UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Classificação de genes em hibridação genómica comparativa de estirpes de *Streptococcus pneumoniae*

Liliana Sofia Mendonça Cardoso

MESTRADO EM BIOESTATÍSTICA

2009

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Classificação de genes em hibridação genómica comparativa de estirpes de *Streptococcus pneumoniae*

Liliana Sofia Mendonça Cardoso

MESTRADO EM BIOESTATÍSTICA

Dissertação orientada pela Prof.ª Marília Cristina de Sousa Antunes e pelo Prof. Francisco Rodrigues Pinto

2009

Agradecimentos

A todos os que de alguma forma contribuíram para a realização desta dissertação de Mestrado deixo aqui os meus reconhecidos agradecimentos.

Aos meus orientadores por toda a ajuda, pelo incentivo e sugestões feitas ao longo deste projecto e, ainda pela disponibilidade que sempre demonstraram durante este percurso.

Ao corpo docente do Departamento de Estatística e Investigação Operacional da FCUL pelos ensinamentos que me transmitiram, os quais considero que constituem uma mais-valia para o meu futuro, contribuindo para o meu crescente interesse e gosto pela Estatística.

Aos meus pais por todo o esforço e sacrifício que fizeram ao longo da minha formação académica para me proporcionarem todas as oportunidades que não tiveram.

Ao meu namorado pelo seu apoio, incentivo e paciência importantes para a realização desta tese.

Algumas palavras de agradecimento também para o Professor Doutor Mário Ramirez, director da Unidade de Microbiologia Molecular e Infecção do Instituto de Medicina Molecular pela oportunidade dada de trabalhar no projecto de investigação "*Caracterização microbiológica e molecular de Streptococcus patogénicos e sua importância na profilaxia e terapêutica das infecções humanas em Portugal*" (bolsa de investigação Ref. IMM/BI/12-2008), que conduziu à escolha do tema desta dissertação.

Resumo

O *Streptococcus pneumoniae* é uma espécie bacteriana responsável por várias infecções no Homem e possui no seu conteúdo genómico uma vasta diversidade. De entre as suas diversas linhagens genéticas é possível identificar aquelas que estão associadas a uma maior virulência através da presença ou ausência de genes específicos. A hibridação genómica comparativa em *microarrays* é uma tecnologia que examina a semelhança genómica entre organismos e permite a busca em larga escala de genes determinantes da virulência bacteriana. A comparação genómica entre estirpes de *Streptococcus pneumoniae* com genoma sequenciado (amostra de controlo) e estirpes com genoma ainda não sequenciado (amostra de teste) permite detectar os genes que são comuns às duas amostras (genes presentes na amostra de teste) e aqueles que são específicos à amostra de controlo (genes ausentes da amostra de teste).

Nesta dissertação foram usados o algoritmo EM e o classificador bayesiano (ambos baseados em modelos de mistura) com o objectivo de se encontrar uma metodologia que, através desta comparação, permita classificar os genes em presentes ou ausentes na amostra de teste.

Bons resultados foram alcançados com o uso do classificador bayesiano após pré-classificação obtida pelo algoritmo EM, usando como dados o rácio entre as intensidades da amostra de controlo e de teste, sem transformação logarítmica nem *normalização*. Corrigindo o rácio das intensidades de acordo com o número de estirpes de controlo identificado por cada *spot*, os resultados anteriores foram ainda melhorados .

Palavras-chave: *microarray*, hibridação genómica, algoritmo EM, classificador bayesiano, *Streptococcus pneumoniae*.

Abstract

Streptococcus pneumoniae is a bacterial microorganism accountable for many types of human infections with highly diverse genomic content. Among its several genetic lineages it is possible to identify a few which are associated with increased virulence due to the presence (or absence) of specific genes. Comparative genomic hybridization using microarrays is a technology used to examine genomic similarity among organisms and allows for a high throughput search for the genes responsible for bacterial virulence. Genomic comparison between sequenced strains (the reference sample) and unsequenced strains (the test sample) of *Streptococcus pneumoniae* allows the detection of genes which are common to both strains (genes that are present in the test sample) and genes which are specific to the reference sample (genes that are absent from the test sample).

In this thesis we used the EM algorithm and a Bayesian classifier (both based on mixture models) with the purpose of finding an optimal strategy to classify genes as present (or absent) in the test sample.

Good results were obtained with the Bayesian classifier (following a previous classification by the EM algorithm), using as data the ratio of reference strain intensities compared to test strain intensities without logarithmic transformation or normalization. By correcting the intensities ratio accordingly with the number of reference strains identified by each spot, we can improve the previous results.

Keywords: microarray, genomic hybridization, EM algorithm, Bayesian classifier, *Streptococcus pneumoniae*.

Conteúdo

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

O *Streptococcus pneumoniae* é uma espécie bacteriana responsável por várias infecções no Homem, umas de carácter mais grave, como a pneumonia e a meningite, e outras menos severas como a sinusite e a otite média, e para além disso integra a flora comensal que coloniza a nasofaringe humana mas sem consequências para o hospedeiro. Esta espécie bacteriana possui no seu conteúdo genómico uma vasta diversidade revelada pelo comportamento heterogéneo das suas diversas linhagens genéticas e pelo facto da virulência[1] não estar igualmente distribuída dentro da população. Dada a existência desta diversidade, é possível identificar linhagens genéticas associadas a uma maior virulência através da presença ou ausência de genes específicos.

Tecnologias genómicas, como a hibridação genómica comparativa em *microarrays* (*comparative genomic hybridization*, ou CGH) são utilizadas para examinar a semelhança genómica entre organismos (como por exemplo em Behr et al. (1999) e Björkholm et al. (2001)) e permitem a busca em larga escala de genes determinantes da virulência bacteriana. Segundo Snipen et al. (2006) este tipo de tecnologia tem sido uma ferramenta bastante utilizada em comparação genómica providenciando uma forma rápida de genomotipar[2] estirpes bacterianas.

A aplicação desta tecnologia pode revelar-se de grande importância pois,

[1] A virulência corresponde à capacidade relativa de um microrganismo causar danos num hospedeiro susceptível.

[2] Genomotipar (*genomotyping*) consiste na identificação dos genes presentes no genoma de um indivíduo pertencente a uma determinada espécie. Note-se que genomotipar é diferente de genotipar. Genotipar consiste na determinação do genótipo de um indivíduo pertencente a uma determinada espécie, sendo que genótipo é a constituição genética de um indivíduo referente a uma característica em estudo.

uma vez que a comparação genómica leva à descoberta de novos determinantes de virulência, irá contribuir para um melhor conhecimento da patogénese pneumocócica resultando em melhores diagnósticos e no desenho de novas estratégias terapêuticas ou de vacinação, posição que é defendida em Lucchini et al. (2001) e Repsilber et al. (2005).

1.1 Hibridação genómica comparativa em *microarrays*

Antes de mais é essencial elucidar o leitor acerca da tecnologia de *microarrays* que começou a ser utilizada na última década do século XX, com o objectivo de quantificar em larga escala a concentração celular de ARN mensageiro (ARNm). Esta tecnologia é bastante sofisticada e altamente interdisciplinar e produz quantidades elevadas de dados cuja gestão e análise coloca desafios importantes a Estatísticos, Informáticos e Matemáticos. Em Silva (2003) um *microarray* é definido como sendo "essencialmente um suporte de vidro (como uma transparência) onde são colocadas em locais fixos e individuais (*spots*) várias porções de uma mesma sequência de ADN (por exemplo, correspondente a um determinado gene)" e um só *microarray* pode conter milhares de *spots*.

Segundo Newton et al. (2006), a forma mais popular de usar esta tecnologia é na comparação dos níveis de expressão dos genes (a qual é feita ao nível do ARNm) em duas ou mais amostras diferentes, como por exemplo, em estado saudável e doente - estudos de expressão genética. No entanto, os *microarrays* são também utilizados em CGH mas neste caso a comparação é feita ao nível do ADN genómico em vez de ser ao nível do ARN mensageiro, estudando-se o conteúdo de genes[3] em vez da expressão do gene[4].

A ideia básica por detrás da tecnologia CGH consiste em construir *microarrays* de genomas sequenciados e anotados em que a amostra de controlo, constituída por uma ou mais estirpes totalmente sequenciadas, é comparada a um conjunto de estirpes não sequenciadas da mesma, ou relativamente próxima, espécie bacteriana (amostra de teste). O objectivo é detectar as semelhanças e as diferenças no conteúdo genómico através da caracterização das amostras de teste e de controlo tendo em conta os genes comuns às duas amostras e aqueles que são específicos à amostra de controlo.

[3] O estudo do conteúdo de genes diz respeito à identificação da lista de genes presentes num dado genoma.

[4] O estudo da expressão do gene diz respeito à quantificação do ARNm de cada gene que está a ser expresso pela célula num dado momento e condição fisiológica.

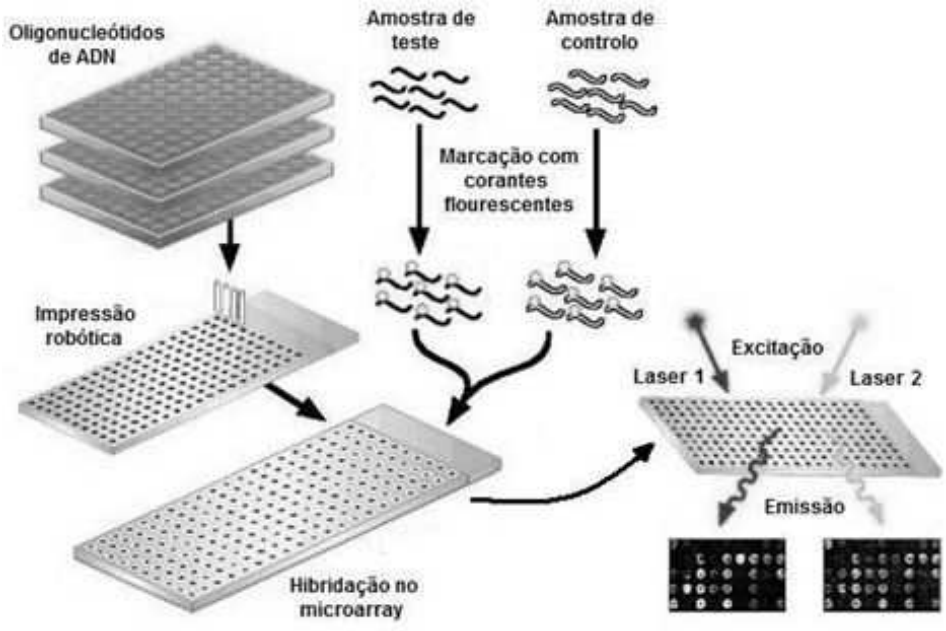

Figura 1.1: Esquema da construção de um *microarray* de duas cores

Em experiências de CGH com *microarrays* de duas cores (*two-color microarrays* ou *two-channel microarrays*) são hibridadas no *microarray* duas amostras - amostra de teste e amostra de controlo - para que possam ser comparadas (ver Figura 1.1). A amostra de controlo é tipicamente ADN de uma ou mais estirpes totalmente sequenciadas, geralmente as mesmas com as quais o *microarray* foi construído e a amostra de teste é constituída por ADN de uma estirpe não sequenciada da mesma espécie bacteriana. Cada amostra é marcada com um corante fluorescente sendo geralmente utilizados o vermelho (*Cy5*), que emite fluorescência de 670nm de comprimento de onda (correspondendo à parte vermelha do espectro luminoso), e o verde (*Cy3*), que emite 570nm de comprimento de onda (correspondendo à parte verde do espectro luminoso). Durante a hibridação, sequências que existam nas amostras que sejam complementares aos oligonucleótidos presentes nos spots do *microarray* vão ligar-se preferencialmente a estes. Depois da hibridação o *microarray* é introduzido num *scanner* o qual, através de um laser (com comprimentos de onda diferentes), vai excitar cada *spot* do *microarray*, e são então efectuadas as medições das intensidades fluorescentes para cada cor.

Segundo Snipen et al. (2009), os dados provenientes deste tipo de experiências são qualitativamente diferentes daqueles que são obtidos em estudos de expressão genética, onde as intensidades devem ser consideradas como contínuas devido à abundância dinâmica de ARNm. Em estudos CGH

de espécies bacterianas, as diferenças nas intensidades das fluorescências devem-se essencialmente a diferenças na composição das sequências, e as intensidades são usadas para identificar quais os genes que estão presentes ou ausentes (ou com sequência divergente) na amostra de teste. Todavia, um dos problemas críticos enfrentados na interpretação de dados CGH em *microarrays* é na classificação desses genes como presentes ou ausentes. Um gene presente tem, em teoria, intensidades de fluorescência iguais nas amostras de teste e de controlo enquanto um gene ausente terá apenas verdadeira intensidade na amostra de controlo.

1.2 Metodologias aplicadas na classificação dos genes

Várias abordagens têm sido aplicadas para resolver o problema da classificação dos genes em presentes ou ausentes.

A maioria das análises a dados de experiências CGH foca-se no rácio das intensidades, $R_i = T_i/C_i$ (ou no logaritmo do rácio, $LR_i = \log_2(\frac{T_i}{C_i})$), onde T_i representa a intensidade da amostra de teste e C_i a intensidade da amostra de controlo para o gene i ($i = 1, ..., n$). Segundo Dean e Raftery (2005), esta quantidade foi usada primeiramente para determinar genes diferencialmente expressos através da "regra de 2", segundo a qual os genes com rácio das intensidades maior que 2 ou menor que 1/2 eram considerados diferencialmente expressos. De um modo geral, os estudos CGH utilizam o rácio das intensidades ou o logaritmo do rácio para ordenar os genes e classificá-los, em ausentes ou presentes, através de um ponto de corte escolhido empiricamente ou recorrendo a estatísticas calculadas com base na distribuição de LR. Ordenando os genes por ordem crescente de LR aqueles que estiverem no topo da lista são, em princípio, bons candidatos a genes ausentes.

Em alguns dos estudos CGH, o método utilizado para resolver o problema de classificação dos genes consistiu, à semelhança da "regra de 2", na escolha de um ponte de corte constante (Dziejman et al. (2002), Murray et al. (2001), Salama et al. (2000) e Dorrell et al. (2001)). Por exemplo em Dorrell et al. (2001) e em Salama et al. (2000), os genes com LR maior que -1.0 (ou 0.5, no caso do rácio R) são considerados presentes. Segundo Kim et al. (2002) este ponto de corte é determinado empiricamente através da comparação da estirpe de referência com outra similar, a qual se sabe ter em falta certos elementos genéticos. No entanto, nem sempre está disponível para todos os organismos uma estirpe assim, o que impede a determinação empírica do ponto de corte e, para além disso, pequenos desvios no valor do ponto de corte podem levar a classificar erradamente muitos genes. Um método mais robusto e preciso foi descrito em Kim et al. (2002), baseando

a escolha do ponto de corte na forma da distribuição de LR, a qual é geralmente enviesada à esquerda dependendo o tamanho da cauda do número de genes ausentes. O ponto de corte que divide os genes em dois grupos é assim calculado independentemente para cada conjunto de dados de hibridação (para cada *microarray*) e é determinado tendo em conta a variação na composição genómica das estirpes analisadas e a qualidade da hibridação.

Outro método é sugerido em Repsilber et al. (2005) o qual, tendo em conta a representação gráfica de M vs. A (*M-A plot*)[5], propõe uma rotação dos dados de 63.4 graus no sentido anti-horário antes de ordenar os genes e determinar o ponto de corte, abordagem que parece melhorar a proporção de genes ausentes bem classificados.

Os modelos de mistura são uma ferramenta que também tem sido utilizada para resolver o problema da classificação dos genes. Estes modelos já foram usados em dados de estudos de expressão genética (Dean e Raftery (2005), e Antunes e Sousa (2008)) e em dados CGH (Snipen et al. (2006) e Feten et al. (2007)), a fim de modelar o rácio (ou o logaritmo do rácio) das intensidades tendo em conta que o conjunto de dados se refere a subpopulações diferentes, normalmente duas (uma subpopulação referente aos genes presentes e outra referente aos genes ausentes), sendo assim cada subpopulação modelada pela respectiva distribuição. Esta estratégia permite obter estimativas de probabilidade a *posteriori* de cada gene pertencer a cada subpopulação, possibilitando assim a classificação dos genes.

Algumas abordagens bayesianas têm também sido aplicadas. Em Lönnstedt e Speed (2002) os genes são ordenados em relação à probabilidade a *posteriori* de cada gene ser diferencialmente expresso através do cálculo do logaritmo das chances a *posteriori*. No entanto, este processo não produz qualquer ponto de corte e geralmente são seleccionados como diferencialmente expressos os genes do top 50, 100 ou 150 da lista obtida.

Dado o interesse em classificar os genes de uma forma mais fiável e rigorosa outras técnicas foram desenvolvidas. Em Kendziorski et al. (2003) é proposto um modelo bayesiano empírico baseado num modelo de mistura hierárquico. Duas parametrizações distintas são consideradas: uma em que o rácio das intensidades observado é modelado pela distribuição Gama e o correspondente parâmetro de escala tem também distribuição Gama e outra em que o logaritmo do rácio é modelado pela distribuição Normal e a distribuição a *priori* para a média é também Normal. Através destes modelos são obtidas as probabilidades a *posteriori* com as quais é feita inferência acerca da expressão diferencial.

No entanto, apesar de mais fiável, também esta abordagem não permite

[5]M-A plot é a representação gráfica do logaritmo do rácio das intensidades, M=LR, contra a média do logaritmo das intensidades das amostras de teste e de controlo, A=$(1/2)(\sum_i \log_2 T_i + \sum_i \log_2 C_i)$.

obter um ponto de corte capaz de dividir os genes em grupos distintos. Esse objectivo é conseguido em Lewin et al. (2007), e Antunes e Sousa (2008) onde são propostos dois modelos bayesianos hierárquicos diferentes para detectar genes diferencialmente expressos através dos quais é possivel obter uma regra de classificação.

1.3 Estrutura da dissertação

Neste trabalho é examinada a semelhança genómica entre amostras de estirpes sequenciadas da bactéria *Streptococcus pneumoniae* e são por isso conhecidos os genes comuns e os genes específicos entre as amostras de estirpes comparadas. Desta forma, o objectivo do projecto consiste em encontrar uma metodologia eficaz na identificação de genes presentes e ausentes na amostra em estudo (amostra de teste) comparando-a com a amostra de controlo, que se possa utilizar quando a amostra em estudo é desconhecida, ou seja, é constituída por uma estirpe bacteriana não sequenciada.

Com este propósito, no capítulo 3, são estudados e analisados do ponto de vista teórico alguns métodos que, com base na modelação do rácio das intensidades, irão permitir classificar cada gene da amostra em estudo em presente ou ausente. Os métodos estudados serão depois aplicados aos dados disponíveis, os quais são descritos pormenorizadamente no capítulo 2, e a eficácia de cada método é avaliada, o que permitirá tecer algumas conclusões acerca da qualidade dos métodos.

Em alguns estudos, os *microarrays* são construídos com base em mais do que uma estirpe bacteriana sequenciada (Pinto et al. (2008) e Salama et al. (2000)) e, neste caso, a escolha para a amostra de controlo pode ser ADN de uma única estirpe ou uma mistura de ADN das estirpes sobre as quais o *microarray* foi construído. Quando a amostra de controlo é constituída por uma mistura de ADN de diferentes estirpes, os *microarrays* onde são hibridadas as amostras, contêm sequências que podem identificar genes presentes numa só estirpe, em mais do que uma ou até em todas as estirpes que constituem a amostra de controlo. No capítulo 4 será abordada esta questão e procurar-se-á encontrar uma forma de classificar os genes que tenha em conta o facto da amostra de controlo ser composta por uma mistura de estirpes, podendo-se eventualmente assim conseguir aumentar a eficácia do método de classificação.

Por fim, as conclusões finais bem como as perspectivas futuras no âmbito da classificação de genes em hibridação genómica comparativa são apresentadas no capítulo 5.

CAPÍTULO 2

Descrição detalhada dos dados

Os dados que vão ser analisados neste projecto foram obtidos a partir de quatro experiências de hibridação genómica comparativa em *microarrays* com o propósito de comparar ADN de duas amostras de estirpes de *Streptococcus pneumoniae*.

As amostras que estão a ser comparadas são a amostra de teste constituída por ADN da estirpe R6 e a amostra de controlo constituída por uma mistura de ADN das três estirpes de *Streptococcus pneumoniae* já sequenciadas, R6, G54 e TIGR4 (MIX). Cada amostra de ADN foi marcada com um corante fluorescente, vermelho (Cy5) e verde (Cy3), e depois hibridada no *microarray* , competindo para se ligar ao oligonucleótido complementar (ver tabela 2.1). Finalmente, o slide (ou *microarray*) é introduzido num *scanner* que possui um laser que permite medir as intensidades de cada canal (R e G, respectivamente, vermelho e verde). A intensidade será tanto maior quanto maior for a afinidade entre as sequências complementares ou maior a concentração das sequências nas amostras.

Para cada slide temos a seguinte informação resultante da análise de imagem realizada pelo Feature Extraction 9.1[1]:

- Intensidades medianas para cada um dos dois canais;
- Intensidades médias para cada um dos dois canais;
- Intensidades normalizadas para cada um dos dois canais.

[1] O Feature Extraction é um *software* de análise de imagem de *microarrays* produzido por *Agilent Technologies (Palo Alto, CA)*.

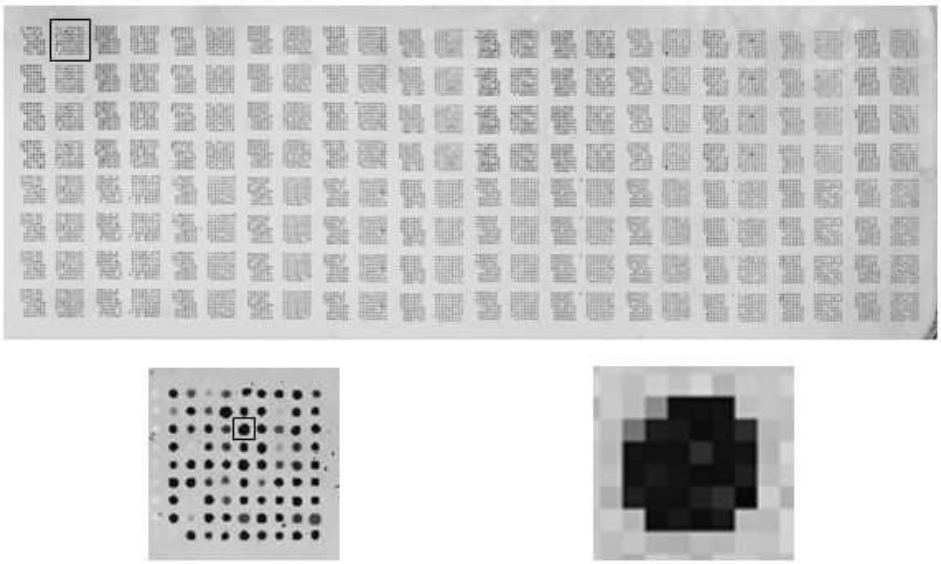

Figura 2.1: Imagens de um *microarray* produzidas depois deste ser introduzido no *scanner*: em cima podemos ver todo o *microarray* e a elevada quantidade de *spots* que o constitui; em baixo, à esquerda, apresenta-se a ampliação do fragmento assinalado no *microarray*; em baixo, à direita, vê-se a ampliação do *spot* assinalado no fragmento do *microarray*.

A Figura 2.1 ajuda-nos a ter uma ideia do que estas intensidades significam. A imagem ampliada do *spot* do *microarray* exposta na Figura 2.1 indica-nos que, aquando da leitura das intensidades fluorescentes para cada cor, é armanezada uma imagem computorizada de todo o *microarray*, em que cada *spot* é composto por várias porções mais pequenas (pixéis[2]), correspondendo a cada uma destas porções um valor de intensidade para a cor vermelha e outro para a cor verde. A intensidade fluorescente de cada *spot* pode então ser obtida calculando-se a média ou a mediana das intensidades de cada uma das porções que compõe o *spot*, dando assim origem às intensidades apresentadas nas primeiras quatro colunas da Tabela 2.2. Outra possibilidade, mais complexa, passa por calcular as intensidades corrigidas (*inten.corr*) de cada *spot* i através das intensidades médias (*intmed*) da seguinte forma:

$$inten.corr_i = intmed.spot_i - min(intmed.spot_i).$$

Às intensidades corrigidas é depois aplicado um modelo espacial para eliminar a tendência e os efeitos espaciais que possam existir e às intensi-

[2]De uma forma simples, um pixel pode ser definido como a menor unidade que forma uma imagem digital.

dades daí resultantes é ainda removido o enviesamento através de uma *normalização loess* (ver secção 3.2 para uma explicação desta *normalização*), obtendo-se as intensidades normalizadas (*processed*) expostas nas últimas duas colunas da Tabela 2.2.

	RedChannel - Cy5	*GreenChanel - Cy3*
slide 61	R6	MIX
slide 65	MIX	R6
slide 67	MIX	R6
slide 73	R6	MIX

Tabela 2.1: Desenho dos *microarrays*: identificação das amostras que foram marcadas com corante fluorescente em cada uma das quatro hibridações.

Cada slide contém 17280 *spots* e a cada *spot* corresponde um gene (existem, no entanto, alguns *spots* vazios ou obsoletos[3]). Cada gene tem quatro réplicas em cada slide e existem 3137 genes diferentes representados no *microarray*. Como as estirpes utilizadas nas amostras têm o seu genoma sequenciado, conhecem-se os genes presentes (e ausentes) nessas estirpes, informação que também está contida nos dados. As tabelas 2.2 e 2.3 ilustram a forma como os dados estão organizados e a informação neles contida.

GeneName	*gMedian*	*rMedian*	*gMean*	*rMean*	*gProcessed*	*rProcessed*
SP1528	203	82	208	83,6	7,38	19,4
Obsolete	223,5	97,5	401	483	146	548
SP2043	211	103	519	681	276	703
SPN21003	227	105	824	953	636	863
Obsolete	229,5	119	800	1630	659	1440
SP1031	211,5	83	214	85,6	6,61	21,7
Empty	—	—	—	—	—	—
Empty	—	—	—	—	—	—
spr0952	1396	6383	1120	5270	1130	4340
spr1701	1074,5	2954,5	1420	4460	1500	3640
SPN09154	5352,5	10344,5	4080	7810	5190	5800
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabela 2.2: Exemplo das intensidades de fluorescência de um dos quatro slides

[3]Os *spots* obsoletos contêm oligonucleótidos que os fabricantes vieram a descobrir não ser os melhores identificadores de um dado gene.

GeneName	*Strain.R6*
SP1528	0
Obsolete	0
SP2043	0
SPN21003	0
Obsolete	1
SP1031	0
Empty	-
Empty	-
spr0952	1
spr1701	1
SPN09154	1
⋮	⋮

Tabela 2.3: Exemplo da informação de presença ou ausência dos genes

Sendo assim, ao contrário do que sucede na maioria das experiências deste tipo em que se pretende classificar os genes da amostra de teste, em ausentes ou presentes, comparando-os com uma amostra de controlo constituída por uma ou mais estirpes já sequenciadas, esta experiência servirá de ferramenta para encontrar um bom método de classificação dos genes que possa ser usado numa situação em que não se conheça a classificação dos genes da amostra de teste.

Para avaliar a qualidade/eficácia de um determinado método é utilizada a exactidão (*accuracy*) que se vai designar por *Acc* e se calcula da seguinte forma:

$$Acc = \frac{VP^1 + VN^2}{\text{n.º de genes no slide}},$$

e que corresponde à proporção de resultados correctos, ou seja, à proporção de genes bem classificados. Esta medida pode ser tomada como uma estimativa da probabilidade de classificação correcta em experiências semelhantes em que a classificação da amostra de teste não seja conhecida.

Note-se que em estudos CGH em *microarrays* não existe preferência em obter maiores proporções de genes presentes bem classificados ou maiores proporções de genes ausentes bem classificados, o que se pretende é obter uma proporção de genes (quer sejam presentes ou ausentes) correctamente classificados o mais elevada possível, sendo por isso analisada fundamentalmente a exactidão.

[1]Verdadeiros Positivos: Genes presentes bem classificados
[2]Verdadeiros Negativos: Genes ausentes bem classificados

Para além disso, dada a vasta informação que se tem para cada gene, há ainda que construir uma medida informativa da intensidade que represente cada gene da amostra.

2.1 Análise exploratória dos dados

Para facilitar a distinção entre os genes ausentes e os genes presentes é conveniente construir uma medida da intensidade que separe bem estes dois grupos.

A medida mais utilizada na literatura referente à análise de dados provenientes de *microarrays* é o logaritmo do rácio, $LR = \log_2(\frac{intensidade\ do\ teste}{intensidade\ do\ controlo})$, facto que se deve à capacidade do logaritmo em simetrizar os dados que são muitas vezes enviesados à direita, permitindo assim satisfazer o requisito de normalidade, um pressuposto frequente para aplicação de diversas metodologias. Quando o LR é baixo isso indica que a intensidade do teste é fraca e portanto os genes que apresentarem valores pequenos de LR serão mais provavelmente genes ausentes (Snipen et al., 2006). Para além desta medida também têm sido aplicadas algumas medidas mais complexas envolvendo LR (Repsilber et al., 2005).

Apesar do logaritmo do rácio ser a medida mais frequente existem estudos em que foram adoptadas outras estratégias, veja-se por exemplo Snipen et al. (2009). Neste projecto irá dar-se especial atenção ao rácio das intensidades, $RI = \frac{intensidade\ do\ controlo}{intensidade\ do\ teste}$, por ser a medida que produz menos alterações nos dados e por se pretender aplicar outras distribuições, nomeadamente, a distribuição Gama, a qual constitui um modelo bastante flexível para dados estritamente positivos e permite acomodar valor elevados que são usuais nos genes ausentes (Antunes e Sousa, 2008). Segundo esta medida, os genes com valores mais elevados de RI são então bons candidatos a genes ausentes.

Experimentou-se ainda utilizar outra medida, LRI, definida da seguinte forma $LRI = \log_2(\frac{intensidade\ do\ controlo}{intensidade\ do\ teste}) + c$, em que c é uma constante tal que $LRI > 0$ para cada gene. Esta medida é uma alternativa à anterior pois também pode ser modelada pela distribuição Gama.

Para se ter uma medida da intensidade para cada gene é conveniente, uma vez que cada gene tem quatro réplicas, compactar a informação relativa a cada gene. Duas alternativas surgem: para um mesmo slide calcular a média ou a mediana das medidas da intensidade para cada gene. Para além disso ainda existem quatro slides, ou seja cada gene (para além de ter quatro réplicas dentro de cada slide) possui ainda valores de intensidade em cada um dos quatro slides.

Tendo ainda em conta a informação de que se dispõe para cada slide existem duas alternativas para calcular as medidas da intensidade:

- Utilizar as intensidades médias;
- Utilizar as intensidades medianas.

Para comparar as várias alternativas possíveis para o cálculo da medida da intensidade efectuaram-se diversos *box-plots*. Nas figuras 2.2 e 2.3 podemos ver os gráficos referentes ao cálculo de RI com as intensidades médias e medianas respectivamente.

Da análise dos gráficos conclui-se que o rácio das intensidades das várias alternativas possíveis se distribuem de forma idêntica e portanto a informação dada por cada uma das alternativas é muito semelhante. A mesma conclusão é obtida com as outras medidas de intensidade referidas.

A escolha para o cálculo da medida de intensidade acabou por recair sobre a média dos valores da medida considerada (R, LR, RI ou LRI) que correspondem ao mesmo gene (ou seja, aplicando-se a média dentro de cada slide e também entre os slides) utilizando-se para calcular essa medida para cada gene as intensidades médias.

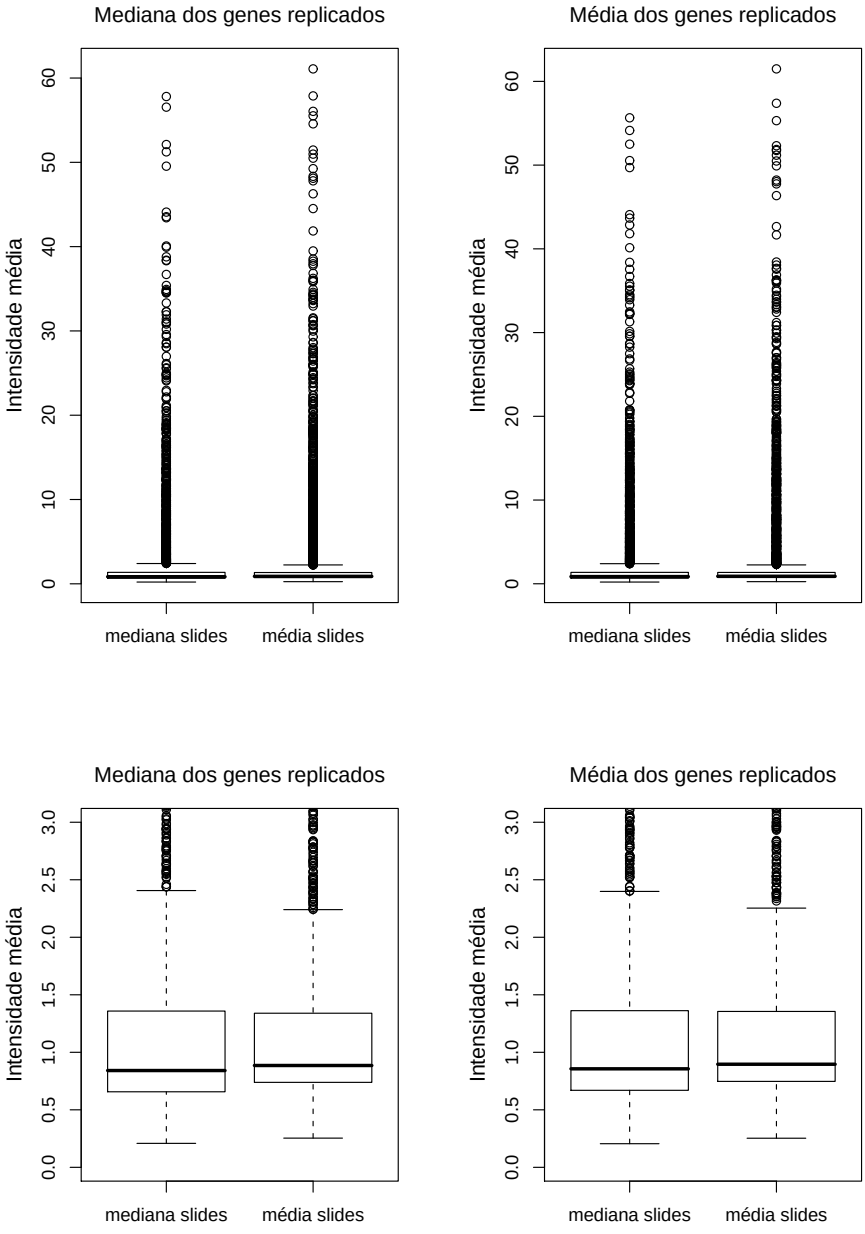

Figura 2.2: *Box-plots* da intensidade média: em cima apresentam-se os *box-plots* do rácio das intensidades (RI) para todo o domínio de RI, em baixo pode-se ver mais pormenorizadamente a parte central desses mesmos *box-plots*.

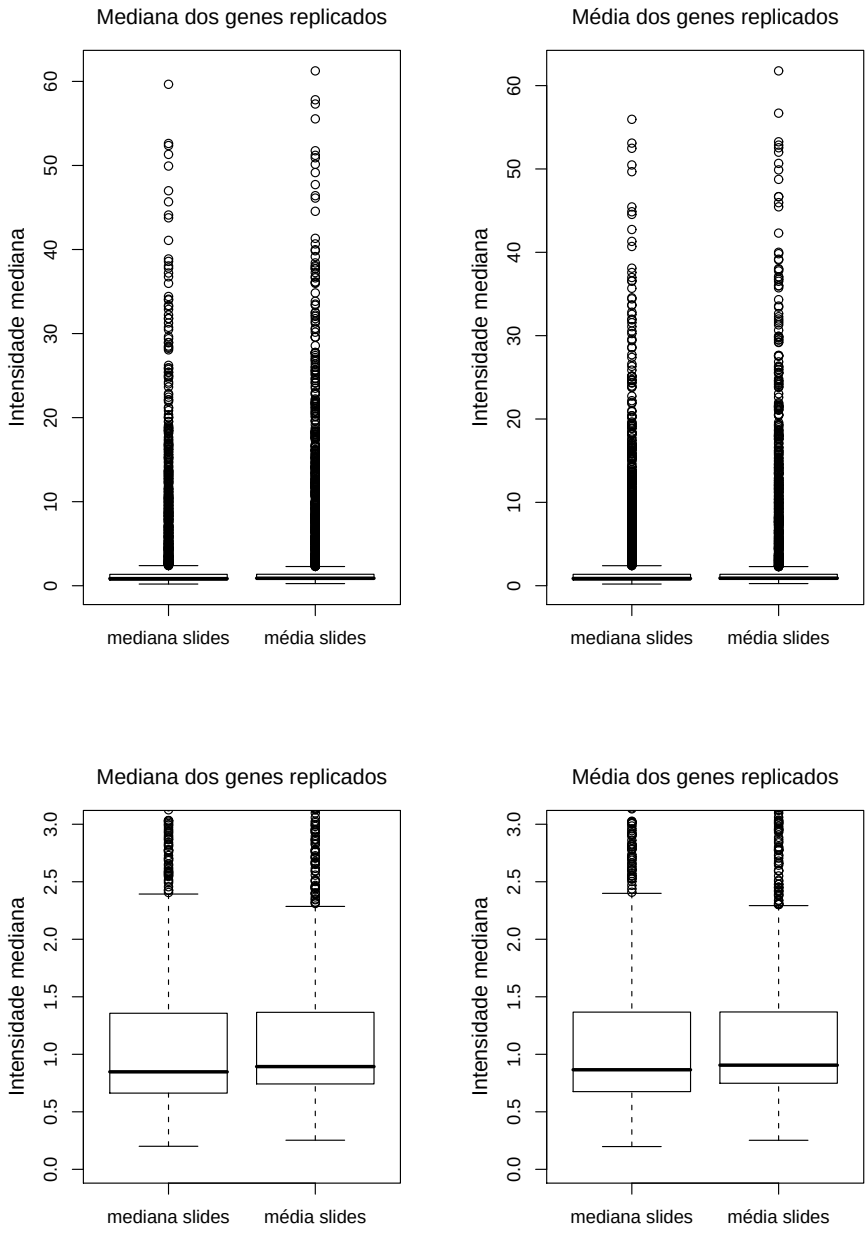

Figura 2.3: *Box-plots* da intensidade mediana: em cima apresentam-se os *box-plots* do rácio das intensidades (RI) para todo o domínio de RI, em baixo pode-se ver mais pormenorizadamente a parte central desses mesmos *box-plots*.

CAPÍTULO 3

Métodos de classificação de genes em CGH

Neste capítulo descrevem-se, do ponto de vista teórico, algumas metodologias através das quais, com base na modelação do rácio das intensidades, se consiga de alguma forma classificar em presente ou ausente cada gene da amostra.

Depois de uma abordagem teórica, os métodos descritos serão aplicados aos dados disponíveis e a sua qualidade será avaliada recorrendo ao cálculo da exactidão através da comparação com a informação presente nos dados a este respeito.

Por último serão feitas algumas considerações e conclusões acerca dos modelos e métodos aplicados a fim de se poder decidir qual o método mais eficaz na classificação dos genes.

3.1 Modelação e métodos de classificação

3.1.1 Modelos de Mistura

Na presença de um conjunto de dados que contém informação vinda de grupos diferentes (grupo dos genes ausentes e dos genes presentes) é comum trabalhar-se com modelos de mistura. Um modelo de mistura é uma combinação convexa de distribuições de probabilidade que permite modelar conjuntos de dados originários de diferentes subpopulações, em que cada subpopulação é modelada pela sua própria distribuição, sendo portanto uma ferramenta útil em contextos como o do presente trabalho.

Considere-se que para cada gene apenas se conhece o respectivo rácio das intensidades e não se tem qualquer informação acerca da sua classificação,

ou seja, não se sabe qual a distribuição correspondente. Este caso pode ser visto como um problema de dados omissos e o algoritmo EM afigura-se como uma ferramenta eficaz na estimação dos parâmetros do modelo e na classificação de cada gene num dos dois grupos distintos (classificação não supervisionada).

Considere-se então o seguinte conjunto de dados $\{x_1, ..., x_n\}$ constituído pelos rácios das intensidades dos n genes da amostra e represente-se por G_A o grupo dos genes ausentes e por G_P o grupo dos genes presentes. De uma forma genérica, todo o conjunto de dados será modelado do seguinte modo:

$$f(x) = \pi_A f(x|\boldsymbol{\theta}_A) + \pi_P f(x|\boldsymbol{\theta}_P) \tag{3.1}$$

em que,

- $\pi_A = P[X_i \in G_A]$ e $\pi_P = P[X_i \in G_P] = 1 - \pi_A$;
- $f(x|\boldsymbol{\theta}_k)$ é a função de densidade de probabilidade do rácio das intensidades para o grupo k, $k = A, P$;
- $\boldsymbol{\theta}_k$ é o vector de parâmetros associado à distribuição do grupo k, $k = A, P$.

O objectivo é estimar as probabilidades *a posteriori* de pertença a cada grupo, para cada gene, e classificá-los no grupo que apresentar uma probabilidade maior.

De seguida apresentam-se os modelos de mistura utilizados no âmbito deste trabalho.

Modelo de mistura Normal-Uniforme

O modelo de mistura Normal-Uniforme proposto em Dean e Raftery (2005) para detectar genes diferencialmente expressos pode também ser utilizado neste contexto. O modelo em causa é aplicado ao logaritmo do rácio das intensidades, considerando-se que são modelados pela distribuição Normal os genes que apresentam, em média, LR próximo de 0, ou seja, os genes presentes enquanto que os genes ausentes são modelados pela distribuição Uniforme por apresentarem valores de LR mais distantes do grupo anterior. Isto é,

$$X|_{X \in G_P} \sim Normal(\mu, \sigma^2), \quad \text{com } f(x|\boldsymbol{\theta}_k) = f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2},$$
$$\mu \in \mathbb{R}, \sigma^2 > 0, x \in \mathbb{R}.$$

e

$$X|_{X \in G_A} \sim Uniforme(a, b), \quad \text{com } f(x|\boldsymbol{\theta}_k) = f(x|a, b) = \frac{1}{b-a}, \quad a < x < b.$$

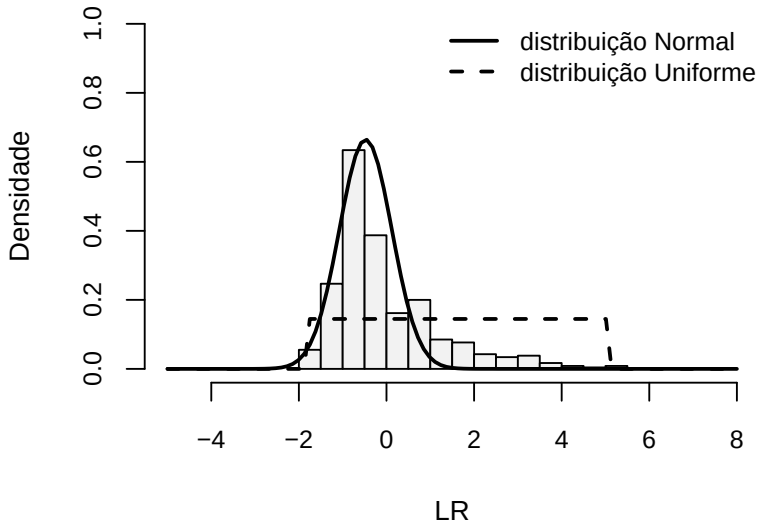

Figura 3.1: Histograma de LR, logaritmo do rácio das intensidades, com sobreposição das densidades das distribuições Normal e Uniforme.

Modelo de mistura Gama-Gama

Neste modelo de mistura segue-se uma abordagem semelhante à tomada em Antunes e Sousa (2008), considerando-se que o rácio das intensidades condicional ao grupo k ($k = A, P$) segue uma distribuição Gama, ou seja,

$$X|_{X\in G_P} \sim Gama(\alpha_P, \beta_P), \quad \text{com } f(x|\boldsymbol{\theta}_P) = f(x|\alpha_P, \beta_P) = \frac{{\beta_P}^{\alpha_P}}{\Gamma(\alpha_P)} x^{\alpha_P - 1} e^{-x\beta_P},$$

$$x > 0, \quad \alpha_P > 0, \quad \beta_P > 0.$$

e

$$X|_{X\in G_A} \sim Gama(\alpha_A, \beta_A), \quad \text{com } f(x|\boldsymbol{\theta}_A) = f(x|\alpha_A, \beta_A) = \frac{{\beta_A}^{\alpha_A}}{\Gamma(\alpha_A)} x^{\alpha_A - 1} e^{-x\beta_A},$$

$$x > 0, \quad \alpha_A > 0, \quad \beta_A > 0.$$

Veja-se a Figura 3.2 para uma ilustração deste modelo de mistura.

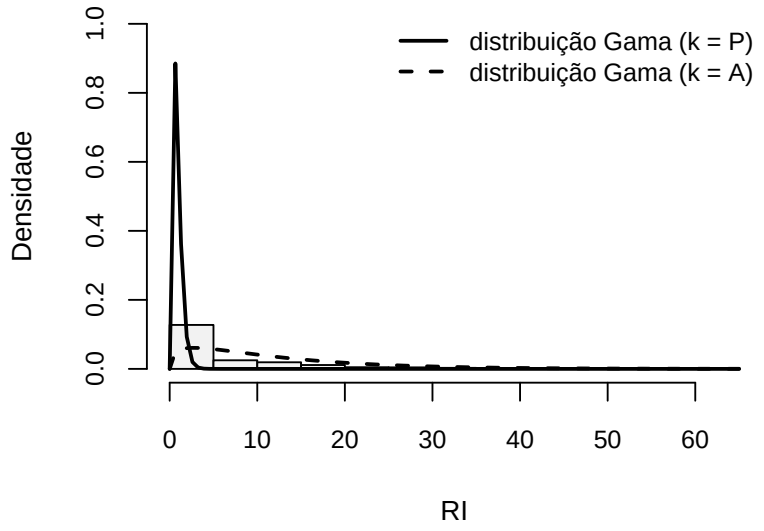

Figura 3.2: Histograma de RI, rácio das intensidades, com sobreposição das densidades das distribuições Gama.

3.1.2 Algoritmo EM

O algoritmo EM, ou algoritmo de Estimação-Maximização, é um método iterativo que permite localizar o máximo de uma função e é geralmente usado para maximizar a estimativa da verosimilhança quando a estrutura dos dados sugere a existência de dados omissos ou incompletos. Este método foi formalizado em Dempster et al. (1977), publicação em que também foi providenciada a prova de convergência do método. Na referência indicada é apresentada toda a teoria, aplicações e exemplos em que o algoritmo EM é uma ferramenta importante.

De seguida irá apresentar-se uma breve descrição teórica do algoritmo baseada em Dempster et al. (1977) e um exemplo da sua aplicação a modelos de mistura.

Considere-se que $X = (x_1, x_2, ..., x_n)$ representa o conjunto de dados observados de dimensão n e suponha-se que existe um conjunto finito de estados de dimensão R, sendo que a cada elemento x_i da amostra está associado um estado não observado. Desta forma existe um conjunto de dados $Z = (z_1, z_2, ..., z_n)$ não observados em que cada elemento z_i é um vector de dimensão R constituído por zeros excepto num componente, em que é 1, o que indica a existência de associação entre o r-ésimo estado e o elemento x_i observado. Considere-se ainda que $\boldsymbol{\theta}$ representa o vector de parâmetros

de interesse. Suponhamos que queremos localizar o máximo de uma função $L(\boldsymbol{\theta}|x)$. Sejam:

- $L(\boldsymbol{\theta}|x, Z)$ a função ampliada com dados não observados Z;
- $\boldsymbol{\theta}^k$ uma aproximação para os parâmetros do vector $\boldsymbol{\theta}$;
- $p_Z(z|\boldsymbol{\theta}^k, x)$ a distribuição de Z condicional a x calculada para $\boldsymbol{\theta} = \boldsymbol{\theta}^k$.

O algoritmo é iniciado com o vector $\boldsymbol{\theta}^k$ que contém as estimativas iniciais para os parâmetros que constituem o vector $\boldsymbol{\theta}$, e repetem-se as duas etapas seguintes até à convergência:

- **Passo de Estimação (Passo E)**:
  Consiste em calcular

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^k) = \int_Z \log[L(\boldsymbol{\theta}|x, Z)] p_Z(z|\boldsymbol{\theta}^k, x) dz$$

  isto é, o valor esperado de $\log[L(\boldsymbol{\theta}|x, Z)]$ com respeito à distribuição de Z condicional a $X = x$.

- **Passo de Maximização (Passo M)**:
  Consiste em maximizar a função $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^k)$ encontrada no passo E com respeito a $\boldsymbol{\theta}$, obtendo-se uma nova aproximação para $\boldsymbol{\theta}$, nomeadamente, $\boldsymbol{\theta}^{k+1}$.

- Repetem-se os passos E e M com esta nova aproximação.

O processo termina quando $||\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k||$, ou $|Q(\boldsymbol{\theta}^{k+1}, \boldsymbol{\theta}^k) - Q(\boldsymbol{\theta}^k, \boldsymbol{\theta}^k)|$, for suficientemente pequeno.

É importante referir que é possível provar que:

1. O algoritmo EM permite, em cada iteração, aumentar $L(\boldsymbol{\theta}|x)$, isto é, $L(\boldsymbol{\theta}^{k+1}|x) \geq L(\boldsymbol{\theta}^k|x)$, com igualdade se e só se $Q(\boldsymbol{\theta}^{k+1}, \boldsymbol{\theta}^k) = Q(\boldsymbol{\theta}^k, \boldsymbol{\theta}^k)$.

2. Se a sucessão de $\{\boldsymbol{\theta}^k\}$ convergir, então converge para um ponto de estacionaridade de $L(\boldsymbol{\theta}|x)$.

O algoritmo EM é um método comummente utilizado na estimação de modelos de mistura.

Um modelo de mistura permite modelar conjuntos de dados oriundos de uma população composta por várias subpopulações diferentes. Cada uma das subpopulações é modelada pela sua própria distribuição, mas não se sabe

a subpopulação a que cada elemento do conjunto de dados pertence. O algoritmo EM é a ferramenta usualmente utilizada para estimar os parâmetros do modelo e identificar quais os elementos do conjunto de dados que pertencem a cada subpopulação.

Para que se torne mais claro o funcionamento do algoritmo EM tome-se um exemplo particular em que todo o conjunto de dados, constituído pelos elementos $\{x_1, x_2, ..., x_n\}$, pode ser modelado por uma mistura de duas distribuições Normais, ou seja

$$\begin{aligned} f(x) &= a_0 h_0(x|\mu_0, \sigma) + a_1 h_1(x|\mu_1, \sigma) \\ &= a_0 h_0(x|\mu_0, \sigma) + (1 - a_0) h_1(x|\mu_1, \sigma) \end{aligned} \tag{3.2}$$

em que,

- a_0 é o coeficiente de mistura, $0 < a_1 < 1$;
- $h_0(x|\mu_0, \sigma)$ é a função de densidade de probabilidade de uma distribuição $N(\mu_0, \sigma)$;
- $h_1(x|\mu_1, \sigma)$ é a função de densidade de probabilidade de uma distribuição $N(\mu_1, \sigma)$;
- σ é constante e conhecido.

Neste caso o conjunto de dados omissos, Z, diz respeito a associação existente entre as duas distribuições Normais e os elementos x_i observados. Tem-se então que:

- $Z_i \sim Bernoulli(p), \qquad 0 < p < 1, \qquad i = 1, ..., n;$
- $X|_{Z_i=j} \sim N(\mu_j, \sigma), \qquad j = 0, 1, \qquad i = 1, ..., n.$

Se não houvessem dados omissos, o conjunto de dados era constituído pelos pares $(x_i, z_i), ..., (x_n, z_n)$.

No passo de Estimação (passo E) é calculado o valor esperado da log-verosimilhança completa com respeito à distribuição de Z condicional a $X =$

x e neste exemplo a log-verosimilhança completa é dada por:

$$\begin{aligned}
\log L(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{z}) &= \log f(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta}) = \log[\prod_{i=1}^{n} f(x_i,z_i|\boldsymbol{\theta})] \\
&= \sum_{i=1}^{n} \log[f(x_i|z_i,\boldsymbol{\theta}) f(z_i|\boldsymbol{\theta})] \\
&= \sum_{i=1}^{n} \log[h_0(x_i|\boldsymbol{\theta})^{(1-z_i)} h_1(x_i|\boldsymbol{\theta})^{z_i} (1-p)^{(1-z_i)} p^{z_i}] \\
&= \sum_{i=1}^{n} (1-z_i)\log h_0(x_i|\boldsymbol{\theta}) + \sum_{i=1}^{n} z_i \log h_1(x_i|\boldsymbol{\theta}) + \\
&+ \sum_{i=1}^{n} (1-z_i)\log(1-p) + \sum_{i=1}^{n} z_i \log p \qquad (3.3)
\end{aligned}$$

Como se pode ver, a log-verosimilhança completa é linear nos componentes z_i e, portanto, neste passo E é apenas necessário calcular as estimativas dos componentes z_i dado X e as correntes estimativas dos parâmetros. As estimativas dos componentes z_i dado X correspondem simplesmente às probabilidades condicionais de cada elemento do conjunto de dados, x_i, pertencer a cada um dos r-ésimos estados, ou seja, pertencer a cada uma das subpopulações Normais.

No passo de Maximização (passo M) calculam-se as novas estimativas para os parâmetros do modelo usando as probabilidades obtidas no passo E. Com estas novas estimativas volta-se novamente ao passo E para recalcular as probabilidades de pertença. Este processo é repetido até que não haja alterações nos parâmetros do modelo de mistura.

Dado que o algoritmo EM é um processo iterativo é necessário inicializar os parâmetros do modelo, considere-se então $\hat{\boldsymbol{\theta}}^{(0)} = \left[\hat{a}_0^{(0)}, \hat{\mu}_0^{(0)}, \hat{\mu}_1^{(0)}\right]$ o vector dos valores iniciais para os parâmetros do modelo. A k-ésima iteração do algoritmo EM é então obtida da seguinte forma:

- **Passo de Estimação (Passo E)**
  Para cada elemento i, $i = 1, ..., n$, do conjunto de dados a probabilidade de pertença à subpopulação $N(\mu_0, \sigma)$ é calculada da seguinte forma:

$$y_{0,i}^{(k)} = \frac{\hat{a}_0^{(k-1)} h(x_i|\hat{\mu}_0^{(k-1)},\sigma)}{\hat{a}_0^{(k-1)} h(x_i|\hat{\mu}_0^{(k-1)},\sigma) + (1-\hat{a}_0^{(k-1)}) h(x_i|\hat{\mu}_1^{(k-1)},\sigma)}.$$

  A probabilidade de pertença à subpopulação $N(\mu_1, \sigma)$ é $y_{1,i}^{(k)} = 1 - y_{0,i}^{(k)}$.

- **Passo de Maximização (Passo M)**
  As estimativas para o coeficiente de mistura e para os valores esperados μ_0 e μ_1 são obtidas da seguinte forma:

$$\begin{aligned}
\hat{a}_0^{(k)} &= \frac{\sum_{i=1}^{n} y_{1,i}^{(k)}}{n} \\
\hat{\mu}_0^{(k)} &= \frac{\sum_{i=1}^{n} x_i \cdot y_{0,i}^{(k)}}{\sum_{i=1}^{n} y_{0,i}^{(k)}} \\
\hat{\mu}_1^{(k)} &= \frac{\sum_{i=1}^{n} x_i \cdot y_{1,i}^{(k)}}{\sum_{i=1}^{n} y_{1,i}^{(k)}}
\end{aligned}$$

Retomando os modelos de mistura descritos em 3.1.1, apresenta-se de seguida a aplicação do algoritmo EM a estes modelos, com o objectivo de estimar os parâmetros desconhecidos associados a cada modelo de mistura.

Modelo de mistura Normal-Uniforme

No modelo de mistura Normal-Uniforme os parâmetros desconhecidos são μ, σ^2, a e b. Estes são estimados pelo algoritmo EM, o qual a cada iteração, no passo E, determina quais os genes que se classificam como presentes e ausentes e, no passo M, calcula as estimativas para os parâmetros do modelo usando a classificação do passo E. As estimativas dos parâmetros obtidas no passo M permitem reclassificar os genes no passo E da iteração seguinte. Este processo é repetido até que a diferença entre os parâmetros do modelo em duas iterações consecutivas seja tão pequena quanto se queira.

A j-ésima iteração do algoritmo EM é dada por:

- Passo de Estimação (Expectation step):
  Para cada gene i, $i = 1, ..., n$, a probabilidade de pertença ao grupo G_k, $k = A, P$ é calculada da seguinte forma:

$$\hat{z}_{k,i}^{(j)} = \frac{\hat{\pi}_k^{(j-1)} f(x_i|\hat{a}, \hat{b})}{\hat{\pi}_A^{(j-1)} f(x_i|\hat{a}, \hat{b}) + \hat{\pi}_P^{(j-1)} f(x_i|\hat{\mu}^{(j-1)}, \hat{\sigma}^{2(j-1)})},$$

  onde $\hat{a} = min\{x_i, i = 1, ..., n\}$ e $\hat{b} = max\{x_i, i = 1, ..., n\}$ são as estimativas de máxima verosimilhança de a e b, e $\hat{\pi}_P^{(j-1)} = 1 - \hat{\pi}_A^{(j-1)}$.
  Naturalmente tem-se que $\hat{z}_{A,i}^{(j)} = 1 - \hat{z}_{P,i}^{(j)}$.

- Passo de Maximização (Maximization step):

$$\begin{aligned}
\hat{\pi}_A^{(j)} &= \frac{\sum_{i=1}^n \hat{z}_{A,i}^{(j)}}{n} \quad \text{e} \quad \hat{\pi}_P^{(j)} = 1 - \hat{\pi}_A^{(j)} \\
\hat{\mu}^{(j)} &= \frac{\sum_{i=1}^n x_i \cdot \hat{z}_{P,i}^{(j)}}{\sum_{i=1}^n \hat{z}_{P,i}^{(j)}} \\
\hat{\sigma}^{2(j)} &= \frac{\sum_{i=1}^n (x_i - \hat{\mu}^{(j)})^2 \cdot \hat{z}_{P,i}^{(j)}}{\sum_{i=1}^n \hat{z}_{P,i}^{(j)}}
\end{aligned}$$

Na primeira iteração do algoritmo é necessário inicializar os valores de $\hat{z}_{k,i}^{(0)}, k = A, P$, o que foi feito da seguinte forma:

- $\hat{z}_{A,i}^{(0)} = 1$, se $|\frac{x_i - \bar{x}}{s}| > 2$, $i = 1, ..., n$, onde $\bar{x}$ e s são, respectivamente, a média e o desvio-padrão dos rácios das intensidades (gene ausente)[1];
- Caso contrário $\hat{z}_{A,i}^{(0)} = 0$, $i = 1, ..., n$ (gene presente);
- $\hat{z}_{P,i}^{(0)} = 1 - \hat{z}_{A,i}^{(0)}$.

Modelo de mistura Gama-Gama

No modelo de mistura Gama-Gama os parâmetros desconhecidos são α_k, β_k e π_k, em que $k = A, P$. Os parâmetros são estimados pelo algoritmo EM, o qual é inicializado com probabilidades de pertença calculadas de acordo com os quantis do rácio das intensidades considerando-se como genes ausentes aqueles que apresentarem valores mais elevados deste rácio.

A j-ésima iteração do algoritmo EM é dada por:

- Passo de Estimação (Expectation step):
  Para cada gene i, $i = 1, ..., n$, a probabilidade de pertença ao grupo G_k, $k = A, P$ é calculada da seguinte forma:

$$\hat{y}_{k,i}^{(j)} = \frac{\hat{\pi}_k^{(j-1)} f(x_i | \hat{\alpha}_k^{(j-1)}, \hat{\beta}_k^{(j-1)})}{\sum_{l=A,P} \hat{\pi}_l^{(j-1)} f(x_i | \hat{\alpha}_l^{(j-1)}, \hat{\beta}_l^{(j-1)})}.$$

  Tendo-se que $\hat{y}_{A,i}^{(j)} = 1 - \hat{y}_{P,i}^{(j)}$.

[1] No caso do presente trabalho seria mais adequado $\hat{z}_{A,i}^{(0)} = 1$, se $\frac{x_i - \bar{x}}{s} > 2$, no entanto o algoritmo EM é robusto e, portanto, converge.

- Passo de Maximização (Maximization step):
  Neste passo, $\hat{\alpha}_k^{(j)}$ e $\hat{\beta}_k^{(j)}$ são calculadas usando o método dos momentos, sendo necessário calcular os primeiros dois momentos empíricos para cada grupo, $m_{1,k}^{(j)}$ e $m_{2,k}^{(j)}$ respectivamente.

$$
\begin{aligned}
m_{1,k}^{(j)} &= \frac{\sum_{i=1}^{n} x_i \cdot \hat{y}_{k,i}^{(j)}}{\sum_{i=1}^{n} \hat{y}_{k,i}^{(j)}}, \quad m_{2,k}^{(j)} = \frac{\sum_{i=1}^{n} x_i^2 \cdot \hat{y}_{k,i}^{(j)}}{\sum_{i=1}^{n} \hat{y}_{k,i}^{(j)}}, \quad k = A, P. \\
\hat{\alpha}_k^{(j)} &= \frac{(m_{1,k}^{(j)})^2}{m_{2,k}^{(j)} - (m_{1,k}^{(j)})^2}, \quad k = A, P. \\
\hat{\beta}_k^{(j)} &= \frac{m_{1,k}^{(j)}}{m_{2,k}^{(j)} - (m_{1,k}^{(j)})^2}, \quad k = A, P. \\
\hat{\pi}_k^{(j)} &= \frac{\sum_{i=1}^{n} \hat{y}_{k,i}^{(j)}}{n}, \quad k = A, P.
\end{aligned}
$$

3.1.3 Classificador bayesiano

Inferência bayesiana

Segundo Paulino et al. (2003) "a semente para a abordagem bayesiana a problemas de inferência foi lançada por Richard Price quando em 1763 publicou a obra póstuma do Rev. Thomas Bayes intitulada - *An Essay Toward Solving a Problem in the Doctrine of Chances*".

A inferência bayesiana é um tipo de inferência estatística, na qual todas as formas de incerteza são expressas em termos de probabilidade. De acordo com a filosofia bayesiana, o parâmetro de interesse θ (ou o vector dos parâmetros de interesse $\boldsymbol{\theta}$) é tomado como uma variável aleatória (ou vector aleatório), enquanto que, do ponto de vista do modelo clássico, o parâmetro de interesse θ (ou o vector dos parâmetros de interesse $\boldsymbol{\theta}$) é desconhecido mas fixo.

A abordagem bayesiana para um problema de inferência inicia-se com a formulação de um modelo que se espera ser adequado para descrever uma dada situação de interesse. De seguida, a informação inicial (ou *a priori*) que se tem do parâmetro de interesse θ é traduzida formalmente através de uma distribuição de probabilidade, geralmente subjectiva, para θ designada por distribuição *a priori*, seja $h(\theta)$ (quando se tem um vector aleatório de parâmetros de interesse, é formalizada uma distribuição *a priori* para cada parâmetro). A informação *a priori* inclui juízos ou experiências individuais

que representam o ponto de vista e os conhecimentos de um dado investigador para o problema em estudo (de acordo com um princípio de coerência ou consistência), antes da amostra ser observada. Na posse da informação amostral, ou seja, após a amostra ser observada (suponha-se que se observa $X = x$), é aplicado o Teorema de Bayes para densidades com o objectivo de se obter a distribuição *a posteriori* para θ, seja $h(\theta|x)$, a qual tem em conta a distribuição *a priori* e os dados x.

$$h(\theta|x) \quad = \quad \frac{f(x|\theta)h(\theta)}{f(x)} = \frac{f(x|\theta)h(\theta)}{\int_\Theta f(x|\theta)h(\theta)d\theta}, \quad \theta \in \Theta \tag{3.4}$$

Veja-se que a posição inicial do investigador, descrita por $h(\theta)$, modificou-se com a informação amostral contida nos dados x, passando a nova posição a traduzir-se por $h(\theta|x)$.

Como na expressão da distribuição *a posteriori* (3.4), o denominador $f(x)$ é a designada distribuição preditiva[2], a qual não depende de parâmetro θ, podemos reescrevê-la:

$$h(\theta|x) \propto f(x|\theta)h(\theta) \tag{3.5}$$

Quando encarado como função de θ dado o valor de x observado (fixo), o factor $f(x|\theta)$ corresponde à função de verosimilhança e, portanto, pode-se escrever de forma simbólica que:

distribuição a posteriori $\propto$ *verosimilhança x distribuição a priori*

A função de verosimilhança incorpora a informação amostral e pode ser vista como o meio através do qual os dados, x, transformam o conhecimento a priori sobre θ.

Em resumo, segundo Paulino et al. (2003), "a distribuição a *posteriori* constitui um elemento fundamental que serve de base ao desenvolvimento de toda a inferência bayesiana", pois "incorpora, por via do Teorema da Bayes, toda a informação disponível sobre o parâmetro (informação inicial + informação da experiência ou da amostra)".

[2] A distribuição preditiva permite predizer observações futuras e pode ser calculada através das distribuições *a posteriori*.

Classificador bayesiano

Fazendo uso da estatística bayesiana é possível seguir uma abordagem diferente para obter a classificação dos genes (Antunes e Sousa, 2008). O classificador bayesiano pressupõe a existência de uma variável aleatória, X, a qual assume um comportamento diferente em cada grupo em que a população se subdivide. Neste contexto, X representa o rácio das intensidades e a população encontra-se dividida em dois grupos: grupo dos genes ausentes e grupo dos genes presentes.

Este método é inicializado com uma pré-classificação dos genes baseada em conhecimento a priori, a qual é representada numa variável categórica, T, que assume os valores 0 (grupo dos genes ausentes) e 1 (grupo dos genes presentes). Desta forma, sob um contexto de classificação supervisionada, haverá idealmente um conjunto de treino, ou seja, um conjunto de genes para os quais se sabe a classificação, que vai servir para calcular as probabilidades preditivas condicionais. É com base nesta distribuição preditiva condicional que é calculada a regra de classificação a qual vai ser aplicada aos restantes genes da amostra.

Considere-se então o conjunto de dados $D = \{(x_1, t_1), (x_2, t_2), ..., (x_n, t_n)\}$ em que x_i representa o rácio das intensidades para o i-ésimo gene e t_i a realização da variável categórica T. As distribuições consideradas foram as seguintes:

- $T \sim Bernoulli(\pi)$, em que $\pi = P[T = 1]$ e $1 - \pi = P[T = 0]$;
- $X|_{T=j} \sim Gama(\alpha_j, \beta_j)$, em que $j = 0, 1$.

As distribuições a *priori* são:

- $\pi \sim Beta(b_0, b_1)$;
- $\beta_j \sim Gama(g, h)$.

O parâmetro α_j é considerado conhecido (sendo na prática estimado a partir dos dados e introduzido no modelo de forma a permitir uma forma fechada para as distribuições preditivas) e os hiperparâmetros do modelo são b_0, b_1, g e h.

As distribuições a *posteriori* são:[3]

- $\pi|_D \sim Beta(n_0 + b_0, n_1 + b_1)$;
- $\beta_j|_D \sim Gama(G_j, H_j)$, em que $G_j = n_j\alpha_j + g$ e $H_j = \sum_{k=1}^{n_j} x_{jk} + h$.

[3] As distribuições a *posteriori* bem como a distribuição preditiva condicional de T foram calculadas em Antunes e Sousa (2008).

Distribuição preditiva condicional de T

A distribuição preditiva condicional de T dado x é dada por:[3]

$$P[T=j|D,x] = \frac{p[x|D,T=j]P[T=j|D]}{p[x|D]} \tag{3.6}$$

em que cada uma das probabilidades corresponde a:

$$\begin{aligned} p[x|D,T=j] &= \int_0^\infty p[x|D,T=j,\alpha_j,\beta_j]p[\beta_j|D]d\beta_j \\ &= \frac{\Gamma(\alpha_j+G_j)}{\Gamma(\alpha_j)\Gamma(G_j)}\frac{x^{\alpha_j-1}{H_j}^{G_j}}{(x+H_j)^{\alpha_j+G_j}} \qquad (3.7) \\ P[T=j|D] &= \int_0^1 P[T=j]p[\pi|D]d\pi \\ &= \frac{n_j+b_j}{n+b_0+b_1} \qquad (3.8) \\ p[x|D] &= \sum_{j=0}^{1} p[x|D,T=j]P[T=j|D] \qquad (3.9) \end{aligned}$$

As funções de probabilidade preditiva condicional (3.6) mostram como a probabilidade de um gene, para o qual $X = x$, pertencer ao grupo j evolui em função de x. Estas funções podem ser usadas para classificar os genes, sendo que um determinado gene com rácio das intensidades igual a x pertencerá ao grupo j se:

$$j = \arg\max_j \{P[T=j|D,x], j=0,1\}. \tag{3.10}$$

O ponto de corte que irá permitir classificar os genes em dois grupos distintos obtém-se resolvendo a equação:

$$\begin{aligned} P[T=0|D,x] &= P[T=1|D,x] \Leftrightarrow \\ &\Leftrightarrow P[T=1|D,x] - P[T=0|D,x] = 0. \end{aligned} \tag{3.11}$$

Seja r a solução da equação anterior, os genes serão atribuídos a cada grupo da seguinte forma:

- Se $x_i > r$ o gene i é ausente;

- Se $x_i \leq r$ o gene i é presente.

[3] As distribuições a *posteriori* bem como a distribuição preditiva condicional de T foram calculadas em Antunes e Sousa (2008).

Como já foi referido, o classificador bayesiano assenta no pressuposto de que dos n genes que constituem a amostra se conhece, de alguma forma, a classificação de m genes (conjunto de treino), ou seja, para esses m genes sabemos se pertencem ao grupo dos genes ausentes ou presentes enquanto que para os restantes $n - m$ genes não temos essa informação.

Assim na aplicação do classificador bayesiano há que considerar dois cenários distintos relativamente ao conhecimento da classificação dos genes do conjunto de treino:

- quando advém de alguma informação biológica (classificação supervisionada);
- quando advém de alguma pré-classificação, por exemplo a obtida pelo algoritmo EM ou outra informação que não a classificação exacta (classificação não supervisionada).

É com esta informação dos m genes que o classificador vai calcular as probabilidades preditivas de T dado x as quais permitirão obter uma regra de classificação. A regra de classificação obtida é depois utilizada para classificar os $n - m$ genes restantes. A dimensão do conjunto de treino (m genes) pode variar e em 3.2 é feito um estudo do efeito da variação do valor de m na qualidade do desempenho do classificador.

3.2 Aplicação prática e resultados

A aplicação dos métodos e modelos teóricos (bem como a análise exploratória dos dados) foi feita no *software R*, versão 2.8.1, fazendo uso de algumas bibliotecas, de funções já existentes e de outras criadas de raíz consoante as necessidades encontradas. No apêndice B estão as rotinas criadas no *software R* que foram utilizadas ao longo do trabalho.

Numa primeira abordagem ao problema em estudo aplicou-se o algoritmo EM com o modelo de mistura Normal-Uniforme, disponível na biblioteca NUDGE[4] (neste caso, através da função *nudge1*), obtendo-se como resultados as probabilidades de pertença a cada grupo para cada gene. Cada gene foi classificado no grupo que apresentou uma probabilidade maior e esta classificação foi comparada com a classificação real. Os resultados desta comparação e as estimativas dos parâmetros do modelo podem ser observadas na Tabela 3.1.

O modelo que está disponível na biblioteca NUDGE tem por base dados logaritmizados e *normalizados* (*LR norm.*). Por *normalização* entende-se neste contexto um processo de dois passos descrito cuidadosamente em Dean e Raftery (2005). Resumidamente, este processo consiste num primeiro passo em estimar o desvio médio absoluto (*running mean absolute deviation*) através de uma regressão *loess*[5] do valor absoluto do logaritmo do rácio normalizado pela média sobre o logaritmo do produto das intensidades, $log(R*G)$, e depois num segundo passo em calcular as estimativas para as intensidades dividindo o logaritmo do rácio normalizado pela média pelo desvio médio absoluto estimado na regressão do primeiro passo. Segundo a referência indicada, este processo torna o LR dos genes diferencialmente expressos aproximadamente normal e homocedástico.

Este modelo foi ligeiramente modificado de forma a poderem ser utilizados dados não *normalizados* tendo sido utilizada como medida de intensidade o rácio das intensidades (RI) em vez do logaritmo, sendo agora os resultados melhores do que os obtidos pelo modelo anterior (tabela 3.1).

Na Tabela 3.1 apresentam-se também os resultados do modelo de mistura Normal-Uniforme com a utilização das medidas de intensidade LRI

[4]NUDGE (*Normal Uniform Differential Gene Expression*) é uma biblioteca construída a fim de aplicar o modelo de detecção de genes diferencialmente expressos desenvolvido em Dean e Raftery (2005).

[5]Método de regressão (não linear) local que consiste numa generalização do método de médias móveis. O valor predito para cada observação é obtido através do ajuste de uma regressão linear local ponderada, em que os pesos são tanto menores quanto maior for a distância à observação de interesse. Com a ligação dos valores preditos é possível obter uma curva suave, cujo grau de suavização é definido por um parâmetro (geralmente denominado de *bandwidth*) que determina a fracção dos dados a ser utilizada para o cálculo da regressão local. O objectivo passa por gerar uma estimativa o mais suave possível sem, no entanto, distorcer a relação de dependência entre as variáveis em análise.

(referida em 2.1) e o inverso de RI, $R = \frac{intensidade\ do\ teste}{intensidade\ do\ controlo}$.

Há ainda a possibilidade de utilizar as intensidades normalizadas (*processed*) em vez das intensidades médias no cálculo da medida de intensidade. Como o modelo com a medida RI obteve melhores resultados, aplicou-se este modelo novamente, mas agora com estas intensidades normalizadas (linha RI norm., Tabela 3.1) para ver se se verificavam algumas melhorias, no entanto os resultados foram semelhantes.

Modelo	Desempenho do modelo					Estimativas dos parâmetros			
aplicado	VP[1]	VN[2]	FP[3]	FN[4]	Acc	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{a}$	$\hat{b}$
LR norm.	2309	238	535	55	0,812	-0,22	0,86	-3,03	5,70
RI	2240	491	282	124	0,871	0,86	0,29	0,25	61,49
RI norm.	2268	462	311	96	0,870	0,85	0,34	0,16	124,14
LRI	1941	546	227	423	0,793	1,85	0,23	0,60	6,06
R	2128	1	772	236	0,679	1,12	0,51	0,02	4,64

Tabela 3.1: Desempenho e estimativas dos parâmetros para o modelo de mistura Normal-Uniforme

Numa segunda abordagem foi aplicado o algoritmo EM com o modelo de mistura Gama-Gama, o qual foi programado de forma semelhante ao algoritmo EM com o modelo Normal-Uniforme disponível na biblioteca NUDGE, mas sem efectuar qualquer *normalização* nos dados. Os passos de Estimação e de Maximização do algoritmo EM são descritos em 3.1.2 na parte referente ao modelo de mistura Gama-Gama. O algoritmo EM mostrou-se consistente relativamente à escolha do quantil utilizado para inicializar as probabilidades de pertença, contudo, utilizou-se o quantil 0.5 pois este é o menos informativo e a escolha que se aproxima mais da realidade, em que pouco ou nada se conhece sobre a classificação dos genes.

O desempenho e as estimativas dos parâmetros deste modelo apresentam-se na Tabela 3.2.

Alternativamente à utilização do algoritmo EM foi usado o classificador bayesiano. O classificador foi construído num contexto de classificação supervisionada, em que se pressupõe o conhecimento da classificação de alguns genes (conjunto de treino de dimensão m), sendo este conhecimento utilizado para a construção do classificador.

[1]Verdadeiros Positivos: Genes presentes bem classificados
[2]Verdadeiros Negativos: Genes ausentes bem classificados
[3]Falsos Positivos: Genes presentes mal classificados
[4]Falsos Negativos: Genes ausentes mal classificados

Modelo aplicado	Desempenho do modelo					Estimativas dos parâmetros			
	VP[1]	VN[2]	FP[3]	FN[4]	*Acc*	$\hat{\alpha}_P$	$\hat{\beta}_P$	$\hat{\alpha}_A$	$\hat{\beta}_A$
RI	2225	506	267	139	0,871	9,85	11,43	0,80	0,09
RI norm.	2256	470	303	108	0,869	6,76	7,99	0,65	0,05
LRI	1622	607	166	742	0,711	125,83	63,36	4,34	1,72
R	928	137	636	1436	0,340	1,59	1,35	47,84	35,89

Tabela 3.2: Desempenho e estimativas dos parâmetros para o modelo de mistura Gama-Gama

Para se avaliar a consistência e o desempenho do classificador bayesiano geraram-se 1000 reamostragens de valores x, rácio das intensidades, de dimensão m e para cada reamostragem:

- calculou-se o ponto de corte (daqui resultando 1000 pontos de corte);
- com o ponto de corte classificaram-se os $n - m$ genes e calculou-se a exactidão com estes genes;
- a regra foi também aplicada a todos os genes (incluindo os m genes do conjunto de treino) e calculou-se a exactidão com todos os genes.

O procedimento acima descrito foi aplicado para diferentes valores de m e para os dois cenários diferentes em relação ao conhecimento dos m genes descritos em 3.1.3:

- **Classificação supervisionada** - quando o conhecimento da classificação dos genes do conjunto de treino advém de alguma informação biológica.

  Neste caso, a informação dos m genes é a classificação real dos genes, ou seja aquela que está parcialmente exposta na Tabela 2.3. Ter o conhecimento desta classificação é uma situação plausível uma vez que se pode ter acesso à classificação real dos genes através de técnicas moleculares, como o PCR (*Polymerase Chain Reaction*), ou recorrendo a estudos anteriores na literatura.
  O desempenho do classificador encontra-se ilustrado nos gráficos da Figura 3.3. No primeiro gráfico podemos ver a evolução do ponto de corte consoante a dimensão do conjunto de treino, o qual acaba por estabilizar por volta de 4.4 ou 4.5 do valor do rácio, a partir de

[1]Verdadeiros Positivos: Genes presentes bem classificados
[2]Verdadeiros Negativos: Genes ausentes bem classificados
[3]Falsos Positivos: Genes presentes mal classificados
[4]Falsos Negativos: Genes ausentes mal classificados

um conjunto de treino entre 700 a 800 genes. No segundo e terceiro gráficos vê-se a evolução da exactidão calculada, respectivamente, com os genes fora do conjunto de treino e com todos os genes. Acerca da exactidão pode-se dizer que ela é muito estável rondando, em média, valores muito satisfatórios entre 0.85 e 0.87.
Note-se que para conjuntos de treino de pequena dimensão a exactidão é, em média (ou mediana), mais elevada mas a dispersão é também maior.

- **Classificação não supervisionada** - quando se desconhece a classificação verdadeira dos genes do conjunto de treino sendo esta substituída por outra informação (experiências anteriores semelhantes ou resultado de um procedimento de classificação, como por exemplo, o algoritmo EM).

  Neste cenário não é conhecida a classificação real dos genes e, portanto, a informação dos m genes pode ser a classificação que resulte da aplicação do modelo de mistura Normal-Uniforme ou do modelo Gama-Gama via algoritmo EM. De forma análoga ao cenário de classificação supervisionada efectuaram-se os mesmos gráficos para estes dois casos (ver figuras 3.4 e 3.5).
  No caso em que a pré-classificação é obtida pelo algoritmo EM usando o modelo Normal-Uniforme podemos observar a partir da Figura 3.4 que o ponto de corte varia muito pouco consoante a dimensão do conjunto de treino, principalmente a partir de 400 genes, atingindo um valor de aproximadamente 1.88 referente ao rácio das intensidades. Este modelo apresenta bons resultados ao nível da exactidão sobretudo porque esta é extremamente estável rondando, em média, valores entre 0.87 e 0.875.
  Quando é utilizado o modelo Gama-Gama, o ponto de corte evolui também de forma estável e atinge 1.8 do valor do rácio a partir de 400 genes. O comportamento da exactidão à medida que a dimensão do conjunto de treino aumenta revela um bom desempenho do modelo. A exactidão é, em termos médios, estável e alcança valores entre 0.87 e 0.874, a um nível muito semelhante ao do modelo Normal-Uniforme. Note-se que, de forma idêntica ao cenário da classificação supervisionada, também neste cenário a exactidão é, em média (ou mediana), mais elevada para conjuntos de treino de pequena dimensão, mas a dispersão é também maior.

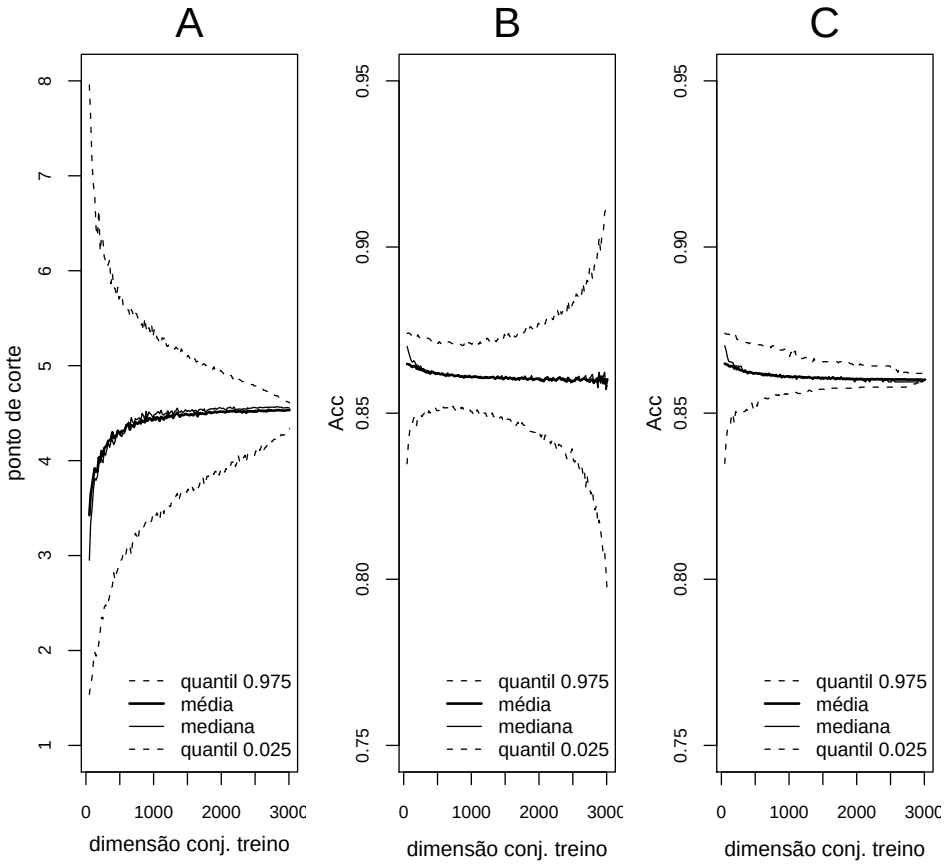

Figura 3.3: Ponto de corte e desempenho do classificador bayesiano consoante a dimensão do conjunto de treino: cenário referente à classificação supervisionada. Em A podemos ver a evolução do ponto de corte calculado pelo classificador, em B está registada a exactidão calculada sem ter em conta a informação dos genes do conjunto de treino e em C regista-se a exactidão calculada com base em todos os genes incluindo os do conjunto de treino. Nos três gráficos é apresentado um sumário de estatísticas (quantil 0.025, média, mediana, quantil 0.975) do ponto de corte e das exactidões das 1000 reamostragens do rácio das intensidades efectuadas, para cada dimensão do conjunto de treino considerada.

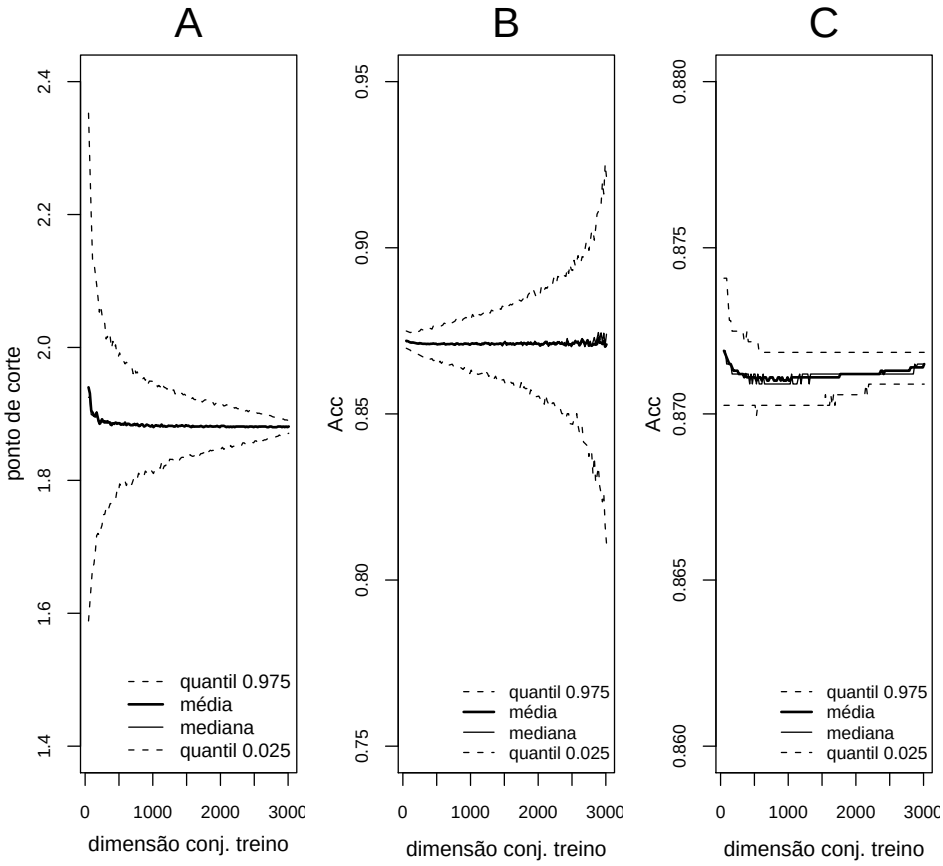

Figura 3.4: Ponto de corte e desempenho do classificador bayesiano consoante a dimensão do conjunto de treino: cenário referente à classificação não supervisionada (modelo de mistura Normal-Uniforme). Em A podemos ver a evolução do ponto de corte calculado pelo classificador, em B está registada a exactidão calculada sem ter em conta a informação dos genes do conjunto de treino e em C regista-se a exactidão calculada com base em todos os genes incluindo os do conjunto de treino. Nos três gráficos é apresentado um sumário de estatísticas (quantil 0.025, média, mediana, quantil 0.975) do ponto de corte e das exactidões das 1000 reamostragens do rácio das intensidades efectuadas, para cada dimensão do conjunto de treino considerada.

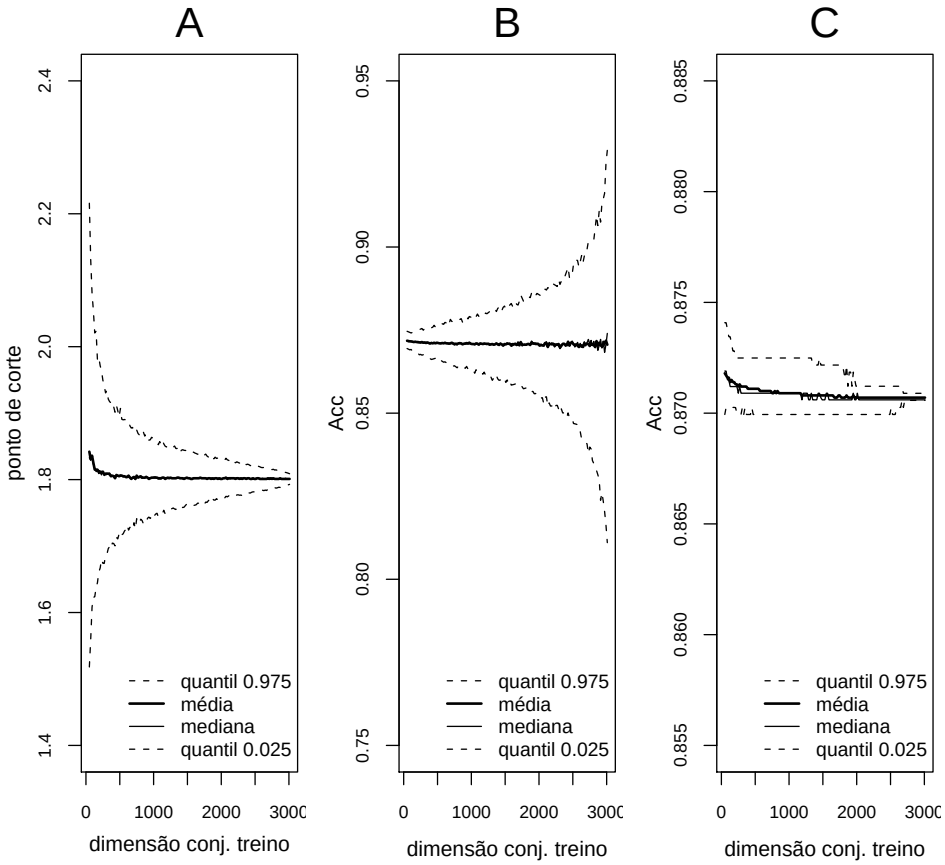

Figura 3.5: Ponto de corte e desempenho do classificador bayesiano consoante a dimensão do conjunto de treino: cenário referente à classificação não supervisionada (modelo de mistura Gama-Gama). Em A podemos ver a evolução do ponto de corte calculado pelo classificador, em B está registada a exactidão calculada sem ter em conta a informação dos genes do conjunto de treino e em C regista-se a exactidão calculada com base em todos os genes incluindo os do conjunto de treino. Nos três gráficos é apresentado um sumário de estatísticas (quantil 0.025, média, mediana, quantil 0.975) do ponto de corte e das exactidões das 1000 reamostragens do rácio das intensidades efectuadas, para cada dimensão do conjunto de treino considerada.

Se analisarmos o caso extremo em que o conjunto de treino é constituído por todos os genes podemos construir, à semelhança do que se fez para o algoritmo EM, a Tabela 3.3 a fim de se poder comparar o desempenho dos diferentes cenários envolvidos.

Classificação genes do conj. de treino	Ponto de corte	Desempenho do classificador				
		VP[1]	VN[2]	FP[3]	FN[4]	*Acc*
Informação biológica	4,5357	2324	374	399	40	0,8601
Normal-Uniforme	1,8805	2243	491	282	121	0,8715
Gama-Gama	1,8013	2236	496	277	128	0,8709

Tabela 3.3: Ponto de corte e desempenho do classificador bayesiano para diferentes cenários da classificação do conjunto de treino de dimensão máxima

Esta tabela é comparável com as tabelas 3.1 e 3.2 e permite confrontar os resultados obtidos pelo classificador bayesiano com os obtidos pelo algoritmo EM. A comparação é feita essencialmente ao nível da exactidão e um resumo dos seus valores para o caso extremo pode ser visto na tabela que se segue, com a apresentação dos métodos, algoritmo EM e classificador bayesiano, e dos modelos, Normal-Uniforme (NU) e Gama-Gama (GG) aplicados.

	Métodos				
	Algoritmo EM		Classificador bayesiano	Algoritmo EM e Classif. bayesiano	
Modelos ajustados	NU	GG		NU	GG
	0.8706	0.8706	0.8601	0.8715	0.8709

Tabela 3.4: Tabela-resumo com a exactidão das metodologias aplicadas, utilizando o rácio *RI*

Através da Tabela 3.4 pode-se observar, relativamente às metodologias, que com a utilização da combinação do algoritmo EM com o classificador bayesiano se consegue atingir melhores exactidões do que com a utilização isolada do algoritmo EM e do classificador bayesiano (classificação supervisionada). Em relação aos modelos aplicados, estes apresentam valores para a exactidão muito semelhantes, levando o modelo de mistura Normal-Uniforme uma pequena vantagem, observada com a combinação do algoritmo EM e o classificador bayesiano.

[1]Verdadeiros Positivos: Genes presentes bem classificados
[2]Verdadeiros Negativos: Genes ausentes bem classificados
[3]Falsos Positivos: Genes presentes mal classificados
[4]Falsos Negativos: Genes ausentes mal classificados

3.3 Algumas conclusões e considerações

No que se refere às metodologias aplicadas podemos dizer relativamente ao algoritmo EM usado na estimação dos modelos de mistura Normal-Uniforme e Gama-Gama que a utilização do rácio RI, como medida de intensidade, é a escolha que produz melhores resultados em termos de exactidão.

Por um lado, o modelo de mistura Normal-Uniforme proposto em Dean e Raftery (2005) [*LR norm.*] não é, neste caso, o mais adequado, pois apresenta uma exactidão mais baixa, o que indica que a transformação logarítimica não não se adequa aos dados em questão. Aliás, o mesmo se pode concluir se compararmos os resultados da medida RI com a medida LRI.

Por outro lado, a utilização das intensidades normalizadas no cálculo de RI, ou seja, a utilização de *RI norm.* também não introduziu melhorias em termos da exactidão da classificação e, portanto, não se recomenda a sua utilização pois as transformações a que são sujeitas as intensidades podem introduzir muito ruído nos dados.

Usando os mesmos critérios de análise, o rácio R (inverso de RI) também não é a melhor escolha e a sua utilização teve apenas como objectivo mostrar que o rácio RI é mais indicado. Aliás, os níveis de exactidão de R são os mais baixos de todas as medidas de intensidade experimentadas.

Assim, o nível máximo de exactidão atingido com a utilização do algoritmo EM e dos modelos de mistura foi de 0.871, alcançado quer pelo modelo Normal-Uniforme quer pelo modelo Gama-Gama, com o uso da medida de intensidade RI, o que significa que estes modelos conseguem classificar de forma correcta 87,1% dos genes que constituem a amostra.

Relativamente ao classificador bayesiano podemos concluir que quando este é usado sozinho, ou seja, quando se está perante um cenário em que o conhecimento da classificação dos genes do conjunto de treino advém de alguma informação biológica, ele consegue classificar correctamente cerca de 86% dos genes. Este resultado é ligeiramente pior que os resultados que se obtêm se o classificador for usado em conjunto com o algoritmo EM. Estes dois métodos são usados em conjunto quando não se tem informação biológica sobre a classificação dos genes e assim é utilizado o algoritmo EM a fim de se produzir uma pré-classificação dos genes do conjunto de treino, classificação esta que é depois usada no classificador bayesiano. Neste cenário é possível obter duas pré-classificações diferentes: uma que advém da aplicação do modelo de mistura Normal-Uniforme e outra que advém do modelo Gama-Gama, apresentando ambas níveis muito semelhantes de exactidão (aproximadamente 0.87).

É também de notar que o número de genes do conjunto de treino para o qual a exactidão estabiliza é muito elevado, inviabilizando o uso de outras

técnicas moleculares (por exemplo, PCR) com o objectivo de melhorar a classificação.

Se compararmos as diferentes metodologias aplicadas, quando se utiliza RI (ver Tabela 3.4), podemos observar que a combinação do algoritmo EM com o classificador bayesiano é a metodologia que apresenta melhores resultados em termos da exactidão. Embora os níveis de exactidão obtidos pelos métodos não sejam muito díspares, há que ter em atenção que o resultado gerado por cada metodologia é diferente. O algoritmo EM gera probabilidades de cada gene pertencer à subpopulação dos genes presentes e à subpopulação dos genes ausentes enquanto que o classificador bayesiano gera uma regra de classificação, ou seja, encontra um ponto de corte, em termos de RI, acima do qual os genes se classificam como ausentes e abaixo como presentes.

CAPÍTULO 4

Classificação de genes em CGH: uma correcção para controlo de estirpes múltiplas

À medida que as sequências genómicas de múltiplas estirpes da mesma espécie bacteriana se tornam conhecidas, torna-se possível construir *microarrays* com base em mais do que uma estirpe. É, por exemplo, o caso da bactéria *Streptococcus pneumoniae* que tem já três estirpes sequenciadas: R6, G54 e TIGR4.

Nas experiências que têm vindo a ser analisadas ao longo do trabalho, os *microarrays* são construídos com base nestas três estirpes e a amostra de controlo pode ser constituída por uma mistura de ADN das mesmas três estirpes. Sendo assim os *microarrays*, onde são hibridadas as amostras (amostra de controlo e de teste), contêm sequências que identificam genes presentes numa estirpe (grupo 1), em duas (grupo 2) ou nas três estirpes (grupo 3) cujo genoma está sequenciado. Desta forma, é de esperar que a intensidade de fluorescência de um determinado gene fique condicionada pelo número de estirpes em que o gene está presente.

Em Pinto et al. (2008) é feito um estudo com base nas mesmas experiências CGH que têm vindo a ser analisadas e, seguindo a mesma linha de pensamento dessa publicação, é de esperar que a intensidade do gene na amostra de teste seja proporcional à intensidade desse mesmo gene na amostra de controlo, variando a razão de proporcionalidade consoante o número de estirpes em que o gene está presente. Segundo a publicação referida espera-se que o rácio das intensidades, $RI = \frac{intensidade\ do\ controlo}{intensidade\ do\ teste}$, seja conforme a seguir se apresenta, consoante um determinado gene esteja presente ou ausente da amostra de teste:

- Quando o **gene i está presente na amostra de teste** espera-se que:

$$RI_i \approx \begin{cases} 1/3, & \text{gene i presente numa estirpe (grupo 1)} \\ 2/3, & \text{gene i presente em duas estirpes (grupo 2)} \\ 3/3, & \text{gene i presente nas três estirpes (grupo 3)} \end{cases}$$

- Quando o **gene i está ausente da amostra de teste** espera-se que RI assuma valores muito elevados, sendo tanto maior quantas mais forem as estirpes sequenciadas em que o gene está presente.

Perante isto, duas abordagens se podem tomar: analisar cada classe de genes (ou seja, os grupos 1, 2 e 3) em separado ou analisar todos os genes em conjunto mas tendo em conta que se espera que RI assuma valores diferentes para cada grupo de genes. Uma vez que, ao longo deste trabalho se têm analisado todos os genes em conjunto, também aqui se irá optar por essa via, para que depois se possam comparar os resultados que se vão obter com aqueles que já foram obtidos.

4.1 Estratégias adoptadas

Dado que se optou por analisar todos os genes em conjunto vamos, numa primeira estratégia, tranformar a variável RI multiplicando-a pelos factores, 3, 3/2 ou 1, consonte cada gene i esteja presente, respectivamente, numa estirpe, em duas estirpes ou nas três estirpes sequenciadas, obtendo-se a nova variável, $RI1$, correspondente ao rácio das intensidades transformadas. De seguida aplicaram-se os métodos, algoritmo EM e classificador bayesiano, descritos em 3.1, com o objectivo de se verificar se com este novo rácio a exactidão regista alterações. Os resultados obtidos constam da Tabela 4.2.

Numa segunda estratégia, e para se tentar melhorar as exactidões obtidas com $RI1$, procurou-se estimar a razão de proporcionalidade entre as intensidades da amostra de teste e de controlo e assim foi ajustado, a cada grupo de genes, o seguinte modelo de regressão linear simples:

$$intens.teste_{ik} = \beta_k \times intens.controlo_{ik}, \quad \text{em que}$$

$$i = 1, ..., n \quad (i\text{-ésimo gene}) \quad \text{e} \quad k = 1, 2, 3 \quad (k\text{-ésimo grupo})$$

As estimativas de $\hat{\beta}_k$, ou seja, $\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\beta}_3$, correspondem às estimativas dos factores de interesse que vão multiplicar, respectivamente, pelo rácio das intensidades dos genes do grupo 1, do grupo 2 e do grupo 3, obtendo-se depois a nova variável $RI2$, a qual vai ser também sujeita à aplicação dos métodos já referidos.

Tentou-se ainda uma regressão linear diferente recorrendo a uma variável adicional designada por *identidade*. Esta variável tem em conta a identidade da sequência existente no genoma de cada estirpe que constitui a amostra de teste ou de controlo com a sequência exacta que está na base do *microarray*. Se a sequência que existe numa estirpe da amostra de teste ou de controlo é exactamente igual à sequência existente no *microarray*, então a identidade é 70, pois o *microarray* foi construído com oligonucleótidos de 70 nucleótidos, se a sequência que existe na estirpe não for exactamente igual à que está no *microarray*, então a identidade é inferior a 70. A variável *identidade* incluída na regressão é obtida calculando a média das identidades entre a sequência presente no *microarray* e cada uma das três estirpes sequenciadas R6, G54 e TIGR4, uma vez que o controlo é composto por uma mistura com partes iguais de ADN destas três estirpes. A regressão linear foi construída de forma a que o rácio das intensidades, RI, pudesse ser explicado através desta variável *identidade*, sendo ajustada uma regressão a cada grupo de genes, ou seja:

$$RI_{ik} = \delta_k \times identidade_{ik}, \quad \text{em que}$$

$$i = 1, ..., n \quad (i\text{-ésimo gene}) \quad \text{e} \quad k = 1, 2, 3 \quad (k\text{-ésimo grupo})$$

Da mesma forma que anteriormente, as estimativas $\hat{\delta}_1$, $\hat{\delta}_2$ e $\hat{\delta}_3$ das razões de proporcionalidade vão multiplicar, respectivamente, pelo rácio das intensidades dos genes do grupo 1, do grupo 2 e do grupo 3, obtendo-se a nova variável $RI3$, a qual será também sujeita à aplicação dos métodos habituais.

Antes de avançar, uma última palavra para o facto de nas regressões apresentadas não se utilizar o termo constante. A escolha da não utilização da constante pode justificar-se pelo facto de estarmos apenas interessados em estimar uma razão de proporcionalidade entre as intensidades das amostras de teste e controlo. Contudo, a constante foi utilizada inicialmente mas nem sempre as variáveis explicativas (*intens.controlo* ou *identidade*) das regressões lineares eram estatisticamente significativas assim, face aos resultados, optou-se por utilizar regressões sem constante.

4.2 Resultados e comentários

As estimativas dos parâmetros dos dois conjuntos de modelos de regressão linear simples ajustados apresentam-se na Tabela 4.1 e os resultados obtidos, em termos de exactidão, da aplicação das metodologias referidas, algoritmo EM e classificador bayesiano, constam da Tabela 4.2.

No que diz respeito às estimativas dos parâmetros utilizados nos modelos de regressão linear pode-se dizer relativamente às estimativas dos parâmetros $\hat{\beta}_k$ que, apesar de diferentes, não estão muito longe dos valores 3, 3/2 e 1, inicialmente propostos e relativamente às estimativas dos parâmetros $\hat{\delta}_k$ que,

embora sejam pequenas, o efeito da variável *identidade* no rácio RI é estatisticamente significativo. As estimativas dos parâmetros $\hat{\delta}_k$ são pequenas porque a variável *identidade* toma valores mais elevados do que RI.

Estimativas dos parâmetros	Classes de genes		
	grupo 1	grupo 2	grupo 3
$\hat{\beta}_{ik}$	2.2359	1.4279	1.2080
$\hat{\delta}_{ik}$	0.0066	0.0153	0.0157

Tabela 4.1: Estimativas dos parâmetros das regressões lineares para cada classe de genes

Rácio das intensidades	Métodos				
	Algoritmo EM		Classificador bayesiano	Algoritmo EM e Classif. bayesiano	
(modelos)	NU	GG		NU	GG
RI	0.8706	0.8706	0.8601	0.8715	0.8709
RI1	0.9362	0.9209	0.8887	0.7536	0.7536
RI2	0.9133	0.9142	0.8785	0.9133	0.9149
RI3	0.8435	0.7785	0.8170	0.8438	0.8333

Tabela 4.2: Tabela-resumo com a exactidão das metodologias aplicadas, recorrendo a diferentes rácios de intensidade

Em relação aos resultados apresentados na Tabela 4.2, é possível elaborar alguns comentários e conclusões, por um lado, acerca do desempenho dos métodos e modelos aplicados e, por outro, acerca do facto de se esperar que a intensidade de fluorescência de um gene esteja condicionada pelo número de estirpes em que o gene está presente.

Relativamente aos métodos utilizados verifica-se que a combinação do algoritmo EM com o classificador bayesiano melhora, em quase todos os casos, as exactidões obtidas quer pela utilização do classificador bayesiano no cenário de classificação supervisionada, quer pela utilização isolada do algoritmo EM, independentemente do modelo de mistura escolhido. Existe apenas um caso em que a combinação dos métodos não é a melhor, que foi com a utilização de $RI1$. Neste caso, a exactidão resultante da combinação dos dois métodos foi de 0.7536, o que corresponde à proporção de genes presentes na amostra. Este resultado deve-se à elevada proporção de genes indicados como presentes resultante da pré-classificação obtida pelo algoritmo EM e ao facto do classificador bayesiano, perante esta situação, não ter conseguido encontrar um ponto de corte. Desta forma todos os genes foram classificados como presentes, o que significa que apenas foi possível classificar correctamente 75.36% dos genes. No entanto, quando se aplica só

o algoritmo EM ao rácio $RI1$, este apresenta bons resultados atingindo-se exactidões de cerca de 0.94. O algoritmo EM pode ser visto como uma alternativa a usar quando a combinação dos dois métodos não resulta.

Veja-se o gráfico da Figura 4.1 que regista o comportamento do classificador bayesiano consoante a proporção de genes indicados como presentes na pré-classificação. O classificador bayesiano atinge exactidões maiores quando a proporção de genes indicados como presentes na pré-classificação se encontra próxima da verdadeira proporção de genes presentes na amostra de teste. Quando a proporção de genes presentes na pré-classificação é muito elevada (quando é maior do que 0.85), a exactidão diminui drasticamente. Quando esta proporção é pequena (quando se afasta muito da verdadeira) o classificador bayesiano tende também a apresentar exactidões não muito boas, isto porque na pré-classificação há um grande número de genes indicados como ausentes, mas que na realidade são presentes. A partir de uma proporção de aproximadamente 0.27 de genes indicados como presentes na pré-classificação, o classificador bayesiano não consegue obter um ponto de corte e classifica todos os genes como presentes (pois a proporção de genes presentes na amostra de teste é elevada), daí a exactidão atingir valores na ordem de 0.7536.

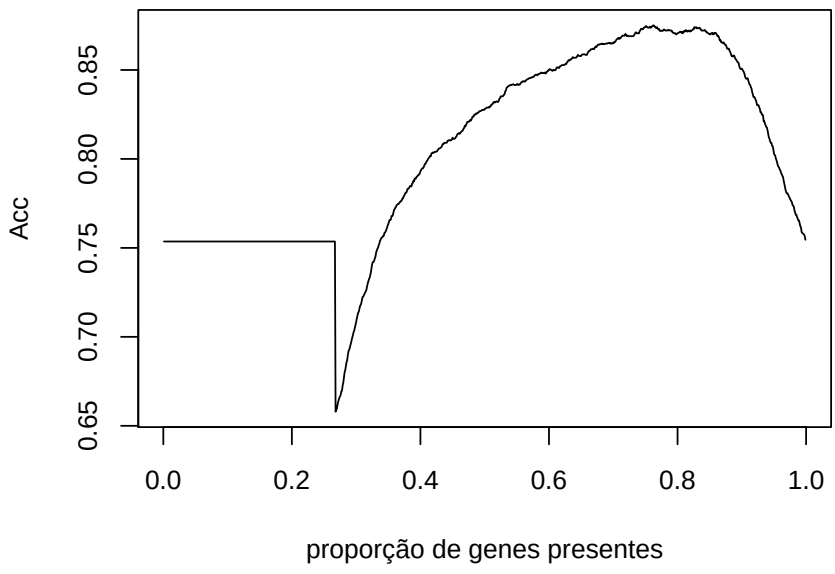

Figura 4.1: Comportamento do classificador bayesiano, em termos de exactidão, consoante a proporção de genes presentes indicados como na pré-classificação.

Quanto aos modelos de mistura aplicados, as diferenças nas exactidões entre os modelos Normal-Uniforme (NU) e Gama-Gama (GG) são muito pequenas, quer com a utilização isolada do algoritmo EM quer com a combinação dos dois métodos.

Se se tiver em conta que as intensidades de um gene na amostra de teste e de controlo são proporcionais, verifica-se que é possível atingir níveis de exactidão mais elevados do que quando não se tem em conta esse facto (caso de RI), o que acontece quando se utiliza o rácio das intensidades $RI2$ (que tem em conta as estimativas $\hat{\beta}_k$ das razões de proporcionalidade). No entanto, quando se tenta estimar as razões de proporcionalidade através da variável *identidade*, ou seja quando se utiliza $RI3$ como rácio das intensidades na aplicação das metodologias habituais, as exactidões obtidas são piores.

Em conclusão, não existe um caso que seja o melhor, em termos de exactidão, em todos os métodos utilizados, contudo aquele em que se utiliza o rácio $RI2$, apresenta sempre resultados melhores do que quando se utiliza o rácio RI, isto leva a crer que é boa prática ter em conta, na análise de experiências CGH em *microarrays*, o facto da amostra de controlo ser composta por uma mistura de estirpes, esperando-se que as intensidades de fluorescência da amostra de teste e controlo, de um determinado gene, sejam proporcionais, consoante o número de estirpes da amostra de controlo em que esse gene está presente.

CAPÍTULO 5

Conclusões finais

O crescente interesse pela compreensão dos factores que influenciam o desenvolvimento de várias patologias e pela procura de melhores diagnósticos e de novas estratégias terapêuticas tem sido ao longo dos últimos anos o grande catalisador de avanços científicos e tecnológicos. A hibridação genómica comparativa em *microarrays* é uma dessas tecnologias, e tem sido utilizada para examinar a semelhança genómica entre organismos, comparando estirpes com genoma sequenciado (amostra de controlo) a outras que ainda não foram sequenciadas (amostra de teste). A comparação genómica entre as estirpes permite detectar os genes que são comuns às duas amostras (genes presentes na amostra de teste) e aqueles que são específicos à amostra de controlo (genes ausentes na amostra de teste).

A análise deste tipo de dados implica uma estratégia para detectar os genes presentes e ausentes na amostra (estirpe) de teste e, por isso, muitas abordagens têm sido aplicadas para resolver este problema, cada uma com as suas vantagens e desvantagens.

Tendo por base modelos de mistura, este trabalho recorreu, por um lado, ao algoritmo EM que permitiu obter estimativas da probabilidade a posteriori de cada gene pertencer à subpopulação dos genes presentes e à dos genes ausentes e, por outro lado, ao classificador bayesiano, um método que permitiu obter uma regra de classificação dos genes, e ainda a uma combinação destes dois métodos usando dados de uma experiência de validação do método em que a verdadeira classificação é conhecida. Para cada metodologia aplicada foi calculada a exactidão a fim de se comparar a proporção de genes correctamente classificados obtida por cada um dos métodos.

Em relação à medida de intensidade utilizada na aplicação dos métodos

referidos, o trabalho desenvolvido mostrou que o rácio das intensidades, $RI = \frac{intensidade\ do\ controlo}{intensidade\ do\ teste}$, é uma boa opção face ao mais popular LR. Também se observou que a *normalização* por métodos típicos de análise de *microarrays* de expressão não conduziu a melhorias nos resultados de CGH em *microarrays*, conseguindo-se exactidões na ordem dos 0.87 (ou seja, conseguindo-se classificar correctamente cerca de 87% dos genes).

Do trabalho realizado resultou que, num caso prático, o algoritmo EM ou a combinação do algoritmo EM com o classificador bayesiano (se da combinação não resultar que todos os genes sejam classificados numa só classe) são ferramentas úteis na identificação de genes presentes na amostra de teste (estirpe não sequenciada), pois estes métodos não necessitam de qualquer informação adicional acerca da classificação dos genes. A escolha dos modelos de mistura a utilizar terá de ter um pouco em conta o comportamento dos dados, mas os resultados obtidos não mostraram grandes diferenças na exactidão entre os modelos Normal-Uniforme e Gama-Gama. Ainda assim tentou verificar-se, usando testes de ajustamento (nomeadamente teste de Kolmogorov-Smirnov e teste do Qui-quadrado à bondade do ajustamento), se a adequabilidade de cada modelo estaria associada a um melhor desempenho. Destes testes resultou que os dois modelos não são um bom ajustamento, embora o modelo Gama-Gama se porte ligeiramente melhor. No entanto, isto apenas significa que os modelos probabilísticos são descrições pobres da realidade (como acontece geralmente) mas, na prática, acabam por ser muitas vezes úteis. Veja-se que mesmo não sendo bons ajustamentos, os modelos permitem atingir proporções de genes classificados correctamente muito satisfatórias.

Nas experiências CGH analisadas, a amostra de controlo é constituída por uma mistura de estirpes e, neste caso, espera-se que a intensidade de fluorescência de um determinado gene fique condicionada pelo número de estirpes em que o gene está presente. Este tema foi tratado no trabalho com base na estimação das razões de proporcionalidade que se pensam existir entre as intensidades das amostras de teste e de controlo através de regressões lineares. Da aplicação dos métodos referidos resulta que, quando se tem em conta esta proporcionalidade, é possível melhorar a exactidão dos métodos e consegue-se, assim, aumentar a percentagem de genes correctamente classificados para cerca de 94%.

Não obstante os resultados obtidos, a principal conclusão que se pode retirar da pesquisa bibliográfica efectuada e do trabalho desenvolvido é que não existe um método perfeito, ou seja, um método que classifique correctamente todos os genes da amostra em estudo, nem tão pouco uma receita única que se possa seguir para resolver o problema da classificação dos genes. Acima de tudo há que ter uma boa noção do comportamento dos dados em questão recorrendo-se para tal a uma prévia análise exploratória dos dados e a uma estreita colaboração entre investigadores das áreas de estatístisca e biologia.

APÊNDICE A

Glossário

Ácido desoxirribonucleico (ADN) - cadeias de moléculas portadoras da informação hereditária. É um ácido nucleico constituído por nucleótidos formados por uma base nitrogenada (adenina, guanina, timina ou citosina), uma pentose (desoxirribose) e um grupo fosfato. As bases nitrogenadas unem-se, através de ligações de hidrogénio, a bases de outros nucleótidos formando uma estrutura de dupla hélice (modelo proposto por Watson e Crick em 1953). Nas cadeias de nucleótidos, a adenina (A) une-se à timina (T) e a guanina (G) à citosina (C).

Ácido ribonucleico (ARN) - é uma macromolécula formada por nucleótidos de adenina (A), guanina (G), citosina (C) e uracilo (U), sendo a pentose no ARN sempre a ribose. A molécula de ARN é formada apenas por uma cadeia de nucleótidos e o seu comprimento é muito inferior ao comprimento de uma molécula de ADN.

ARN mensageiro (ARNm) - molécula do ácido ribonucleico que contém informação necessária para construir uma proteína.

Bactéria - organismo unicelular procariota (em que o material genético não está encerrado por uma membrana nuclear). A célula bacteriana possui como qualquer célula viva, um genoma, um citoplasma (totalidade da área intra-celular), uma membrana plasmática (estrutura que delimita a célula), uma parede celular, e nalguns casos possui ainda uma cápsula externa.

Cromossoma - estrutura filamentosa onde se encontra o material com informação hereditária na célula.

Estirpe - população de células descendentes de uma única célula ancestral.

Gene - unidade fundamental do material genético. Os genes são a mais pequena porção de um cromossoma e constituem os moldes de todas as moléculas de ARN.

Genoma - conteúdo genético total de uma célula.

Hibridação - detecção ou caracterização de uma determinada sequência de um ácido nucleico numa amostra biológica com respeito à sua habilidade de se ligar a uma molécula de ácido nucleico complementar através das ligações de hidrogénio (de acordo com emparelhamentos A-T e G-C).

Nucleótido - é uma molécula simples constituída por uma base nitrogenada, adenina (A), guanina (G), citosina (C), timina (T) ou uracilo (U), unida a um açúcar (pentose), ribose ou desoxirribose, e a um grupo fosfato.

Oligonucleótido - cadeia pequena de ácidos nucleicos. Os oligonucleótidos podem ser, por exemplo, compostos por ADN tal como aqueles que são frequentemente usados no procedimento *Polymerase Chain Reaction*, onde se costumam designar por *primers*.

Polymerase Chain Reaction (PCR) - método que permite a multiplicação de fragmentos de uma cadeia dupla de ADN. Esta técnica foi introduzida em 1985 e utiliza ao longo do processo de multiplicação duas porções de ADN (*primers*), nucleótidos e uma enzima chamada de ADN polimerase. O processo desenrola-se da seguinte forma: a dupla cadeia de ADN é separada, por acção do calor, em duas cadeias simples, depois a enzima ADN polimerase actua de forma a adicionar às duas cadeias simples separadas os nucleótidos necessários para formar duas cadeias duplas de ADN. Este processo é repetido continuamente permitindo obter em algumas horas milhões de cópias do fragmento inicial.

Sequenciação - determinação da ordem ou sequência dos elementos de uma molécula (por exemplo, nucleótidos do ADN).

APÊNDICE B

Rotinas criadas no sofware R

1. A função *nudge1.sn* é uma adaptação da função *nudge1* da biblioteca NUDGE e implementa o modelo de mistura Normal-Uniforme sem efectuar a *normalização* nos dados. Os argumentos da função são:

   - *ratio* - um vector (ou uma matriz quando se está a trabalhar com mais do que um *microarray*) com o rácio das intensidades para todos os genes;
   - n - o número de colunas do argumento *ratio*;
   - *FUN* - a função aplicada ao argumento *ratio* quando este é uma matriz. Por defeito *FUN=mean*, o que significa que é aplicada a média entre as colunas da matriz;
   - z - uma matriz com duas colunas que contém informação sobre a classificação dos genes, a primeira coluna é a indicatriz de genes presentes e a segunda é a indicatriz de genes ausentes. Por defeito $z=NULL$, ou seja, parte-se do princípio que não se tem informação sobre a classificação dos genes;
   - *tol* - a tolerância utilizada no algoritmo EM. Quando a diferença da log-verosimilhança entre duas iterações consecutivas é inferior a *tol* está preenchido o critério principal de paragem do algoritmo;
   - *iterlim* - o número máximo de iterações permitidas até à convergência do algoritmo EM. Por defeito $iterlim = 500$, ou seja, se depois de 500 iterações o critério de paragem anterior não for atingido, então o processo de estimação termina e diz-se que o algoritmo não convergiu.

```
nudge1.sn=function(ratio,n,FUN=mean,z=NULL,tol=1e-05,
iterlim = 500)
{
   if(n == 1) X=ratio else X=apply(ratio, 1, FUN)
   n=length(X)
   if (is.null(z)) {
      z=matrix(0,n,2)
      m=mean(X)
      s=sqrt(var(X))
      d=abs((X-m)/s)>2
      z[, 2]=d^2
      z[, 1]=1-z[, 2]
   }
   p=sum(z[, 1])/n
   muhat=sum(z[, 1] * X)/sum(z[, 1])
   sigma2hat=sum(z[, 1] * (X - muhat)^2)/sum(z[, 1])
   sigmahat=sqrt(sigma2hat)
   llike=c(0, 100)
   criterion=abs(llike[1] - llike[2])
   iter=0
   while ((criterion > tol) & (iter < iterlim)) {
      iter=iter + 1
      z[, 1]=(p * dnorm(X, muhat, sigmahat))/
       ((p * dnorm(X, muhat, sigmahat)) +
       ((1 - p) * dunif(X, min(X), max(X))))
      z[, 2]=1 - z[, 1]
      p=sum(z[, 1])/n
      muhat=sum(z[, 1] * X)/sum(z[, 1])
      sigma2hat=sum(z[, 1] * (X - muhat)^2)/sum(z[, 1])
      sigmahat=sqrt(sigma2hat)
      loglike=sum(log((p * dnorm(X, muhat, sigmahat)) +
                 ((1 - p) * dunif(X, min(X), max(X)))))
      llike[2]=llike[1]
      llike[1]=loglike
      criterion=abs(llike[1] - llike[2])
   }
   colnames(z) = c("Prob. gene presente", "Prob. gene ausente")
   lista=list(pdiff = z[, 2], Ratio = X, mu = muhat,
        sigma = sigmahat, mixprob = p, a = min(X), b = max(X),
        loglike = loglike, iter = iter)
   lista
}
```

O que a função nos dá é uma lista que contém:

- A probabilidade de cada ser gene ser classificado como ausente (*pdiff*);
- O rácio das intensidades (controlo/teste) para cada gene (*Ratio*);
- As estimativas dos parâmetros do modelo (*mu*, *sigma*, a e b) e a probabilidade de mistura (*mixprob*);
- O máximo da função log-verosimilhança (*loglike*);
- O número de iterações necessário para a convergência do algoritmo EM (*iter*).

2. A função *modmistgama* implementa o modelo de mistura Gama-Gama (com a utilização do algoritmo EM) descrito em 3.1.1. Os argumentos da função são:

   - *intens* - um vector (ou uma matriz quando se está a trabalhar com mais do que um *microarray*) com o rácio das intensidades para todos os genes;
   - n - o número de colunas do argumento *intens*;
   - *FUN* - a função aplicada ao argumento *intens* quando este é uma matriz. Por defeito *FUN=mean*, o que significa que é aplicada a média entre as colunas da matriz;
   - y - uma matriz com duas colunas que contém informação sobre a classificação dos genes semelhante à matriz z da função *nudge1.sn*. Por defeito supõe-se que esta matriz não é conhecida;
   - *delta* - o quantil utilizado para obter uma primeira classificação, usado quando a matriz y não é fornecida. Por defeito $delta = 0.5$, ou seja, um gene tem a mesma probabilidade de se classificar como ausente e como presente;
   - *tol* e *iterlim* têm o mesmo significado do que na função *nudge1.sn*.

```
modmistgama=function(intens,n,FUN=mean,y=NULL,delta=0.5,tol=1e-05,
iterlim=500)
{
   if(n == 1) X=intens else X=apply(intens, 1, FUN)
   n=length(X)
   if (is.null(y)) {
      y=matrix(0, n, 2)
      q=quantile(X,1-delta)
      r=(X > q) #ausentes
      y[,2]=r^2 #ausentes
```

```
      y[,1]=1-y[,2] #presentes
   }
   p=sum(y[, 1])/n #proporção de presentes
   m1=numeric(2); m2=numeric(2)
   alfahat=numeric(2); betahat=numeric(2)
   for(i in 1:2) {
      m1[i]=sum(y[, i] * X)/sum(y[, i])
      m2[i]=sum(y[, i] * (X^2))/sum(y[, i])
      alfahat[i]=(m1[i]^2)/(m2[i]-(m1[i]^2))
      betahat[i]=m1[i]/(m2[i]-(m1[i]^2))
   }
   llike=c(0, 100); criterion=abs(llike[1] - llike[2])
   iter=0
   while ((criterion > tol) & (iter < iterlim)) {
      iter=iter+1
      y[, 1]=(p*dgamma(X,alfahat[1],betahat[1],1/betahat[1]))/
           ((p*dgamma(X,alfahat[1],betahat[1],1/betahat[1]))+
           ((1-p)*dgamma(X,alfahat[2],betahat[2],1/betahat[2])))
      y[, 2]=1-y[, 1]
      p=sum(y[, 1])/n #actualiza a proporção de presentes
      for(i in 1:2) {
         m1[i]=sum(y[, i] * X)/sum(y[, i])
         m2[i]=sum(y[, i] * (X^2))/sum(y[, i])
         alfahat[i]=(m1[i]^2)/(m2[i]-(m1[i]^2))
         betahat[i]=m1[i]/(m2[i]-(m1[i]^2))
      }
      loglike=sum(log((p*dgamma(X,alfahat[1],betahat[1],
            1/betahat[1])) + ((1 - p) * dgamma(X,alfahat[2],
            betahat[2],1/betahat[2]))))
      llike[2]=llike[1];  llike[1]=loglike
      criterion=abs(llike[1]-llike[2])
   }
   colnames(y)=c("Prob. gene presente", "Prob. gene ausente")
   rownames(y)=rownames(intens)
   pre=c(alfahat[1],betahat[1]);  names(pre)=c("alfa", "beta")
   aus=c(alfahat[2],betahat[2]);  names(aus)=c("alfa", "beta")
   lista=list(pdiff = y[, 2], intensities = X, dist.pres = pre,
   dist.aus = aus, mixprob = p, loglike = loglike, iter = iter)
   lista
}
```

Os parâmetros de saída da função *modmistgama* são, com as devidas adaptações, muito semelhantes aos da função *nudge1.sn*.

3. A função *classbayes.k* foi criada para classificar os genes, tendo-se como conhecimento inicial uma determinada pré-classificação e corresponde ao classificador bayesiano descrito em 3.1.3 e requer o uso da biblioteca *rootSolve*. Os argumentos da função são:

   - m - a dimensão do conjunto de treino;
   - D - uma matriz com duas colunas, a primeira correspondendo ao rácio das intensidades de cada gene e a segunda com a pré-classificação dos genes;
   - k - número de reamostragens desejadas. Ao longo do trabalho utilizou-se $k = 1000$.

```
classbayes.k=function(m,D,k) {
   g=0; h=0; b0=0; b1=0
   ng=m
   numet0=function(x)
      exp(-lbeta(a0,G0)-log(x)+log(n0+b0)-log(ng+b0+b1)+
          G0*log(H0/(x+H0))+a0*log(x/(x+H0)))
   numet1=function(x)
      exp(-lbeta(a1,G1)-log(x)+log(n1+b1)-log(ng+b0+b1)+
          G1*log(H1/(x+H1))+a1*log(x/(x+H1)))
   predt0=function(x) numet0(x)/(numet0(x)+numet1(x))
   predt1=function(x) numet1(x)/(numet0(x)+numet1(x))
   ft=function(x) predt1(x)-predt0(x)

   m1=numeric(); m2=numeric()
   data=mat.or.vec(k,4)
   colnames(data)=c("root1","root2","acc","acc.todos")

   i=0
   while(i<k) {
      amostra=sample(seq(1,nrow(D)),m)
      n0=m-sum(D[amostra,2])
      n1=sum(D[amostra,2])
      while(n0<2 | n1<2) {
         amostra=sample(seq(1,nrow(D),m)
         n0=m-sum(D[amostra,2])
         n1=sum(D[amostra,2])
      }
      Ds=D[amostra,]
      x=Ds[,1]; t=Ds[,2]
      n1=sum(t); n0=m-n1
```

```
    m1[2]=Ds[,2]%*%Ds[,1]/n1
    m1[1]=(1-Ds[,2])%*%Ds[,1]/n0
    m2[2]=Ds[,2]%*%(Ds[,1]^2)/n1
    m2[1]=(1-Ds[,2])%*%(Ds[,1]^2)/n0
    a0=(m1[1]^2)/(m2[1]-(m1[1]^2))

    a1=(m1[2]^2)/(m2[2]-(m1[2]^2))
    G0=n0*a0+g
    G1=n1*a1+g
    H0=as.numeric(Ds[,1]%*%(1-Ds[,2]))
    H1=as.numeric(Ds[,1]%*%Ds[,2])

    estim=numeric(nrow(D)-m)
    estim.todos=numeric(nrow(D))
    roots=uniroot.all(ft,c(min(x),max(x)))
    lenrt=length(roots)
    if(lenrt==0) {
       root1=NA; root2=NA
       if(ft(min(x)>0)) {
          estim=rep(1,nrow(D)-m)
          estim.todos=rep(1,nrow(D))
       }
    } else {
       if(lenrt==1) root1=roots else {
          estims=mat.or.vec(nrow(D)-m,lenrt)
          estims.tot=mat.or.vec(nrow(D),lenrt)
          accs=numeric(lenrt)
          accs.tot=numeric(lenrt)
          for(j in 1:lenrt) {
             estims[,j]=ifelse(D[-amostra,1]>=roots[j],0,1)
             estims.tot[,j]=ifelse(D[,1]>=roots[j],0,1)
             accs[j]=SSA(isR[-amostra],estims[,j])$ssa["Acc"]
             accs.tot[j]=SSA(isR,estims.tot[,j])$ssa["Acc"]
          }
          root1=roots[which.max(accs.tot)]
          root2=roots[which.min(accs.tot)]
       }
       estim=ifelse(D[-amostra,1]>=root1,0,1)
       estim.todos=ifelse(D[,1]>=root1,0,1)
    }
  acc=SSA(isR[-amostra],estim)$ssa["Acc"]
  acc.todos=SSA(isR,estim.todos)$ssa["Acc"]
  data[i+1,1]=root1
```

```
   data[i+1,2]=ifelse(lenrt==1,root1,root2)
   data[i+1,3]=as.numeric(acc)
   data[i+1,4]=as.numeric(acc.todos)

   # PS: Em root1 está sempre a raíz da equação que maximiza
   #    a accuracy. Quando há duas raízes a outra é root2.

   i=i+1
   }
   data
}
```

A função *classbayes.k* produz uma matriz com k linhas que contém a informação dos pontos de corte (*root1* e *root2*) e dos valores de exactidão (*acc*) obtidos em cada reamostragem.

4. A função *classbayes2.k* foi criada com o objectivo de se comparar os resultados do classificador bayesiano utilizando duas pré-classificações diferentes, como por exemplo, a classificação supervisionada e a classificação não supervisionada. Os argumentos desta função são os mesmos do que os da função *classbayes.k*, à excepção da tabela D que contém aqui três colunas, devido às duas pré-classificações requeridas.

```
classbayes2.k=function(m,D,k) {
   g=0; h=0; b0=0; b1=0
   ng=m
   numet0=function(x)
      exp(-lbeta(a0,G0)-log(x)+log(n0+b0)-log(ng+b0+b1)+
       G0*log(H0/(x+H0))+a0*log(x/(x+H0)))
numet1=function(x)
   exp(-lbeta(a1,G1)-log(x)+log(n1+b1)-log(ng+b0+b1)+
       G1*log(H1/(x+H1))+a1*log(x/(x+H1)))
 predt0=function(x) numet0(x)/(numet0(x)+numet1(x))
 predt1=function(x) numet1(x)/(numet0(x)+numet1(x))
 ft=function(x) predt1(x)-predt0(x)

 numet0.alt=function(x)
    exp(-lbeta(a0.alt,G0.alt)-log(x)+log(n0.alt+b0)-
        log(ng+b0+b1)+G0.alt*log(H0.alt/(x+H0.alt))+
        a0.alt*log(x/(x+H0.alt)))
 numet1.alt=function(x)
    exp(-lbeta(a1.alt,G1.alt)-log(x)+log(n1.alt+b1)-
        log(ng+b0+b1)+G1.alt*log(H1.alt/(x+H1.alt))+
```

```
       a1.alt*log(x/(x+H1.alt)))
predt0.alt=function(x)
           numet0.alt(x)/(numet0.alt(x)+numet1.alt(x))
predt1.alt=function(x)
           numet1.alt(x)/(numet0.alt(x)+numet1.alt(x))
ft.alt=function(x) predt1.alt(x)-predt0.alt(x)

m1=numeric(); m2=numeric()
m1.alt=numeric(); m2.alt=numeric()
data=mat.or.vec(k,4)
colnames(data)=c("root1","root2","acc","acc.todos")
data.alt=mat.or.vec(k,4)
colnames(data.alt)=c("root1","root2","acc","acc.todos")

i=0
while(i<k) {
   amostra=sample(seq(1,nrow(D)),m)
   n0=m-sum(D[amostra,2])
   n1=sum(D[amostra,2])
   n0.alt=m-sum(D[amostra,3])
   n1.alt=sum(D[amostra,3])
   while((n0<2 | n1<2) | (n0.alt<2 | n1.alt<2)) {
      amostra=sample(seq(1,nrow(D)),m)
      n0=m-sum(D[amostra,2])
      n1=sum(D[amostra,2])
      n0.alt=m-sum(D[amostra,3])
      n1.alt=sum(D[amostra,3])
   }
   Ds=D[amostra,]
   x=Ds[,1]; t=Ds[,2]; t.alt=Ds[,3]
   n1=sum(t); n0=m-n1;
   n1.alt=sum(t.alt); n0.alt=m-n1.alt

   m1[2]=Ds[,2]%*%Ds[,1]/n1
   m1[1]=(1-Ds[,2])%*%Ds[,1]/n0
   m2[2]=Ds[,2]%*%(Ds[,1]^2)/n1
   m2[1]=(1-Ds[,2])%*%(Ds[,1]^2)/n0
   a0=(m1[1]^2)/(m2[1]-(m1[1]^2))
   a1=(m1[2]^2)/(m2[2]-(m1[2]^2))
   G0=n0*a0+g
   G1=n1*a1+g
   H0=as.numeric(Ds[,1]%*%(1-Ds[,2]))
   H1=as.numeric(Ds[,1]%*%Ds[,2])
```

```
m1.alt[2]=Ds[,3]%*%Ds[,1]/n1.alt
m1.alt[1]=(1-Ds[,3])%*%Ds[,1]/n0.alt
m2.alt[2]=Ds[,3]%*%(Ds[,1]^2)/n1.alt
m2.alt[1]=(1-Ds[,3])%*%(Ds[,1]^2)/n0.alt
a0.alt=(m1.alt[1]^2)/(m2.alt[1]-(m1.alt[1]^2))
a1.alt=(m1.alt[2]^2)/(m2.alt[2]-(m1.alt[2]^2))
G0.alt=n0.alt*a0.alt+g
G1.alt=n1.alt*a1.alt+g
H0.alt=as.numeric(Ds[,1]%*%(1-Ds[,3]))
H1.alt=as.numeric(Ds[,1]%*%Ds[,3])

# pré-classificação 1 (supervisionada)
estim=numeric(nrow(D)-m)
estim.todos=numeric(nrow(D))
roots=uniroot.all(ft,c(min(x),max(x)))
lenrt=length(roots)
if(lenrt==0) {
   root1=NA
   root2=NA
   if(ft(min(x)>0)) {
      estim=rep(1,nrow(D)-m)
      estim.todos=rep(1,nrow(D))
   }
} else {
   if(lenrt==1) root1=roots else {
      estims=mat.or.vec(nrow(D)-m,lenrt)
      estims.tot=mat.or.vec(nrow(D),lenrt)
      accs=numeric(lenrt)
      accs.tot=numeric(lenrt)
      for(j in 1:lenrt) {
         estims[,j]=ifelse(D[-amostra,1]>=roots[j],0,1)
         estims.tot[,j]=ifelse(D[,1]>=roots[j],0,1)
         accs[j]=SSA(isR[-amostra],estims[,j])$ssa["Acc"]
         accs.tot[j]=SSA(isR,estims.tot[,j])$ssa["Acc"]
      }
      root1=roots[which.max(accs)]
      root2=roots[which.min(accs)]
   }
   estim=ifelse(D[-amostra,1]>=root1,0,1)
   estim.todos=ifelse(D[,1]>=root1,0,1)
}
acc=SSA(isR[-amostra],estim)$ssa["Acc"]
acc.todos=SSA(isR,estim.todos)$ssa["Acc"]
data[i+1,1]=root1
```

```
data[i+1,2]=ifelse(lenrt==1,root1,root2)
data[i+1,3]=as.numeric(acc)
data[i+1,4]=as.numeric(acc.todos)

# pré-classificação 2 (não supervisionada)
estim=numeric(nrow(D)-m)
estim.todos=numeric(nrow(D))
roots=uniroot.all(ft.alt,c(min(x),max(x)))
lenrt=length(roots)
if(lenrt==0) {
   root1=NA
   root2=NA
   if(ft(min(x)>0)) {
      estim=rep(1,nrow(D)-m)
      estim.todos=rep(1,nrow(D))
   }
} else {
   if(lenrt==1) root1=roots else {
      estims=mat.or.vec(nrow(D)-m,lenrt)
      estims.tot=mat.or.vec(nrow(D),lenrt)
      accs=numeric(lenrt)
      accs.tot=numeric(lenrt)
      for(j in 1:lenrt) {
         estims[,j]=ifelse(D[-amostra,1]>=roots[j],0,1)
         estims.tot[,j]=ifelse(D[,1]>=roots[j],0,1)
         accs[j]=SSA(isR[-amostra],estims[,j])$ssa["Acc"]
         accs.tot[j]=SSA(isR,estims.tot[,j])$ssa["Acc"]
      }
      root1=roots[which.max(accs)]
      root2=roots[which.min(accs)]
   }
   estim=ifelse(D[-amostra,1]>=root1,0,1)
   estim.todos=ifelse(D[,1]>=root1,0,1)
}
acc=SSA(isR[-amostra],estim)$ssa["Acc"]
acc.todos=SSA(isR,estim.todos)$ssa["Acc"]
data.alt[i+1,1]=root1
data.alt[i+1,2]=ifelse(lenrt==1,root1,root2)
data.alt[i+1,3]=as.numeric(acc)
data.alt[i+1,4]=as.numeric(acc.todos)

# PS: Em root1 está sempre a raíz da equação que maximiza
#    a accuracy. Quando há duas raízes a outra é root2.
```

```
    i=i+1
 }
 ans=list(data,data.alt)
 names(ans)=c("isR","init")
 ans
}
```

O resultado da função *classbayes2.k* é muito semelhante ao da função *classbayes.k* com a diferença de que são geradas duas tabelas uma para cada pré-classificação. Esta função requer também a biblioteca *rootSolve*.

5. A função *SSA* calcula a *especifidade*, a *sensibilidade*, a *exactidão*, verdadeiros positivos, verdadeiros negativos, falsos positivos e os falsos negativos. É utilizada dentro das funções *classbayes.k* e *classbayes2.k*.

```
SSA=function(real,estimada) {
   TP=numeric(); TN=numeric()
   FP=numeric(); FN=numeric()
   TP=(estimada==1)+(real==1)
   TN=(estimada==0)+(real==0)
   FP=(estimada==1)+(real==0)
   FN=(estimada==0)+(real==1)
   nTP=sum(TP==2,na.rm=T)
   nTN=sum(TN==2,na.rm=T)
   nFP=sum(FP==2,na.rm=T)
   nFN=sum(FN==2,na.rm=T)
   data1=data.frame(TP=nTP, TN=nTN, FP=nFP, FN=nFN)
   data2=data.frame(Sn = nTP/(nTP+nFN), Sp = nTN/(nTN+nFP),
                    Acc = (nTP+nTN)/(nTP+nTN+nFP+nFN))
   values=list(tab=data1, ssa=data2)
   values
}
```

Bibliografia

Antunes M. and Sousa L. **Bayesian Classification and Non-Bayesian Label Estimation via EM Algorithm to Identify Differentially Expressed Genes: a Comparative Study**. *Biometrical Journal* **50**, 5, 824-836. (2008).

Behr M.A., Wilson M.A., Gill W.P., Salamon H., Schoolnik G.K., Rane S. and Small P.M. **Comparative genomics of BCG vaccines by whole-genome DNA microarray**. *Science* **284**, 15201523. (1999).

Björkholm B., Lundin A., Sillén A., Guillemin K., Salama N., Rubio C., Gordon J.I., Falk P. and Engstrand L. **Comparison of genetic divergence and fitness between two subclones of Helicobacter pylori**. *Infection and Immunity* **69**, 78327838. (2001).

Carlos Daniel Paulino, M. Antónia Amaral Turkman, Bento Murteira. **Estatística Bayesiana**. *Fundação Calouste Gulbenkian. Lisboa.* (2003).

Cleveland W.S. **Robust Locally Weighted Regression and Smoothing Scatterplots**. *American Statistical Association* **74**:368, 829-836. (1979).

Dean N. and Raftery A.E. **Normal uniform mixture differential gene expression detection for cDNA microarrays**. *Genome Biology* **6**:173. (2005).

Dempster A.P., Laird N.M. and Rubin D.B. **Maximum likelihood from incomplete data via the EM algorithm (with discussion)**. *Journal of the Royal Statistical Society. Series B (Methodological)*. Vol. **39**, No. **1**, pp. 1-38 (1977).

Dorrell N., Mangan J.A., Laing K.G., Hinds J., Linton D., Al-Ghusein H., Barrell B.G., Parkhill J., Stoker N.G., Karlyshev A.V., Butcher P.D. and

Wren B.W. **Whole Genome Comparison of Campylobacter jejuni Human Isolates Using a Low-Cost Microarray Reveals Extensive Genetic Diversity**. *Genome Research* **11**:1706-1715. (2001)

Dziejman M., Balon E., Boyd D., Fraser C.M., Heidelberg J.F. and Mekalanos J.J. **Comparative genomic analysis of Vibrio cholerae: genes that correlate with cholera endemic and pandemic disease**. *Proceedings of the National Academy of Sciences of the United States of America* **99**:1556-1561. (2002).

Feten G., Almøy T., Snipen L., Aakra Å., Nyquist O.L. and Aastveit A.H. **Mixture Models as a Method to Find Present and Divergent Genes in Comparative Genomic Hybridization Studies on Bacteria**. *Biometrical Journal* **49**, 2, 242-258. (2007).

Geoffrey L. Zubay. **Biochemistry**. *WCB/McGraw-Hill*. **Fourth Edition**. (1998).

James N. Thompson, Jr., Jenna J. Hellack, Gerald Braver, David S. Durica. **Primer of Genetic Analysis - A problems approach**. *Cambridge University Press*. **Second Edition**. (1997).

Kendziorski C.M., Newton M.A., Lan H. and Gould M.N. **On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles**. *Statistics in Medicine* **22**, 3899-3914. (2003).

Kim-Ann Do, Peter Müller, Marina Vannucci. **Bayesian Inference for Gene Expression and Proteomics**. *Cambridge University Press*. (2006).

Kim C.C., Joyce E.A., Chan K. and Falkom S. **Improved analytical methods for microarray-based genome-composition analysis**. *Genome Biology* **3**:11. (2002).

Lewin A., Bochkina N. and Richardson S. **Fully Bayesian Mixture Model for Differential Gene Expression: Simulations and Model Checks**. *Statistical Applications in Genetics and Molecular Biology* **6**(1), article 36. (2007).

Lönnstedt I. and Speed T. **Replicated Microarray Data**. *Statistica Sinica* **12**, 31-46. (2002).

Lucchini S., Thompson A. and Hinton J.C.D. **Microarrays for microbiologists**. *Microbiology* **147**, 1403-1414. (2001).

Luís Miguel Almeida da Silva. **Selecção de variáveis em microarrays de ADN**. *Tese submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Estatística*. Maio de 2003.

Michael T. Madigan, John M. Martinko and Jack Parker. **Brock Biology of Microorganisms**. *Prentice Hall International, Inc.* **Eighth Edition**. (1997).

Murray A.E., Lies D., Li G., Nealson K., Zhou J. and Tiedje J.M. **DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes**. *Proceedings of the National Academy of Sciences of the United States of America* **98**:9853-9858. (2001).

Newton R., Hinds J. and Wernisch L. **A Hidden Markov model web application for analysing bacterial genomotyping DNA microarray experiments**. *Applied Bioinformatics*, **5**(4), 211-218. (2006).

Pinto F.R., Aguiar S.I., Melo-Cristino J. and Ramirez M. **Optimal control and analysis of two-color genomotyping experiments using bacterial multistrain arrays**. *BMC Genomics*, **9**:230. (2008).

Repsilber D., Mira A., Lindroos H., Andersson S. and Ziegler A. **Data Rotation Improves Genomotyping Efficiency**. *Biometrical Journal* **47**, 4, 585-598. (2005).

Salama N., Guillemin K., McDaniel T.K., Sherlock G., Tompkins L. and Falkow S. **A whole-genome microarray reveals genetic diversity among Helicobacter pylori strains**. *Proceedings of the National Academy of Sciences of the United States of America* **97**:14668-14673. (2000).

Snipen L., Repsilber D., Nyquist L., Ziegler A., Aakra Å. and Aastveit A. **Detection of divergent genes in microbial aCGH experiments**. *BMC Bioinformatics* **7**:181. (2006).

Snipen L., Nyquist L., Solheim M., Aakra Å. and Nes I. **Improved analysis of bacterial CGH data beyond the log-ratio paradigm**. *BMC Bioinformatics* **10**:91. (2009).

The R project for Statistical Computing. http://www.r-project.org/.