

# Métodos de clustering em dados de Microarray

## Introdução à biologia computacional

Clerton Ribeiro de A. Filho

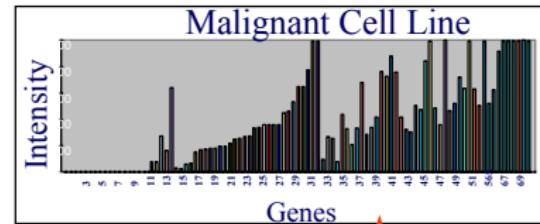
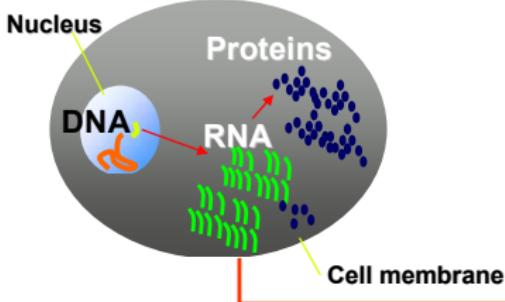
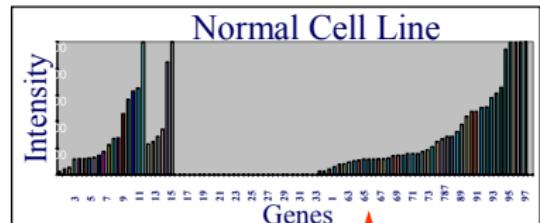
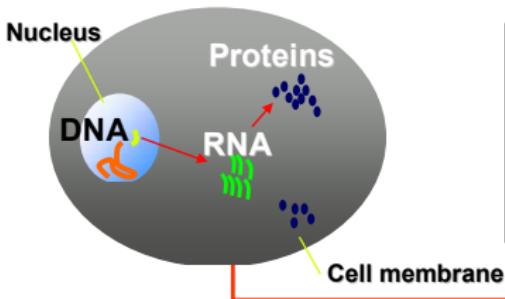
Universidade Federal de Pernambuco (UFPE)

11 de junho de 2010

- Microarray;
- Clustering;
- Agrupamento hierárquico;
- SOM (Self-Organizing Maps);
- K-means;
- PAM (Partitioning Around Medoids);
- Modelo de Misturas;
- Número de  $K$  clusters;
- Rand Index;
- P-value;
- Comparação entre métodos;
- Discussão.

- Mede a quantidade de moléculas de mRNA codificados por cada gene;
  - O número de cópias de moléculas de mRNA em uma célula em particular é um bom indicador da expressão dos genes (número de proteínas);
- Genes são diferentemente expressados em:
  - Diferentes tipos celulares (células musculares, fibroblastos, ...);
  - Condições de ambiente (choque térmico, privação de nutrientes, ...);
  - Fases de desenvolvimento (décimo dia embrionário,...)
  - Estágios do ciclo celular (fase G1,...)
  - Estágios de doenças (células com tumores, células infectadas por vírus, ...)

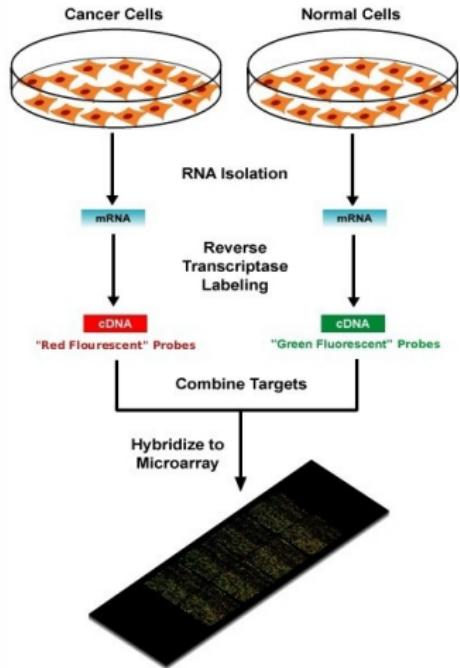
## Gene expression = RNA “volume”



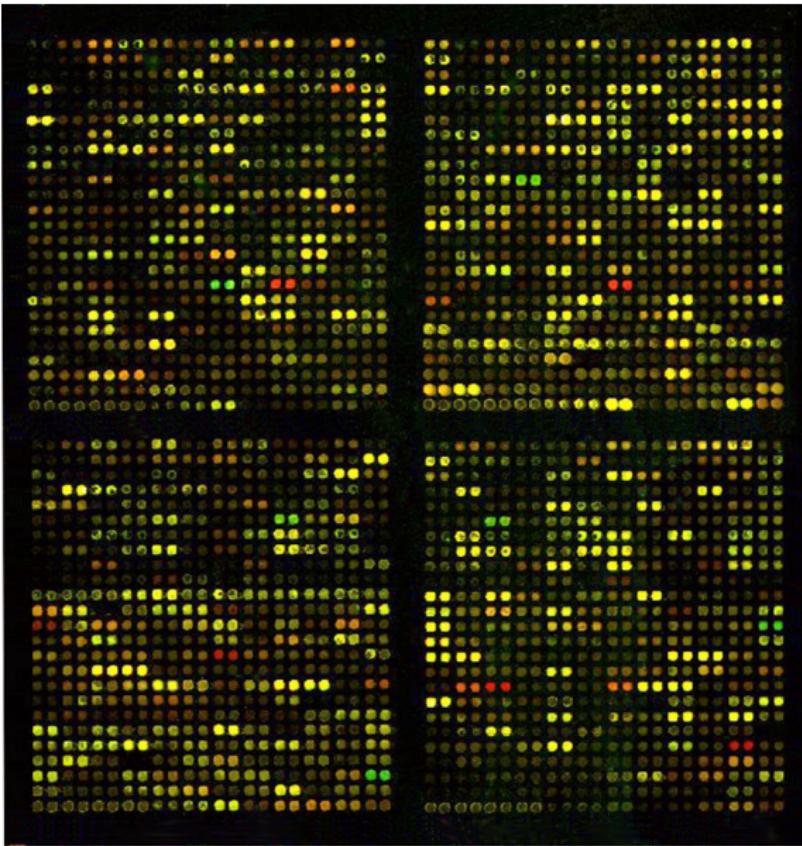
# Microarray

## Experimento com microarrays:

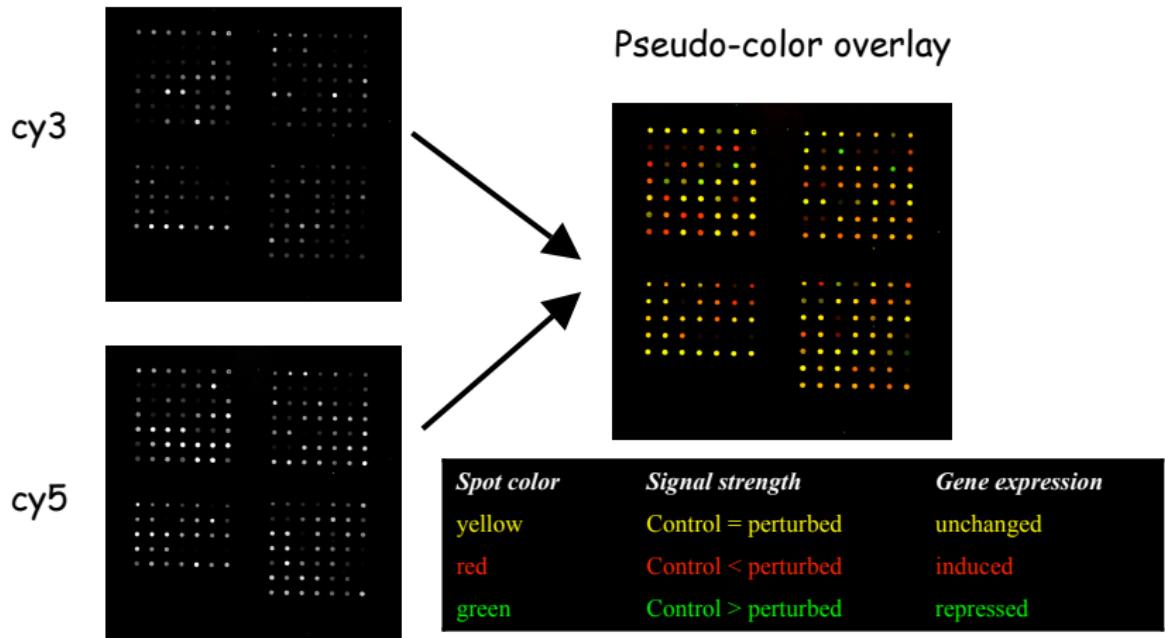
- ① Isolar o mRNA;
- ② Fazer a transcrição reversa do mRNA para o cDNA;
- ③ Rotular o cDNA incorporando nucleotídeos fosforecentes;
- ④ Hibridizar o cDNA no microarray;
- ⑤ "Scanear" o microarray com um scanner laser focal;
- ⑥ Analisar os dados.



# Microarray

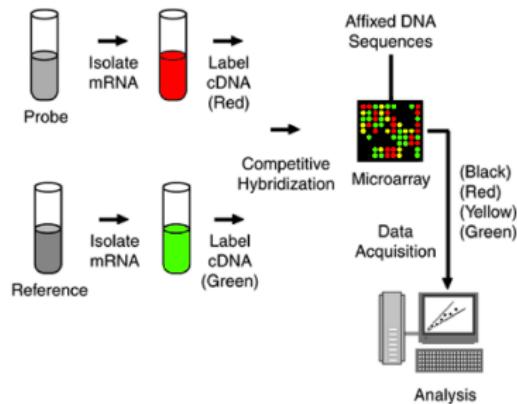


## cDNA microarray scanning



# Microarray

- Alta escala (genomas completos);
- Baixo custo;
- Experimentos comparativos;
- Comparando os sinais das amostras pode-se entender como os genes delas se diferem.



# Microarray

Applications Places System 26 °C

National Center for Bi... http://www.ncbi.nlm.nih.gov/ Other Bookmarks

Tutorials NCBI Resources How To My NCBI Sign In

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

More about the NCBI | Mission | Organization | Research | RSS

**3D Structures**  
Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, mutations, ligands, binding partners, and associated biotransformations.

1 2 3 4

**Popular Resources**

- BLAST
- Bookshelf
- Gene
- Genome
- Nucleotide
- OMIM
- Protein
- PubChem
- PubMed
- PubMed Central
- SNP

**NCBI News**

Education resource information in the May NCBI News 07 Jun 2010

May NCBI News is available.

OMIM's new look, Epigenomics in April NCBI News 10 May 2010

The April NCBI News Issue is now available.

NIH Roadmap Epigenomics Project data in GEO database 22 Apr 2010

GEO's Roadmap Epigenomics Project Data Listings page allows

More...

http://www.ncbi.nlm.nih.gov/#

# Microarray

Applications Places System 26 °C Thu Jun 10, 4:56 PM craf

Entrez cross-database... Tutorials Other Bookmarks

http://www.ncbi.nlm.nih.gov/gquery?term=gene+expression+leukemia

NCBI

PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases gene expression leukemia GO Clear Help

Result counts displayed in gray indicate one or more terms not found

24100 PubMed: biomedical literature citations and abstracts	89 Books: online books
19898 PubMed Central: free, full text journal articles	888 OMIM: online Mendelian Inheritance in Man
2 Site Search: NCBI web and FTP sites	none OMIA: online Mendelian Inheritance in Animals
2869 Nucleotide: Core subset of nucleotide sequence records	7 dbGaP: genotype and phenotype
12174 EST: Expressed Sequence Tag records	8 UniGene: gene-oriented clusters of transcript sequences
1063 GSS: Genome Survey Sequence records	5 CDD: conserved protein domain database
2804 Protein: sequence database	529 3D Domains: domains from Entrez Structure
none Genome: whole genome sequences	none UniSTS: markers and mapping data
247 Structure: three-dimensional macromolecular structures	none PopSet: population study data sets
none Taxonomy: organisms in GenBank	196552 GEO Profiles: expression and molecular abundance profiles
none SNP: single nucleotide polymorphism	514 GEO DataSets: experimental sets of GEO data
none dbVar: Genomic structural variation	271 Cancer Chromosomes: cytogenetic data
892 Gene: gene-centered information	432 PubChem BioAssay: bioactivity screens of chemical substances
none SRA: Sequence Read Archive	3 PubChem Compound: unique small molecule chemical structures
16 BioSystems: Pathways and systems of interacting molecules	83 PubChem Substance: deposited chemical substance records
1 HomoloGene: eukaryotic homology groups	none Protein Clusters: a collection of related protein sequences

Go to GEO DataSets Results Page

# Microarray

Applications Places System 26 °C Thu Jun 10, 5:02 PM craf

gene expression leuk... Other Bookmarks

Tutorials My NCBI Sign In [Register]

**GEO DATASETS Gene Expression Omnibus**

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search GEO DataSets for gene expression leukemia Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort By Send to

All: 514 DataSets: 53 Platforms: 14 Series: 447

Items 1 - 20 of 514 Page 1 of 26 Next

**1: GDS2926 record: Megakaryocytic differentiation: time course [ Homo sapiens ]**

Summary: Temporal analysis of phorbol ester-treated CHRF-288-11 megakaryoblastic cells induced to undergo megakaryocytic (Mk) differentiation and primary Mk (PrfMk) cells derived from cytokine-treated CD34+ peripheral blood cells. Results provide insight into molecular mechanisms underlying megakaryopoiesis.

Parent Platform: GPL887

Reference Series: GSE8914

Type: Expression profiling by array, log e ratio

Subsets: 4 agent, 2 cell line, 13 time sets.

Samples: 77

GSM87962: CHRF\_Expt3\_DMSO\_4d\_rep1  
GSM87963: CHRF\_Expt3\_DMSO\_4d\_rep2  
GSM87983: CHRF\_Expt4\_DMSO\_4d\_rep1  
GSM87984: CHRF\_Expt4\_DMSO\_7d\_rep1  
GSM87961: CHRF\_Expt3\_DMSO\_12d\_rep1  
GSM87970: CHRF\_Expt3\_PMA\_1h\_rep1

**2: GDS2908 record: T-cell prolymphocytic leukemia with Inv(14)(q11q32) [ Homo sapiens ]**

Summary: Analysis of T-cell prolymphocytic leukemia (T-PLL) cells with the cytogenetic abnormality inv(14)(q11q32). T-PLL is a rare aggressive lymphoma derived from mature T cells. Results provide insight into the pathogenesis of T-PLL.

Parent Platform: GPL96

Reference Series: GSE5788

Type: Expression profiling by array, count

Subsets: 2 disease state, 2 genotype/variation sets.

Supplementary Files: CEL download...

Samples: 14

GSM135264: T-cell prolymphocytic leukemia, inv(14)(q11q32), case 1  
GSM135265: T-cell prolymphocytic leukemia, inv(14)(q11q32), case 2

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8914>

**Top Organisms [Tree]**

- Homo sapiens (386)
- Mus musculus (138)
- Rattus norvegicus (5)
- Human herpesvirus 8 (2)
- Human herpesvirus 4 (2)
- Simian virus 40 (2)
- Macaca mulatta (1)
- Canis lupus familiaris (1)
- Murid herpesvirus 4 (1)
- Human herpesvirus 5 (1)
- Murid herpesvirus 1 (1)
- Human herpesvirus 1 (1)
- BK polyomavirus (1)
- JC polyomavirus (1)
- Human immunodeficiency virus 1 (1)

Recent activity Turn Off Clear

gene expression leukemia (514)  
gene expression lymphoma (269)  
Rattus norvegicus chemokine (C-X-C motif)

# Microarray

Applications Places System 26 °C Thu Jun 10, 5:02 PM craf

GEO Accession viewer http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8914 Other Bookmarks

Tutorials

NCBI GEO Gene Expression Omnibus

HOME | SEARCH | SITE MAP NCBI > GEO > Accession Display

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSE8914

**Series GSE8914** Query DataSets for GSE8914

Status Public on Aug 30, 2007  
Title CHRF-288-11 and primary human megakaryocytic cell cultures provide novel insights into lineage-specific differentiation  
Organism Homo sapiens  
Experiment type Expression profiling by array  
Summary This SuperSeries is composed of the following subset Series:  
GSE3838: Temporal expression profile of CHRF-288 cell line after phorbol ester stimulation  
GSE3839: Temporal expression profile of megakaryocytic differentiation primary CD34+ cell culture  
Keywords: SuperSeries

Overall design Refer to individual Series

Citation(s) Fuhrken PG, Chen C, Miller WM, Papoutsakis ET. Comparative, genome-scale transcriptional analysis of CHRF-288-11 and primary human megakaryocytic cell cultures provides novel insights into lineage-specific differentiation. *Exp Hematol* 2007 Mar;35(3):476-489. PMID: 17309828

Submission date Aug 30, 2007  
Last update date Sep 04, 2007  
Contact name Eleftherios T Papoutsakis  
E-mail(s) papoutsakis@dtb.udel.edu  
Organization name University of Delaware  
Department Chemical Engineering  
Street address 15 Innovation Way  
City Newark  
State/province DE  
ZIP/Postal code 19711  
Country USA

Platforms (1) GPL887 Agilent-012097 Human 1A Microarray (V2) G4110B (Feature Number version)

Samples (77) GSM87961 CHRF Expt3 DMSO 12d rep1

# Microarray

Applications Places System 26 °C Thu Jun 10, 5:03 PM craf

GEO Accession viewer Other Bookmarks

Tutorials

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8914>

CD34+ cell culture  
Keywords: SuperSeries

Overall design Refer to individual Series

Citation(s) Fuhrken PG, Chen C, Miller WM, Papoutsakis ET. Comparative, genome-scale transcriptional analysis of CHRF-288-11 and primary human megakaryocytic cell cultures provides novel insights into lineage-specific differentiation. *Exp Hematol* 2007 Mar;35(3):476-489. PMID: 17309828

Submission date Aug 30, 2007  
Last update date Sep 04, 2007  
Contact name Eleftherios T Papoutsakis  
E-mail(s) papoutsakis@dtb.udel.edu  
Organization name University of Delaware  
Department Chemical Engineering  
Street address 15 Innovation Way  
City Newark  
State/province DE  
ZIP/Postal code 19711  
Country USA

Platforms (1) [GPL887](#) Agilent-012097 Human 1A Microarray (V2) G4110B (Feature Number version)

Samples (77) [GSM87961](#) CHRF\_Expt3\_DMSO\_12d\_rep1  
[GSM87962](#) CHRF\_Expt3\_DMSO\_4d\_rep1  
[GSM87963](#) CHRF\_Expt3\_DMSO\_4d\_rep2

This SuperSeries is composed of the following SubSeries:  
[GSE3838](#) Temporal expression profile of CHRF-288 cell line after phorbol ester stimulation  
[GSE3839](#) Temporal expression profile of megakaryocytic differentiation primary CD34+ cell culture

**Download family**

	Format
SOFT formatted family file(s)	SOFT
MINIML formatted family file(s)	MINIML
<a href="#">Series Matrix File(s)</a>	TXT

**Supplementary data files not provided**

Raw data included within Sample table  
Processed data included within Sample table

NLM | NIH | GEO Help | Disclaimer | Section 508 |

vascript:OpenFTP('ftp://ftp.ncbi.nih.gov/pub/geo/DATA/SeriesMatrix/GSE8914/')

# Microarray

Applications Places System G 26 °C Thu Jun 10, 5:06 PM craf

GSE8914\_series\_matrix.txt (~/.cache/fr-RKmIRB) - gedit

File Edit View Search Tools Documents Help

Open Save Undo Redo Cut Copy Paste Find Replace

GSE8914\_series\_matrix.txt

lseries matrix_table begin																							
"ID_REF"	"GSM87961"	"GSM87962"	"GSM87963"	"GSM87964"	"GSM87965"	"GSM87966"	"GSM87967"	"GSM87968"	"GSM87969"	"GSM87970"	"GSM87971"	"GSM87972"	"GSM87973"	"GSM87974"	"GSM87975"	"GSM87976"	"GSM87977"						
1	NULL																						
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL						
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL						
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL						
2	NULL																						
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL						
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL						
3	NULL	1.5255	NULL	NULL	0.6698	1.0706	NULL	NULL	0.0989	NULL	NULL	0.9732	NULL	1.1811	1.0984	NULL	0.6523						
NULL	0.6214	0.9713	NULL	0.7141	NULL	-0.498	1.0451	1.1033	1.373	NULL	0.4686	1.1651	NULL	0.7291	1.3577	0.924	0.632						
1.0081	0.7418	NULL	1.2429	NULL	0.8568	1.2371	NULL	0.32	NULL	-0.7559	NULL	NULL	-0.3339	0.9978	0.4473	1.1619	NULL	0.6597					
1.0083	0.4599	0.2932	0.716	0.9087	0.4992	NULL	1.1433	1.2674	1.3242	1.0953													
4	NULL	NULL	NULL	NULL	-1.106	-0.6901	NULL	NULL	-0.3923	NULL	NULL	0.1399	NULL	-0.3419	-0.3022	NULL	-0.723	-0.0817	NULL				
NULL	NULL	-0.6222	NULL	NULL	NULL	-1.11	-1.2487	-1.3314	-0.7638	NULL	NULL	-1.0293	NULL	NULL	0.3362	NULL	-0.1593	NULL					
NULL	-0.5258	-0.497	-0.257	-0.5535	NULL	0.2952	NULL	0.6742	NULL	-0.6204	NULL	-0.6317	NULL	NULL	NULL	-0.5831	-0.7774	0.3396	NULL				
0.2189	-0.6184	-1.8832	-0.9266	NULL	-1.1067	-1.8707	0.0829	NULL	-0.525	NULL	0.2192	0.3557	-0.4566	-0.2936	NULL	0.2582	-0.1554	-0.0894	-0.1823	NULL			
5	NULL	0.1663	NULL	NULL	-0.4484	0.1925	NULL	0.3069	NULL	0.0735	NULL	0.199	NULL	0.9562	NULL	0.4261	0.2286	NULL	0.3977	0.0851	NULL		
NULL	-0.0718	0.0189	0.4756	0.3765	NULL	NULL	-0.2016	0.1867	0.2429	-0.1853	0.2132	0.2192	0.3557	-0.4566	-0.2936	NULL	0.2582	-0.1554	-0.0894	-0.1823	NULL		
0.1325	0.5122	0.546	NULL	0.1016	0.5127	NULL	0.4885	0.4201	0.2428	NULL	NULL	NULL	-0.1286	0.3795	0.4134	-0.0749	0.2641	-0.2344	0.6565	-0.0541	0.2577		
0.3022	0.5142	0.7484	NULL	0.1052	0.2309	NULL	NULL	NULL	NULL	NULL													
6	NULL	NULL	NULL	NULL	NULL																		
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL		
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL		
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL		
7	NULL	NULL	NULL	NULL	NULL																		
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL		
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL		
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL		
8	NULL	-0.0438	NULL	NULL	-0.5893	-0.2633	NULL	-0.0066	NULL	0.1519	NULL	-0.4997	NULL	0.2408	NULL	0.0022	0.1684	-0.1239	-0.2295	0.0681	-0.1652	NULL	
NULL	-0.3788	-0.439	NULL	-0.1471	NULL	-0.2355	0.012	-0.2117	0.0129	-0.4557	-0.2879	-0.364	-0.4987	-0.1555	NULL	-0.5948	-0.3968	-0.2934	-0.4392	-0.4078	-0.4393	NULL	
NULL	-0.3225	-0.4352	NULL	-0.3372	NULL	-0.5752	-0.2442	NULL	-0.0594	NULL	-0.6133	NULL	-0.9218	-0.5746	-0.4585	-0.4476	-0.6541	-0.7048	-0.2156	0.0991	NULL	NULL	
0.6318	-0.6003	-0.5099	-0.2652	0.4227	-0.319	0.5536	-0.4566	NULL	-0.1619	0.6876	0.4537	0.4593	0.549	0.4789	NULL								
9	NULL	0.0997	NULL	NULL	-0.2549	-0.3352	NULL	-0.287	NULL	-0.1839	NULL	-0.4992	0.0331	0.0936	NULL	-0.3933	-0.2905	-0.5431	-0.1333	0.2278	-0.3096	-0.0018	
NULL	0.036	0.2346	NULL	-0.3616	NULL	0.0063	-0.3605	-0.218	-0.0188	-0.1665	-0.3234	-0.2948	NULL	-0.5085	NULL	-0.1619	0.6876	0.4537	0.4593	0.549	0.4789	NULL	
0.532	0.3254	NULL	0.5716	-0.0813	0.3506	0.2774	0.3367	NULL	0.3645	0.178	0.2829	NULL	NULL	0.1268	0.0528	0.3548	0.124	0.288	0.5024	0.0828	0.0423	NULL	
0.155	0.0177	0.2195	0.121	0.2529	NULL	-0.453	0.2986	NULL	-0.0858	-0.1868	-0.1779	-0.1515											
10	NULL	NULL	NULL	-0.5443	-0.266	NULL	0.388	NULL	-0.5124	NULL	NULL	0.7233	NULL	NULL	-0.1114	-0.4995	-0.0893	-0.6327	NULL	NULL	0.481	NULL	
NULL	-0.5078	-0.104	NULL	-0.2452	NULL	0.0289	-0.1408	-0.4676	-0.0183	0.8339	-0.1627	-0.6211	-0.2179	-0.6941	NULL	0.0318	0.5317	-0.0296	-0.1937	NULL	-0.5179	0.2072	NULL
NULL	-0.131	NULL	0.6873	NULL	NULL	NULL	NULL	NULL	NULL	-0.453	0.2986	NULL	-0.0858	-0.1868	-0.1779	-0.1515	NULL						

# Microarray

Alguns dados públicos, normalizados, de microarray com expressão de câncer:

Dataset	Classes	n	C	d
Alizadeh-v2	DLBCL(42), FL(9), CLL(11)	62	3	4022
Alizadeh-v3	DLBCL1(21), DLBCL2(21), FL(9), CLL(11)	62	4	4022
Armstrong-v1	ALL(24), MLL(48)	72	2	12582
Armstrong-v2	ALL(24), MLL(20), AML(28)	72	3	12582
Chen	HCC(104), liver(75)	179	2	22699
Golub-v1	ALL(47), AML(25)	72	2	7129
Golub-v2	ALL-B(38), ALL-T(9), AML(25)	72	3	7129
Nutt-v2	CG(14), NG(14)	28	2	12625
Nutt-v3	CO(7), NO(15)	22	2	12625
Yeoh-v1	T-ALL(43), B-ALL(205)	248	2	12625

[http://algorithmics.molgen.mpg.de/Static/Supplements/  
CompCancer/datasets.htm](http://algorithmics.molgen.mpg.de/Static/Supplements/CompCancer/datasets.htm)

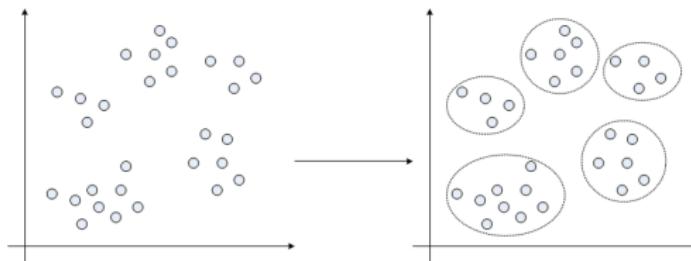
## O que é clustering?

- Clustering é um método de aprendizado não-supervisionado;
- Um cluster é uma coleção de objetos que são similares segundo um critério pré-estabelecido;
- Clustering é um processo de agrupar objetos similares em grupos;
- O objetivo do clustering é determinar um agrupamento intrínseco dado um conjunto de dados sem rótulos de classes.

# Clustering

Possíveis aplicações:

- Redução de dimensionalidade:
  - Retirar dados que são homogêneos (similares);
- Achar clusters "naturais";
- Achar novos grupos, novas relações entre os dados;
- Detecção de outliers.



Com microarrays, queremos encontrar genes co-expresos!

- Genes que participam do mesmo processo biológico;
- Simplificar a análise dos dados agrupando genes com padrões de expressão similares;
- Redução da dimensionalidade:
  - Ao invés de analisarmos 10.000 níveis de expressão, analisamos a expressão de 25 clusters, por exemplo!

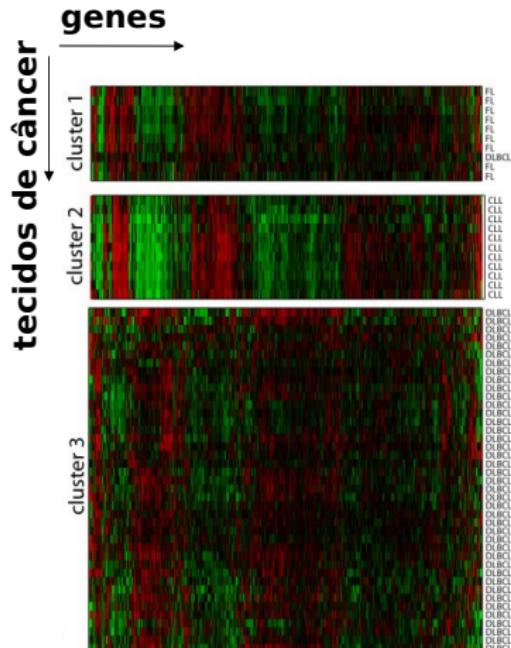
# Clustering

Desafios:

- Poucos amostras e muitos genes;
- Microarray são ruidosos;
- Sub-tipos desconhecidos de doenças:
  - Clustering permitiu encontrar novos tipos de linfomas (*Alizadeth et al.*)

Métodos mais usados em microarrays:

- Agrupamento hierárquico;
- K-means;
- Partition Around Medoids (PAM);
- Self-Organizing Maps (SOM);
- Modelos de misturas.



# Agrupamento hierárquico

- Organiza os dados em árvores (dendogramas), geralmente binárias;
- Usada principalmente para visualização e análise exploratória dos dados;
- Pode ser:
  - Aglomerativo: bottom-up
  - Divisivo: top-down
- A abordagem aglomerativa é mais usada:
  - Muitas heurísticas foram propostas para o problema, mas muitos violam o conceito de monotonicidade tornando difícil construir dendogramas!

# Agrupamento hierárquico

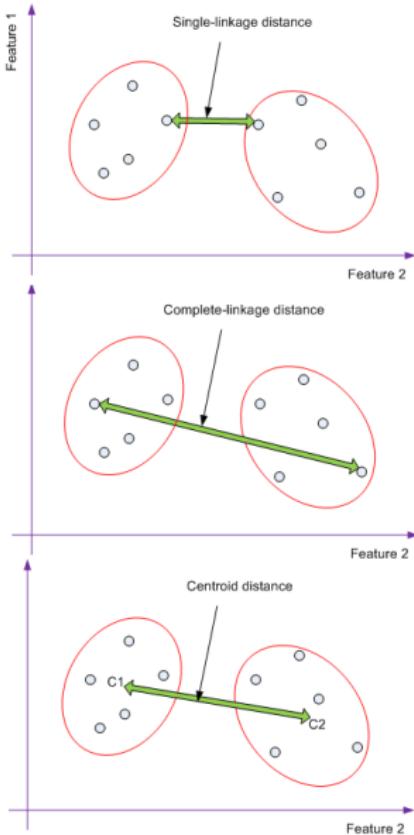
Algoritmo do agrupamento hierárquico aglomerativo:

- ① Inicio: Cada objeto é um grupo
- ② Iterar:
  - ① Escolher os dois grupos mais similares;
  - ② Juntar os grupos.
- ③ Retornar a árvore quando houver um só grupo.

# Agrupamento hierárquico

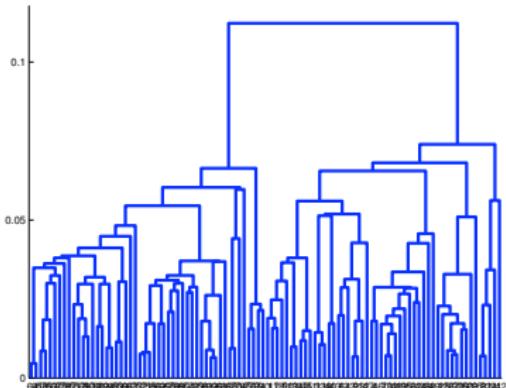
Medidas de similaridade:

- Single-linkage;
  - **Menor** distância de qualquer membro de um cluster para qualquer membro de outro;
- Complete-linkage;
  - **Maior** distância de qualquer membro de um cluster para qualquer membro de outro;
- Average-linkage;
  - Distância entre o **centróide** de um cluster e outro.

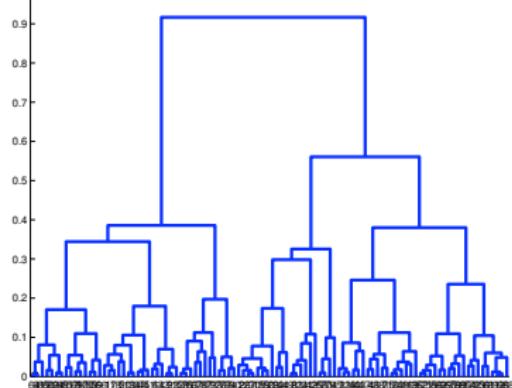


# Agrupamento hierárquico

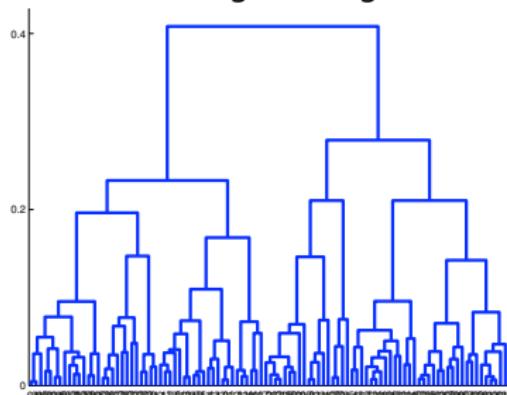
Single-linkage



Complete-linkage



Average-linkage



# Agrupamento hierárquico

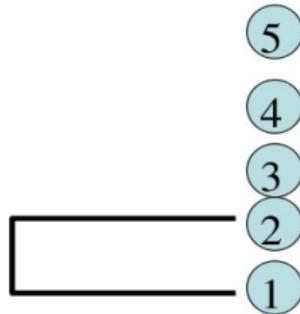
Exemplo: Single-linkage (I)

$$\begin{array}{cc} \begin{matrix} & 1 & 2 & 3 & 4 & 5 \\ 1 & \left[ \begin{matrix} 0 & & & & \\ 2 & 0 & & & \\ 3 & 6 & 3 & 0 & \\ 4 & 10 & 9 & 7 & 0 \\ 5 & 9 & 8 & 5 & 4 & 0 \end{matrix} \right] & \longrightarrow & \begin{matrix} (1,2) & 3 & 4 & 5 \\ (1,2) & \left[ \begin{matrix} 0 & & & \\ 3 & 3 & 0 & \\ 4 & 9 & 7 & 0 \\ 5 & 8 & 5 & 4 & 0 \end{matrix} \right] \end{matrix} \end{array}$$

$$d_{(1,2),3} = \min\{d_{1,3}, d_{2,3}\} = \min\{6,3\} = 3$$

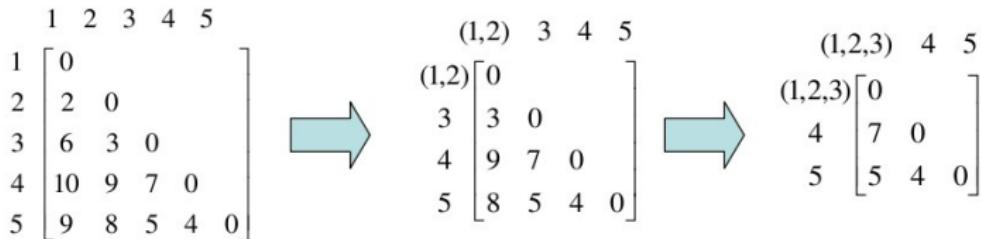
$$d_{(1,2),4} = \min\{d_{1,4}, d_{2,4}\} = \min\{10,9\} = 9$$

$$d_{(1,2),5} = \min\{d_{1,5}, d_{2,5}\} = \min\{9,8\} = 8$$



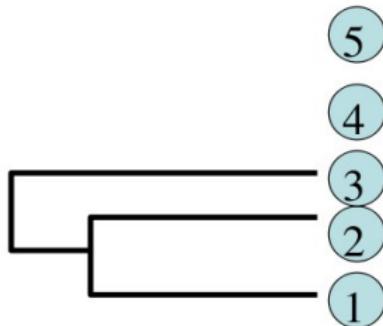
# Agrupamento hierárquico

Exemplo: Single-linkage (II)



$$d_{(1,2,3),4} = \min\{d_{(1,2),4}, d_{3,4}\} = \min\{9, 7\} = 7$$

$$d_{(1,2,3),5} = \min\{d_{(1,2),5}, d_{3,5}\} = \min\{8, 5\} = 5$$



# Agrupamento hierárquico

Exemplo: Single-linkage (III)

1	2	3	4	5
1	0			
2	2	0		
3	6	3	0	
4	10	9	7	0
5	9	8	5	4
				0

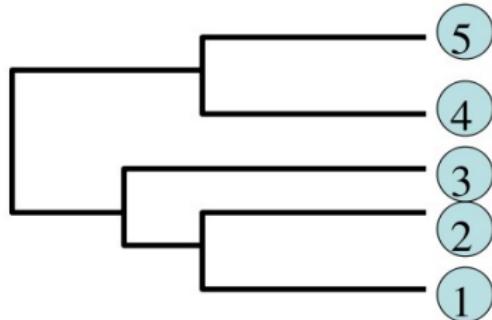


(1,2)	3	4	5	
(1,2)	0			
3	3	0		
4	9	7	0	
5	8	5	4	0



(1,2,3)	4	5		
(1,2,3)	0			
4	7	0		
5	5	4	0	

$$d_{(1,2,3),(4,5)} = \min\{d_{(1,2,3),4}, d_{(1,2,3),5}\} = 5$$



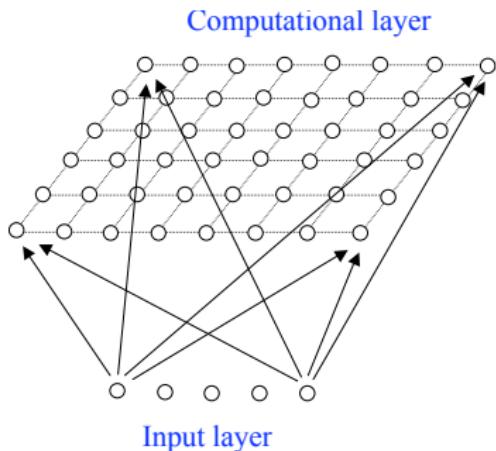
# SOM (Self-Organizing Maps)

- Rede neural com topologia dinâmica;
- Gera um mapeamento de um espaço de alta dimensão para uma estrutura de baixa dimensão;
- Regiões com alta densidade de sinal são representados pela alta densidade dos nós;
- Alguns tipos de SOM:
  - Kohonen: mais conhecida;
  - Growing cell structures (GCS);
  - Growing neural gas (GNG);

# SOM (Self-Organizing Maps)

Estrutura:

- Duas camadas de unidades:
  - Entrada:  $n$  unidades;
  - Saída:  $m$  unidades;
- A todas as unidades da camada de entrada se conectam com todas da de saída. Essas conexões têm pesos;
- Conexões laterais na camada de saída entre os nós:
  - Definido de acordo com a topologia;
  - Não possuem pesos, mas são usados no algoritmo para atualizar os pesos.
- Todo nó da camada de entrada é ligado a todos os nós da camada de saída;



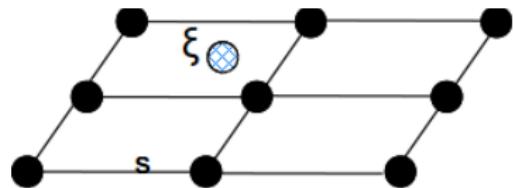
# SOM (Self-Organizing Maps)

O algoritmo é composto pelas seguintes fases:

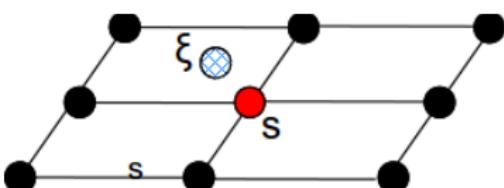
- ① Inicialização:
  - Todos os pesos são inicializados com pequenos valores aleatórios;
- ② Competição:
  - Para cada padrão de entrada, cada neurônio computa a distância entre ele e o padrão. O neurônio que obtiver a menor distância é o vencedor.
- ③ Adaptação:
  - O neurônio vencedor define o alcance topológico dos neurônios vizinhos;
- ④ Colaboração:
  - O aprendizado é estimulado para os neurônios vizinhos e diminuido para os restantes.

# SOM (Self-Organizing Maps)

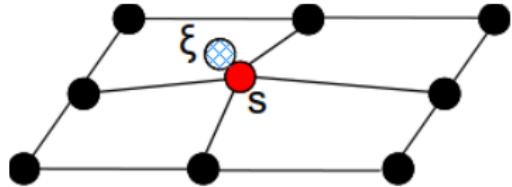
Inicialização



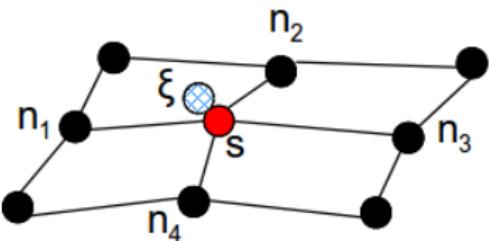
Competição



Adaptação



Colaboração



# K-means

- Um dos algoritmos mais simples para clustering;
- Só é necessário definir a quantidades de clusters desejados ( $K$ ) e os centroides (mesmo número de  $K$ )
- Os centroides tem que ser cuidadosamente escolhidos para não cair em ótimos locais;
- Tenta minimizar a função objetivo:

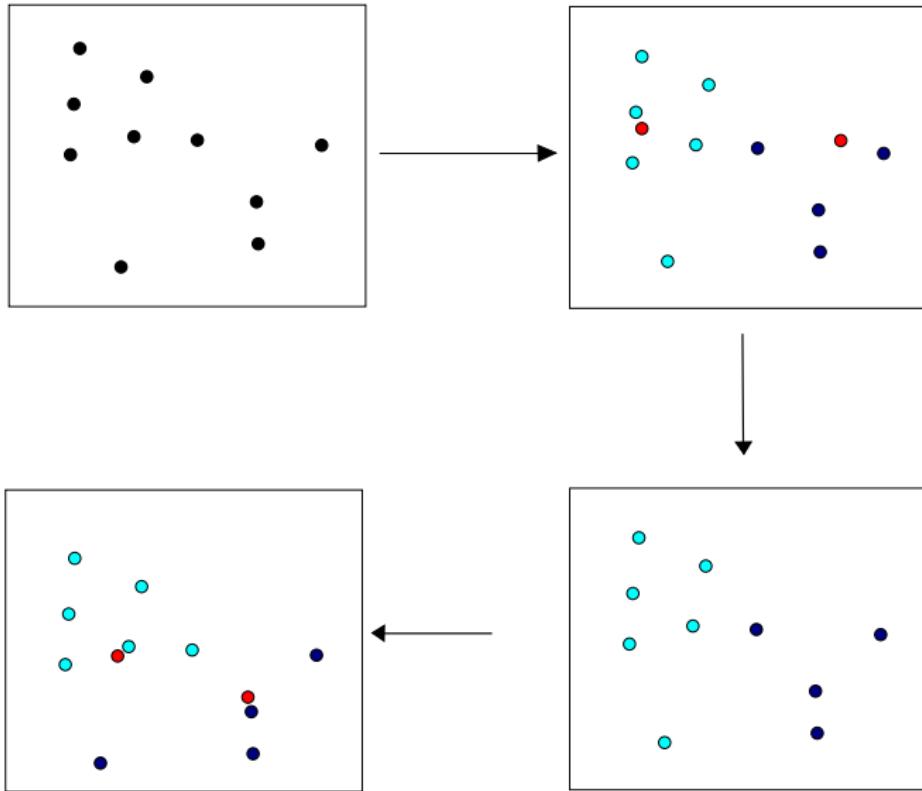
$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2, \quad (1)$$

onde  $x_i$  é um dos dados e  $c$  é um centroide

## Algoritmo:

- ① Inserir o  $K$  pontos aleatoriamente no espaço representados pelos objetos que serão agrupados. Esses pontos são os centróides iniciais;
- ② Associe o objeto para o grupo que contém o centróide mais próximo;
- ③ Quando todos os objetos forem associados, recalcule a posição dos centróides;
- ④ Repita os passos 2 e 3 até que os centróides não se alterem.

# K-means



# PAM (Partitioning Around Medoids)

- É uma versão mais flexível do K-means;
- Medoids são objetos representativos do cluster que pertencem ao conjunto de dados;
  - O centróide é um dos dados.
- O algoritmo é parecido com o do K-means;
- A função objetivo a ser minimizada é:

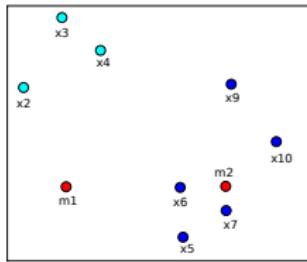
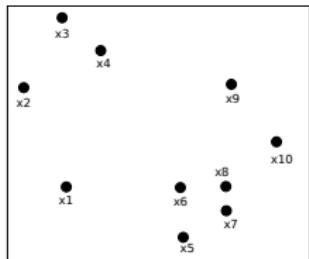
$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - \tilde{x}_j\|^2, \quad (2)$$

onde  $x_i$  é um dos dados e  $\tilde{x}_j \in \{x_1, \dots, x_n\}$

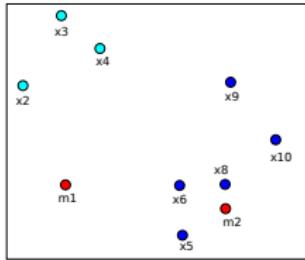
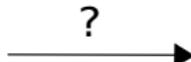
Algoritmo:

- 1 Inicialização: aleatoriamente selecionar  $K$  de  $n$  objetos nos dados para serem os medoides;
- 2 Associar cada objeto do dados com o medoide mais próximo;
- 3 Para cada medoide  $m$ :
  - 1 Para cada objeto não-medoide  $x$ :
    - 1 Troque  $m$  por  $x$  e calcule o custo total da configuração;
- 4 Selecione a configuração com o menor custo;
- 5 Repetir os passos 2 a 5 até não haver mudanças no medoide.

# PAM (Partitioning Around Medoids)



$$\text{custo total} = \{d(m1,x2)+d(m1,x3) + d(m1,x4)\} + \\ \{d(m2,x5) + d(m2,x6) + d(m2,x7) + d(m2,x9) + d(m2,x10)\}$$



$$\text{custo total} = \{d(m1,x2)+d(m1,x3) + d(m1,x4)\} + \\ \{d(m2,x5) + d(m2,x6) + d(m2,x8) + d(m2,x9) + d(m2,x10)\}$$

# Modelo de Misturas

- Abordagem estatística;
- Cada cluster é matematicamente representado por uma distribuição paramétrica.
  - Gaussiana (continua);
  - Poisson (discreta).
- Todo o conjunto de dados é modelado por uma mistura dessas distribuições;
- Cada distribuição usada para modelar um cluster é frequentemente referido como um componente da distribuição.
- Queremos saber se, dado um cluster, qual a probabilidade de um objeto do conjunto de testes pertencer ao cluster;

- Considerando  $\pi_c$  como a probabilidade a priori no cluster  $c$ , pelo Teorema de Bayes temos:

$$\mathbf{P}[y_i = c|x_i] = \frac{\pi_c \mathbf{P}[x_i|y_i = c]}{\sum_{c'=1}^C \pi_{c'} \mathbf{P}[x_i|y_i = c']} \quad (3)$$

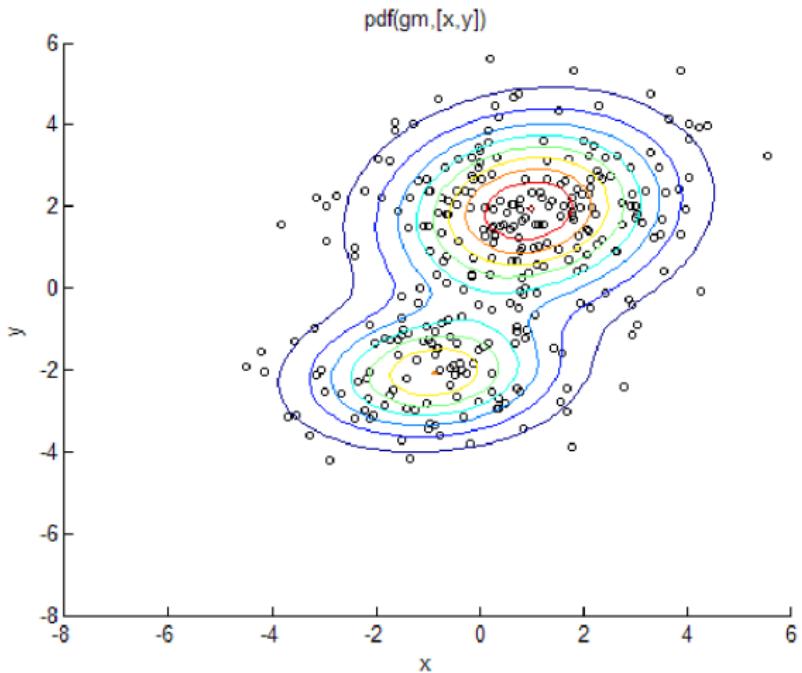
- Como não sabemos a função densidade de probabilidade do cluster, podemos definir como uma distribuição Gaussiana multivariada:

$$\mathbf{P}[x_i|y_i = c, \theta_c] = \frac{1}{\sqrt{(2\pi)^d |\Sigma_c|}} \exp^{\frac{1}{2}(x_i - \mu_c)^\top \Sigma_c^{-1} (x_i - \mu_c)}, \quad (4)$$

onde  $\theta_c$  são os parâmetros  $(\mu_c, \Sigma_c)$ , que são respectivamente, a média e a matriz de covariância do cluster.

- Temos que inicializar  $\mu_c$  e  $\Sigma_c$  para todo cluster  $c$ ;
- $\pi_c$  é inicializado pela proporção dos dados no cluster;
- Os parâmetros são estimados pelo critério de máxima verossimilhança usando o algoritmo EM (Expectation Maximization);
  - $\mu_c$ ,  $\Sigma_c$  são re-estimados em cada iteração do EM (M-step);
  - $\pi_c$  também é estimado novamente (E-step);
- Uma prática comum é usar o k-means para estabelecer os primeiros parâmetros da inicialização;
- Para diminuir a complexidade do modelo, para todos os componentes, geralmente assume-se:
  - Probabilidades a priori iguais;
  - Matrizes de covariância idênticas.

# Modelo de Misturas



- Vários métodos propostos na literatura;
  - Milligan e Cooper (1985) comparou mais de 30 métodos distintos;
- Estimar o número exato de  $K$  é muito difícil;
  - Com dados de microarray é ainda mais complicado;
  - A interação entre os genes de um organismo é tão complexo que a definição de clusters e o exato número de  $K$  é vago;
- A abordagem geralmente usada é usar vários valores de  $K$  e analisar os clusters.

- Um dos critérios utilizados para se comparar clusters é baseado em quão bem o clustering reproduz clusters reais já conhecidos;
  - Em dados simulados podemos conhecer a real estrutura do cluster;
- A performance do clustering pode ser medido pela similaridade entre a partição resultante e a real;
- Mede a similaridade entre duas partições;
- Foi criada em 1971 por Rand, e melhorada por Hubert and Arabie em 1985;
- Rand Index pode ser definido como a proporção de pares de genes concordantes em duas partições dentre todas as partições de genes possível;

- O Rand Index (1971):

- $a$  é o número de pares de objetos pertencentes ao cluster calculado e o cluster real;
- $b$  é o número de pares de objetos que estão no mesmo cluster real e em diferentes clusters calculado;
- $c$  é o número de pares de objetos que estão no mesmo cluster calculado e em diferentes clusters reais;
- $d$  é o número de pares de objetos em clusters diferentes, em ambas as partições;
- Assim temos:

$$RandIndex = \frac{a + d}{a + b + c + d}, \quad (5)$$

- O resultado do Rand Index é um numero entre 0 e 1:
  - Quando as duas partições concordam perfeitamente o valor do Rand Index é 1.

- O Adjusted Rand Index (1985):
  - Assume um modelo aleatório usando a distribuição hipergeométrica;
  - O Adjusted Rand Index tem como resultado 0 quando as partições concordam perfeitamente;
  - Assim, podemos calcular o Adjusted Rand Index pela fórmula:

$$\frac{\sum_{i,j} \binom{n_{ij}}{2} - [\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}] / \binom{n}{2}}{\frac{1}{2} [\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}] - [\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}] / \binom{n}{2}}, \quad (6)$$

onde  $n_{ij}$  é o número de objetos que estão na classe  $u_i$  e no cluster  $v_j$ . E  $n_{i.}$  e  $n_{.j}$  representam, respectivamente, o número de objetos da classe  $u_i$  e  $v_j$ .

# Rand Index

- Pra facilitar, construimos uma tabela de contingência, como por exemplo:

Class/Cluster	$v_1$	$v_2$	$v_3$	Total
$u_1$	1	1	0	2
$u_2$	1	2	1	4
$u_3$	0	0	4	4
Total	2	3	5	$n = 10$

# Rand Index

- Com a tabela de contingência, fica mais simples calcular o Adjusted Rand Index:
  - $\sum_{i,j} \binom{n_{ij}}{2} = \binom{2}{2} + \binom{4}{2} = 1 + 6 = 7$
  - $\sum_i \binom{n_{i.}}{2} = \binom{2}{2} + \binom{2}{2} + \binom{4}{2} = 1 + 6 + 6 = 13$
  - $\sum_j \binom{n_{.j}}{2} = \binom{2}{2} + \binom{3}{2} + \binom{5}{2} = 1 + 3 + 10 = 14$
  - $\binom{n}{2} = \binom{10}{2} = 45$
- Assim, o calculo do Adjusted Rand Index, segundo a Eq. 6, temos:

- $$AdjustedRandIndex = \frac{7 - 14 * 13/45}{(14 + 13)/2 - 14 * 13/45} = 0.313$$

# P-value

- O p-value mede a consistencia entre um resultado no experimento e uma explicação por "pura chance" dos resultados;
  - Supondo uma moeda justa, em um experimento, joga-se 20 vezes a moeda e consegue-se 14 caras. Isso foi coincidencia ou o resultado do experimento está correto?
  - Qual a probabilidade do experimento se repetir dado que sabemos que a chance de obetermos cara é de 0.5?
  - O p-value calcula essa probabilidade. No caso, a probabilidade de termos 14 caras ao jogarmos a moeda justa 20 vezes é de 0.115.
- Definimos um limiar  $\delta$  para aceitar ou não o p-value. Geralmente, usa-se  $\delta \geq 0.05$  para rejeitar.

- No caso de dados de microarray com um total de  $G$  genes, consideramos um conjunto de  $m$  genes que sabemos que pertencem a uma categoria funcional "F";
- No cluster com  $D$  genes, temos  $h$  genes pertencentes a categoria funcional "F";
- O P-value pode ser calculado com a seguinte fórmula:

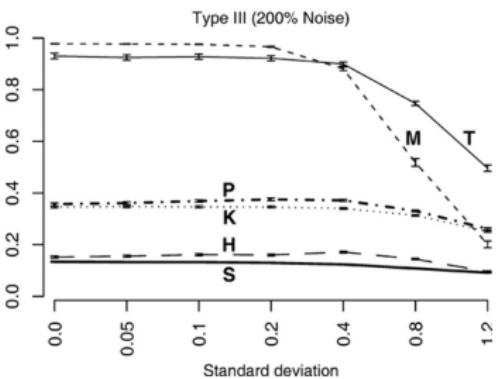
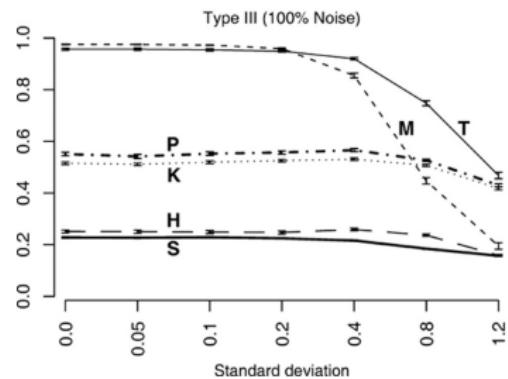
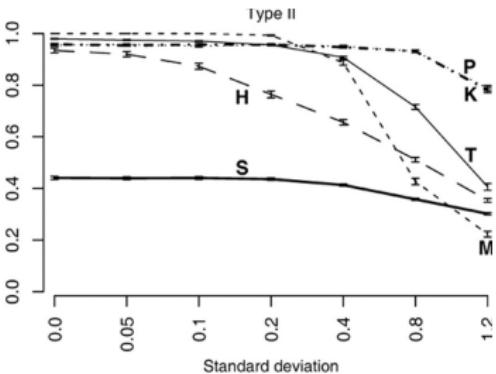
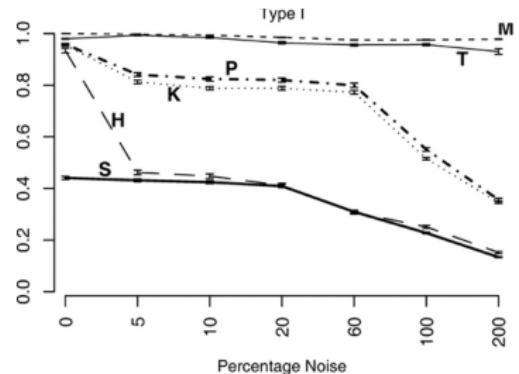
$$P[X \geq h] = 1 - \sum_{i=0}^{h-1} \binom{D}{i} \binom{G-D}{m-i} / \binom{G}{m} \quad (7)$$

# Comparação entre os métodos

Para comparar com o Adjusted Rand Index foram simulados 3 tipos de dados com  $K = 15$ .

Scattered genes (%)	SD for random normal error added to each gene						
	0	0.05	0.1	0.2	0.4	0.8	1.2
0	✓ (I)	✓ (II)					
5	✓ (I)						
10	✓ (I)						
20	✓ (I)						
60	✓ (I)						
100	✓ (I)	✓ (III)					
200	✓ (I)	✓ (III)					

# Comparação entre os métodos

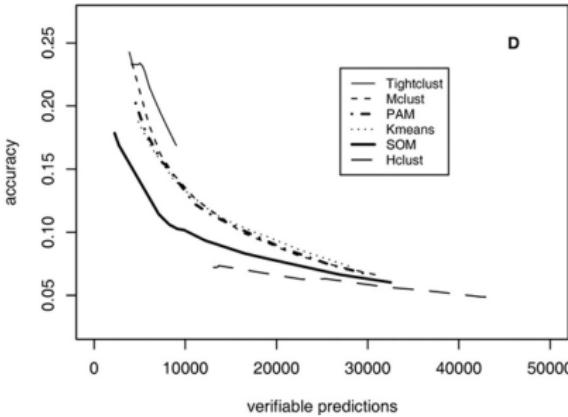
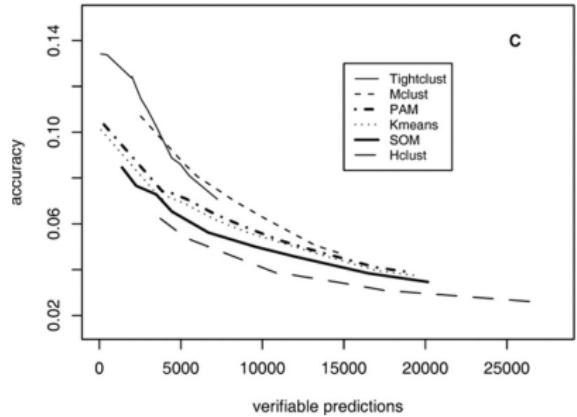
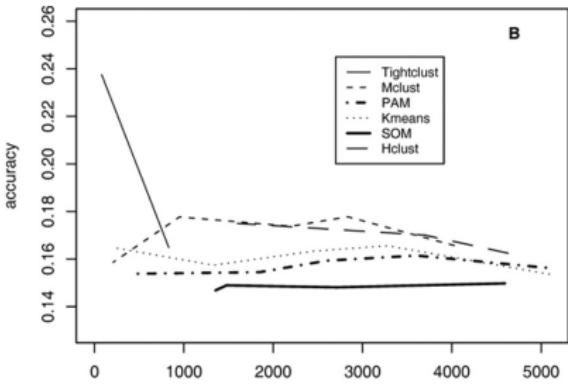
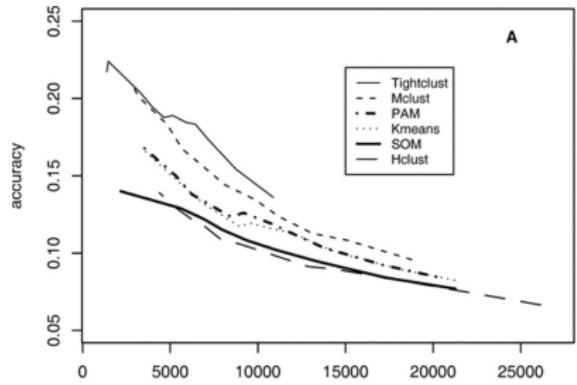


# Comparação entre os métodos

Para comparar a acurácia do métodos com os P-value, foram usados os seguintes dados:

Organism, sample perturbation and reference	Dimension of data	Annotation
Yeast; cell cycle; Spellman <i>et al.</i> , 1998	1663 genes x 77 samples	M/G1 boundary; Late G1, SCB regulated; Late G1, MCB regulated; S-phase; S/G2-phase; G2/M-phase
Yeast; environmental changes; Causton <i>et al.</i> , 2001	1744 genes x 45 samples	response to stress
Human; cell cycle; Whitefield <i>et al.</i> , 2002	2570 genes x 114 samples	G1/S; S; G2; G2/M; histone genes
human; lung cancer; Bhattacharjee <i>et al.</i> , 2001	1920 genes x 203 samples	Keratin; metallothionein; melanoma antigen family; major histocompatibility complex (MHC); interferon; immunoglobulin heavy constant; G antigen; collagen

# Comparação entre os métodos



- Tanto com os dados simulados quanto com os reais, o modelo de misturas e o tight clustering apresentaram desempenho superiores aos demais;
  - Ambos os métodos não colocam os "outliers" nos clusters;
  - Os demais métodos sempre colocam os "outliers" nos clusters;
- K-means e PAM tiveram performances similares;
  - Resultado esperado, pois são métodos semelhantes;
  - Performance pior que o modelo de misturas e o tight clustering;
  - Performance melhor do que SOM e Agrupamento Hierárquico;
- SOM e Agrupamento Hierárquico tiveram as piores performances
  - Melhor visualização dos clusters em detrimento da performance;

# Discussão

- Baseado nos experimentos, podemos inferir que Modelo de Misturas e Tight clustering são os melhores métodos para realizar clustering em dados de expressão gênica;
- Segundo o autor, agrupamento hierárquico e SOM são os métodos mais populares em vários estudos biológicos;
  - Se o objetivo principal é identificar genes marcadores e visualização é o objetivo secundário, o uso de SOM e agrupamento hierárquico é desencorajado;
- Novos métodos de clustering são elaborados constantemente, o problema é fazer a avaliação/comparação dentre modelos de forma justa;

# Dúvidas



- Anbupalam Thalamuthu, Indranil Mukhopadhyay, Xiaojing Zheng, and George C. Tseng. **Evaluation and comparison of gene clustering methods in microarray analysis Bioinformatics;**
- K. Y. Yeung, and W. L. Ruzzo. **Principal component analysis for clustering gene expression data;**
- D'haeseleer P. **How does gene expression clustering work?**
- Marcilio CP de Souto, Ivan G Costa, Daniel SA de Araujo, Teresa B Ludermir and Alexander Schliep. **Clustering cancer gene expression data: a comparative study**

- Ronald A. Thisted. **What is a P-value? Evaluation and comparison of gene clustering methods in microarray analysis Bioinformatics;**
- Richard O. Duda, Peter E. Hart, David G. **Pattern Classification**
- Haykin, S. **Redes Neurais**
- Slides de Ivan Gesteira, Aluízio Araujo, R. Palaniappan, Andrew W. Moore, Jia Li, Pankaj K. Agarwal, David Blei, Todd Lowe e John A. Bullinaria;

# Métodos de clustering em dados de Microarray

## Introdução à biologia computacional

Clerton Ribeiro de A. Filho

Universidade Federal de Pernambuco (UFPE)

11 de junho de 2010