Data Mining in Bioinformatics
Day 8: Clustering in Bioinformatics
Clustering Gene Expression Data

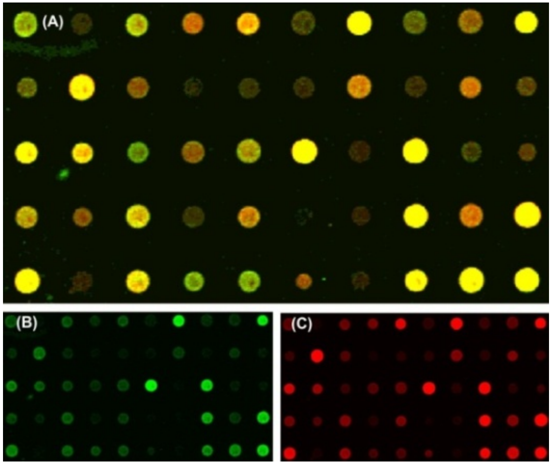
Chloé-Agathe Azencott & Karsten Borgwardt

February 10 to February 21, 2014

Machine Learning & Computational Biology Research Group
Max Planck Institutes Tübingen and
Eberhard Karls Universität Tübingen

Gene expression data

Microarray technology



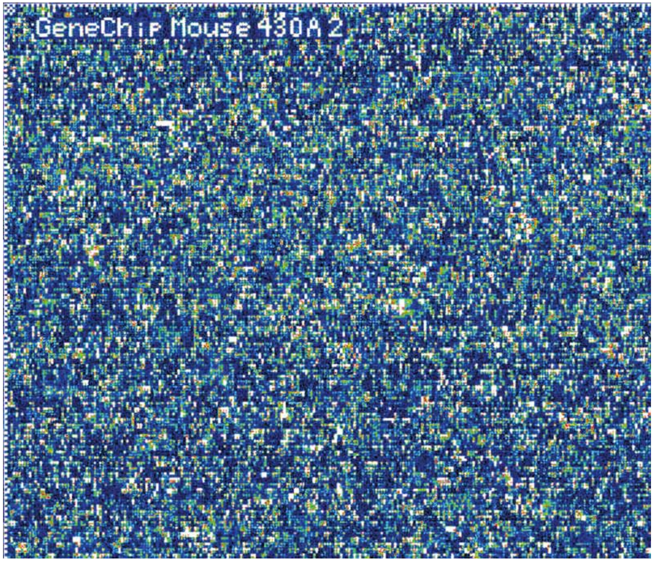
- High density arrays
- **Probes** (or “reporters”, “oligos”)

- Detect probe-target hybridization
  - Fluorescence, chemiluminescence
  - E.g. Cyanine dyes: Cy3 (green) / Cy5 (red)

Gene expression data

Data

- $X : n \times m$ matrix
- n genes
- m experiments:
  - conditions
  - time points
  - tissues
  - patients
  - cell lines



Clustering gene expression data

Group samples

- Group together tissues that are similarly affected by a disease
- Group together patients that are similarly affected by a disease

Group genes

- Group together functionally related genes
- Group together genes that are similarly affected by a disease
- Group together genes that respond similarly to an experimental condition

Clustering gene expression data

Applications

- Build regulatory networks
- Discover subtypes of a disease
- Infer unknown gene function
- Reduce dimensionality

Popularity

- Pubmed hits: $33\,548$ for “microarray AND clustering”, $79\,201$ for “"gene expression" AND clustering”
- Toolboxes: MatArray, Cluster3, GeneCluster, Bioconductor, GEO tools, $\ldots$

Pre-processing

Pre-filtering

- Eliminate poorly expressed genes
- Eliminate genes whose expression remains constant

Missing values

- Ignore
- Replace with random numbers
- Impute
  - Continuity of time series
  - Values for similar genes

Pre-processing

Normalization

- $log_2(ratio)$
  particularly for time series
- $log_2(Cy5/Cy3)$
  $\rightarrow$ induction and repression have opposite signs
- variance normalization
- differential expression

Distances

Euclidean distance

Distance between gene x and y, given n samples

(or distance between samples x and y, given n genes)

$$d(x, y) = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2}$$

Emphasis: **shape**

Pearson's correlation

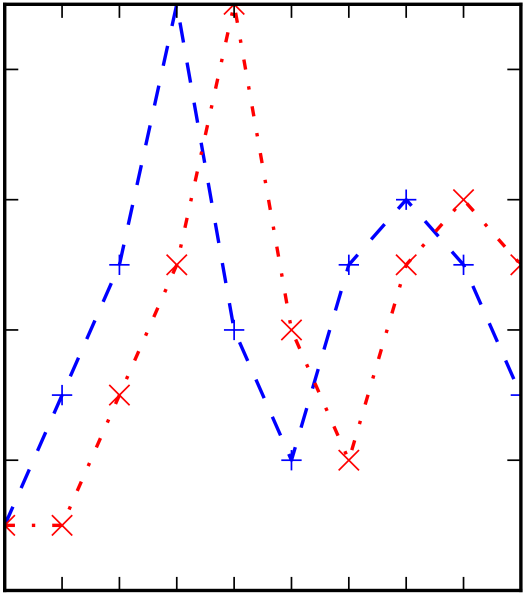
Correlation between gene x and y, given n samples

(or correlation between samples x and y, given n genes)

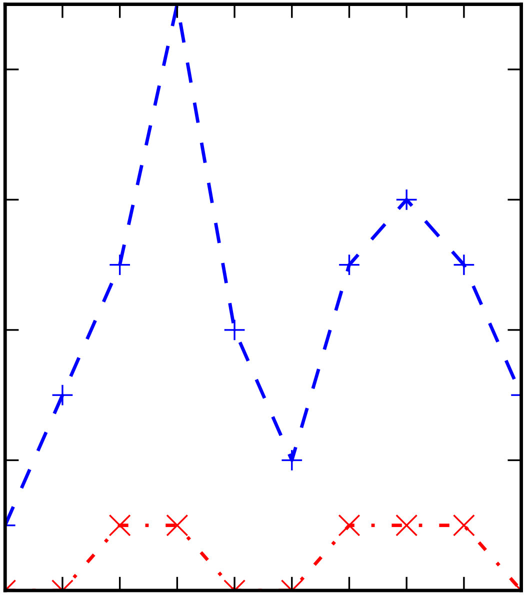
$$\rho(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Emphasis: **magnitude**

Distances



- $d = 8.25$
- $\rho = 0.33$



- $d = 13.27$
- $\rho = 0.79$

Clustering evaluation

Clusters shape

- Cluster **tightness (homogeneity)**

$$\sum_{i=1}^{k} \underbrace{\frac{1}{|C_i|} \sum_{x \in C_i} d(x, \mu_i)}_{T_i}$$

- Cluster **separation**

$$\sum_{i=1}^{k} \sum_{j=i+1}^{k} \underbrace{d(\mu_i, \mu_j)}_{S_{i,j}}$$

- Davies-Bouldin index

$$D_i := \max_{j:j \neq i} \frac{T_i + T_j}{S_{i,j}} \qquad DB := \frac{1}{k} \sum_{i=1}^{k} D_i$$

Clustering evaluation

Clusters stability

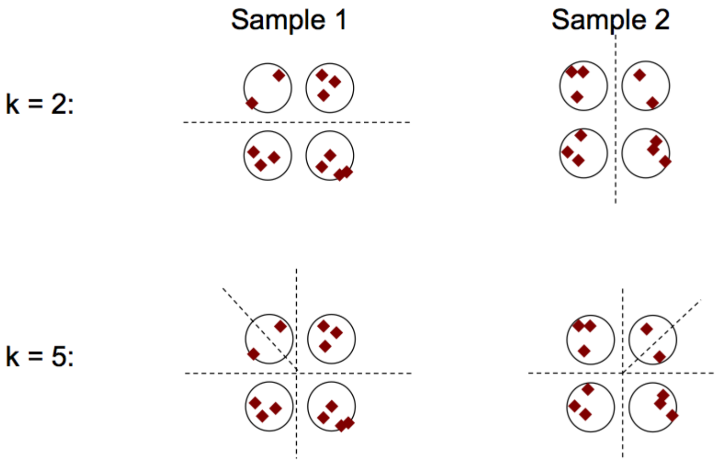


image from [von Luxburg, 2009]

Does the solution change if we perturb the data?

- Bootstrap
- Add noise

Quality of clustering

The Gene Ontology

> *"The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner"*

- **Cellular Component**: where in the cell a gene acts
- **Molecular Function**: function(s) carried out by a gene product
- **Biological Process**: biological phenomena the gene is involved in (e.g. cell cycle, DNA replication, limb formation)
- Hierarchical organization ("is a", "is part of")

Quality of clustering

GO enrichment analysis: TANGO

[Tanay, 2003]

- Are there more genes from a given GO class in a given cluster than expected by chance?
- Assume genes sampled from the **hypergeometric** distribution

$$Pr(|C \cap G| \geq t) = 1 - \sum_{i=1}^{t} \frac{\binom{|G|}{i}\binom{n-|G|}{|C|-i}}{\binom{n}{|C|}}$$

- Correct for **multiple hypothesis testing**
  - Bonferroni too conservative (dependencies between GO groups)
  - Empirical computation of the null distribution

Quality of clustering

Gene Set enrichment analysis (GSEA)

[Subramanian et al., 2005]

- Use correlation to a phenotype y
- Rank genes according to the correlation ρ_i of their expression to $y \rightarrow L = \{g_1, g_2, \ldots, g_n\}$
- $P_{hit}(C,i) = \sum_{j:j \leq i, g_j \in C} \frac{|\rho_j|}{\sum_{g_j \in C} |\rho_j|}$
- $P_{miss}(C,i) = \sum_{j:j \leq i, g_j \notin C} \frac{1}{n-|C|}$
- **Enrichment score**: $ES(C) = \max_i |P_{hit}(C,i) - P_{miss}(C,i)|$

Hierarchical clustering

Linkage

- **single linkage**: $d(A,B) = \min_{x \in A, y \in B} d(x,y)$
- **complete linkage**: $d(A,B) = \max_{x \in A, y \in B} d(x,y)$
- **average (arithmetic) linkage**:
  $d(A,B) = \sum_{x \in A, y \in B} d(x,y)/|A||B|$
  also called UPGMA
  (Unweighted Pair Group Method with Arithmetic Mean)
- **average (centroid) linkage**:
  $d(A,B) = d(\sum_{x \in A} x/|A|, \sum_{y \in B} y/|B|)$
  also called UPGMC
  (Unweighted Pair-Group Method using Centroids)

Hierarchical clustering

Construction

- **Agglomerative** approach (bottom-up)

  Start with every element in its own cluster, then iteratively join nearby clusters

- **Divisive** approach (top-down)

  Start with a single cluster containing all elements, then recursively divide it into smaller clusters

Hierarchical clustering

Advantages

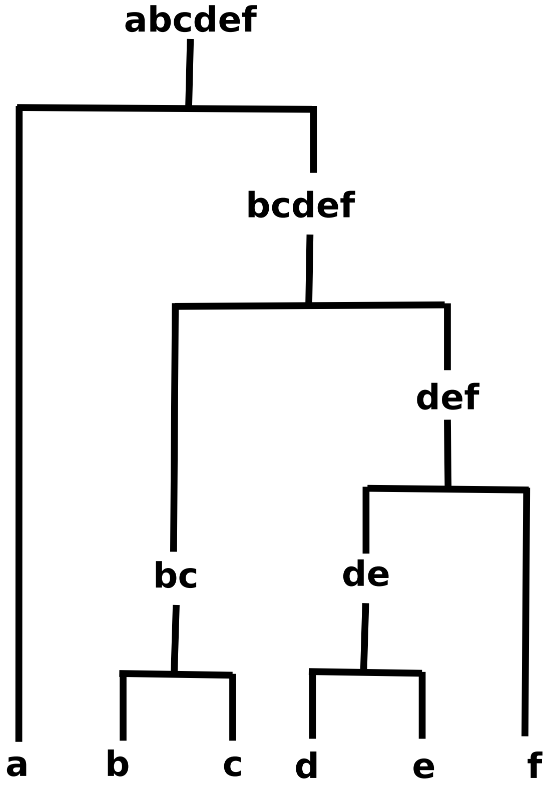
- Does not require to set the number of clusters
- Good interpretability

Drawbacks

- Computationally intensive $\mathcal{O}(n^2 log\ n^2)$
- Hard to decide at which level of the hierarchy to stop
- Lack of robustness
- Risk of locking accidental features (local decisions)

Hierarchical clustering

Dendrograms

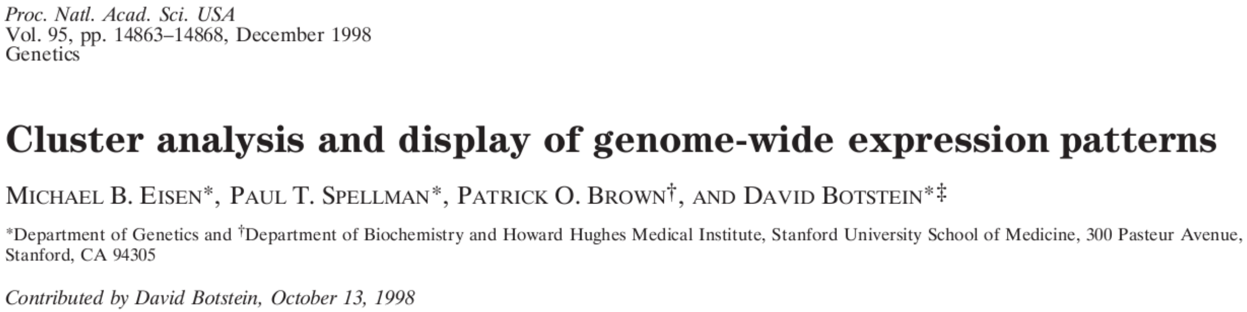


In biology

- Phylogenetic trees
- Sequences analysis
  infer the evolutionary history of sequences being compared

Hierarchical clustering

[Eisen et al., 1998]



Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

Cluster analysis and display of genome-wide expression patterns

Michael B. Eisen*, Paul T. Spellman*, Patrick O. Brown†, and David Botstein*‡

*Department of Genetics and †Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305

Contributed by David Botstein, October 13, 1998

Motivation

- Arrange genes according to similarity in pattern of gene expression
- Graphical display of output
- Efficient grouping of genes of similar functions

Hierarchical clustering

[Eisen et al., 1998]

Data

- *Saccharomyces cerevisiae*:
  - DNA microarrays containing all ORFs
  - Diauxic shift; mitotic cell division cycle; sporulation; temperature and reducing shocks
- Human
  - $9\,800$ cDNAs representing $\sim 8\,600$ transcripts
  - fibroblasts stimulated with serum following serum starvation

Data pre-processing

Cy5 (red) and Cy3 (green) fluorescences $\rightarrow log_2(\text{Cy5}/\text{Cy3})$

Hierarchical clustering

[Eisen et al., 1998]

Methods

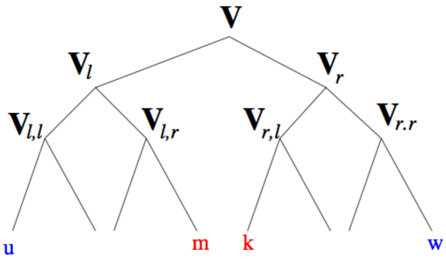
- Distance: Pearson's correlation
- Pairwise average-linkage cluster analysis
- Ordering of elements:
  - Ideally: such that adjacent elements have maximal similarity (impractical)
  - In practice: rank genes by average gene expression, chromosomal position

Hierarchical clustering

[Bar-Joseph et al., 2001]

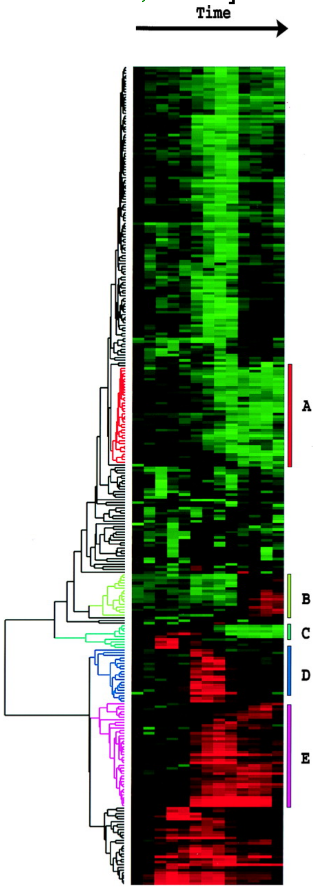
Fast optimal leaf ordering for hierarchical clustering

- n leaves → $2^n - 1$ possible ordering
- Goal: maximize the sum of similarities of adjacent leaves in the ordering
- Recursively find, for a node v, the cost $\mathcal{C}(v, u_l, u_r)$ of the optimal ordering rooted at v with left-most leaf u_l and right-most leaf u_r
- Work bottom up: $\mathcal{C}(v, u, w) = \mathcal{C}(v_l, u, m) + \mathcal{C}(v_r, k, w) + \sigma(m, k)$, where $\sigma(m, k)$ is the similarity between m and k
- $\mathcal{O}(n^4)$ time, $\mathcal{O}(n^2)$ space
- Early termination → $\mathcal{O}(n^3)$



Hierarchical clustering

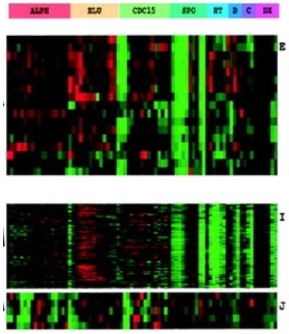
[Eisen et al., 1998]



- Genes “represent” more than a mere cluster together
- Genes of similar function cluster together
  - cluster A: cholesterol biosyntehsis
  - cluster B: cell cycle
  - cluster C: immediate-early response
  - cluster D: signaling and angiogenesis
  - cluster E: tissue remodeling and wound healing

Hierarchical clustering

[Eisen et al., 1998]



- cluster E: genes encoding glycolytic enzymes share a function but are not members of large protein complexes
- cluster J: mini-chromosomoe maintenance DNA replication complex

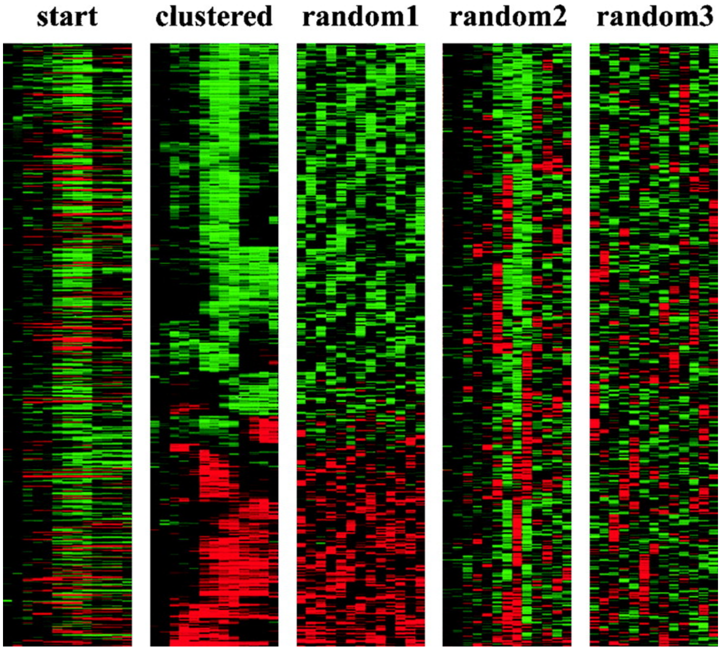
- cluster I: 126 genes strongly down-regulated in response to stress 112 of those encode ribosomal proteins Yeast responds to favorable growth conditions by increasing the production of ribosome, through transcriptional regulation of genes encoding ribosomal proteins

Hierarchical clustering

[Eisen et al., 1998]

Validation

Randomized data does not cluster



Hierarchical clustering

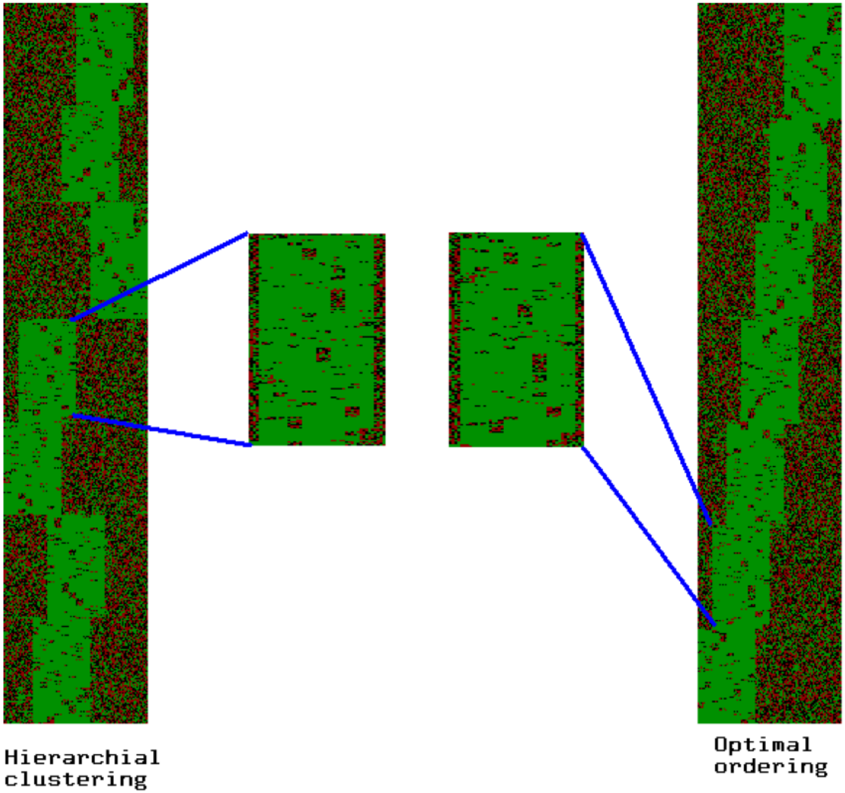
[Eisen et al., 1998]

Conclusions

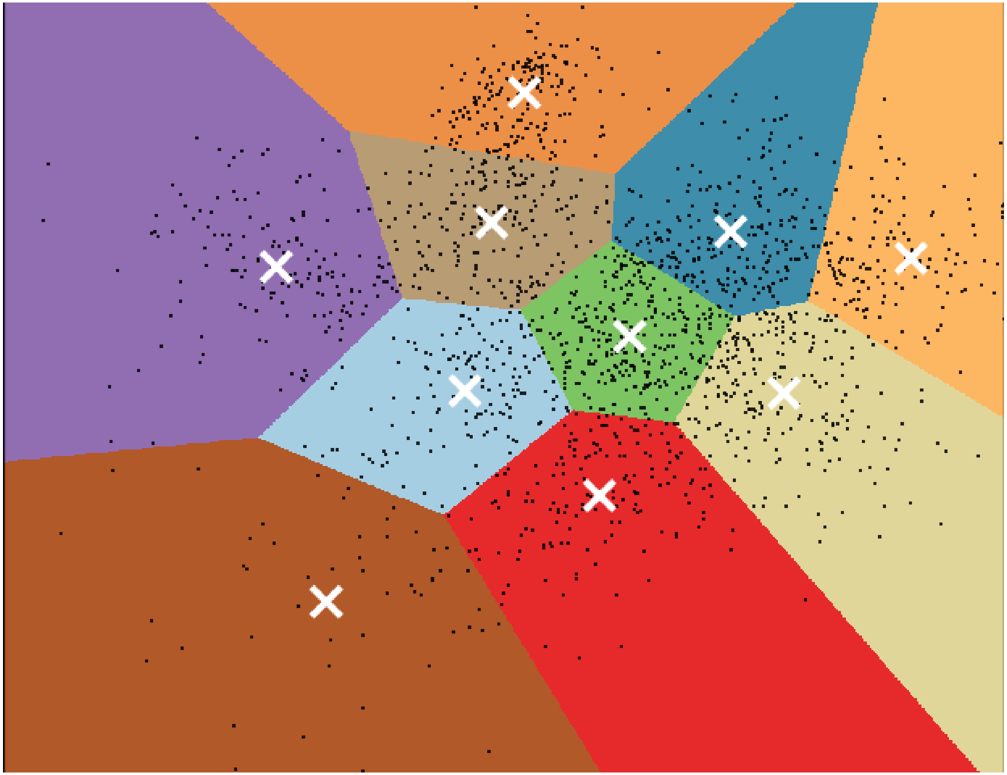
- Hierarchical clustering of gene expression data groups together genes that are known to have similar functions
- Gene expression clusters reflect biological processes
- Coexpression data can be used to infer the function of new / poorly characterized genes

Hierarchical clustering

[Bar-Joseph et al., 2001]



K-means clustering



source: scikit-learn.org

K-means clustering

Advantages

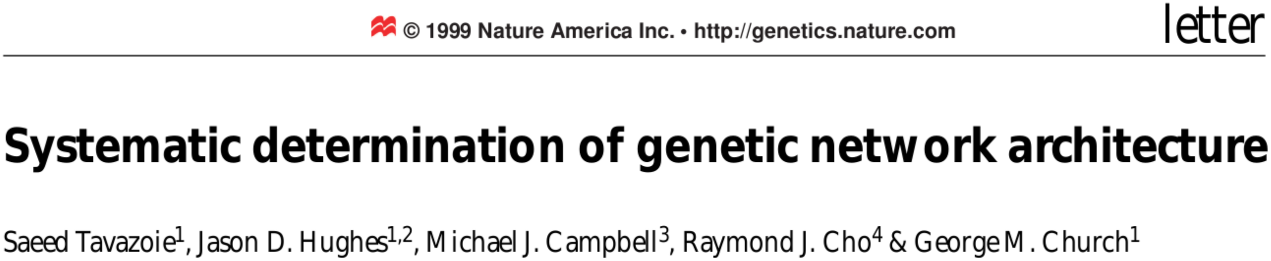
- Relatively efficient $\mathcal{O}(ntk)$
  n objects, k clusters, t iterations
- Easily implementable

Drawbacks

- Need to specify k ahead of time
- Sensitive to noise and outliers
- Clusters are forced to have convex shapes
  (kernel k-means can be a solution)
- Results depend on the initial, random partition (k-means++ can be a solution)

K-means clustering

[Tavazoie et al., 1999]



© 1999 Nature America Inc. • http://genetics.nature.com

letter

Systematic determination of genetic network architecture

Saeed Tavazoie[1], Jason D. Hughes[1,2], Michael J. Campbell[3], Raymond J. Cho[4] & George M. Church[1]

Motivation

- Use whole-genome mRNA data to identify transcriptional regulatory sub-networks in yeast
- Systematic approach, minimally biased to previous knowledge
- An upstream DNA sequence pattern common to all mRNAs in a cluster is a candidate *cis*-regulatory element

K-means clustering

[Tavazoie et al., 1999]

Data

- Oligonucleotide microarrays, $6\,220$ mRNA species
- 15 time points across two cell cycles

Data pre-processing

- variance-normalization
- keep the most variable $3\,000$ ORFs

K-means clustering

[Tavazoie et al., 1999]

Methods

- k-means, $k = 30 \to$ 49–186 ORFs per cluster
- cluster labeling:
  - map the genes to 199 functional categories (MIPS[a] database)
  - compute p-values of observing frequencies of genes in particular functional classes

    cumulative hypergeometric probability distribution for finding at least k ORFs (g total) from a single functional category (size f) in a cluster of size n

$$P = 1 - \sum_{i=1}^{k} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}}$$

  - correct for 199 tests

[a] Martinsried Institute of Protein Science

K-means clustering

[Tavazoie et al., 1999]

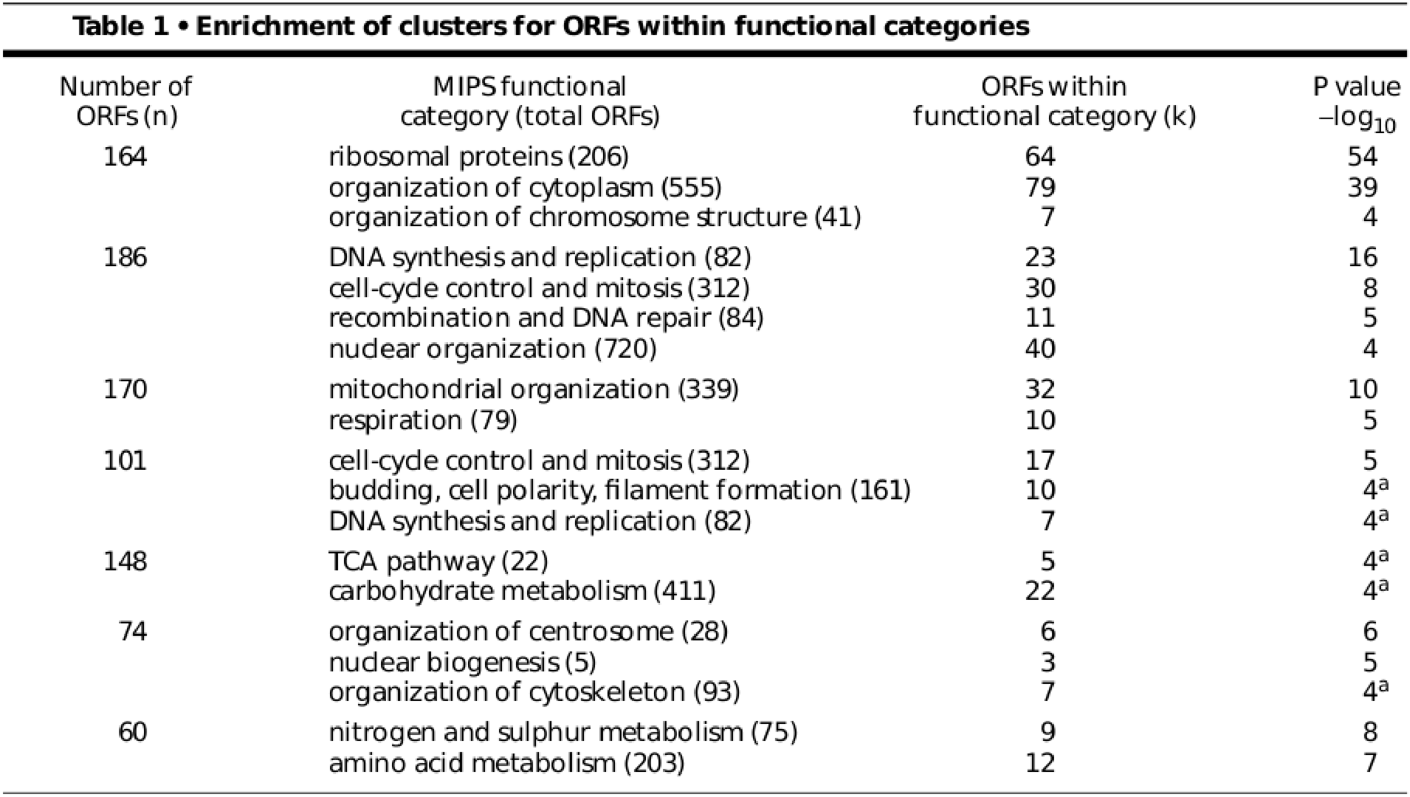
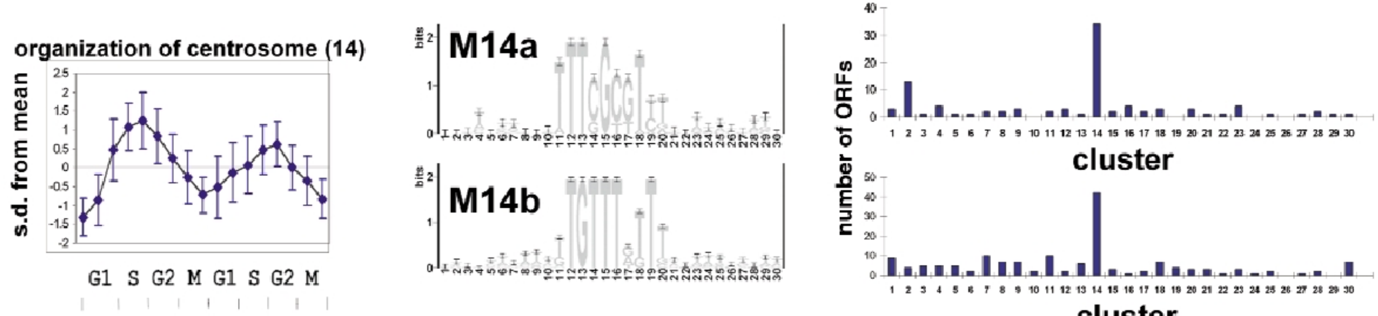


Table 1 • Enrichment of clusters for ORFs within functional categories

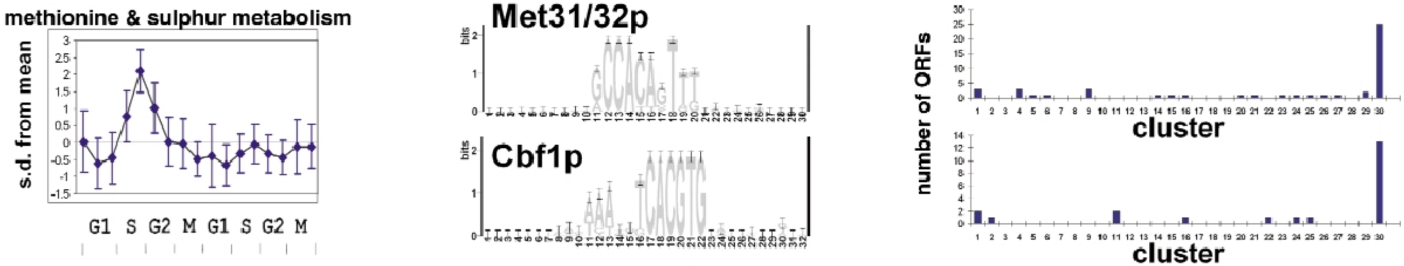
Number of ORFs (n)	MIPS functional category (total ORFs)	ORFs within functional category (k)	P value $-\log_{10}$
164	ribosomal proteins (206)	64	54
	organization of cytoplasm (555)	79	39
	organization of chromosome structure (41)	7	4
186	DNA synthesis and replication (82)	23	16
	cell-cycle control and mitosis (312)	30	8
	recombination and DNA repair (84)	11	5
	nuclear organization (720)	40	4
170	mitochondrial organization (339)	32	10
	respiration (79)	10	5
101	cell-cycle control and mitosis (312)	17	5
	budding, cell polarity, filament formation (161)	10	4[a]
	DNA synthesis and replication (82)	7	4[a]
148	TCA pathway (22)	5	4[a]
	carbohydrate metabolism (411)	22	4[a]
74	organization of centrosome (28)	6	6
	nuclear biogenesis (5)	3	5
	organization of cytoskeleton (93)	7	4[a]
60	nitrogen and sulphur metabolism (75)	9	8
	amino acid metabolism (203)	12	7

K-means clustering

[Tavazoie et al., 1999]



Periodic cluster



Aperiodic cluster

K-means clustering

[Tavazoie et al., 1999]

Conclusions

- Clusters with significant functional enrichment tend to be tighter (mean Euclidean distance)
- Tighter clusters tend to have significant upstream motifs
- Discovered new regulons

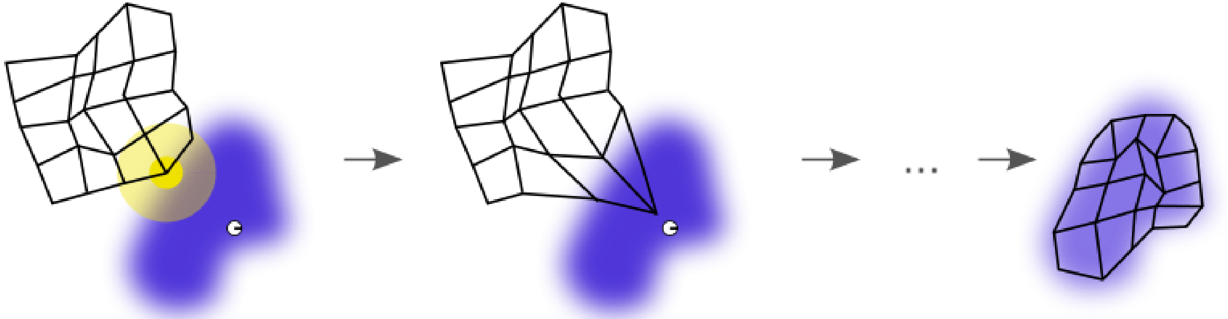
Self-organizing maps

- a.k.a. **Kohonen networks**
- Impose partial structure on the clusters
- Start from a geometry of nodes $\{N_1, N_2, \ldots, N_k\}$
  E.g. grids, rings, lines
- At each iteration, randomly select a data point P, and move the nodes towards P.
- The nodes closest to P move the most, and the nodes furthest from P move the least.

$$f^{(t+1)}(N) = f^{(t)}(N) + \tau(t, d(N, N_P))(P - f^{(t)}(N)) \qquad N_P: \text{ node closest to } P$$

- The **learning rate** τ decreases with t and the distance from N_P to N

Self-organizing maps



Source: Wikimedia Commons – Mcld

Self-organizing maps

Advantages

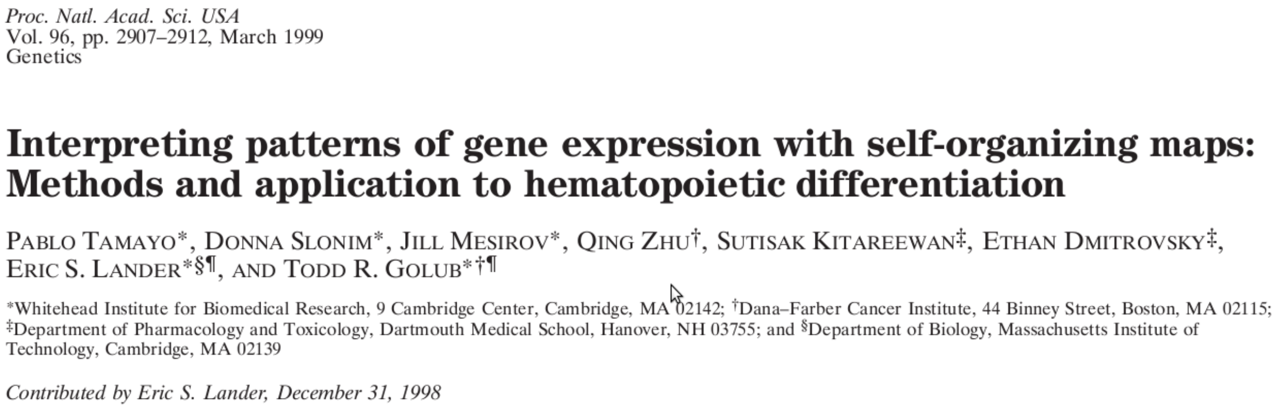
- Can impose partial structure
- Visualization

Drawbacks

- Multiple parameters to set
- Need to set an initial geometry

Self-organizing maps

[Tamayo et al., 1999]



Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 2907–2912, March 1999
Genetics

Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation

PABLO TAMAYO[*], DONNA SLONIM[*], JILL MESIROV[*], QING ZHU[†], SUTISAK KITAREEWAN[‡], ETHAN DMITROVSKY[‡], ERIC S. LANDER[*§¶], AND TODD R. GOLUB[*†¶]

[*]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142; [†]Dana–Farber Cancer Institute, 44 Binney Street, Boston, MA 02115; [‡]Department of Pharmacology and Toxicology, Dartmouth Medical School, Hanover, NH 03755; and [§]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

Contributed by Eric S. Lander, December 31, 1998

Motivation

- Extract fundamental patterns of gene expression
- Organize the genes into biologically relevant clusters
- Suggest novel hypotheses

Self-organizing maps

[Tamayo et al., 1999]

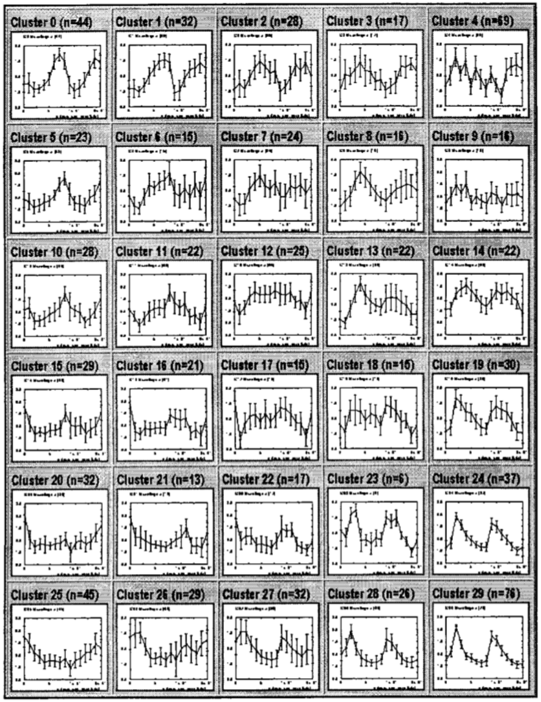
Data

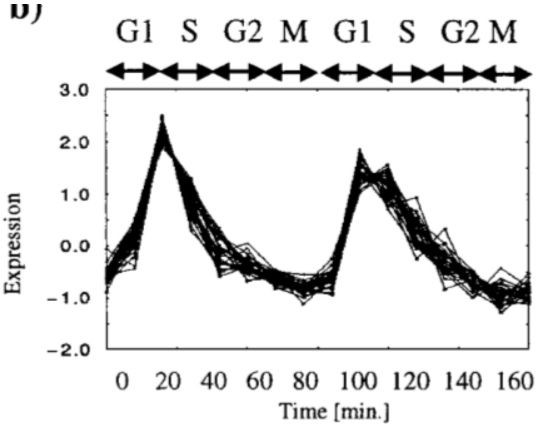
- Yeast
  - $6\,218$ ORFs
  - 2 cell cycles, every 10 minutes
  - SOM: 6×5 grid
- Human
  - Macrophage differentiation in HL-60 cells (myeloid leukemia cell line)
  - $5\,223$ genes
  - cells harvested at 0, 0.5, 4 and 24 hours after PMA stimulation
  - SOM: 4×3 grid

Self-organizing maps

[Tamayo et al., 1999]

Results: Yeast



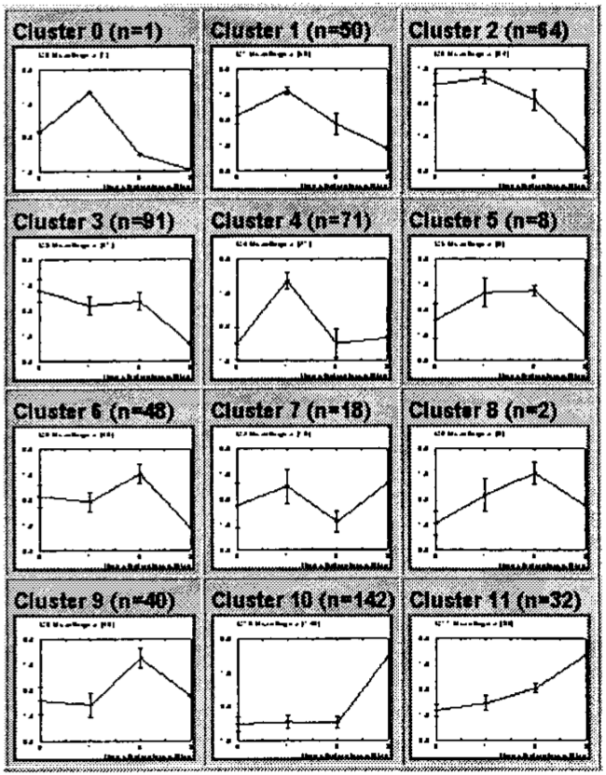


- Periodic behavior
- Adjacent clusters have similar behavior

Self-organizing maps

[Tamayo et al., 1999]

Results: HL-60



Cluster 11:

- gradual induction as cells lose proliferative capacity and acquire hallmarks of the macrophage lineage
- $8/32$ genes *not* expected given current knowledge of hematopoietic differentiation
- 4 of those suggest role of immunophilin-mediated pathway in macrophage differentiation

Self-organizing maps

[Tamayo et al., 1999]

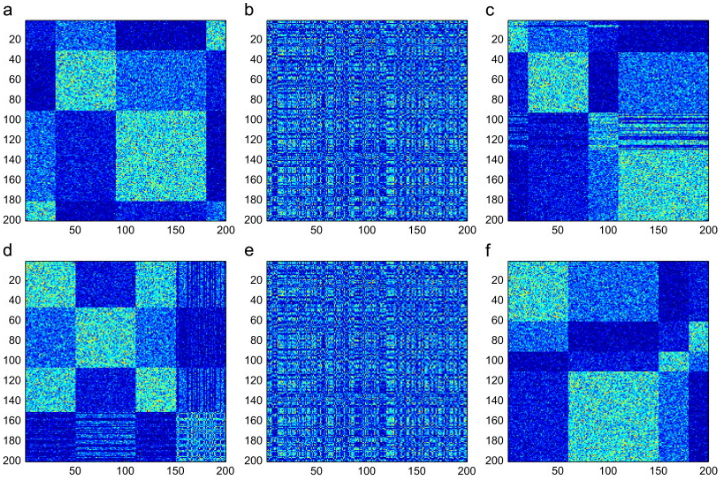
Conclusions

- Extracted the k most prominent patterns to provide an "executive summary"
- Small data, but illustrative:
  - Cell cycle periodicity recovered
  - Genes known to be involved in hematopoietic differentiation recovered
  - New hypotheses generated
- SOMs scale well to larger datasets

Biclustering

Biclustering, co-clustering, two-ways clustering

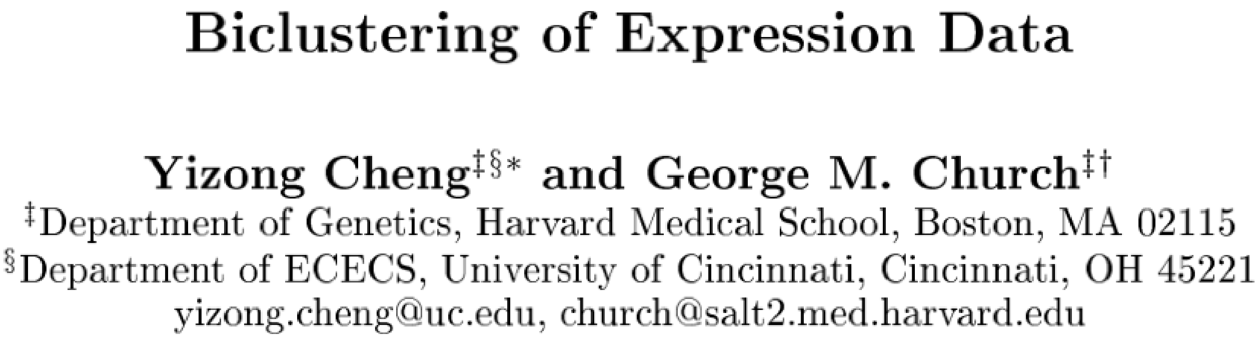
- Find subsets of rows that exhibit similar behaviors across subsets of columns
- **Bicluster**: subset of genes that show similar expression patterns across a subset of conditions/tissues/samples



source: [Yang and Oja, 2012]

Biclustering

[Cheng and Church, 2000]



Biclustering of Expression Data

Yizong Cheng[‡§*] and George M. Church[‡†]
[‡]Department of Genetics, Harvard Medical School, Boston, MA 02115
[§]Department of ECECS, University of Cincinnati, Cincinnati, OH 45221
yizong.cheng@uc.edu, church@salt2.med.harvard.edu

Motivation

- Simultaneous clustering of genes and conditions
- Overlapped grouping

  More appropriate for genes with multiple functions or regulated by multiple factors

Biclustering

[Cheng and Church, 2000]

Algorithm

- Goal: minimize intra-cluster variance
- **Mean Squared Residue**:

$$\text{MSR}(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (x_{ij} - x_{iJ} - x_{Ij} + x_{IJ})^2$$

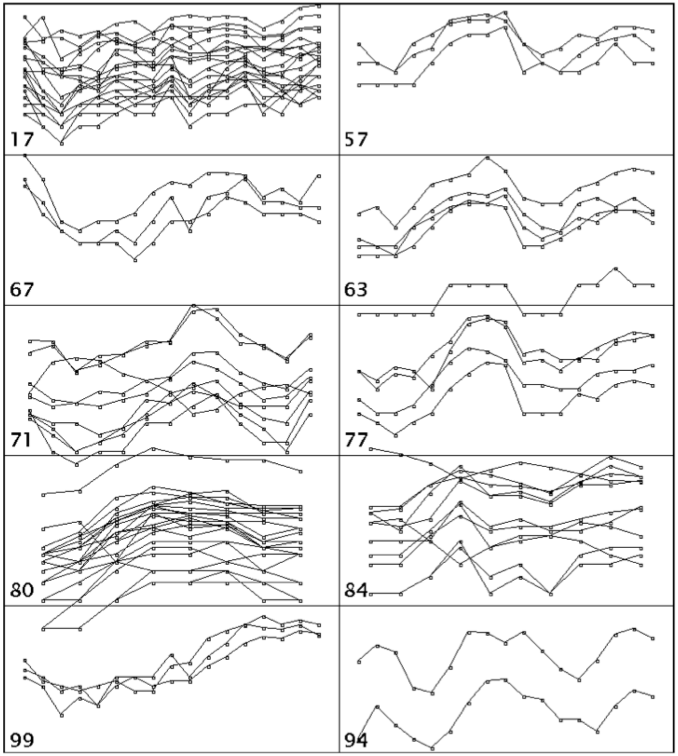
  x_{iJ}, x_{Ij}, x_{IJ}: mean expression values in row i, column j, and over the whole cluster

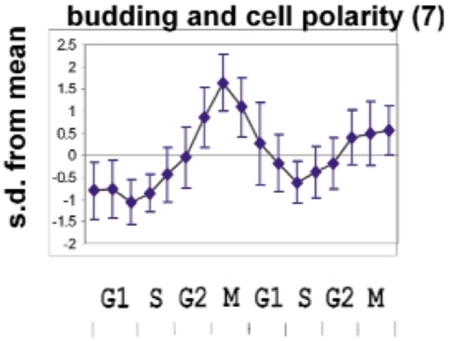
- δ: maximum acceptable MSR
- **Single Node Deletion**: remove rows/columns of X with largest variance $\left(\frac{1}{|J|}\sum_{j \in J}(x_{ij} - x_{iJ} - x_{Ij} + x_{IJ})^2\right)$ until MSR $< \delta$
- **Node Addition**: some rows/columns may be added back without increasing MSR
- **Masking Discovered Biclusters**: replace the corresponding entries by random numbers

Biclustering

[Cheng and Church, 2000]

Results: Yeast





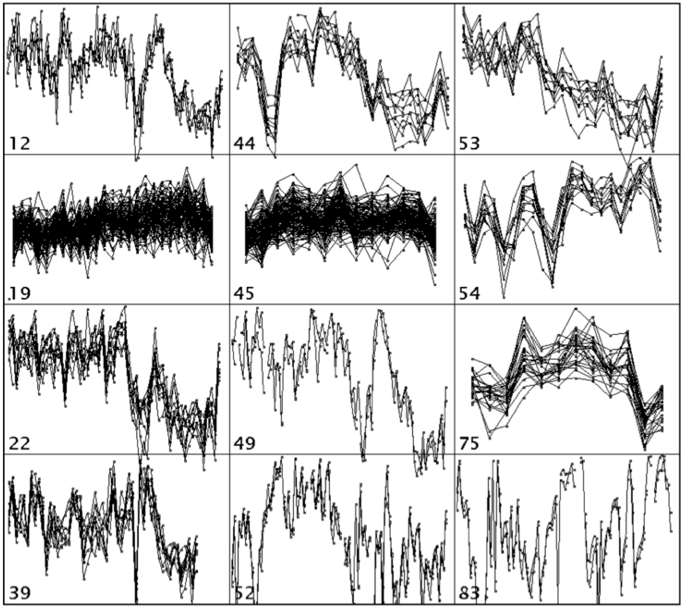
- Biclusters 17, 67, 71, 80, 90 contain genes in clusters 4, 8, 12 of [Tavazoie et al., 1999]
- Biclusters 57, 63, 77, 84, 94 represent cluster 7 of [Tavazoie et al., 1999]

Biclustering

[Cheng and Church, 2000]

Results: Human B-cells

Data: $4\,026$ genes, 96 samples of normal and malignant lymphocytes



Cluster 12: 4 genes, 96 conditions

19: 103, 25
22: 10, 57
39: 9, 51
44:10, 29
45: 127, 13
49: 2, 96
52: 3, 96
53: 11, 25
54: 13, 21
75: 25, 12
83: 2, 96

Biclustering

[Cheng and Church, 2000]

Conclusion

- Biclustering algorithm that does not require computing pairwise similarities between all entries of the expression matrix
- Global fitting
- Automatically drops noisy genes/conditions
- Rows and columns can be included in multiple biclusters

References and further reading

[Bar-Joseph et al., 2001] Bar-Joseph, Z., Gifford, D. K. and Jaakkola, T. S. (2001). Fast optimal leaf ordering for hierarchical clustering. Bioinformatics *17*, S22–S29. 22, 27

[Cheng and Church, 2000] Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In Proceedings of the eighth international conference on intelligent systems for molecular biology vol. 8, pp. 93–103,. 45, 46, 47, 48, 49

[Eisen et al., 1998] Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences *95*, 14863–14868. 19, 20, 21, 23, 24, 25, 26

[Eren et al., 2012] Eren, K., Deveci, M., Küçüktunç, O. and Çatalyürek, U. V. (2012). A comparative analysis of biclustering algorithms for gene expression data. Briefings in Bioinformatics .

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545–15550. 14

[Tamayo et al., 1999] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences *96*, 2907–2912. 39, 40, 41, 42, 43

[Tanay, 2003] Tanay, A. (2003). The TANGO program technical note. `http://acgt.cs.tau.ac.il/papers/TANGO_manual.txt`. 13

[Tavazoie et al., 1999] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., Church, G. M. et al. (1999). Systematic determination of genetic network architecture. Nature genetics *22*, 281–285. 30, 31, 32, 33, 34, 35, 47

[von Luxburg, 2009] von Luxburg, U. (2009). Clustering stability: an overview. Foundations and Trends in Machine Learning *2*, 235–274. 11

[Yang and Oja, 2012] Yang, Z. and Oja, E. (2012). Quadratic nonnegative matrix factorization. Pattern Recognition *45*, 1500–1510. 44