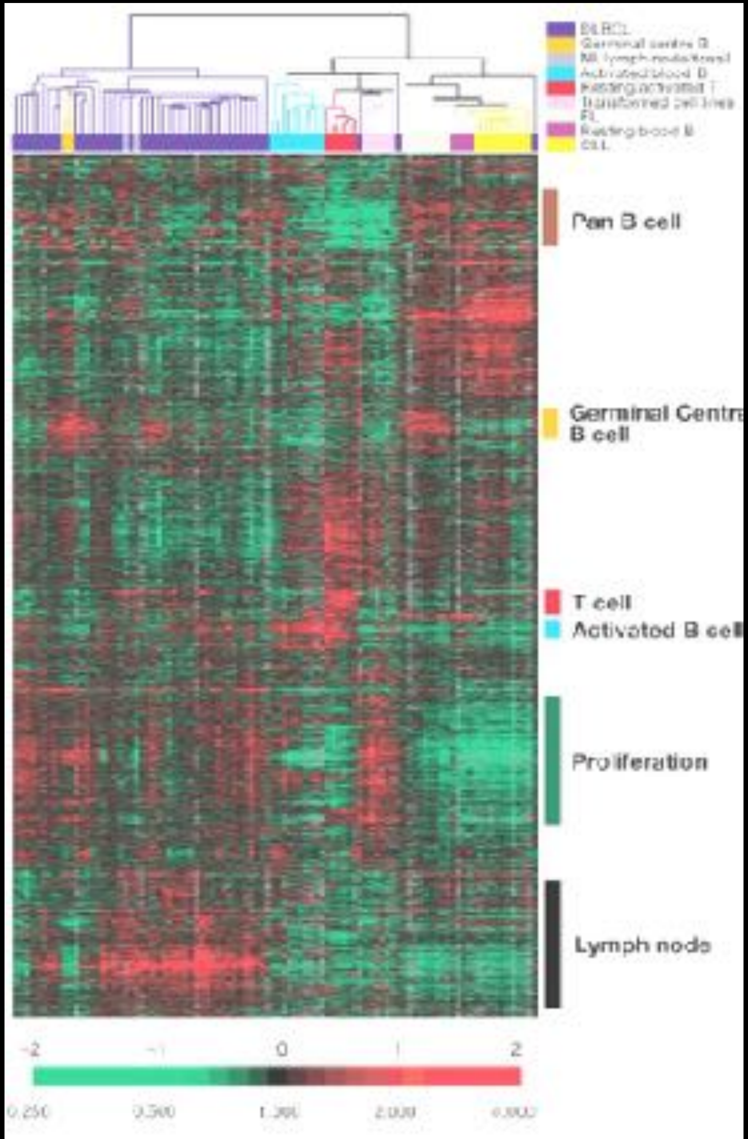
- Model-based (AUTOCLASS, SNOB)
  vs. Non-parametric
- Partitioning (SOMs, PAM, KMEANS)
  vs. Hierarchical
  - Agglomerative (CLUSTER, AGNES)
    - Linkage: single, complete, average
  - Devisive (SOTA, DIANA, TSVQ)
  - Hybrid (HOPACH, MUTUAL CLUSTERS)

Partitioning vs. Hierarchical



Cho *et al.* (1998) Molecular Cell, 2: 65-73

Tamayo *et al.* (1999) PNAS, 96: 2907-2912



Alizadeh *et al.* (2000) Nature, 403: 503-511

How Many Clusters?

- Relevant for both partitioning and hierarchical algorithms (with pruning)
- Level of structure: global vs. detailed
- Two main approaches:
  - Direct Methods (criteria)

    e.g. sums of squares, silhouettes
  - Resampling Methods (testing)

    e.g. Clest, gap statistic, bootstrap

Silhouette-based Criteria

- The silhouette for j’th object (e.g., gene):

$$S_j = \frac{\overset{\text{avg distance own cluster}}{a_j} - \overset{\text{avg distance next closest cluster}}{b_j}}{\max(a_j,b_j)}$$

- Cluster average silhouette: mean S_j per cluster
- Average silhouette: overall mean S_j
- Median split silhouette (MSS): split each cluster and see if silhouettes get smaller

Inference for Clustering

- How reliable and repeatable are cluster results from a single data set?
- Can view output (e.g. gene cluster labels) as a parameter estimate.
- Use resampling methods to estimate the variability of this estimator (since no closed form typically).

Example: Bootstrap cluster memberships

Issues in Cluster Analysis

- Results can be very sensitive to input (i.e., pre-processing and filtering).
- What method fits your application?
  - Distance: capture what "close" means
  - Algorithm: the kind of clusters you seek
- Clustering methods will always return some output, but is it meaningful?
  - Evaluate variability
  - Assess biological relevance
  - Confirm hypotheses with experiments

Clustering Algorithms in R

Package	Functions	Type
`stats`	`kmeans` `hclust`	partitioning divisive
`class`	`SOM`	partitioning

`cluster`	`pam`
	`agnes`
	`diana`
		partitioning
		agglomerative
		divisive

`hopach`	`hopach`	hybrid

Other packages: `cclust`, `e1071`, `flexmix`, `fpc`, `mclust`

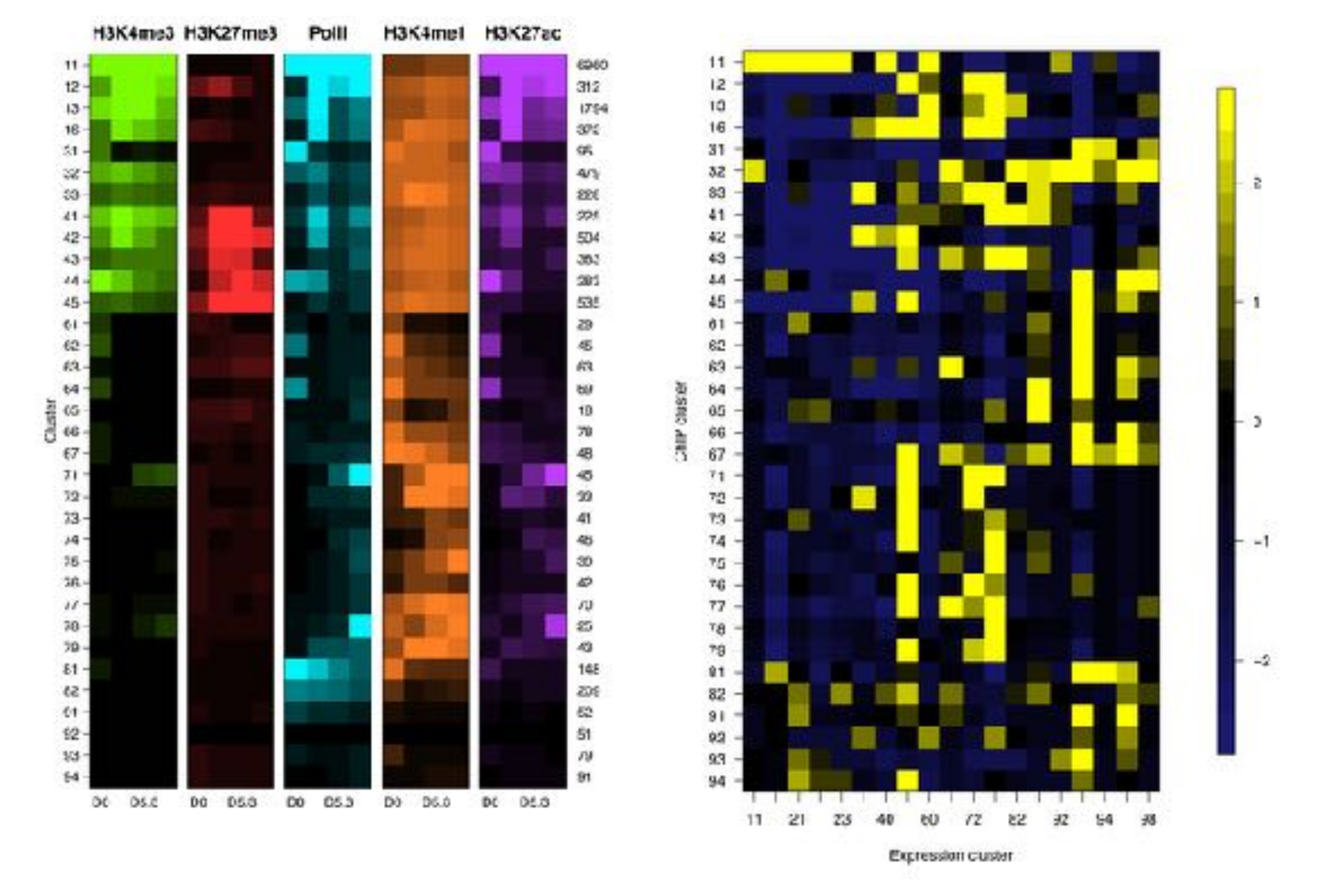
Comparing Clustering

- Any distance measure?
- Reaction to noise: robust vs. efficient
- Results reproducible?
- Biological relevance

Partitioning	Hierarchical
Variety of cluster sizes	Sensible ordering
Overlapping clusters	

Hybrid tree of partitions

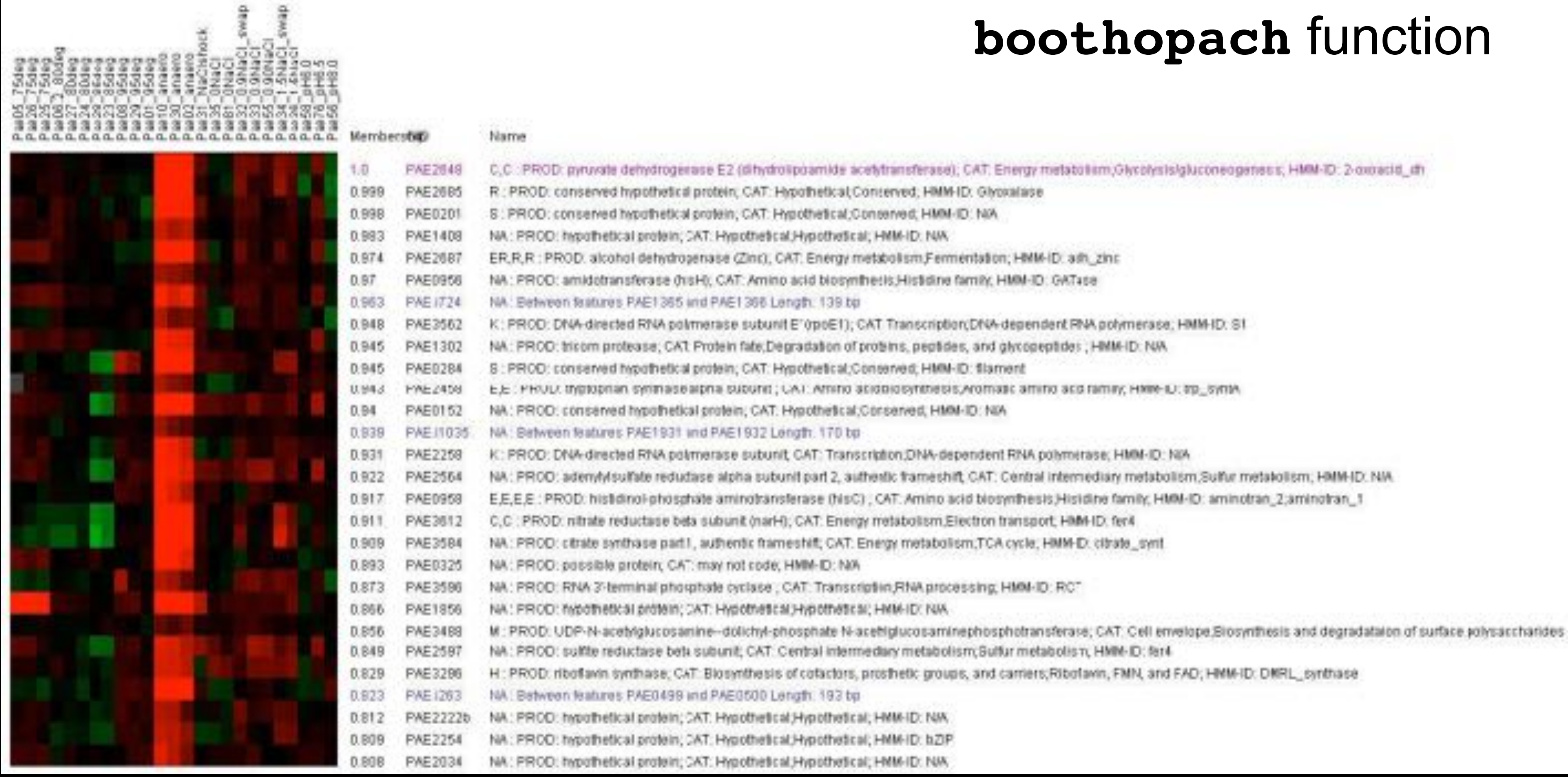
HOPACH: Pruned hierarchical tree produces nested clusters



Wamstad *et al.* (2012) Cell, 151: 206-220

Bootstrap Fuzzy Clustering

`boothopach` function



Membership	ID	Name
1.0	PAE2648	C,C : PROD: pyruvate dehydrogenase E2 (dihydrolipoamide acetyltransferase); CAT: Energy metabolism;Glycolysis/gluconeogenesis; HMM-ID: 2-oxoacid_dh
0.999	PAE2685	R : PROD: conserved hypothetical protein; CAT: Hypothetical;Conserved; HMM-ID: Glyoxalase
0.998	PAE0201	S : PROD: conserved hypothetical protein; CAT: Hypothetical;Conserved; HMM-ID: N/A
0.983	PAE1408	NA : PROD: hypothetical protein; CAT: Hypothetical;Hypothetical; HMM-ID: N/A
0.974	PAE2687	ER,R,R : PROD: alcohol dehydrogenase (Zinc); CAT: Energy metabolism;Fermentation; HMM-ID: adh_zinc
0.97	PAE0956	NA : PROD: amidotransferase (hisH); CAT: Amino acid biosynthesis;Histidine family; HMM-ID: GATase
0.963	PAE I724	NA : Between features PAE1365 and PAE1366 Length: 139 bp
0.948	PAE3562	K : PROD: DNA-directed RNA polmerase subunit E'(rpoE1); CAT Transcription;DNA-dependent RNA polymerase; HMM-ID: S1
0.945	PAE1302	NA : PROD: tricorn protease; CAT Protein fate;Degradation of protins, peptides, and glycopeptides ; HMM-ID: N/A
0.945	PAE0284	S : PROD: conserved hypothetical protein; CAT: Hypothetical;Conserved; HMM-ID: filament
0.943	PAE2459	E,E : PROD: tryptophan synthase alpha subunit ; CAT: Amino acid biosynthesis;Aromatic amino acid family; HMM-ID: trp_syntA
0.94	PAE0152	NA : PROD: conserved hypothetical protein; CAT: Hypothetical;Conserved; HMM-ID: N/A
0.939	PAE I1035	NA : Between features PAE1931 and PAE1932 Length: 170 bp
0.931	PAE2258	K : PROD: DNA-directed RNA polmerase subunit; CAT: Transcription;DNA-dependent RNA polymerase; HMM-ID: N/A
0.922	PAE2564	NA : PROD: adenylylsulfate redudase alpha subunit part 2, authentic frameshift; CAT: Central intermediary metabolism;Sulfur metabolism; HMM-ID: N/A
0.917	PAE0958	E,E,E,E : PROD: histidinol-phosphate aminotransferase (hisC) ; CAT: Amino acid biosynthesis;Histidine family; HMM-ID: aminotran_2;aminotran_1
0.911	PAE3612	C,C : PROD: nitrate reductase beta subunit (narH); CAT: Energy metabolism;Electron transport; HMM-ID: fer4
0.909	PAE3584	NA : PROD: citrate synthase part1, authentic frameshift; CAT: Energy metabolism;TCA cycle; HMM-ID: citrate_synt
0.893	PAE0325	NA : PROD: possible protein; CAT: may not code; HMM-ID: N/A
0.873	PAE3596	NA : PROD: RNA 3'-terminal phosphate cyclase ; CAT: Transcription;RNA processing; HMM-ID: RCT
0.866	PAE1856	NA : PROD: hypothetical protein; CAT: Hypothetical;Hypothetical; HMM-ID: N/A
0.856	PAE3488	M : PROD: UDP-N-acetylglucosamine--dolichyl-phosphate N-acetylglucosaminephosphotransferase; CAT: Cell envelope;Biosynthesis and degradataion of surface polysaccharides
0.849	PAE2597	NA : PROD: sulfite reductase beta subunit; CAT: Central intermediary metabolism;Sulfur metabolism; HMM-ID: fer4
0.829	PAE3286	H : PROD: riboflavin synthase; CAT: Biosynthesis of cofactors, prosthetic groups, and carriers;Riboflavin, FMN, and FAD; HMM-ID: DMRL_synthase
0.823	PAE I263	NA : Between features PAE0499 and PAE0500 Length: 193 bp
0.812	PAE2222b	NA : PROD: hypothetical protein; CAT: Hypothetical;Hypothetical; HMM-ID: N/A
0.809	PAE2254	NA : PROD: hypothetical protein; CAT: Hypothetical;Hypothetical; HMM-ID: bZIP
0.808	PAE2034	NA : PROD: hypothetical protein; CAT: Hypothetical;Hypothetical; HMM-ID: N/A

Pyrobaculum aerophilum array data from Lowe Lab (UCSC)