Introdução ao Reconhecimento de Padrões e aplicações em problemas de Bioinformática

Fabrício M. Lopes
fabricio@utfpr.edu.br

UTFPR-CP
Grupo de Pesquisa em Bioinformática e Reconhecimento de Padrões
bioinfo-cp@utfpr.edu.br

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Curso de Verão - Bioinformática - USP, 2012

Organização

Introdução - Definições

Definição de Reconhecimento de Padrões:

- *"É uma área de pesquisa que tem por objetivo a classificação de objetos (padrões) em um número de categorias ou classes"*, Theodoridis e Koutroumbas [1].
- *"O ato de observar os dados brutos e tomar uma ação baseada na categoria de um padrão"*, Duda et al. [2].

Introdução - Definições

- **Padrão**: é uma entidade, objeto, processo ou evento, vagamente definido, que pode assumir um nome.
- **Classe**: conjunto de padrões que possuem características em comum.
- **Característica** ou **Atributo**: dado extraído de uma amostra por meio de medida e/ou processamento. Em geral são organizadas na forma de um **vetor de características**.
- **Classificação**: atribuir classes para as amostras, baseado em suas características.
- **Ruído**: distorção, falha ou imprecisão que ocorre na aquisição dos dados.

Introdução - Classificadores

- **Classificadores**: utilizados para classificar ou descrever padrões ou objetos a partir de um conjunto de propriedades ou características.

- Existem essencialmente dois casos particulares de reconhecimento de padrões:

- **Classificação supervisionada**.
- **Classificação não supervisionada**.

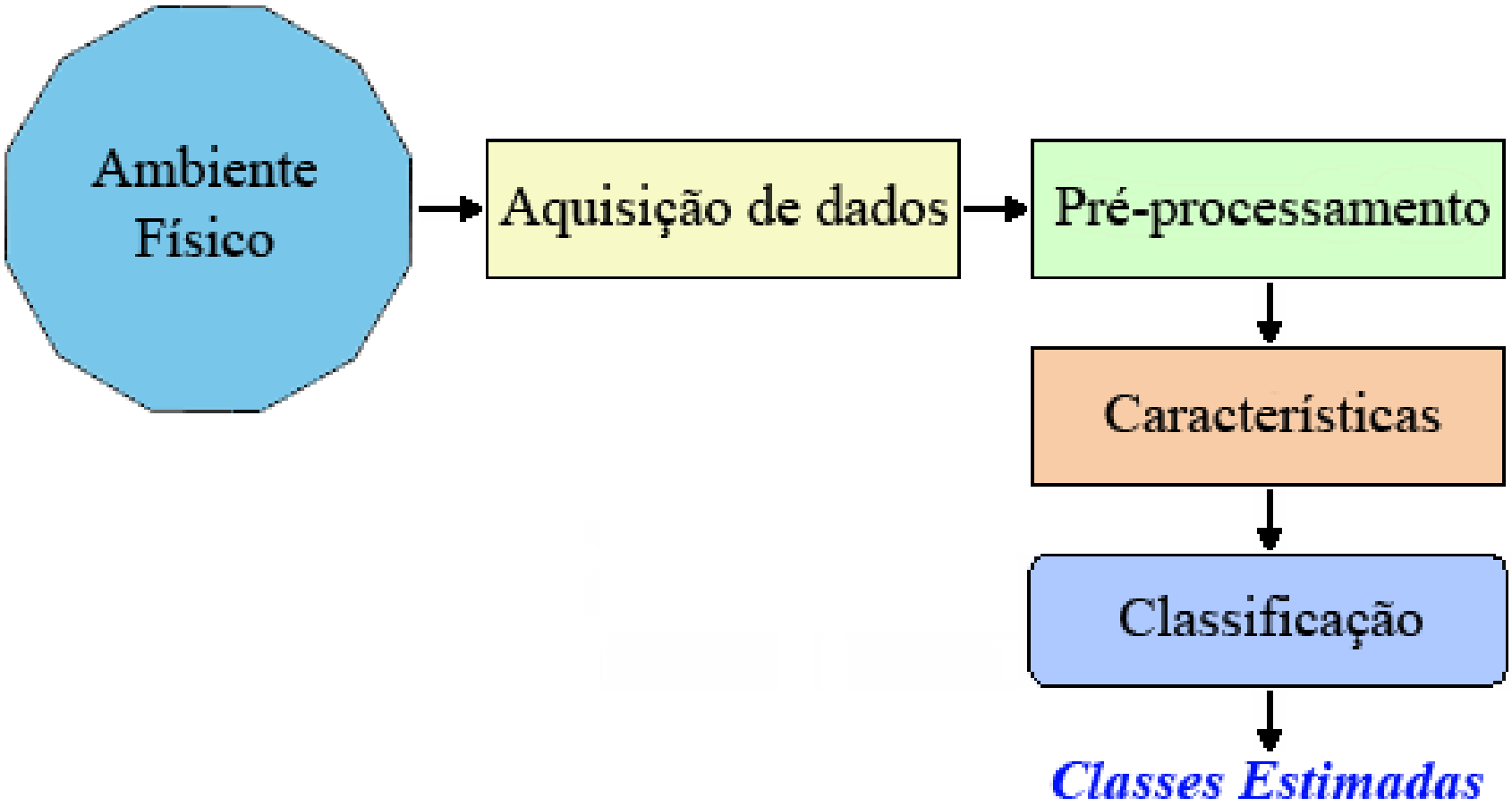
Introdução - Classificação supervisionada

- Selecionam-se amostras representativas para cada uma das classes que se deseja classificar.
- Conhecemos o padrão e classes que estamos procurando.
- Também conhecido como Aprendizado supervisionado.

Introdução - Classificação não supervisionada

- Não conhecemos o padrão, nem o número total de classes a serem encontradas durante a classificação.
- Também conhecido como aprendizado não supervisionado ou análise de agrupamentos (***clusters***).
- O conjunto de dados é particionado em grupos, baseados em características específicas, tais que os pontos dentro de um grupo (cluster) sejam mais similares do que os pontos de outros grupos.
- Pode ajudar compreender funções de muitos genes para os quais não há informações disponíveis, Jiang et al. [3].

Etapas do Reconhecimento de Padrões

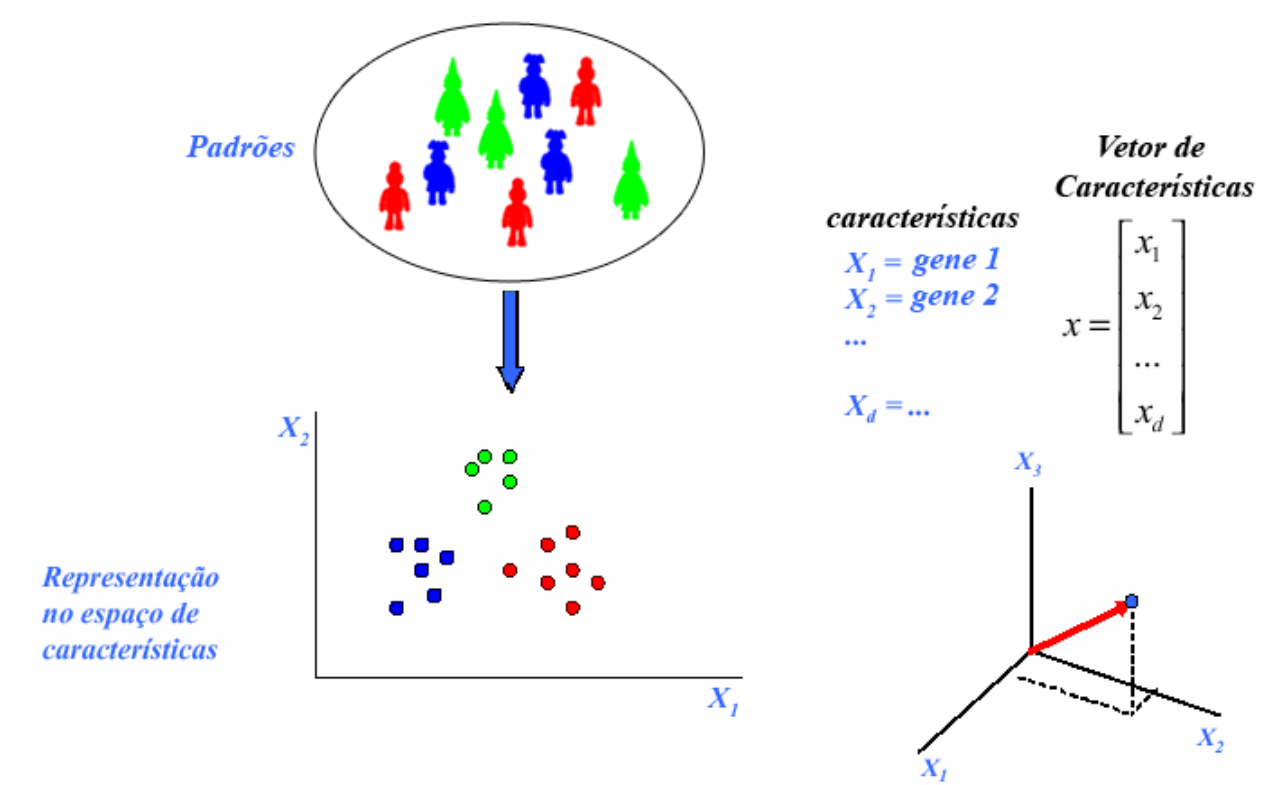


Introdução - Pré-processamento

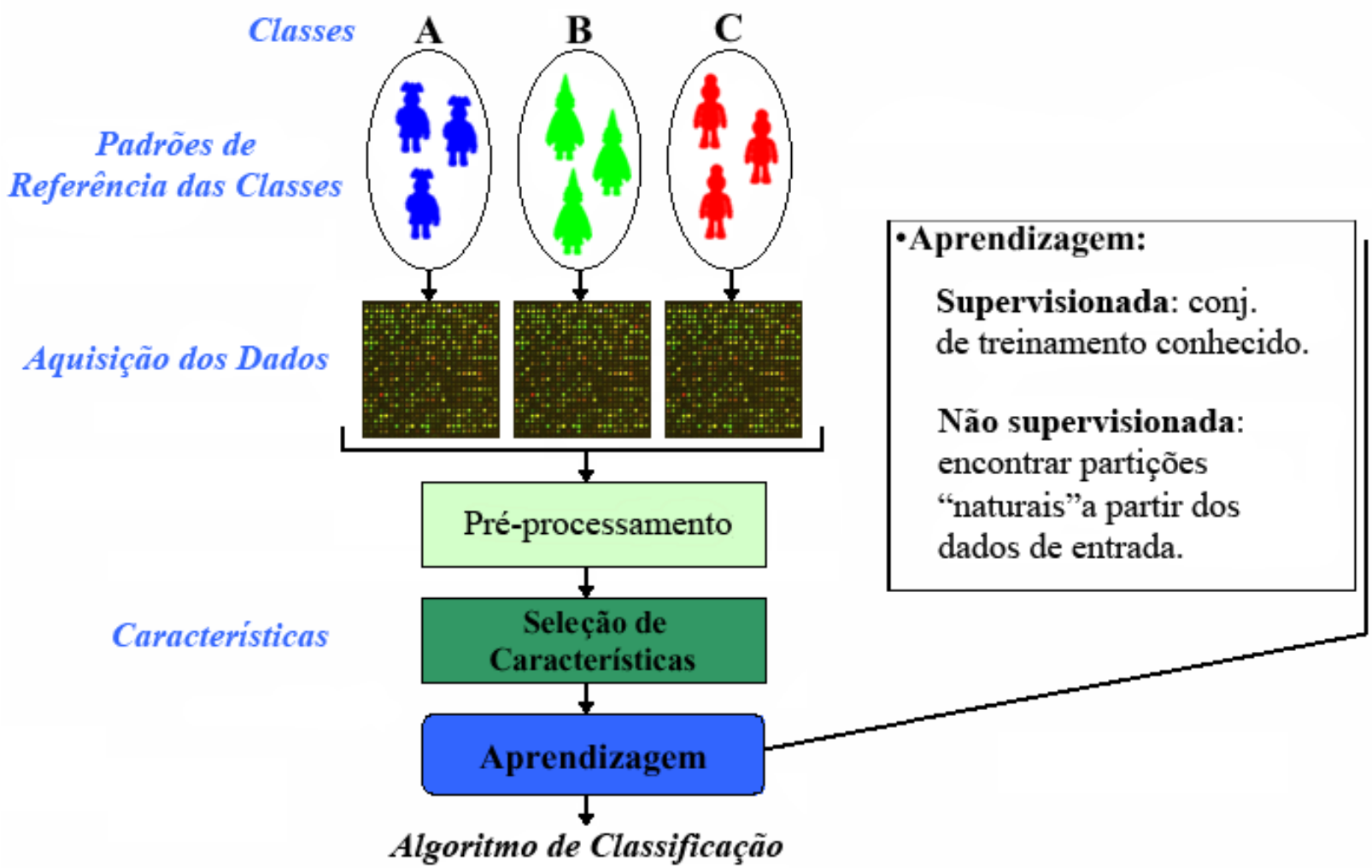
- Genes que apresentam pouca variância
- Genes que apresentam ausência de dados
- Transformação de escala numérica (normalização):
$$X' = \frac{X - \mu}{\sigma^2}$$
- Aplicar escala logarítmica, em geral $\log_2$

Introdução - Características

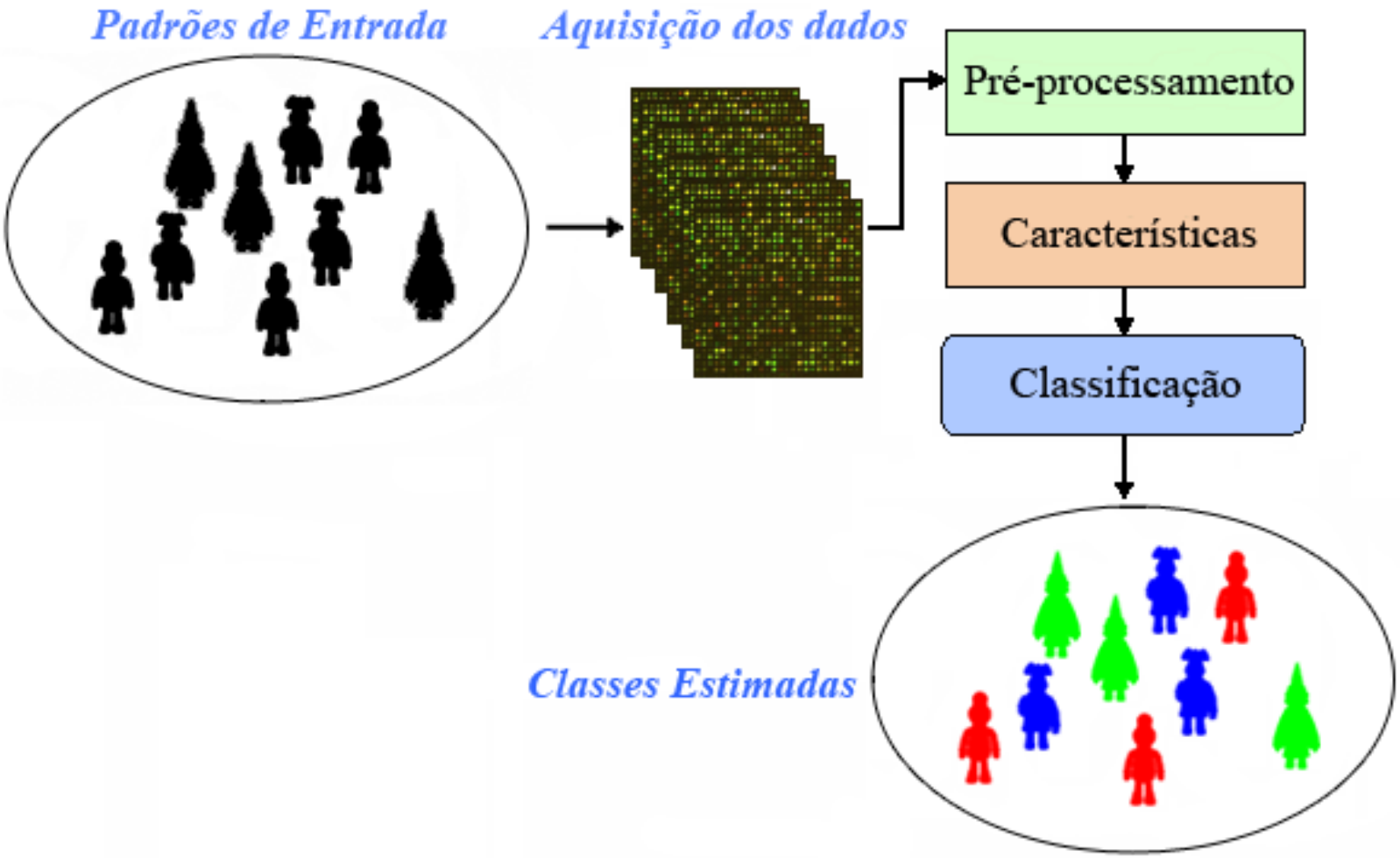
- **Característica** ou **Atributo**: dado extraído de uma amostra por meio de medida e/ou processamento.



Introdução - Treinamento



Introdução - Classificação



Reconhecimento de Padrões - Aplicações

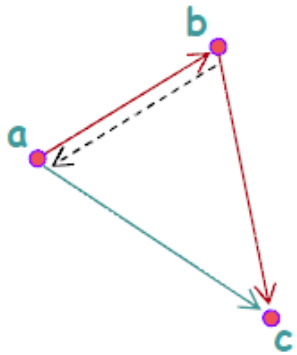
Aplicação	**Padrão de Entrada**	**Classes (saída)**
Reconhecimento óptico de caracteres	imagem de um documento	caracteres/palavras
Busca na internet	documento texto/imagem	categoria semântica
Filtro de e-mails	e-mail	spam/normal
Identificação de pessoas	face, iris, impressão digital	acesso de usuários credenciados
Diagnóstico auxiliado por computador	imagem microscópica	células saudáveis/doentes
Reconhecimento de alvos militares	imagem óptica ou infravermelho	tipo do alvo
Seleção automática de qualidade	imagem em esteira de produção	níveis de qualidade
Análise de sequências de DNA	sequência de DNA	gene conhecido/desconhecido
Estimação de expressão gênica	imagem de microarray	intensidades/classes.
Análise de expressão gênica	expressão gênica	similaridade entre os elementos dos clusters
Inferência de redes gênicas	perfil de expressão temporal	rede de regulação estimada

Distância - Definição

- **Distância** é um número que caracteriza a separação entre dois objetos

- Deve satisfazer os requisitos:
  - Sejam $a, b, c \in S$
  - Ser não negativa: $d(a, b) \geq 0$
  - Ser comutativa: $d(a, b) = d(b, a)$
  - Satisfazer à desigualdade triangular:

    $d(a, b) + d(b, c) \geq d(a, c)$



Distância Euclideana

A distância Euclideana entre dois padrões $x = (x_1, x_2, \ldots, x_d)^t$ e $y = (y_1, y_2, \ldots, y_d)^t$ no espaço $\mathbb{R}^d$ é definida por:

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

Atribui a mesma importância a cada dimensão (característica).

Distância Mahalanobis

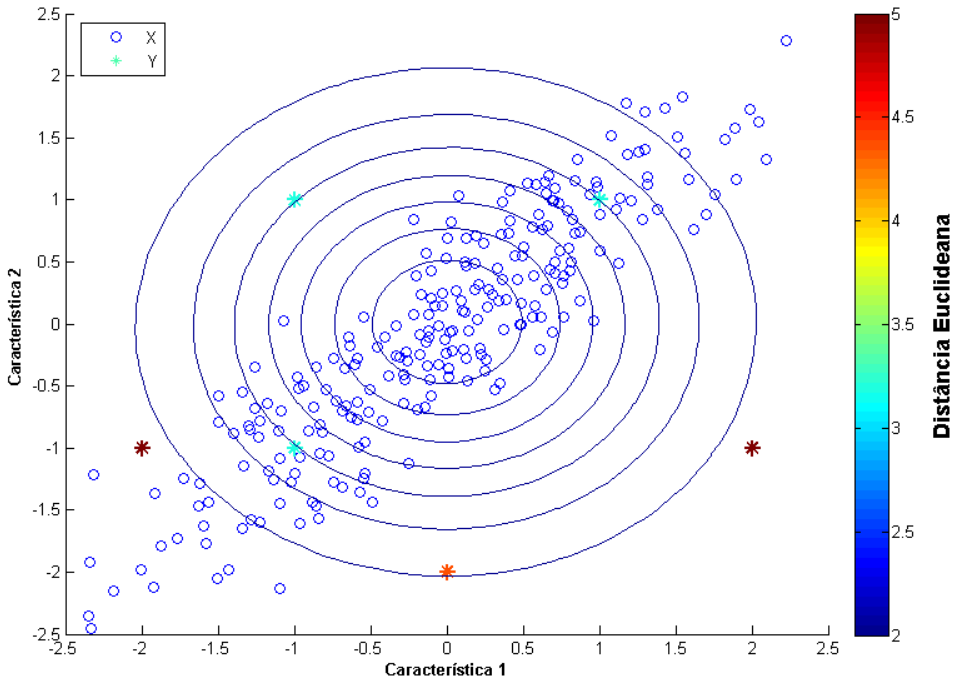
A distância estatística ou distância Mahalanobis [2] entre dois padrões $x = (x_1, x_2, \ldots, x_d)^t$ e $y = (y_1, y_2, \ldots, y_d)^t$ no espaço $\mathbb{R}^d$ é definida por:

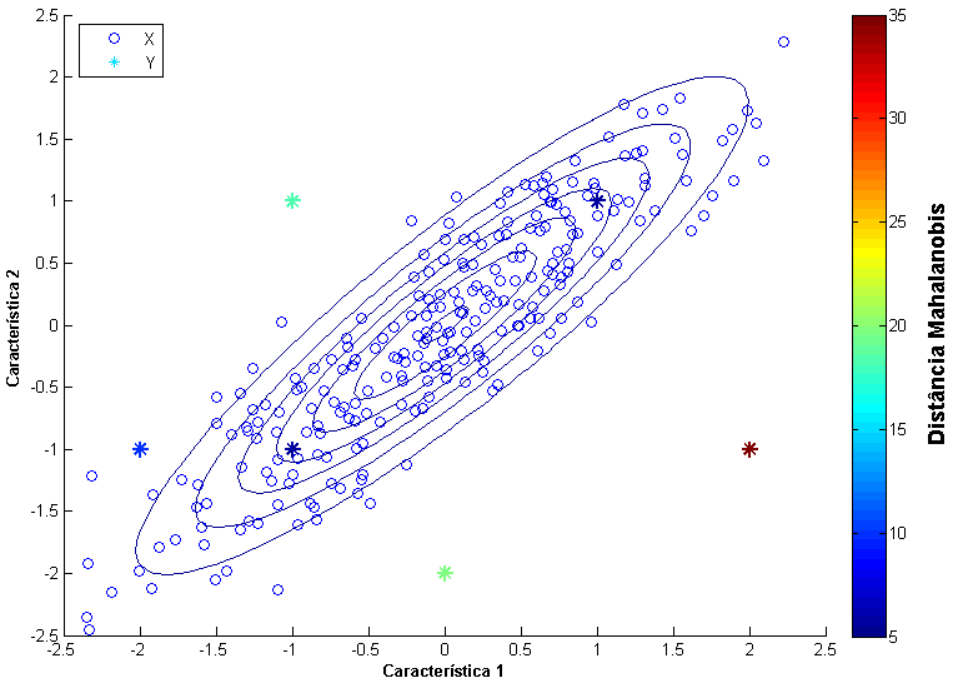
$$d_M(x, y) = \sqrt{(\overrightarrow{x} - \overrightarrow{y})^t \sum^{-1} (\overrightarrow{x} - \overrightarrow{y})}$$

Onde $\sum^{-1}$ é a matriz de covariância das variáveis. Cada elemento de $\sum^{-1}$ é dado por: $\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$

Pode atribuir um peso (importância) diferente para cada dimensão.

Distâncias - Exemplo





Dados usados neste estudo de caso

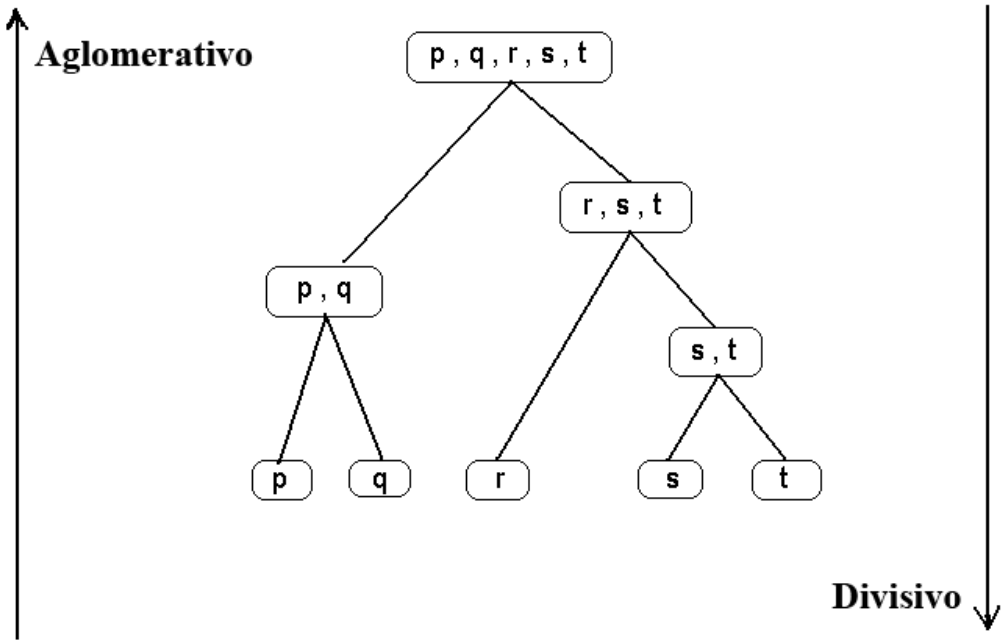
- Dados de microarray para estudo de expressão gênica de levedura, DeRisi et al. [4].
- O conjunto de dados completo pode ser copiado `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28`
- Contendo **6400** genes e **7** amostras temporais.

Pré-processamento dos Dados

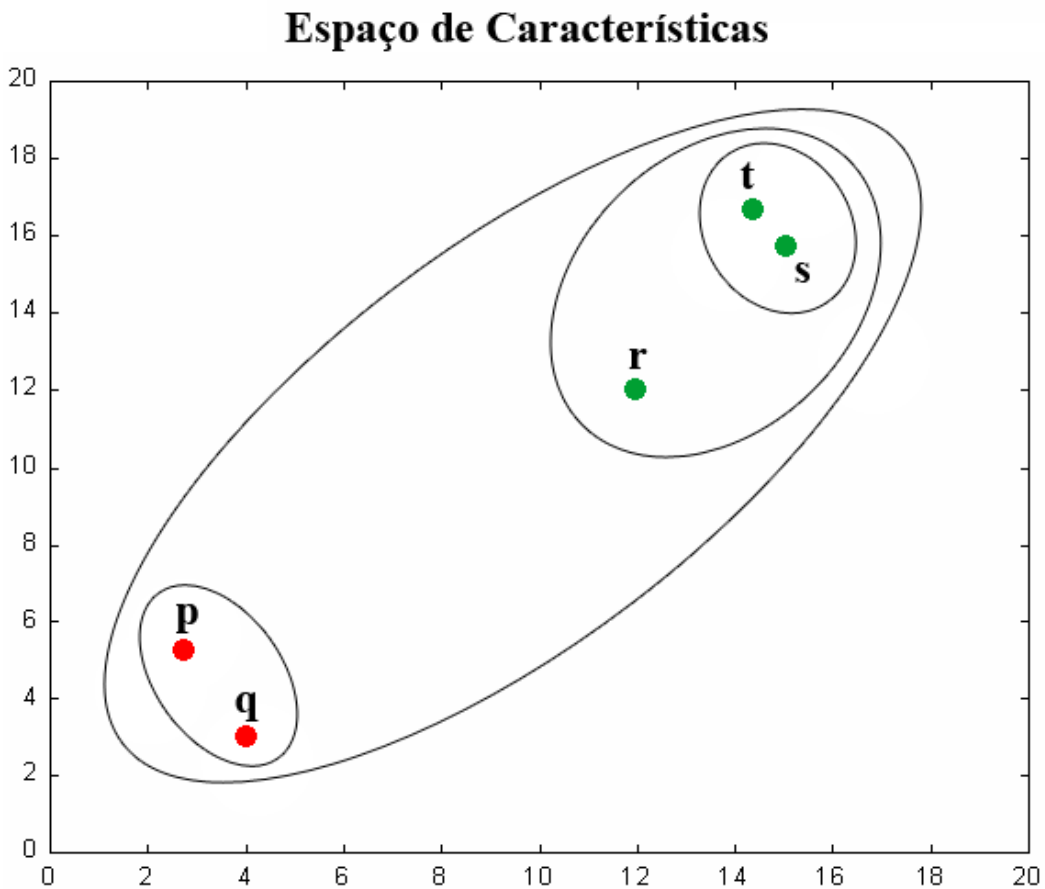
- Aplicada a escala logarítmica $\log_2$
- Genes que apresentam spots do array marcados como vazios (**removidos 108 = 6292**)
- Genes que apresentam pouca variância (10%) (**removidos 629 = 5663**)
- Genes que apresentam baixos valores de expressão (< 3.0) (**removidos 4829 = 834**)
- Após o pré-processamento restaram **834 genes**.

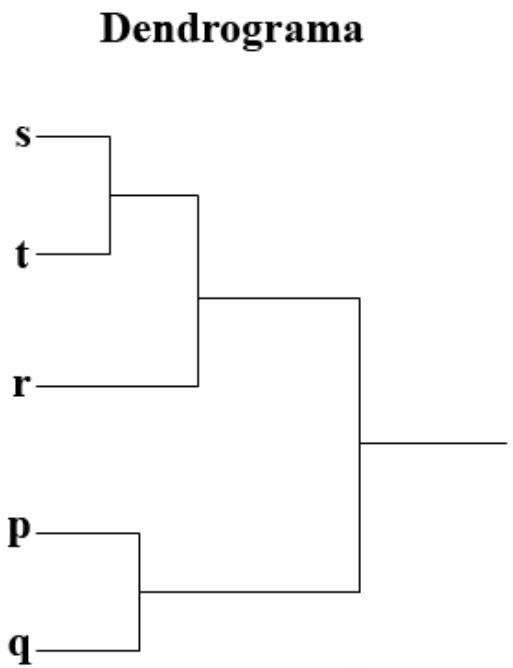
Análise de Agrupamentos

- Considerando os perfis de expressão já pré-processados, iremos procurar por relacionamentos entre os genes.
- Vamos usar o algoritmo de **Agrupamento Hierárquico**



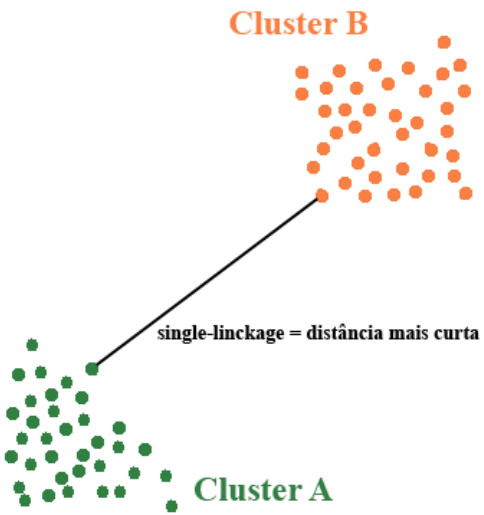
Agrupamento Hierárquico

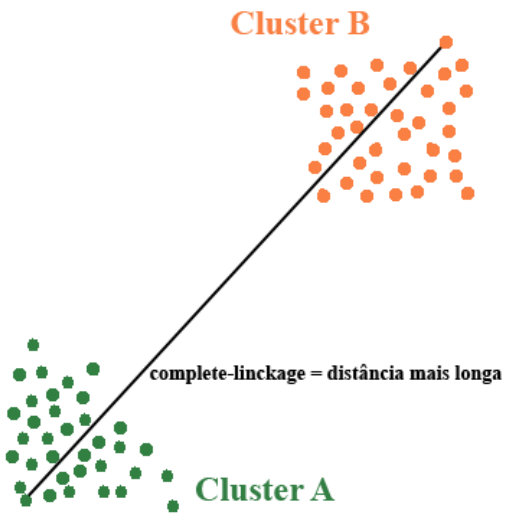


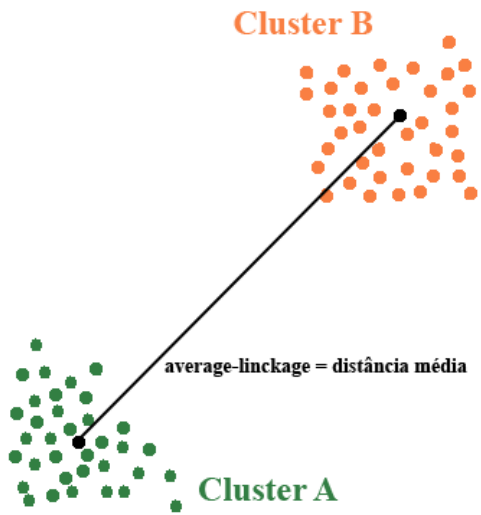


Distâncias entre Clusters - Exemplos

Além da função de similaridade, é preciso escolher como se dá a distância entre um ponto e um cluster.



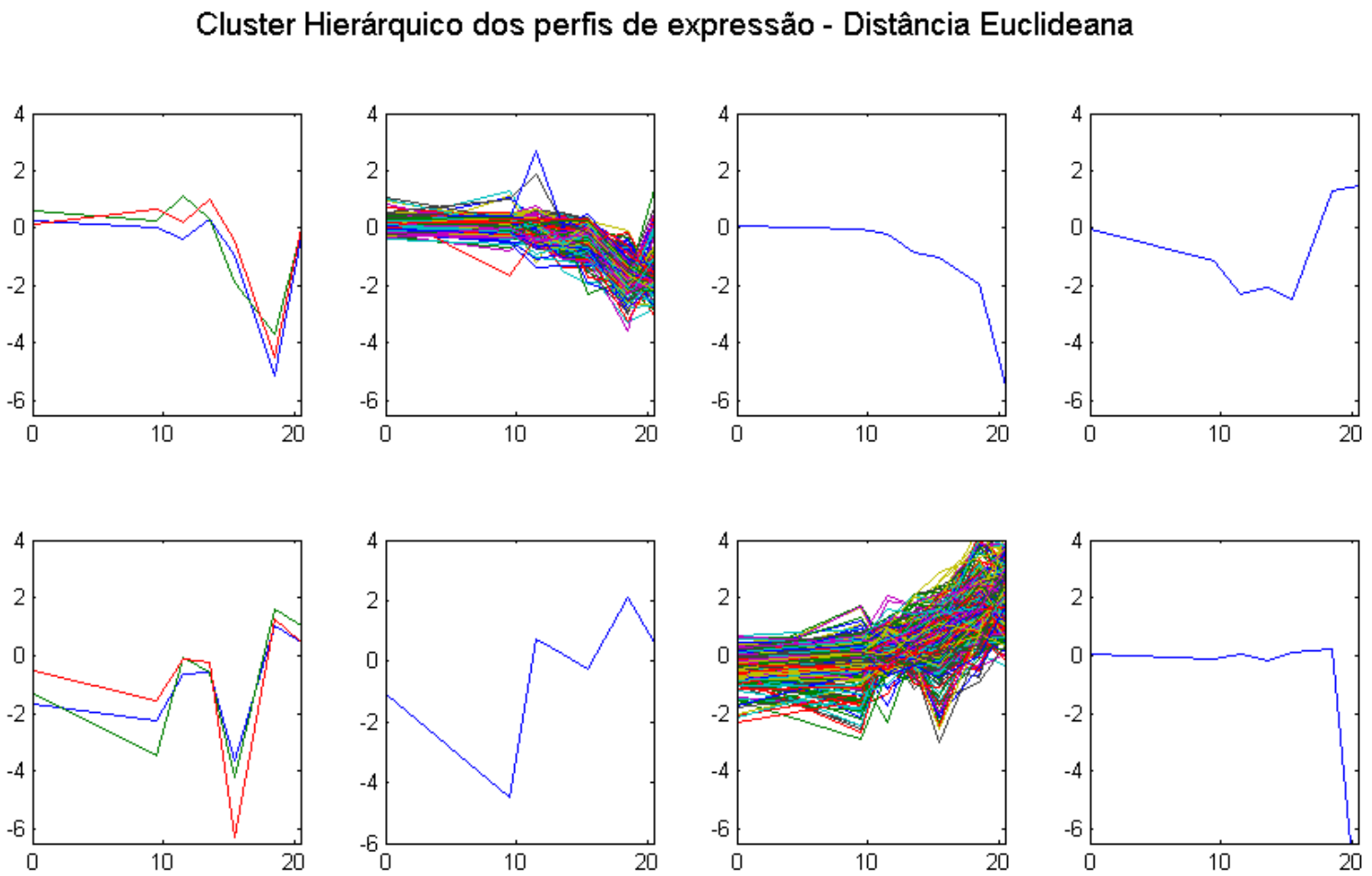




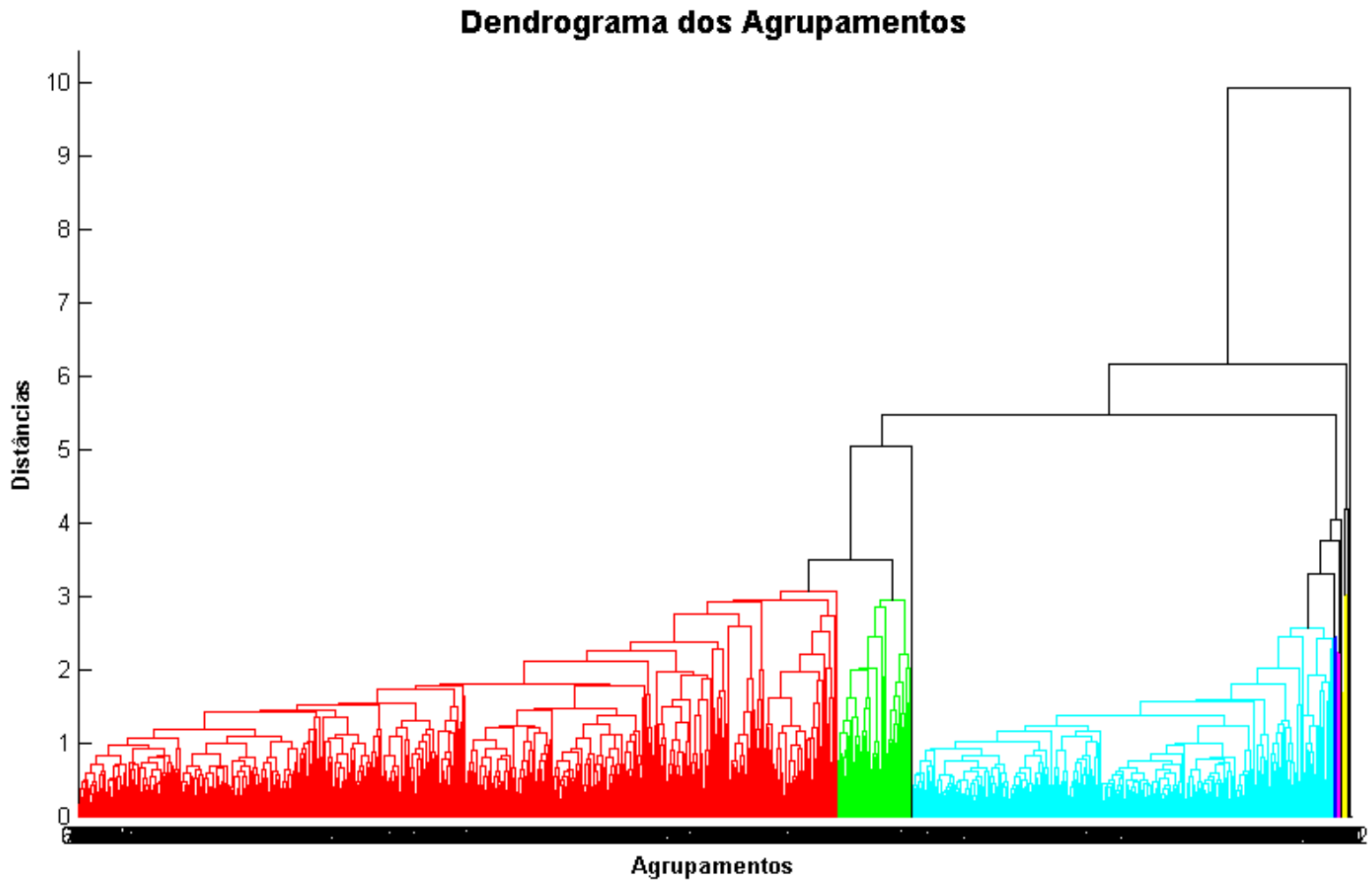
Agrupamento Hierárquico - Distância

- Foi utilizada distância Euclideana entre os genes
- Considerando a média da distância entre o cluster e a nova amostra agrupada (average-linkage)
- Análise inicial identificou **8** agrupamentos (clusters)

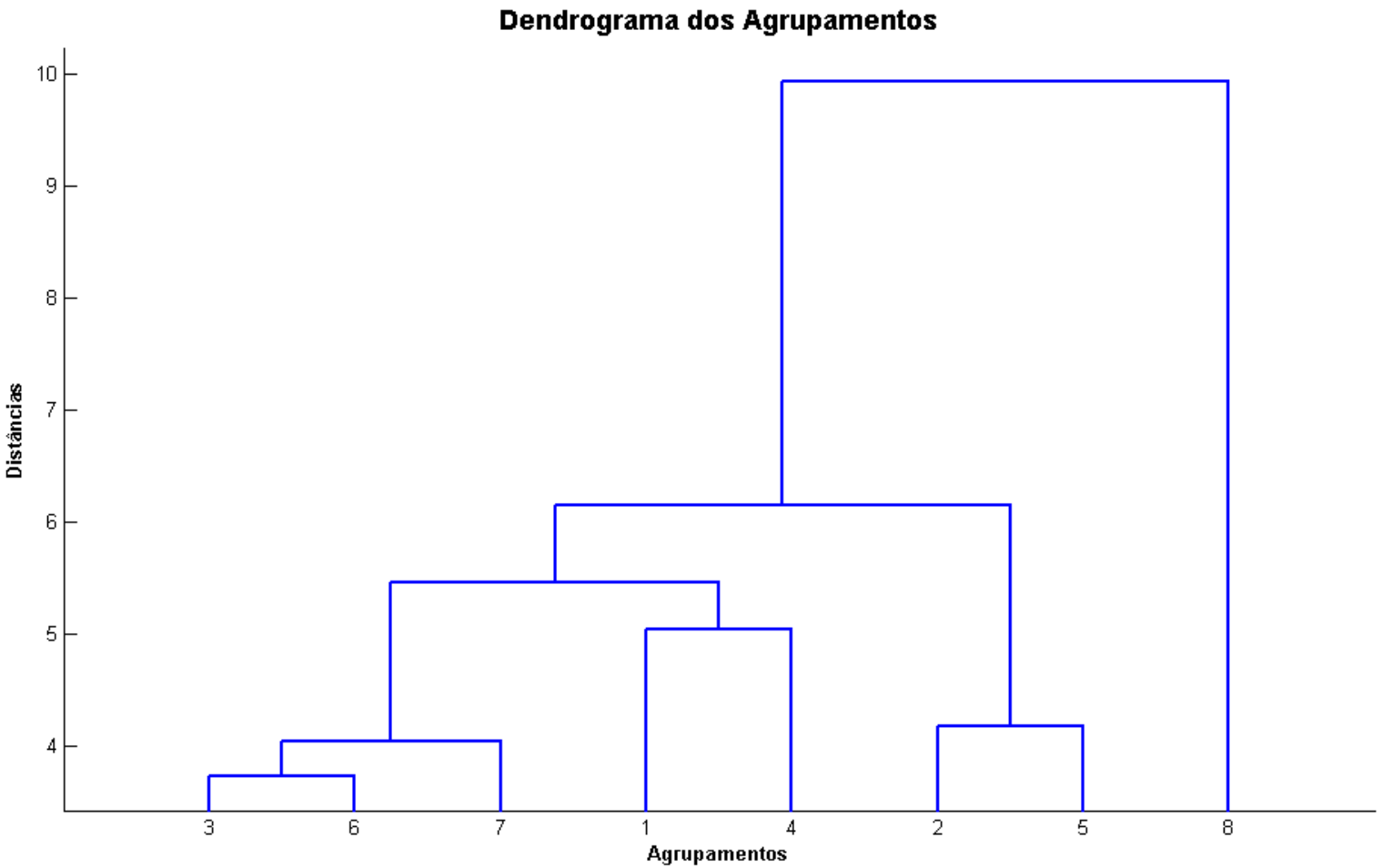
Resultados


Cluster Hierárquico dos perfis de expressão - Distância Euclideana

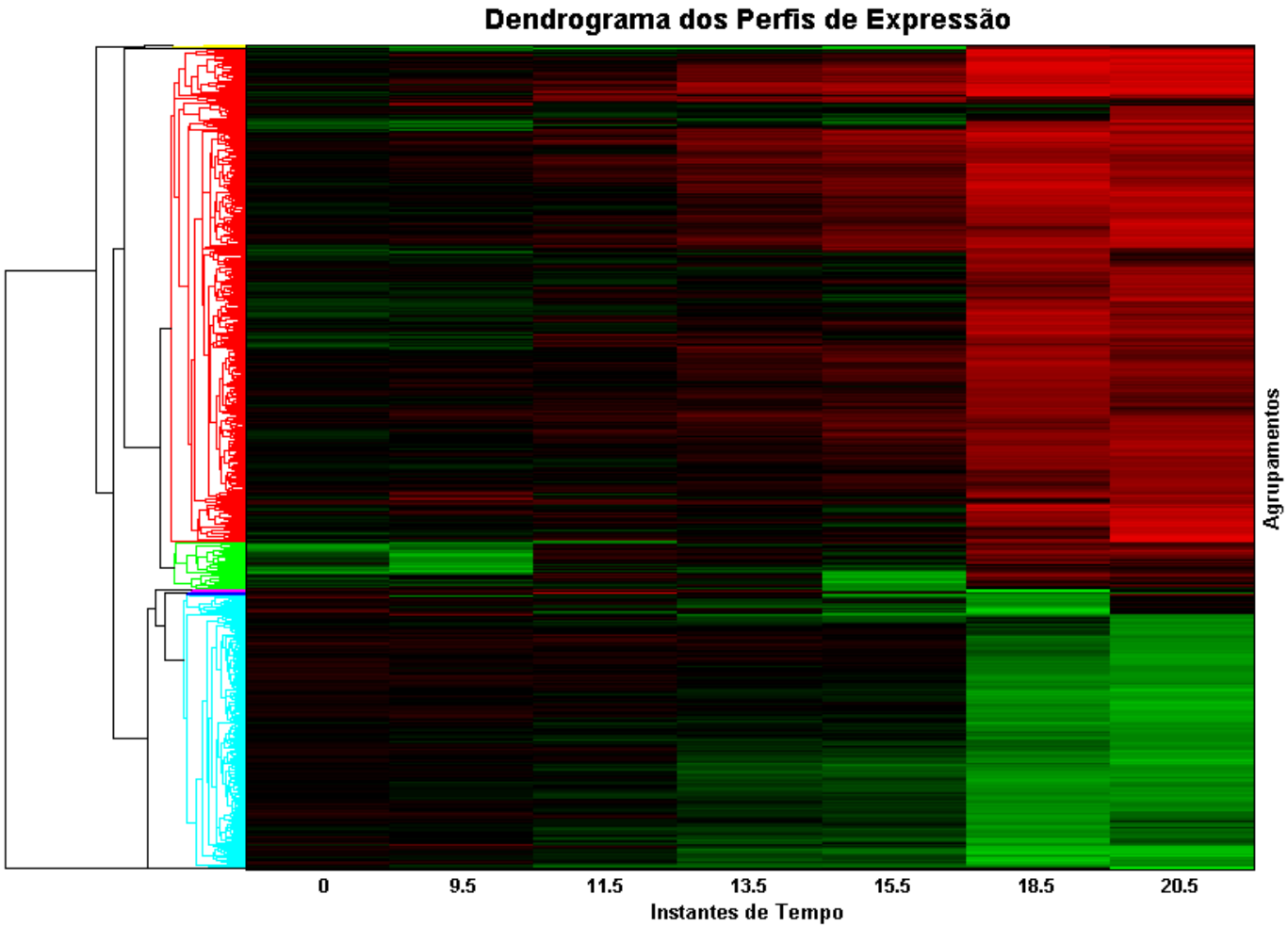
Resultados


Dendrograma dos Agrupamentos
Distâncias
0
1
2
3
4
5
6
7
8
9
10
Agrupamentos

Resultados


Dendrograma dos Agrupamentos
Distâncias
10
9
8
7
6
5
4
3
6
7
1
4
2
5
8
Agrupamentos

Resultados


Dendrograma dos Perfis de Expressão
Agrupamentos
0
9.5
11.5
13.5
15.5
18.5
20.5
Instantes de Tempo

Conclusão - Estudo de Caso 1

- O que escolher para a análise de agrupamentos:
  - Como fazer o pré-processamento dos dados
  - Algoritmo de agrupamento
  - Medida de distância ou similaridade
  - Método de distâncias entre agrupamentos
  - Visualização dos resultados
- Resultados científicos dos dados disponíveis em DeRisi et al. [4]

Dados usados neste estudo de caso

- Iremos usar os mesmos dados pré-processados do estudo de caso 1
- O conjunto de dados completo pode ser copiado `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28`
- Contendo **834** **genes** e **7** amostras temporais
- Os genes foram quantizados em 2 níveis: 0 e 1 (não expressos e expressos)

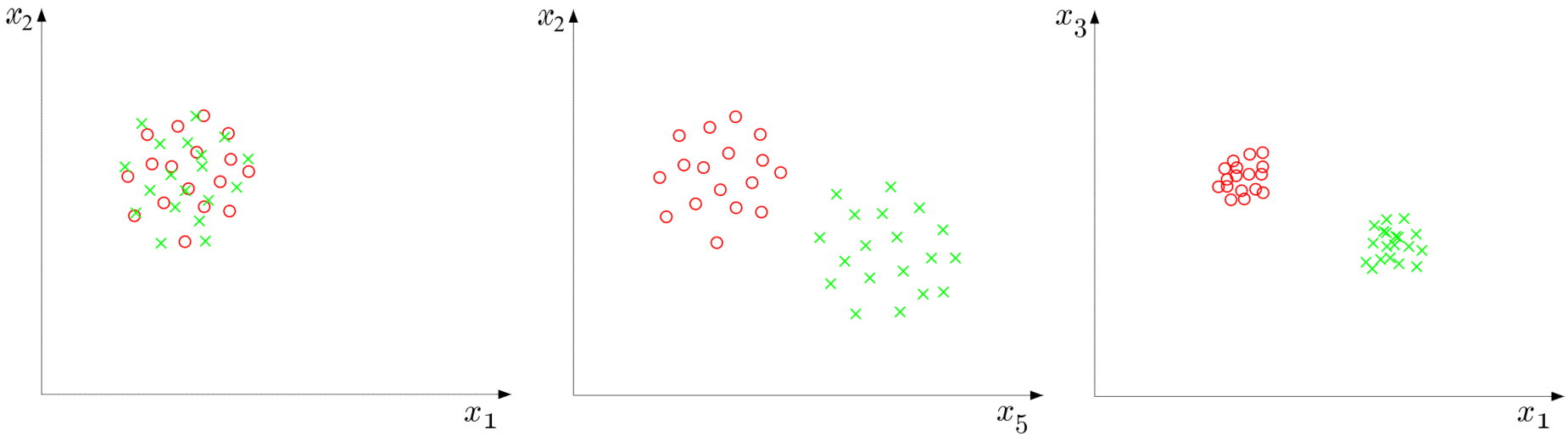
Classificação Supervisionada

- Iremos usar uma abordagem de reconhecimento de padrões conhecida como **Seleção de Características**, como definida por Barrera et al. [5]
- A seleção de características é usada para escolher o "melhor" conjunto de varáveis para a classificação
- É caracterizada por dois elementos principais: **Algoritmo de Busca** e **Função Critério**

Seleção de Características

A idéia de se usar seleção de características é encontrar as características que melhor separem essas classes (conhecidas).

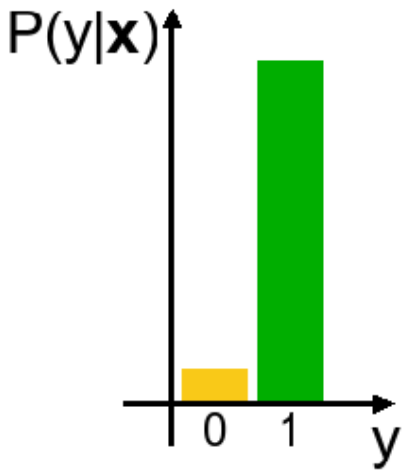
Exemplos:



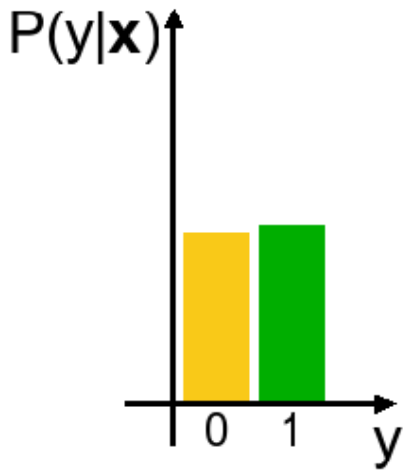
Função Critério

- Entropia condicional média (teoria da informação)
  - $H(Y \mid \mathbf{X}) = \sum_{\mathbf{x} \in \mathbf{X}} P(\mathbf{x})\, H(Y \mid \mathbf{x})$

- Baixos valores de H produzem melhores espaços de características



$H(Y|\mathrm{x})$ baixa



$H(Y|\mathrm{x})$ alta

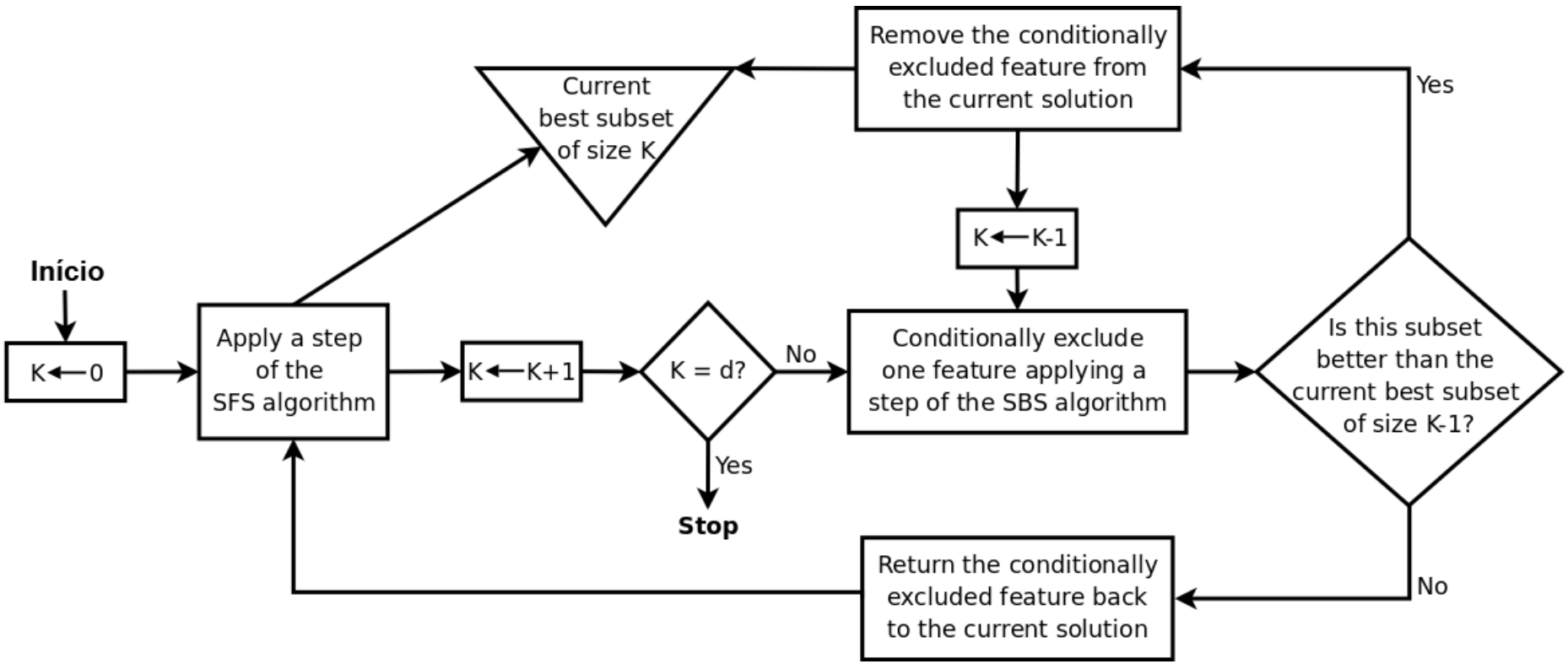
Algoritmo de Busca - SFS e SBS

- Sequential Forward Selection (**SFS**)
  - O conjunto de características inicia vazio e características são incluídas de acordo com a função critério.
- Sequential Backward Selection (**SBS**)
  - O conjunto de características inicia cheio e características são descartadas de acordo com a função critério.
- Apresentam uma desvantagem conhecida como **efeito nesting**

Algoritmo de Busca - SFFS

- Sequential floating forward selection (**SFFS**), Pudil et al. [7]:
  - Os algoritmos SFS e SBS são sucessivamente aplicados
  - A quantidade de características incluída/removida a cada iteração é flutuante, de acordo com a função critério
- Evita o efeito nesting

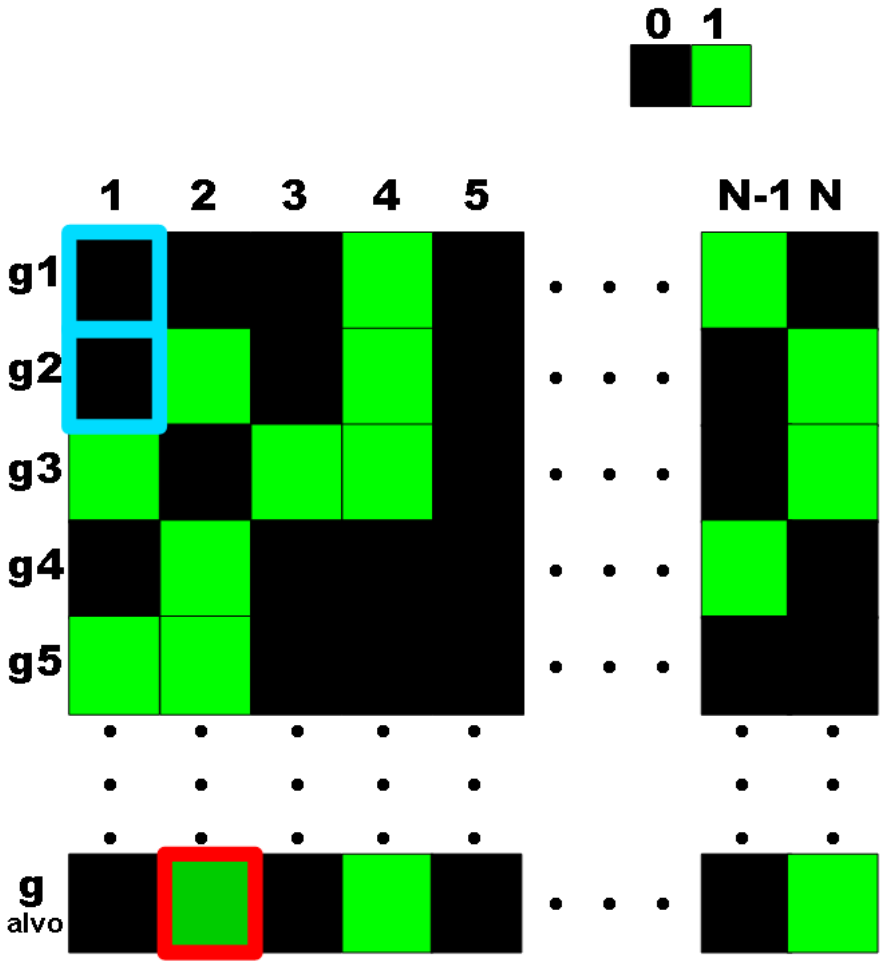
Algoritmo de Busca - Diagrama SFFS



Método de Classificação

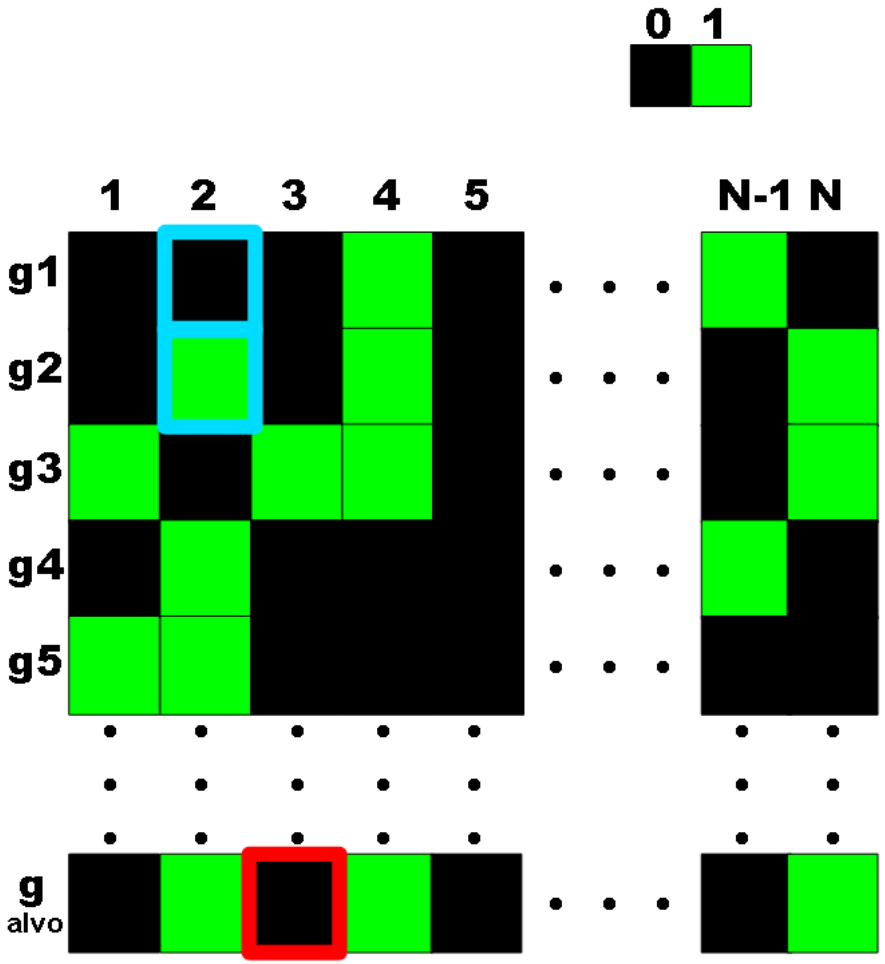
- Escolha de um gene ou conjunto de genes como alvos de interesse (**gene alvo**)
- Observação dos demais genes como possíveis preditores.
- Os possíveis preditores são observados no instante de tempo **t** e o alvo no instante **t+1**.

Método de Classificação - Descrição



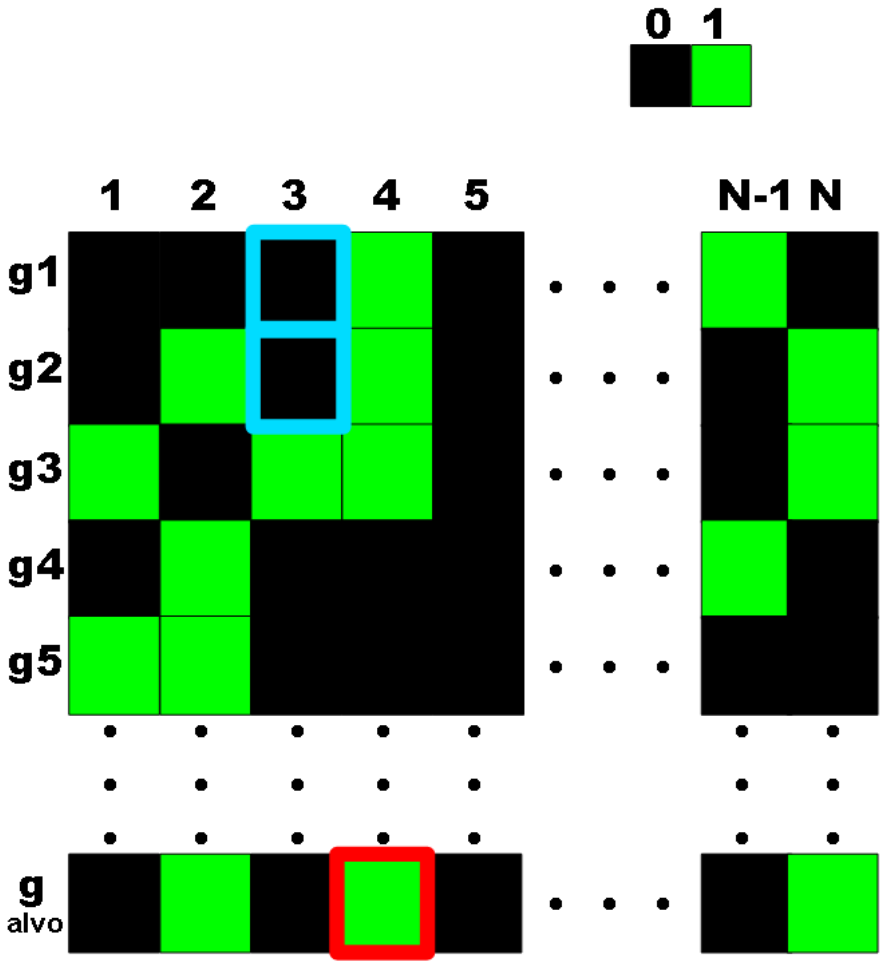
g1	g2	g alvo 0	g alvo 1
0	0	0	1
0	1	0	0
1	0	0	0
1	1	0	0

Método de Classificação - Descrição



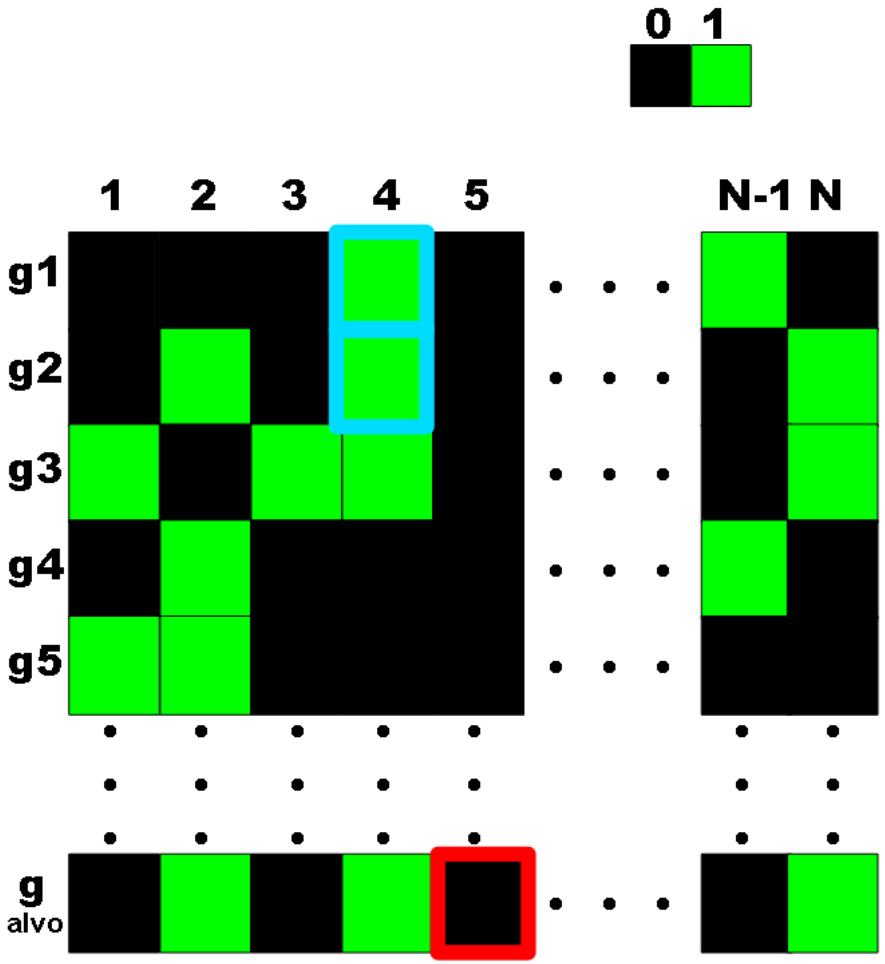
g1	g2	g alvo 0	g alvo 1
0	0	0	1
0	1	1	0
1	0	0	0
1	1	0	0

Método de Classificação - Descrição



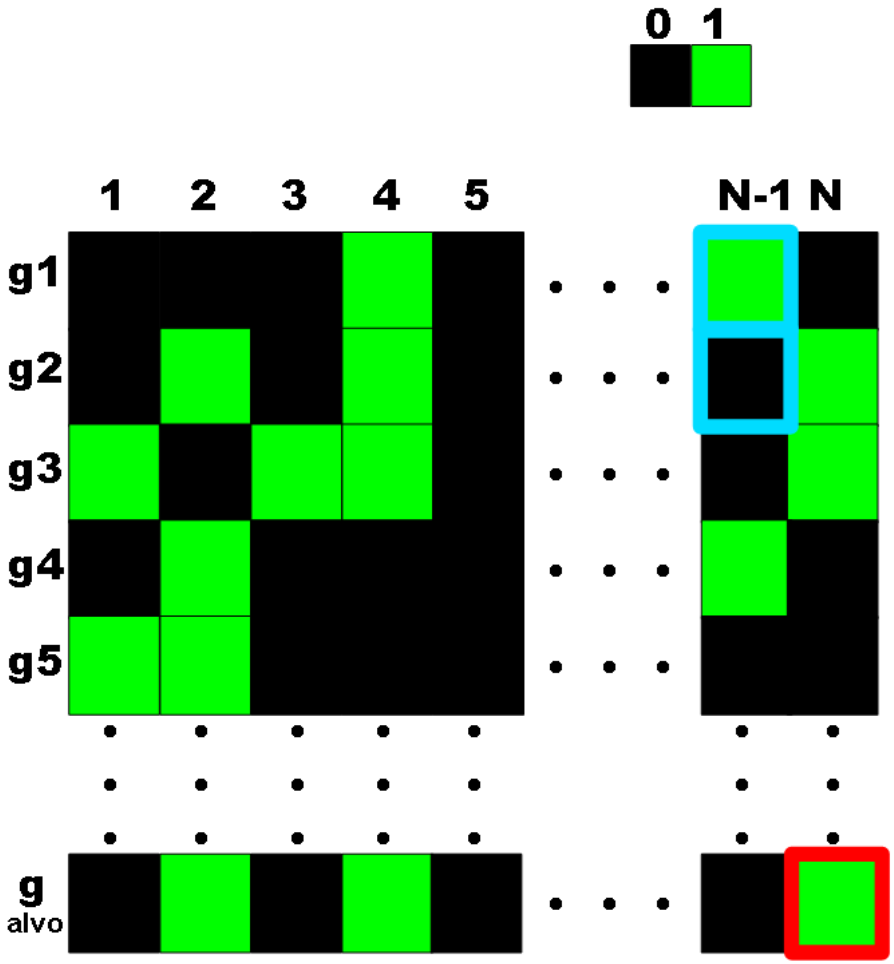
g1	g2	g alvo 0	g alvo 1
0	0	0	2
0	1	1	0
1	0	0	0
1	1	0	0

Método de Classificação - Descrição



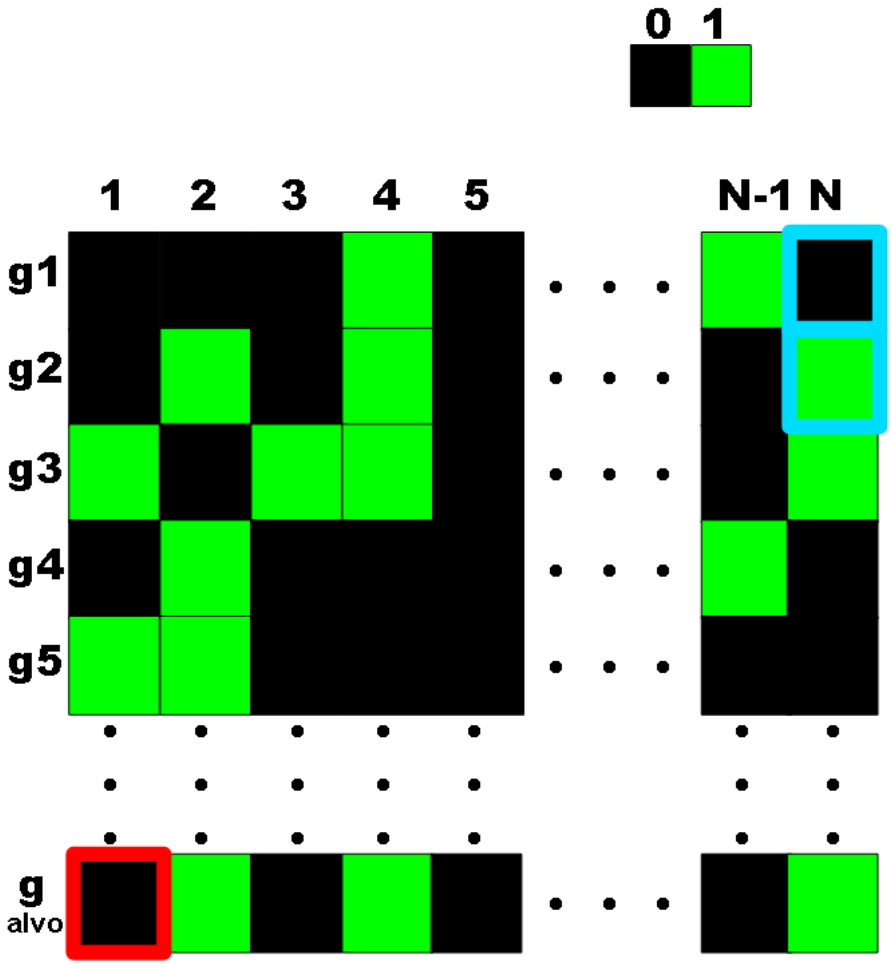
g1	g2	g alvo 0	1
0	0	0	2
0	1	1	0
1	0	0	0
1	1	1	0

Método de Classificação - Descrição



g1	g2	g alvo 0	g alvo 1
0	0	0	9
0	1	7	0
1	0	0	7
1	1	5	0

Método de Classificação - Descrição



g1	g2	g alvo 0	g alvo 1
0	0	0	9
0	1	8	0
1	0	0	7
1	1	5	0

Método de Classificação - Descrição

g1	g2	g alvo 0	g alvo 1
0	0	0	9
0	1	8	0
1	0	0	7
1	1	4	1

Características da dupla (g1,g2)

Entropia **baixa**

Predição **quase perfeita**

Forte candidata possivelmente g1 e g2 serão conectados ao gene alvo

g4	g6	g alvo 0	g alvo 1
0	0	4	5
0	1	3	5
1	0	4	3
1	1	3	2

Características da dupla (g4,g6)

Entropia **alta**

Predição **muito ruim**

Desclassificada

DimReduction - Inferência de Redes

Software que implementa essa abordagem está disponível em `http://code.google.com/p/dimreduction/`.

BMC Bioinformatics

IMPACT FACTOR 3.03

Software

Highly accessed Open Access

Feature selection environment for genomic applications

Fabrício Martins Lopes[1,2], **David Corrêa Martins Jr**[1] and **Roberto M Cesar Jr**[1]

1 Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão 1010, 05508-090, São Paulo-SP, Brazil

2 COINF, Universidade Tecnológica Federal do Paraná, Av. Alberto Carazzai, 1640, 86300-000, Cornélio Procópio-PR, Brazil

author email corresponding author email

BMC Bioinformatics 2008, **9**:451 doi:10.1186/1471-2105-9-451

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2105/9/451

Received: 30 May 2008
Accepted: 22 October 2008
Published: 22 October 2008

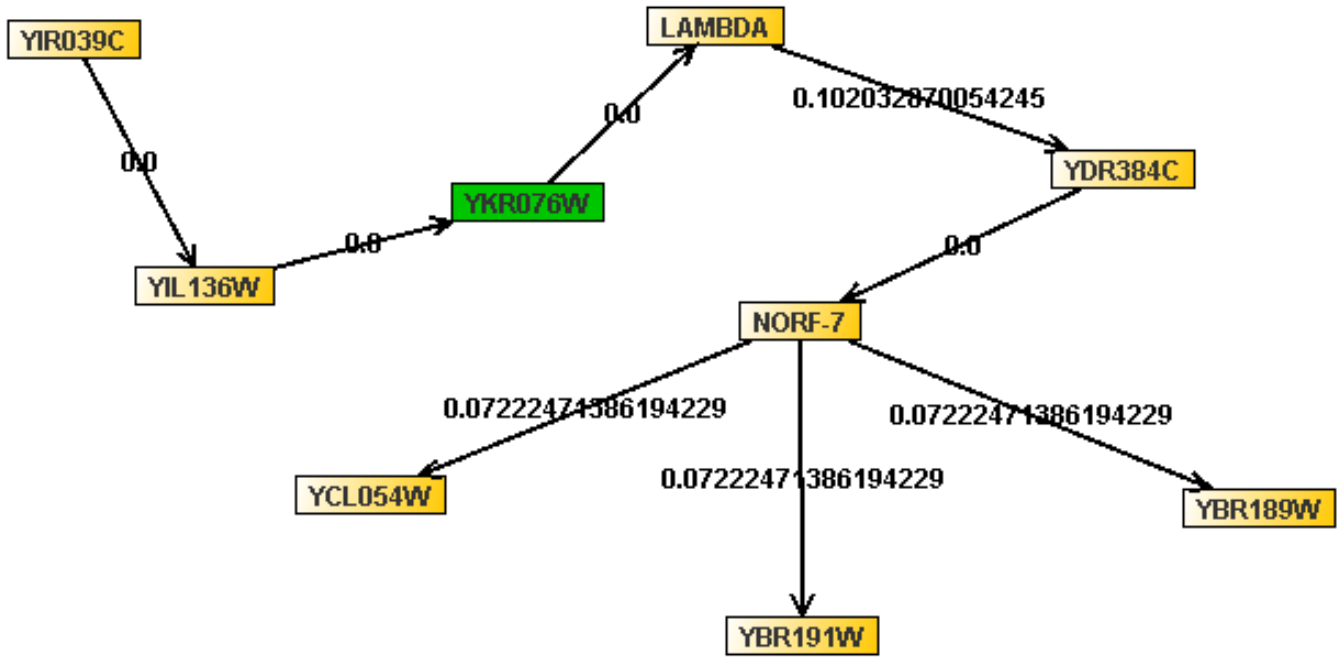
DimReduction - Seleção de Características e Inferência de GRNs

- Usado o aplicativo para seleção de características (DimReduction [6]).
  - pré-processamento dos dados
  - diferentes algoritmos de seleção de características
  - diferentes funções critérios
  - validação cruzada
  - ferramentas de visualização
  - help e documentação
  - software livre

`http://code.google.com/p/dimreduction/` ou
`http://sourceforge.net/projects/dimreduction`

Resultado - Rede Gênica

Foi considerado o gene **YKR076W** como alvo para identificação da rede gênica.
Resultado:



Observações Finais

- Pré-processamento dos dados
- Cuidado ao usar métodos de agrupamento
- Qual classificador, medida de distância e Método de distâncias entre agrupamentos usar?
- Quais genes considerar para análise e classificação?

Bioinfo-CP

Obrigado!

fabricio@utfpr.edu.br

Referências I


S. Theodoridis and K. Koutroumbas
Pattern Recognition.,
Academic Press, 1999.


R. O. Duda and P. E. Hart and D. G. Stork
Pattern Classification.,
Wiley-Interscience, 2000.


D. Jiang and C. Tang and A. Zhang
Cluster Analysis for Gene Expression Data: A Survey.
IEEE Trans. on Knowledge and Data Engineering,
16(11):1370-1386, 2004.

Referências II


J. L. DeRisi and V.R. Iyer, P. O. Brown
Exploring the metabolic and genetic control of gene expression on a genomic scale.
Science, 278(5338):680-6, 1997.


J. Barrera and R. M. Cesar-Jr and et al.
Methods of Microarray Data Analysis V, Constructing probabilistic genetic networks of Plasmodium falciparum from dynamical expression signals of the intraerythrocytic development cycle.
Springer-Verlag, 2006.

Referências III


F. M. Lopes and D. C. Martins-Jr and R. M. Cesar-Jr
Feature selection environment for genomic applications.
BMC Bioinformatics, 9(451), 2008.


P. Pudil and J. Novovičová and J. Kittler
Floating search methods in feature selection.
Pattern Recogn. Lett., 11(15):1119-1125, 1994.