Métodos Estatísticos na Análise de Experimentos de *Microarray*

Elier Broche Cristo

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA OBTENÇÃO DO GRAU DE MESTRE
EM
ESTATÍSTICA

Área de Concentração: **ESTATÍSTICA**
Orientador: **Prof. Dr. Eduardo Jordão Neves**

O autor recebeu apoio financeiro da
FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo
Processo: 01/07082-5

– São Paulo, Dezembro de 2003 –

Métodos Estatísticos na Análise de Experimentos de *Microarray*

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e apresentada por **Elier Broche Cristo**
e aprovada pela Comissão Julgadora.

São Paulo, Dezembro de 2003.

Banca Examinadora:

- **Prof. Dr. Eduardo Jordão Neves** (orientador) (IME-USP)
- **Profa. Dra. Julia Maria Pavan Soler** (IME-USP)
- **Profa. Dra. Nancy Lopes Garcia** (IME-UNICAMP)

*A mis padres,
Inés Cristo Pérez y Juan Antonio Broche Sánchez,
para quienes mis estudios siempre fueron un sueño.*

Agradecimentos

Agradeço a meus pais, Inés e Juan Antonio, por meus estudos sempre terem sido um sonho para eles, e por terem me dado todo o apoio necessário para realizá-los. A meu irmão, Osnel, pela amizade e orientação constante, desde criança, nos meus estudos, assim como pela ajuda na vinda para o Brasil e instalação em São Paulo. A meu orientador, Neves, pela sua dedicação, motivação e orientação na minha formação e pesquisa, assim como pelo agradável clima de trabalho no nosso grupo. Ao David e ao Walter pela ajuda no BIOINFO. Aos demais colegas do grupo de trabalho, Roberto, Gustavo e Silvio, pelo constante intercâmbio e apoio. A Rita e a sua família por terem me recebido neste país como mais um membro da família. Ao Robson, pelas nossas conversas e amizade. A Diana, Juvêncio e Persé pela amizade e pelos momentos compartilhados nas disciplinas. A todos, pela ajuda nestes dois anos, correções do português, orientações, explicações, etc.

Agradeço também à FAPESP, pelo apoio financeiro e técnico. Ao BIOINFO/USP, ao Instituto de Matemática e Estatística e à Universidade de São Paulo. A todas estas instituições, agradeço a oportunidade de dar continuidade à minha formação. Considero extraordinária a existência de tais possibilidades no Brasil para estudantes de outros países.

Sumário

1	Introdução	5
1.1	Tecnologia de <i>Microarray</i>	8
1.2	Análise Estatística de Dados de <i>Microarray</i>	13
2	Normalização	18
2.1	Normalização de lâminas independentes, sem estabilização da variância	20
2.1.1	Normalização pela energia total	20
2.1.2	Normalização com dependência da intensidade	21
2.1.3	Normalização com dependência da agulha	22
2.2	Normalização de lâminas independentes, com estabilização da variância	22
2.3	Normalização para pares de lâminas nas quais foi empregada inversão de corante	23
2.4	Normalização composta	24
2.5	Normalização usando métodos locais e robustos	25
2.5.1	Lowess-Loess, (<i>Locally Weighted Smoothing Scatterplot</i>)	25
2.5.2	Normalização utilizando Loess	29
2.6	Comparações entre normalizações	30
3	Modelo para expressão gênica	31
3.1	Notação do modelo	32
3.2	Correção do <i>background</i>	33
3.3	Análise assintótica da variância	35
3.4	Transformação proposta para estabilizar a variância	35

3.5	Estimação dos parâmetros do modelo	36
3.5.1	Estimação do <i>background</i> usando espaços em branco	36
3.5.2	Estimação do <i>background</i> com réplicas	37
3.5.3	Estimação do <i>background</i> sem réplicas	38
3.5.4	Previsão do erro multiplicativo	38
4	Técnicas de agrupamento	40
4.1	Distância entre dois pontos	41
4.2	Métodos de união	42
4.3	Agrupamento Hierárquico	44
4.4	Agrupamento K -Médias	44
4.5	Agrupamento usando SOM (<i>Self-Organizing Maps</i>)	45
5	Discriminação - Classificação	48
5.1	Classificação usando K -Vizinhos	49
5.2	Discriminante linear de Fisher	50
5.3	Árvores de classificação	51
5.4	Classificação usando Máxima Verossimilhança	52
5.4.1	Estimadores viciados ou não viciados?	55
5.5	Classificação usando <i>Aggregating</i>	57
5.5.1	Geração dos LS perturbados usando <i>Bagging</i> não paramétrico	59
5.5.2	Geração dos LS perturbados usando <i>Bagging</i> paramétrico	59
5.5.3	Geração dos LS perturbados usando <i>Boosting</i>	59
5.6	Alerta no Diagnóstico	60
5.7	Método BE (Busca e Escolha) de clique de genes	63
5.8	Classificação usando RL (Regressão Local)	64
5.9	Comparação dos métodos de classificação	65
6	Análise de Dados de Experimentos de Expressão Gênica	68
6.1	Acompanhamento de experimentos de <i>Microarray</i>	68

6.2	Genes DE (Diferencialmente Expressos)	70
6.3	Análise do conjunto de dados de câncer de mama disponível na <i>web</i>	72
6.4	Análise do conjunto de dados de testes de <i>Microarrays</i> comerciais	74
6.5	Análise do conjunto de dados de câncer de mama gerado no Instituto Ludwig . . .	77
6.6	Análise do conjunto de dados de testes de protocolo bioquímico	78
6.7	Análise do conjunto de dados “Amplificado vs não Amplificado”	81
6.8	Análise do conjunto de dados de câncer de estômago	81
7	Apêndices: Artigos associados à presente dissertação	103
A	Comparative analysis of amplified and nonamplified RNA for hybridization in cDNA microarray	104
B	MOLECULAR CLASSIFIERS FOR GASTRIC CANCER AND PRE-MALIGNANT DISEASES OF THE GASTRIC MUCOSA (in print)	105

Lista de Figuras

1.1	Processos de Transcrição e Tradução na célula [65].	6
1.2	Etapas do processo de <i>Microarray</i> [65].	9
1.3	Resultado da digitalização da imagem da lâmina [32].	10
1.4	Intensidades nos dois canais para um determinado <i>spot</i> [32].	11
1.5	Estimação do sinal e <i>background</i> [32].	12
1.6	Agrupamento de dados de expressão gênica de melanoma [6]; a) Dendograma do agrupamento hierárquico, com um grupo de 19 melanomas no centro; b) gráfico MDS (<i>Multidimensional Scaling</i>) tri-dimensional com as 31 amostras de melanomas cutâneos, apresentando o grupo principal de 19 amostras (azul, dentro do cilindro) e as 12 restantes (ouro, fora do cilindro); c) gráfico do número esperado e observado de genes produzindo um determinado número de classificações erradas para uma partição de 31 melanomas em dois grupos de 12 e 19; os triângulos vermelhos são os grupos observados, as linhas são grupos produzidos aleatoriamente e os círculos são os resultados preditos para a variável aleatória de expressão gênica; d) Introdução de ruído aleatório seguido por cortes horizontais, descendo no dendograma até resultar na obtenção de k grupos para determinar a WADP (<i>Weighted Average Discrepant Pairs</i>) depois da perturbação.	16
2.1	Normalização de uma lâmina do conjunto de dados da Seção 6.6 usando Loess. . .	30

- 5.1 Comparação da eficiência do algoritmo K -Vizinhos em função de k , empregando validação cruzada. Na curva (a) foi usada a distância Euclideana enquanto que em (b), (c), (d) e (e) a medida de similaridade utilizada foi a Correlação. Em (c) e (d) amostras Normais e Gastrite foram consideradas do mesmo tipo. Em (d) e (e) só foram considerados os 18 genes mais DE nos pares de comparações. Já em (e) só foram levados em conta dois tipos de amostras, as Tumoriais e as não Tumoriais. 50
- 5.2 Proposta para alerta no diagnóstico. A primeira coluna representa a descrição das 99 amostras no conjunto de dados da Seção 6.8 organizadas pelos tipos, as outras cinco colunas foram usadas para representar os resultados obtidos referente às comparações. Cada uma das amostras foi classificada usando os melhores trios de genes de cada uma das cinco comparações, sendo o número de trios 100, 17, 20, 52 e 4 da comparação NT, GT, MT, NM e GM respectivamente [Tabela 6.3]. Ao lado da figura é apresentada a escala de cor empregada, a qual representa a frequência com que as amostras são classificadas, do tipo da primeira condição biológica do par, pelos trios de genes de cada comparação (100% = verde e 0% = vermelho). 61
- 5.3 Exemplo de duas populações quaisquer 1 e 2 [35], nas quais uma discriminação linear não é apropriada. Sejam f_1 e f_2 as densidade empíricas das duas populações e R_1, R_2 as referentes regiões de classificação. Também são apresentadas as curvas de contorno, assim como as curvas de separação por classificação linear e a ideal. 65
- 6.1 Genes DE no conjunto de dados da Seção 6.8 nos seis pares de comparações, considerando as quatro condições biológicas: Normal (N), Gastrite (G), Metaplasia (M) e Tumor (T), sendo 28, 21, 22 e 28 as respectivas quantidades de amostras de cada um dos tipos. Os pontos vermelhos e verdes representam os 42 genes com $|\log(pv) * \text{sinal}(\text{fold_change})| > 7$. A cor vermelha foi usada para genes com expressão baixa na primeira condição do par e alta na segunda, enquanto que a cor verde foi usada para genes com expressão alta na primeira condição do par e baixa na segunda. 71
- 6.2 Distribuição empírica do $\log_2(T)$ normalizado referente às 40 lâminas. 73

6.3	Comparação da distribuição empírica do tv do total de genes com a $N(0, 1)$	74
6.4	Histograma da distribuição empírica para z_j	76
6.5	Histograma dos pv do teste <i>Mann-Whitney</i>	76
6.6	Genes mais DE, no eixo x temos a descrição dos genes e entre parênteses é indicado C (clone) ou F (fragmento). Para cada um dos genes é apresentado o <i>boxplot</i> da distribuição empírica do $\log_2(T)$ normalizado, considerando os onze pares de lâminas.	78
6.7	Sinais medidos nas três lâminas, apenas com a correção de <i>background</i>	79
6.8	Distribuições empíricas de M , observadas nos dados brutos (sem normalizar) nas lâminas 1:1, 1:2 e 2:1.	80
6.9	Genes com melhor comportamento monotonicamente crescente nas patologias. Ao lado de cada gene tem-se o pv do teste t referente à comparação Normal-Tumor (só as tumorais tipo Intestinal), e abaixo o pv do teste de <i>Mann-Whitney</i> da mesma comparação. Acima de cada degrau tem-se o pv do teste t na comparação que ele representa. A altura de um degrau é $-\log(pv)$, na qual o pv é o referente ao teste t	83
6.10	Genes com melhor comportamento monotonicamente decrescente nas patologias. Ao lado de cada gene tem-se o pv do teste t referente à comparação Normal-Tumor (só as tumorais tipo Intestinal), e abaixo o pv do teste de <i>Mann-Whitney</i> da mesma comparação. Acima de cada degrau tem-se o pv do teste t na comparação que ele representa. A altura de um degrau é $-\log(pv)$, na qual o pv é o referente ao teste t	84
6.11	Separação realizada nas amostras pelo melhor trio de genes achado na busca exaustiva.	86
6.12	Dendograma para os 18 genes mais DE usando a Correlação como medida de similaridade e Divisão [Seção 4.2], como método de união.	87
6.13	Dendograma para os 2 genes mais DE no Normal-Tumor usando a Correlação como medida de similaridade e Divisão [Seção 4.2], como método de união.	88

6.14	<i>K</i> -Médias para (a) quatro e (b) três grupos respectivamente. São consideradas as 99 amostras, os 18 genes mais DE nos pares de comparações e é usada a Correlação como medida de similaridade.	89
6.15	Disposição das amostras nas topologias do SOM. São consideradas as 99 amostras e os 18 genes mais DE nos pares de comparações.	90
6.16	SOM para topologia (a) Retangular 1x2 e (b) Retangular 1x3 respectivamente. São consideradas as 99 amostras e os 18 genes mais DE nos pares de comparações. . .	91
6.17	SOM para topologia Retangular 1x4. São consideradas as 99 amostras e os 18 genes mais DE nos pares de comparações.	92
6.18	SOM para topologia (a) Retangular 2x2 e (b) Hexagonal 2x2 respectivamente. São consideradas as 99 amostras, os 18 genes mais DE nos pares de comparações. . . .	93

Lista de Tabelas

5.1	Avaliação dos métodos de classificação testados no conjunto de dados da Seção 6.8 usando validação cruzada. Foram considerados os 18 genes mais DE nos pares de comparações, e só duas condições biológicas: Tumoraís e não Tumoraís. No caso do K -Vizinhos a medida de similaridade empregada foi a Correlação. No caso da Regressão Local só foram usados quatro genes, os mais DE de cada um dos pares de comparações. Já no <i>Aggregating</i> não paramétrico o método de classificação usado foi o K -Vizinhos com $k = 5$, enquanto que no <i>Aggregating</i> paramétrico foi assumida a Normal Multivariada como a densidade das distribuições condicionais, da mesma forma que no método de Máxima Verossimilhança.	66
6.1	Informação dos genes mais DE, verificada no NCBI [54] e SOURCE [70].	74
6.2	Resultado obtido do ajuste do modelo linear, explicando a intensidade em um canal pela intensidade no outro, isto nas três lâminas.	80
6.3	Melhores trios de genes achados na busca exaustiva. O valor x na posição (i, j) da tabela indica que na comparação i do total de 8,510,740 classificadores, exatamente x deles classificaram, erroneamente, segundo a validação cruzada, j amostras das 99. Por exemplo, na comparação Normal-Tumor, 1044 dos 8,510,740 não erraram em nenhuma.	85

Resumo

Neste trabalho é proposto um estudo comparativo de alguns métodos de Agrupamento (Hierárquico, K -médias e *Self-Organizing Maps*) e de Classificação (K -Vizinhos, Fisher, Máxima Verossimilhança, *Aggregating* e Regressão Local), os quais são apresentados teoricamente. Tais métodos são testados e comparados em conjuntos de dados reais, gerados com a técnica de *Microarray* [64]. Esta técnica permite mensurar os níveis de expressão de milhares de genes simultaneamente, possibilitando comparações entre amostras de tecidos pelos perfis de expressão. É apresentada uma revisão de conceitos básicos relacionados ao processo de normalização, sendo este uma das primeiras etapas da análise deste tipo de conjunto de dados. Em particular, estivemos interessados em encontrar pequenos grupos de genes que fossem “suficientes” para distinguir amostras em condições biológicas diferentes. Por fim, é proposto um método de busca que, dado os resultados de um experimento envolvendo um grande número de genes, encontra de uma forma eficiente os melhores classificadores.

Abstract

In this work we propose a comparative study of some clustering methods (Hierarchic, K -Means and Self-Organizing Maps) and some classification methods (K -Neighbours, Fisher, Maximum Likelihood, Aggregating and Local Regression), which are presented teoretically. The methods are tested and compared based on the analysis of some real data sets, generated from Microarray experiments [64]. This technique allows for the measurement of expression levels from thousands of genes simultaneously, thus allowing the comparative analysis of sample of tissues in relation to their expression profile. We present a review of basic concepts regarding normalization of microarray data, one of the first steps in microarray analysis. In particular, we were interested in finding small groups of genes that were “sufficient” to identify samples originating from different biological conditions. Finally, a search method is proposed, which will find efficiently the best classifiers from the results of an experiment involving a huge number of genes.

Capítulo 1

Introdução

Um dos principais objetivos da Biologia Molecular atual é identificar os programas de expressão gênica que estão por trás de fenômenos biológicos importantes, como o crescimento e o ciclo celular, a diferenciação e o desenvolvimento, bem como as patologias decorrentes de falhas na execução destes programas [17]. Um objetivo tão amplo envolve, naturalmente, inúmeras etapas, que podem depender bastante do desenvolvimento de novas tecnologias e de abordagens interdisciplinares originais.

Da necessidade de entender como os genes determinam os processos biológicos citados acima, deriva a importância do estudo das diferenças de expressão gênica entre determinadas amostras. Em organismos eucariotos, dos quais derivam os dados analisados neste trabalho, a informação genética está organizada da seguinte forma: o DNA (ácido desoxirribonucléico), localizado no núcleo da célula, contém todas as unidades de informação (genes) que determinam as características do indivíduo (desconsiderando-se a contribuição de genes de organelas, como as mitocôndrias). No núcleo é produzida uma cópia complementar desta informação, conhecida como mRNA (ácido ribonucléico mensageiro), etapa denominada transcrição. O mRNA obtido é transportado do núcleo para o citoplasma depois para o ribossoma, sendo que no ribossoma ocorre a tradução da molécula de mRNA em proteína, etapa esta conhecida como tradução. Este processo é esquematizado na Figura 1.1. Uma forma de se medir diferenças de expressão gênica é quantifi-

1. INTRODUÇÃO

cando os mRNAs na célula ou em tecidos para diferentes estados biológicos.

Figura 1.1: Processos de Transcrição e Tradução na célula [65].

A pesquisa nesta área, na década passada, se caracterizou pelos esforços de seqüenciamento tanto de genomas completos como de ESTs (*Expressed Sequence Tags*). A lista de genomas completamente seqüenciados não pára de crescer, envolvendo vários organismos modelo, e já existem duas versões preliminares do genoma humano. Tudo isto tem gerado grandes volumes de informação que precisam ser armazenados e analisados. Surgiu, a partir deste ponto, um esforço interdisciplinar, a necessidade do desenvolvimento de novas ferramentas analíticas, da qual resultou a Bioinformática. Este termo, Bioinformática, que começou como algo restrito, referindo-se basicamente a aplicações de computação e técnicas estatísticas associadas à montagem e à análise de seqüências genômicas, aplica-se atualmente a um conjunto de idéias e técnicas muito mais amplo e difícil de precisar, situando-se em algum lugar na interface da Biologia Molecular, Matemática, Estatística, Computação e Física.

A evolução dos níveis de expressão gênica de um organismo pode ser comparada à evolução de um complexo sistema dinâmico. Uma abordagem natural procura identificar subsistemas que sejam razoavelmente simples e cuja evolução seja, em primeira aproximação, independente dos

demais [67]. Para que fosse possível comparar quantitativamente os resultados de previsões de modelos matemáticos destes sistemas com a realidade, teríamos que monitorar níveis de expressão de milhares de genes simultaneamente.

A possibilidade de medir de forma rápida e precisa as expressões dos genes de um conjunto de células tem também importantes aplicações na medicina, como por exemplo, em diagnóstico e desenvolvimento de drogas [19, 42, 55, 28]. Uma situação concreta importante, que tem sido bastante estudada, envolve o diagnóstico de câncer. Parece ser freqüente a situação do médico não ter condições de distinguir, com os dados histopatológicos usuais, entre tipos de tumores que resultam em diferentes evoluções e respostas a um determinado tratamento. Dada a agressividade de vários tumores com esta característica, é muito importante que a escolha do tratamento mais adequado seja feita corretamente e o mais cedo possível. Hoje, a possibilidade de se identificar a assinatura de cada tipo específico de patologia pela análise da expressão gênica já é uma realidade [50], [Apêndice B].

Num experimento de expressão gênica, a idéia é mensurar comparativamente os níveis de expressão para um determinado conjunto de genes, os quais podem ser escolhidos com base em conhecimento prévio destes. O experimento começa com a obtenção de amostras de tecidos da região avaliada nos pacientes. A etapa seguinte é a extração de mRNA destas amostras. Para a realização deste tipo de experimento, duas das técnicas mais usadas são *Microarray* e *SAGE* (<http://www.sagepub.com/>). Neste trabalho é introduzida a primeira delas, que foi utilizada na geração dos dados analisados neste trabalho, conforme ilustra a Figura 1.2.

1.1 Tecnologia de *Microarray*

Depois do seqüenciamento de um genoma, a técnica de *Microarray* tem sido uma das mais usadas na geração de conjuntos de dados referentes à expressão gênica. Esta técnica permite medir níveis de expressão de milhares de genes ao mesmo tempo. Atualmente, tem-se dedicado um grande esforço na padronização dos dados gerados a partir deste tipo de experimento, tanto pela sua importância como pelo fato de possibilitar que os resultados obtidos a partir de experimentos diferentes possam ser comparados. Um exemplo destes esforços é a definição de um protocolo comum conhecido como MIAME (*Minimum Information About a Microarray Experiment*) [8]. Outro exemplo que também pode ser colocado é o caso do BASE (*BioArray Software Environment*) [63], um modelo de banco de dados baseado no padrão MIAME, o qual foi desenhado para armazenar informações como: imagens, dados brutos, descrições tanto dos genes como das amostras, normalizações e outros dados relevantes na análise de expressão gênica. Tanto o padrão MIAME como o modelo BASE estão sendo estudados e analisados no nosso grupo de pesquisas, MAIGES (*Mathematical Analysis of Interacting Gene Expression Systems*).

Figura 1.2: Etapas do processo de *Microarray* [65].

A primeira etapa na preparação de um *Microarray* é a obtenção de uma lâmina de vidro, a qual é caracterizada pela sua geometria, definida basicamente pelos seguintes fatores: quantidade de *spots* e quantidade de linhas e colunas de sub-arrays, quantidade de linhas e colunas de *spots* por sub-array. Além disso, para uma lâmina específica, é conhecido o cDNA (ácido desoxirribonucléico complementar, sequência de um determinado gene) fixado em cada uma das posições por um robô que consegue a precisão necessária para separar os *spots*. Estes cDNA são selecionados de um banco de clones, previamente criado. O material fixado na lâmina é conhecido como material fixado ou *probe* [53].

Muitas vezes as quantidades de mRNAs iniciais são insuficientes para a realização dos experimentos, como por exemplo, quando se tem amostras de biópsia. Para contornar este tipo de problema são usadas reações de amplificação de mRNA. Dentro das análises realizadas, foram

comparados os resultados obtidos a partir de amostras originais e amostras obtidas por amplificação [Apêndice A].

Numa lâmina são testadas duas amostras de mRNAs, sendo que uma é marcada por fluorescência com Cy3 e a outra com Cy5, as quais fluorescem nas cores verde e vermelho, respectivamente. É comum ter-se apenas uma amostra sendo testada por lâmina, enquanto que a segunda é uma amostra de referência, comum para todas as lâminas, sendo a amostra de referência geralmente obtida a partir de uma mistura de amostras de mRNAs. Por exemplo, em um experimento testando amostras tumorais, a referência pode ser formada por uma mistura de amostras normais. Depois de ter marcado ambas as amostras, elas são misturadas em proporções iguais, e esta mistura é espalhada pela lâmina.

Neste processo, conhecido como hibridização, ocorre um pareamento das moléculas complementares, a partir do qual em cada um dos *spots* da lâmina, que referencia um certo gene, tem-se as proporções de mRNAs nas duas amostras testadas. O material usado na hibridização é conhecido como material flutuante ou *target* [53]. As proporções poderão ser usadas para estimar e comparar as diferenças nos níveis de expressão dos genes sendo analisados.

Figura 1.3: Resultado da digitalização da imagem da lâmina [32].

Após a hibridização a lâmina é digitalizada usando dois comprimentos de onda diferentes, e as imagens geradas devem ser processadas por um *software* de análise de imagens. A imagem mais conhecida que identifica esta técnica é aquela que mistura os dois canais, como visto na Figura 1.3. Do ponto de vista computacional, porém, tanto essa imagem composta como duas em níveis de cinza, cada qual representando um canal, podem ser usadas para análise. Uma prática bastante

comum, conhecida como *swap*, consiste em inverter o canal atribuído a cada uma das amostras de mRNAs, para compensar, de certa forma, o efeito provocado pelos fluorocromos.

A imagem obtida é processada computacionalmente e o desafio é quantificar um valor de intensidade para cada um dos genes analisados em cada uma das amostras de mRNAs. Existem vários aplicativos com esta finalidade [34, 3, 2, 51, 59, 32], muito embora, até o momento, parte do processo é manual. A partir da imagem para cada um dos *spots* da Figura 1.3 estima-se o sinal de intensidade e ruído em cada um dos canais.

Figura 1.4: Intensidades nos dois canais para um determinado *spot* [32].

Na Figura 1.4 apresentam-se os valores de intensidade por pixel obtidos nos dois canais para um determinado *spot*. A partir destes valores de intensidade pode-se estimar um único valor em cada canal do *spot*. Já a estimação do *background* pode ser particular para cada *spot* ou global. Na Figura 1.5 é apresentado um dos métodos de estimação das intensidades do *spot*, o qual toma o valor médio em cada um dos canais, tanto para os valores de intensidade como para os valores de *background*. Este método é bastante usado, mas outros, mais robustos, podem ser empregados contra valores de intensidade de pixels muito discrepantes dos demais, como é o caso da mediana, por exemplo. A partir deste ponto obtém-se uma tabela de dados que caracteriza a expressão dos

genes avaliados e sobre a qual será realizada toda a análise estatística.

Figura 1.5: Estimação do sinal e *background* [32].

Os detalhes da tecnologia de *Microarray* serão apresentados, para o caso particular de lâminas de vidro. Entretanto, também existem outros tipos de substratos (vidro ou nylon) que podem ser usados para a fixação dos cDNAs, como é o caso das membranas de nylon (que apresenta certas diferenças em relação às lâminas de vidro). As membranas são fisicamente maiores, assim como seus *spots* e a distância entre eles (geometria conhecida na literatura por *cDNA Array*, em vez de *Microarray*), a quantidade de *spots* é em geral menor e as superfícies são porosas, ao contrário das lâminas de vidro que são completamente lisas (rígidas). Este último detalhe é importante, pois o material fixado não fica exposto diretamente como no vidro. Nas membranas de nylon não existe marcação com fluorescência, sendo usado um marcador radioativo, razão pela qual esse tipo de experimento também é conhecido como *Microarray* de um canal, dado que somente uma amostra de mRNA pode ser testada por experimento. Neste trabalho, foram avaliados dados gerados usando os dois tipos de substratos.

O objetivo deste trabalho é analisar e comparar diferentes técnicas e abordagens matemáticas e estatísticas para problemas de análise de dados de experimentos de *Microarray*. Alguns destes problemas são: identificação de genes DE (Diferencialmente Expressos), robustez de *clusters* e

problemas de classificação. Foram usados para a criação das rotinas necessárias pacotes computacionais como R [60], S-Plus [45, 75] e MatLab [49], os quais já têm a maioria dos métodos clássicos implementados. A criação das rotinas foi baseada na experiência de análise de dados de problemas biológicos reais, gerados nas colaborações no âmbito do CAGE e de colaboradores externos.

1.2 Análise Estatística de Dados de *Microarray*

O primeiro passo na análise de dados de *Microarray* é a análise da imagem obtida por *scanner*. É preciso identificar os *spots* e quantificar o material genético expresso nas lâminas. Para localizar os *spots* é feita a segmentação da imagem. Em várias das abordagens iniciais desse problema essa segmentação era feita manualmente, num processo trabalhoso e extremamente propício à introdução de erros cujo efeito global é difícil de avaliar. Em seguida a leitura do sinal de cada *spot*, nos dois canais, deve ser feita. Nos sistemas existentes há inúmeras opções para esta quantificação, sendo que os resultados numéricos finais podem variar muito, conforme a opção. Essas diferentes opções estão associadas a diferentes suposições de cada modelagem com relação aos fenômenos bioquímicos e fontes de ruído. Por exemplo, não está claro como distinguir pixels informativos de não-informativos (ou seja, pixels correspondentes à região com *probe* ou não) e sobre como compensar a existência de ruído. Além disso, usualmente se considera importante apenas razões entre as intensidades nos dois canais, mas não está claro o que isto significa numericamente (razões entre medianas, médias, regressão linear, etc). No entanto, é fato que todas as etapas da análise matemática dos dados de *Microarray* devem ser escolhidas de forma a obter bons estimadores para os níveis de expressão gênica [76].

Dada a importância da etapa de leitura ou quantificação da imagem de *Microarray*, da qual toda a análise posterior depende, este foi um dos primeiros problemas abordados no BIOINFO/USP [2, 32]. Foi desenvolvido o protótipo de um *software* para análise de imagens de *Microarray*, que é utilizado nas imagens geradas no CAGE e por colaboradores externos. Todo o processo de segmentação é automático e há várias opções para quantificação da intensidade nos dois canais,

bem como de sua razão. Tem-se avaliado em colaboração com o prof. Luiz Fernando Reis, do Instituto Ludwig, qual método estima melhor o sinal biológico. Para isso, foi realizado um experimento preliminar [Seção 6.1] com uma lâmina especial preparada com *probes* com diluições controladas e também marcada com fluoróforo. Em seguida, *targets* artificiais, simulando aquelas provenientes das situações biológicas estudadas, foram preparados e hibridizados. O ponto importante aqui é que os sinais biológicos artificiais nas “duas situações biológicas” que estão sendo comparados são conhecidos, pois foram preparados em laboratório independentemente do experimento de *Microarray* e de forma controlada. As imagens dessas lâminas foram usadas para comparar a eficiência de diferentes métodos para análise das imagens e para estimar as quantidades de interesse.

O presente trabalho concentrou-se nas questões estatísticas relacionadas a etapa seguinte na análise dos dados de experimentos de *Microarray*, ou seja, após a análise da imagem. Esta etapa consiste na interpretação de toda a coleção de dados numéricos obtidos das imagens para a determinação dos estimadores adequados. Tipicamente haverá alguns *spots* numa mesma lâmina correspondendo ao mesmo gene e cópias em lâminas diferentes. Por diversos motivos, tais como erros sistemáticos, propriedades da lâmina, quantidade de *probe* fixado em cada *spot*, entre outros, não se espera que as leituras correspondentes sejam iguais, contudo, diferenças muito grandes podem indicar problemas experimentais. Há também efeitos sistemáticos a serem considerados como diferenças nas eficiências dos dois fluoróforos, o qual prova a importância do balanceamento no desenho dos experimentos.

Uma abordagem possível é aquela adotada por Speed T et. al. [22, 20, 79], que trata o problema por etapas, primeiro tentando remover fontes sistemáticas de variação, processo conhecido como normalização, e em seguida proceder análise de agrupamento, classificação, assim como tentar identificar genes cuja expressão variou segundo critérios estatísticos. Há várias questões interessantes sendo consideradas nesta abordagem. Por exemplo, a identificação de genes DE [22, 20, 79] é feita essencialmente por um teste de hipóteses aplicado gene a gene. Para controlar o número de falsos positivos encontrados numa abordagem como esta, o nível descritivo deve ser ajustado. Há

várias possibilidades para efetuar este ajuste [66, 77, 21], de forma que é interessante compará-los e identificar o método mais adequado ao modelo considerado.

Uma vez de posse de uma tabela com as quantificações das intensidades de cada gene, o passo seguinte, em boa parte dos trabalhos da literatura, é usualmente realizar a análise exploratória para extrair informações potencialmente úteis dos dados. A análise de agrupamento é uma das ferramentas disponíveis, que consiste em um método de estatística exploratória para agrupar os dados de acordo com uma medida de similaridade. A similaridade pode ser definida segundo uma medida de distância entre as amostras no espaço p -dimensional, sendo ambas inversamente proporcionais. Existem diferentes algoritmos de agrupamento como K -médiãs, SOM (*Self-Organizing Maps*), Hierárquico, entre outros [33, 25, 36, 73], os quais são descritos, testados em dados reais e comparados neste trabalho. No caso de estudos em séries temporais, por exemplo, o agrupamento pode ser feito para agrupar genes com padrão de expressão semelhante. Desta forma, pretende-se relacionar o padrão de expressão com a função do gene e possivelmente inferir a função de genes desconhecidos pelo seu padrão de expressão. Extremamente importante aqui é tentar avaliar o quanto os grupos obtidos são reais, isto é, fruto de alguma estrutura real presente no fenômeno em observação ou mero artefato, produto da aleatoriedade. Num dos grupos de trabalho do programa de Genoma Funcional do IPAM (*Institute for Pure and Applied Mathematics*) da Universidade de Califórnia houve considerável debate sobre como quantificar a confiança em problemas de agrupamento [6], Figura 1.6, referente a expressões gênicas de melanomas. Para analisar este problema podemos verificar a estabilidade desses aglomerados perante perturbações aleatórias compatíveis com os erros experimentais usando a técnica de *bootstrap* [18, 26, 36, 46], abordada na Seção 5.5. Além de suporte teórico, a análise deste tipo de problema requer um grande esforço computacional.

Figura 1.6: Agrupamento de dados de expressão gênica de melanoma [6]; a) Dendrograma do agrupamento hierárquico, com um grupo de 19 melanomas no centro; b) gráfico MDS (*Multidimensional Scaling*) tri-dimensional com as 31 amostras de melanomas cutâneos, apresentando o grupo principal de 19 amostras (azul, dentro do cilindro) e as 12 restantes (ouro, fora do cilindro); c) gráfico do número esperado e observado de genes produzindo um determinado número de classificações erradas para uma partição de 31 melanomas em dois grupos de 12 e 19; os triângulos vermelhos são os grupos observados, as linhas são grupos produzidos aleatoriamente e os círculos são os resultados preditos para a variável aleatória de expressão gênica; d) Introdução de ruído aleatório seguido por cortes horizontais, descendo no dendrograma até resultar na obtenção de k grupos para determinar a WADP (*Weighted Average Discrepant Pairs*) depois da perturbação.

Um problema estatístico extremamente importante em aplicações da tecnologia de *Microarray* em medicina, principalmente na pesquisa de câncer, é a questão do desenvolvimento de métodos eficientes para *Class discovery* e *Class prediction* [72, 10, 31]. Os dois são problemas de classificação,

estando o primeiro preocupado em identificar diferenças desconhecidas entre classes de objetos (tipos de câncer) descobrindo novas classes, enquanto o segundo procura classificar objetos observados (os dados de expressão de um particular paciente) em diversas categorias conhecidas (diagnóstico do câncer). Esse tipo de problema, seja de aprendizado computacional supervisionado ou não, necessita de considerável tempo para cálculo, o que pode tornar o processo inviável.

Capítulo 2

Normalização

A normalização é geralmente a segunda transformação, depois do $\log_2()$, realizada nos dados provenientes de um experimento de *Microarray* [58]. Nesta etapa é feito o ajuste das intensidades de cada hibridização individualmente ou não, para que comparações posteriores possam ser feitas. Existem várias razões pelas quais a normalização precisa ser feita, como por exemplo: colocação de quantidades diferentes de mRNAs inicial, diferenças na eficiência de detecção entre as duas substâncias fluorescentes usadas (vermelha e verde) e erros sistemáticos ao medir os níveis de expressão.

A escolha de quais métodos de normalização devem ser usados em cada caso particular depende do modelo estatístico adotado. Todas as transformações dos dados têm a finalidade última de gerar melhores estimadores para quantidades de interesse, compensando efeitos sistemáticos e devem ser justificadas pelo modelo. Naturalmente, também o modelo deve ser validado, na medida do possível, a partir dos dados experimentais. Alguns dos métodos de normalização mais usuais serão descritos neste capítulo.

A normalização pode ser feita de diferentes formas, dependendo das condições próprias de cada experimento. Existem três situações: normalização de lâminas independentes, sem estabilização da variância [Seções 2.1] e com estabilização da variância [Seções 2.2], e normalização entre pa-

2. NORMALIZAÇÃO

res de lâminas (no caso que foi empregada a inversão de corante, Seção 2.3). Em cada uma destas situações, precisamos decidir o conjunto de genes que vai ser usado para realizar a normalização. A proporção de genes que se espera ser DE nos canais vermelho e verde, assim como a disponibilidade de seqüências de DNAs de controle são questões que influenciam na decisão. A seguir apresentamos possíveis critérios para a escolha do grupo de gene: (i) todos os genes na lâmina, (ii) genes expressados constantemente e (iii) controles.

(i) Frequentemente, comparações biológicas feitas usando *Microarray* são muito específicas por natureza, ou seja, só alguns genes estão envolvidos, pelo que espera-se que uma pequena proporção dos genes possam ser DE. Conseqüentemente, para os genes restantes, espera-se que tenham expressão constante (com simetria) e possam só ser usados como indicadores da intensidade relativa dos dois canais.

(ii) Em lugar de usar todos os genes na lâmina, pode ser usado um subconjunto pequeno, chamados genes *housekeeping*, assumidos como constante em uma ampla variedade de condições. Embora seja muito difícil identificar o conjunto de genes *housekeeping* que não mudam significativamente sob quaisquer condições, podemos procurar o conjunto de genes *housekeeping* para condições experimentais particulares. A limitação de genes *housekeeping* é que estes têm tendência a ser genes com expressão alta, não representando adequadamente outros genes de interesse.

(iii) É um critério alternativo aos genes *housekeeping* usando controles *spiked*. Um controle *spiked* consiste em colocar dois genes controle, um forte e um fraco, sendo o resto dos genes utilizados para completar a amplitude existente desde um até o outro. No método de controles *spiked*, seqüências de DNAs sintético ou não de um organismo diferente do que está sendo estudado são fixadas na lâmina e incluídas em duas diferentes amostras de mRNAs em quantidades iguais. Espera-se que estas seqüências de controle fixadas tenham as mesmas intensidades nos dois canais e possam ser usadas na normalização.

A notação usada na abordagem dos diferentes métodos de normalização e suas características é a seguinte, sejam R_i e G_i as intensidades com correção do *background* do i -ésimo *spot* nos canais vermelho e verde, N_s é o número de *spots* na lâmina. $T_i = \frac{R_i}{G_i}$, $i \in \{1, 2, \dots, N_s\}$, é a razão das intensidades nos dois canais e a nossa medida de interesse. A seguir detalhamos as três circunstâncias de normalização enumeradas acima.

2.1 Normalização de lâminas independentes, sem estabilização da variância

Neste caso, a normalização é feita em cada lâmina, independentemente das demais, usando somente as intensidades nos dois canais da lâmina analisada.

2.1.1 Normalização pela energia total

Este método assume que as distribuições das intensidades (canais vermelho e verde) são proporcionais, ou seja, $R = kG$ (na qual $k > 0$) e o centro da distribuição do $\log_2(T)$ é deslocado para zero.

$$\log_2(T'_i) = \log_2\left(\frac{R'_i}{G'_i}\right) = \log_2\left(\frac{R_i}{kG_i}\right) = \log_2\left(\frac{R_i}{G_i}\right) - \log_2(k) = \log_2\left(\frac{R_i}{G_i}\right) - c,$$

o valor de c pode ser estimado como segue:

- a) $\hat{c} = \text{Média}\left(\log_2\left(\frac{R_i}{G_i}\right)\right)$, pela média
- b) $\hat{c} = \text{Mediana}\left(\log_2\left(\frac{R_i}{G_i}\right)\right)$, pela mediana
- c) $\hat{c} = \text{MédiaAparada}\left(\log_2\left(\frac{R_i}{G_i}\right)\right)$, pela média aparada.

O mais comum é utilizar a “média” [24, 68, 65], por considerar todos os elementos, porém usar a “mediana” pode ser interessante, por exemplo, quando temos a presença de *outliers* e não queremos incluí-los na nossa função normalizadora, obtendo-se desta forma uma normalização mais robusta.

2.1. NORMALIZAÇÃO DE LÂMINAS INDEPENDENTES, SEM ESTABILIZAÇÃO DA VARIÂNCIA

Já a “média aparada” pode ser usada para excluir uma porcentagem dos pontos extremos, podendo ser interpretada como uma “média” mais robusta.

Esses métodos de normalização são conhecidos como métodos de normalização global e são considerados como parte do pre-processamento dos conjuntos de dados na identificação de genes DE em experimentos de *Microarray* com lâminas independentes. Em alguns aplicativos a constante c é calculada pela média aritmética do $\log_2(T)$, como, por exemplo, no aplicativo *GenePix* [30], sendo estes métodos de normalização global os mais comumente usados. Estas normalizações globais não são adequadas em situações onde a medida de interesse, $M = \log_2(T) = \log_2(R/G) = \log_2(R) - \log_2(G)$, possa depender (fonte de variação sistemática) da média das intensidades nos dois canais $A = \log_2(\sqrt{RG}) = \frac{1}{2}(\log_2(R) + \log_2(G))$, nem em ocasiões nas quais existam padrões geométricos, ou seja, que em algumas regiões da lâminas algum canal ascendeu mais que em outras.

2.1.2 Normalização com dependência da intensidade

Em certas ocasiões a polarização das substâncias fluorescentes fica dependente da intensidade do *spot*, o qual pode ser confirmado no gráfico $M = \log_2(R/G)$ vs $A = \log_2(\sqrt{RG})$. Nestes casos, a normalização global não é indicada, sendo que deve ser usado um método mais apropriado, que considere esta dependência.

$$\log_2(T'_i) = \log_2\left(\frac{R'_i}{G'_i}\right) = \log_2\left(\frac{R_i}{k(A)G_i}\right) = \log_2\left(\frac{R_i}{G_i}\right) - \log_2(k(A)) = \log_2\left(\frac{R_i}{G_i}\right) - c(A),$$

na qual $c(A)$ pode ser estimado através do ajuste Loess [Seção 2.5.1] do gráfico M vs A . A função *loess()* [60] realiza o ajuste local e robusto, o qual não é afetado por uma porcentagem pequena de genes DE e que aparecem como *outliers* no gráfico M vs A . Sendo definido pelo usuário a porcentagem de pontos usados na suavização de cada ponto, definido como 40% no pacote Bioconductor implementado no R [60].

2.1.3 Normalização com dependência da agulha

Cada sub-matriz da lâmina é preenchida por uma agulha na etapa de fixação dos cDNAs no substrato utilizado, e sabe-se que podem existir diferenças entre as agulhas, tais como: comprimento, abertura e deformações depois de horas de uso.

$$\log_2(T'_i) = \log_2\left(\frac{R'_i}{G'_i}\right) = \log_2\left(\frac{R_i}{k_j(A)G_i}\right) = \log_2\left(\frac{R_i}{G_i}\right) - \log_2(k_j(A)) = \log_2\left(\frac{R_i}{G_i}\right) - c_j(A),$$

na qual: $c_j(A)$ pode ser estimado através do ajuste *Loess* do gráfico M vs A da j -ésima grade, $j \in \{1, 2, \dots, I\}$, sendo I o número de agulhas.

2.2 Normalização de lâminas independentes, com estabilização da variância

Depois de realizar a normalização por grupos de *spots* referentes a cada agulha, o centro da distribuição do $\log_2(T)$ para cada um deles fica deslocado para zero. Entretanto, é possível que as distribuições do $\log_2(T)$ tenham variabilidades diferentes, de modo que um ajuste de escala é requerido. Uma possível aproximação é assumir que $(\log_2(T))^{(i)} \sim N(0, a_i^2 \sigma^2)$, ou seja, o $\log_2(T)$ referente aos *spots* preenchidos pela i -ésima agulha tem distribuição Normal com parâmetros especificados acima, sendo σ^2 a variância da distribuição $\log_2(T)$ referente aos *spots* da lâmina toda e a_i^2 é o fator de escala da i -ésima agulha. Para realizar a normalização de escala, precisamos estimar os fatores de escala a_i , $i \in \{1, 2, \dots, I\}$. É lógico esperar que os valores para a_i^2 fiquem próximos de um, fazendo com que os valores para $\log_2(a_i^2)$ fiquem próximos de zero. Então, podemos colocar a seguinte condição de indentificabilidade, $\sum_{i=1}^I \log_2(a_i^2) = 0$, obtendo-se desta forma o estimador de máxima verossimilhança para a_i ,

$$\hat{a}_i = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{\sqrt[I]{\prod_{k=1}^I \sum_{j=1}^{n_i} M_{kj}^2}},$$

2.3. NORMALIZAÇÃO PARA PARES DE LÂMINAS NAS QUAIS FOI EMPREGADA INVERSÃO DE CORANTE

na qual: M_{ij} é o j -ésimo $\log_2(T)$ da i -ésima agulha, $j \in \{1, 2, \dots, n_i\}$, n_i é o número de *spots*, que são preenchidos pela agulha i . Uma alternativa robusta para este estimador pode ser definida usando o MAD (Desvio Absoluto da Mediana),

$$\hat{a}_i = \frac{MAD_i}{\sqrt[I]{\prod_{k=1}^I MAD_k}},$$

na qual $MAD_i = \text{Mediana}_j(|M_{ij} - \text{Mediana}_j(M_{ij})|)$.

Este procedimento assume que uma proporção relativamente pequena de genes serão DE entre as duas amostras de mRNAs. Também assume que a variabilidade da distribuição $\log_2(T)$ seja aproximadamente igual para todas as agulhas. A estatística robusta MAD, da mesma forma que o Loess, não será afetada por uma pequena porcentagem de genes DE, os quais aparecem como *outliers* no gráfico M vs A .

2.3 Normalização para pares de lâminas nas quais foi empregada inversão de corante

Esta normalização é aplicada quando temos experimentos com troca das substâncias fluorescentes, ou seja, são realizadas duas hibridizações para duas amostras de mRNAs, de tal forma que na segunda hibridização os corantes atribuídos aos mRNAs são trocados.

Sejam $\log_2(R^{(1)}/G^{(1)}) - c^{(1)}$ e $\log_2(R^{(2)}/G^{(2)}) - c^{(2)}$ o $\log_2(T)$ normalizado para a primeira e segunda lâmina, nas quais $c^{(1)}$ e $c^{(2)}$ representam as funções de normalização definidas na Seção 2.1.1 para as duas lâminas. Assumindo que estas duas funções são aproximadamente iguais, segue que

$$\begin{aligned} \frac{1}{2}(M^{(1)} - M^{(2)}) &= \frac{1}{2} \log_2 \left(\frac{R^{(1)}G^{(2)}}{G^{(1)}R^{(2)}} \right) = \frac{1}{2} \left(\log_2 \left(\frac{R^{(1)}}{G^{(1)}} \right) + \log_2 \left(\frac{G^{(2)}}{R^{(2)}} \right) \right) \\ &\approx \frac{1}{2} \left(\log_2 \left(\frac{R^{(1)}}{G^{(1)}} \right) - c^{(1)} - \left(\log_2 \left(\frac{R^{(2)}}{G^{(2)}} \right) - c^{(2)} \right) \right). \end{aligned}$$

Desta forma, podemos combinar os níveis de expressão relativa para as duas lâminas sem que as funções de normalização apareçam explicitamente. Este processo é conhecido como auto-normalização e pode ser aplicado a todos os genes, incluindo os DE. Os genes para os quais $\frac{1}{2}(M^{(1)} - M^{(2)}) = \frac{1}{2} \log_2 \left(\frac{R^{(1)} G^{(2)}}{G^{(1)} R^{(2)}} \right) \approx 0$, independem da atribuição dos corantes. Para validar esta consideração, podemos usar um conjunto de genes para os quais se esperam níveis de expressão constantes. Até o momento definimos que,

$$\log_2 \left(\frac{R^{(1)}}{G^{(1)}} \right) - c^{(1)} \approx - \left(\log_2 \left(\frac{R^{(2)}}{G^{(2)}} \right) - c^{(2)} \right),$$

de forma que podemos estimar a função normalizadora,

$$c \approx \frac{1}{2} \left(\log_2 \left(\frac{R^{(1)}}{G^{(1)}} \right) + \log_2 \left(\frac{R^{(2)}}{G^{(2)}} \right) \right) = \frac{1}{2} \log_2 \left(\frac{R^{(1)} R^{(2)}}{G^{(1)} G^{(2)}} \right) = \frac{1}{2} (M^{(1)} + M^{(2)})$$

Na prática, $c_i = c_i(A)$ é estimada pelo ajuste Loess do gráfico $\frac{1}{2}(M^{(1)} + M^{(2)})$ vs $\frac{1}{2}(A^{(1)} + A^{(2)})$, isto é, usando todos os genes.

2.4 Normalização composta

Podem existir casos em que faz sentido pensar em compensar mais de uma fonte de ruído ao mesmo tempo. Nestes casos a Normalização composta é uma alternativa, a qual funciona mediante uma ponderação de efeitos. Um exemplo disto pode ser observado quando queremos compensar o efeito provocado pelas agulhas [Seção 2.1.3]. Para uma determinada agulha, a curva de normalização composta é a média ponderada da curva do Loess [Seção 2.5.1] referente ao MSP (*Microarray Sample Pool*) e a curva do Loess dos genes colocados por ela. Os pesos dependem do número acumulado de genes em diferentes níveis de intensidade de A . A seguir, é apresentado um esboço do procedimento para os *spots* da i -ésima agulha:

- i) Estimar o ajuste do Loess ao gráfico M vs A da i -ésima agulha, $\hat{f}_i(A)$;

ii) Estimar o ajuste do Loess ao gráfico M vs A dos *spots* do MSP, $\hat{g}(A)$;

iii) Calcular a média ponderada, $c_i = \alpha_A \hat{g}(A) + (1 - \alpha_A) \hat{f}_i(A)$,

na qual α_A é a proporção dos genes que ficam abaixo de um determinado valor para A .

2.5 Normalização usando métodos locais e robustos

Esta normalização é proposta para eliminar certa dependência que costuma aparecer no $\log_2(T)$ pela intensidade, principalmente observada em *spots* com intensidade baixa. Inicialmente, precisamos observar o gráfico M vs A . O Loess detecta desvios sistemáticos apresentados pelo gráfico M vs A , corrige-os fazendo regressão local ponderada, como a função A , e subtrai o melhor ajuste calculado para a média de M sob os dados observados experimentalmente.

2.5.1 Lowess-Loess, (*Locally Weighted Smoothing Scatterplot*)

Existem ocasiões, na normalização, onde métodos locais e robustos [16] são necessários. Isto pode ser verificado no gráfico M vs A , o qual pode ser melhorado por suavização. A idéia central de suavização pode variar. Por exemplo, os pontos podem ser agrupados pelos valores de x_i e construir um gráfico de $\bar{y}_i$ vs $\bar{x}_i$ para cada grupo. Outra possibilidade é interpolar o gráfico por sucessivas uniões de pontos com linhas retas e suavizar por convolução usando uma função de peso. A seguir, é apresentado um destes métodos, o Lowess-Loess, o qual foi desenhado para ajustar os dados por suavização,

$$y_i = g(x_i) + \varepsilon_i,$$

na qual: g é a função de suavização, ε_i ($i \in \{1, \dots, n\}$) são variáveis aleatórias independentes com média zero e escala constante. Note que $\{\varepsilon_i\}_{i=1}^n$ representa uma série de ruído branco. Então temos que $\hat{y}_i$, valor ajustado para x_i , é uma estimativa para $g(x_i)$. A consideração de suavização permite que pontos em uma vizinhança (de acordo com f , sendo $100 * f$ a porcentagem do total de pontos usados para suavizar cada ponto, correspondente ao conjunto de pontos mais próximos) de x_i sejam usados para calcular $\hat{y}_i$.

Para uma função de peso $W(x)$, a qual decresce com o aumento da distância, os pesos $w_k(x_i)$ diminuem quando a distância entre x_k e x_i aumenta. Desta forma, os pontos com x próximo de x_i , têm um papel importante na determinação de $\hat{y}_i$, enquanto que pontos mais afastados ficam com um papel secundário. Com o aumento de f ($0 < f \leq 1$), aumenta a vizinhança dos pontos de influência e conseqüentemente tende a aumentar a suavização dos pontos. Seja $d_i(\cdot, \cdot)$ a função de distância empregada, então os pesos são definidos da seguinte forma,

$$w_k(x_i) = W(d_i^{-1}(x_i, x_k)), k \in \{1, \dots, n\}.$$

Note que a função de peso leva em conta o inverso da distância, ou seja, a proximidade, de tal forma que é atribuído um peso maior aos pontos mais próximos. A seguir, o algoritmo de suavização é apresentado:

(1) Calculam-se as estimativas $\hat{\beta}_i(x_i)$, com $i \in \{0, \dots, d\}$, dos parâmetros da regressão polinomial de grau d de y_k em x_k , a qual é ajustada por mínimos quadrados ponderados com peso $w_k(x_i)$ para (x_i, y_i) . Desta forma, $\hat{\beta}_j(x_i)$ são os valores de β_j que minimizam,

$$\sum_{k=1}^n w_k(x_i) (y_k - \beta_0 - \beta_1 x_k - \dots - \beta_d x_k^d)^2.$$

O ponto suavizado em x_i , usando Loess de grau d , é $(x_i, \hat{y}_i)$, na qual $\hat{y}_i$ é o valor ajustado da regressão em x_i . Então,

$$\hat{y}_i = \sum_{j=0}^d \hat{\beta}_j(x_i) x_i^j = \sum_{k=1}^n r_k(x_i) y_k,$$

na qual $r_k(x_i)$ não depende de y_j , $j \in \{1, \dots, n\}$.

(2) Calcula-se o peso robusto do k -ésimo ponto no i -ésimo ponto da seguinte forma:

$$\delta_k^{(i)} = W\left(\frac{e_i}{6s}\right),$$

na qual: W é a função de peso, $e_i = y_i - \hat{y}_i$, o resíduo do atual valor ajustado, $s = \text{Mediana}(|e|)$ e $i, k \in \{1, \dots, n\}$.

(3) Determina-se o novo $\hat{y}_i$ para cada i por ajuste do polinômio de grau d usando mínimos quadrados ponderados com peso $\delta_k w_k(x_i)$ em (x_k, y_k) .

(4) Repete-se (2) e (3) N vezes. Os valores finais para $\hat{y}_i$ são os valores ajustados pelo Lowess robusto.

Note que têm-se cinco parâmetros de interesse: (i) d , o grau do polinômio empregado no ajuste; (ii) f , sendo $100 * f$ a porcentagem do total de pontos usados para suavizar cada ponto, correspondente ao conjunto de pontos mais próximos; (iii) N , o número de vezes que o algoritmo é iterado; (iv) W , a função de peso; (v) $d(\cdot, \cdot)$, a função de distância.

Este método não paramétrico é uma extensão do ajuste local de polinômio e pode ser implementado a um custo computacional baixo. O valor ajustado em x_i é o valor do ajuste polinomial dos dados usando mínimos quadrados ponderados, no qual o peso do ponto (x_k, y_k) é maior se x_i e x_k ficam próximos e menor se ficam afastados, nas quais $k \in \{1, \dots, n\}$. A seguir, são apresentadas as propriedades das funções de peso usadas. Seja W uma função de peso, então,

1. $W(x) > 0$, se $|x| < 1$;
2. $W(x) = 0$, se $|x| \geq 1$;
3. $W(-x) = W(x)$;
4. $W(x)$ é não crescente para $x \geq 0$.

2.5. NORMALIZAÇÃO USANDO MÉTODOS LOCAIS E ROBUSTOS

Note que a função de peso tem suporte no intervalo $(0, 1)$, definindo o conjunto dos pontos que vão ser considerados na suavização de cada um dos pontos. Além disto, é uma função par, com simetria no eixo $x = 0$. A seguir são apresentados alguns exemplos de funções de peso:

a) função de peso biquadrada,

$$B(x) = \begin{cases} (1 - x^2)^2, & \text{se } |x| < 1 \\ 0, & \text{se } |x| \geq 1 \end{cases}.$$

b) função de peso tricúbica,

$$T(x) = \begin{cases} (1 - |x|^3)^3, & \text{se } |x| < 1 \\ 0, & \text{se } |x| \geq 1 \end{cases}.$$

A condição do procedimento é a seguinte: para cada x_i são definidos os pesos $w_k(x_i)$, $k \in \{1, \dots, n\}$ usando a função de peso. Isto é realizado centrando W em x_i e escalando-a de tal forma que os pontos nos quais toma valor zero seja a partir do r -ésimo vizinho mais próximo de x_i . O valor inicial ajustado $\hat{y}_i$, para cada x_i , é calculado pelo ajuste de polinômio de ordem d usando mínimos quadrados ponderados com peso $w_k(x_i)$. Este procedimento para cálculo do valor ajustado inicial é referenciado como Loess. Seja δ_i o conjunto de pesos definidos para o i -ésimo ponto (x_i, y_i) usando os resíduos $(y_i - \hat{y}_i)$, atribuindo pesos pequenos aos resíduos maiores e grandes aos menores. Os novos valores ajustados são calculados, como já vimos anteriormente, trocando-se $w_k(x_i)$ por $\delta_i w_k(x_i)$. O cálculo dos novos pesos e valores ajustados é repetido N vezes. O procedimento completo, incluindo o cálculo inicial e as iterações, é referenciado como Lowess robusto.

2.5.2 Normalização utilizando Loess

Sejam $x_i = A_i$, $y_i = M_i$ e $g(x_i)$ a primeira estimativa para y_i pelo Loess [Seção 2.5.1], representando a dependência que existe entre M e A , nas quais i corresponde com o índice do *spot*, $i \in \{1, \dots, N_s\}$ e N_s é o número de *spots* na lâmina. Usando esta função para cada um dos *spots* da lâmina é realizada a normalização das estimativas obtidas experimentalmente para M ,

$$\begin{aligned}\log_2(T'_i) &= \log_2(T_i) - g(x_i) = \log_2(T_i) - \log_2(2^{g(x_i)}) \\ &= \log_2\left(T_i \frac{1}{2^{g(x_i)}}\right) = \log_2\left(\frac{R_i}{G_i} \frac{1}{2^{g(x_i)}}\right).\end{aligned}$$

Na Figura 2.1 é apresentado um exemplo de lâmina do conjunto de dados da Seção 6.6 normalizada usando Loess.

Figura 2.1: Normalização de uma lâmina do conjunto de dados da Seção 6.6 usando Loess.

2.6 Comparações entre normalizações

Os diferentes métodos de normalização são comparados pela sua habilidade na identificação de genes, dos quais já se conhece que são DE. Um bom método é aquele que cria uma clara diferenciação entre genes DE e genes com expressão constante. Esta diferenciação pode ser checada pelo tv (t-valor), obtido do teste t, assim como pelo pv (p-valor) obtido por algum teste, como exemplo *Mann-Whitney*, t, etc.

Capítulo 3

Modelo para expressão gênica

Neste capítulo é apresentado um modelo para os níveis de expressão gênica [61], que estamos interessados em validar utilizando os conjuntos de dados analisados neste trabalho. As questões de normalização [Capítulo 2] são justificadas pelo modelo e pelas considerações tomadas, de forma que cada um dos métodos de normalização pode ser confrontado ao modelo.

Algumas técnicas estatísticas clássicas freqüentemente assumem que os dados são normalmente distribuídos, como é o caso dos classificadores por Máxima Verossimilhança [Seção 5.4], onde é assumido que as distribuições condicionais para cada um dos tipos de amostra são normais multivariadas. Além disto, também é assumida muitas vezes a homocedasticidade (variâncias constantes, ou seja, não dependem da média dos dados).

Normalmente, dados que não satisfazem estas premissas são transformados de modo que as variáveis transformadas atendam estes quesitos. No caso particular de dados de *Microarray*, que têm uma complicada estrutura de erro, existe uma dependência não linear entre a variância e a média dos dados (réplicas do mesmo gene em *spots* diferentes de uma lâmina). Por esta razão, surge a necessidade de procurar uma transformação que estabilize a variância dos dados de *Microarray* em todo o intervalo da expressão.

Muitas metodologias estatísticas, tais como regressão ou ANOVA [11], assumem que os dados

são normalmente distribuídos ou pelo menos distribuídos simetricamente, com variância constante e independente da média dos dados. No caso que os dados não satisfaçam tais hipóteses podem ser escolhidas técnicas alternativas ou realizar transformações nos dados, sendo a segunda alternativa a mais comumente usada. Os dados de *Microarray*, que medem os níveis de expressão de milhares de genes simultaneamente podem produzir uma quantidade de informação muito grande acerca da biologia também analisada estatisticamente. No entanto, os dados de *Microarray* não se ajustam nas considerações necessárias para uma análise usando técnicas clássicas. A Seção 3.1 traz um modelo para dados de *Microarray* [61].

3.1 Notação do modelo

Uma proposta para o modelo de expressão gênica é a seguinte:

$$y = \alpha + \mu e^{\eta} + \varepsilon ,$$

na qual: y é o nível de expressão bruta observada em um *spot* referente a um canal, α é a média do ruído de fundo (*background*), e μ é o verdadeiro nível de expressão do gene fixado no *spot* para o canal utilizado. As distribuições assumidas para as fontes de erro são as seguintes, $\eta \sim N(0, \sigma_{\eta}^2)$ e $\varepsilon \sim N(0, \sigma_{\varepsilon}^2)$, sendo ε o erro associado ao *background* e que terá maior influência nos genes sub-expressos, enquanto η , o erro multiplicativo que sempre existe, terá maior influência nos genes super-expressos. Neste modelo de efeitos aleatórios misto assume-se que η e ε são independentes.

Podem ser consideradas três condições para y . A primeira é expressão baixa, ou seja, genes sub-expressos, na qual temos $\mu \approx 0$, o que faz com que $y \approx \alpha + \varepsilon$, desta forma $y \sim N(\alpha, \sigma_{\varepsilon}^2)$. Nesta situação encontramos medições que geralmente são descartadas, uma vez que o sinal é menor do que o *background*, o que impede a correção tradicional do *background*, que gera valores negativos, caso em que deve ser aplicada a transformação $\log_2()$. Neste capítulo apresentamos a correção tradicional do *background* e duas alternativas muito interessantes pelo fato de que nenhuma informação é desconsiderada. Na segunda condição, ao contrário da primeira, observam-se

genes super-expressos, embora não estejamos na presença de saturação, na qual o intervalo de valores possíveis é excedido. Os valores para μ são muito grandes, o qual faz com que $y \approx \mu e^\eta$. Note que $e^\eta \sim \text{LogNormal}(0, \sigma_\eta^2)$, logo $\text{Var}(y) \approx \mu^2 \text{Var}(e^\eta)$, na qual,

$$\begin{aligned} \text{Var}(e^\eta) &= \exp(2\mathbb{E}(\eta) + 2\text{Var}(\eta)) - \exp(2\mathbb{E}(\eta) + 2\text{Var}(\eta)) \\ &= e^{2\sigma_\eta^2} - e^{\sigma_\eta^2} = e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1). \end{aligned}$$

Desta forma, temos $\text{Var}(y) = \mu^2 e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)$, que apresenta a dependência existente entre a variância e a média dos dados. No caso do desvio padrão, esta dependência é linear, $DP(y) = \mu \sqrt{e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)}$.

Já vimos os dois casos extremos, genes sub-expressos e genes super-expressos. No terceiro caso, para valores médios de μ , temos que,

$$y = \alpha + \mu e^\eta + \varepsilon, \text{ e neste caso,}$$

$$\text{Var}(y) = \mu^2 \text{Var}(e^\eta) + \sigma_\varepsilon^2, \text{ considerando a independência entre } \eta \text{ e } \varepsilon.$$

No modelo acima, a distribuição do erro da medida muda dependendo de μ , criando a necessidade de se realizar uma transformação no conjunto de dados antes de aplicar as metodologias estatísticas clássicas.

3.2 Correção do *background*

A seguir são apresentados alguns possíveis métodos usados na correção de *background*. Sejam x_{ij}^s e x_{ij}^b as medida referentes ao i -ésimo *spot* no j -ésimo canal para o sinal e o *background*, respectivamente.

a) Método mais comumente usado na literatura [24, 68, 65],

3.2. CORREÇÃO DO *BACKGROUND*

$x_{ij} = x_{ij}^s - x_{ij}^b$, neste caso são descartados os *spots*, com $x_{ij}^s \leq x_{ij}^b$.

b) Método alternativo truncado,

$$x_{ij} = \begin{cases} x_{ij}^s - x_{ij}^b, & \text{se } x_{ij}^s - x_{ij}^b > \delta \\ \delta, & \text{se } x_{ij}^s - x_{ij}^b \leq \delta \end{cases},$$

na qual δ é um parâmetro a ser definido, que indica a diferença mínima entre o sinal e o *background*, abaixo da qual o valor de intensidade será truncado. Este método alternativo resolve o problema de perda de informação, visto em (a), mas implica na mesma estimativa para o sinal em todos os *spots* com sinal menos intenso.

c) Método alternativo [24],

$$x_{ij} = \begin{cases} x_{ij}^s - x_{ij}^b, & \text{se } x_{ij}^s - x_{ij}^b > \delta \\ \delta \exp \left\{ 1 - \frac{x_{ij}^b + \delta}{x_{ij}^s} \right\}, & \text{se } x_{ij}^s - x_{ij}^b \leq \delta \end{cases},$$

na qual δ é um parâmetro a ser definido, que indica a diferença mínima entre o sinal e o *background*, abaixo da qual o valor de intensidade será estimado de forma alternativa.

Note que no método alternativo (c), valores de sinal que não sejam baixos em relação ao *background* são tratados da mesma forma que em (a). Já para valores de sinal próximos do *background*, iguais ou menores, o método funciona de forma diferente, que aproxima uma correção exponencial, decrescendo com a diminuição da diferença $x_{ij}^s - x_{ij}^b$, sendo porém sempre positiva. Desta forma, o método alternativo não despreza nenhuma informação e ainda consegue manter o *rank*, ou seja, a ordenação existente entre eles em cada uma das lâminas.

3.3 Análise assintótica da variância

Depois de realizada a correção do *background*, a transformação mais usada é o $\log_2()$, para a qual vamos analisar o comportamento assintótico da variância. Seja $AV(x)$ a variância assintótica da variável aleatória x , então segue que $AV(\log(\hat{\mu})) = \sigma_\eta^2 + \sigma_\varepsilon^2/\mu^2$, desta forma temos que a variância assintótica do $\log(\hat{\mu})$ é praticamente constante para valores grandes de μ , mas toma valores muito grandes quando μ é pequeno. Isto considerando $\log(\hat{\mu}) = \log(y - \hat{\alpha})$. Note que a transformação $\log_2()$, depois de feita a correção do *background*, tem um bom comportamento para valores de μ grandes, já nos pequenos a variabilidade fica muito aumentada. É por isto que tirar as observações com sinal próximo do *background* pode ser uma boa solução, apesar de estarmos descartando parte da informação. Por outro lado, num experimento com muitas lâminas, próximo de 100, por exemplo, esta solução não será adequada, pois geralmente os *spots* com sinal menos intenso variam entre as lâminas. Desta forma, tirando-os da análise em cada uma das lâminas, pode representar uma perda significativa de informação.

3.4 Transformação proposta para estabilizar a variância

A idéia é procurar a transformação que estabilize a variância em todo o intervalo da expressão nos dados obtidos num experimento de *Microarray*. Seja $g(\cdot)$ uma função de suavização, então teremos para $y \xrightarrow{p} \theta$, $AV(g(y)) = g'(\theta)^2 \text{Var}(y)$,

Deseja-se achar a transformação $g(\cdot)$ para $y = \alpha + \mu e^\eta + \varepsilon$, tal que $AV(g(y))$ seja constante. Desta forma temos que,

$AV(g(y)) = g'(\mu + \alpha)^2 \text{Var}(y) = k$, na qual k é uma determinada constante, segue que,

$$(g'(\mu + \alpha))^2 = \frac{k}{\text{Var}(y)} = \frac{k}{\mu^2 \text{Var}(e^\eta) + \sigma_\varepsilon^2}$$

$$\iff g'(\mu + \alpha) = \frac{\sqrt{k}}{\sqrt{\mu^2 \text{Var}(e^\eta) + \sigma_\varepsilon^2}}$$

$$\Leftrightarrow g'(y) = \frac{\sqrt{k}}{\sqrt{(y - \alpha)^2 \text{Var}(e^\eta) + \sigma_\varepsilon^2}}, \text{ na qual } y = \mu + \alpha$$

$$\Leftrightarrow \int g'(y) dy = \int \frac{\sqrt{k}}{\sqrt{(y - \alpha)^2 \text{Var}(e^\eta) + \sigma_\varepsilon^2}} dy.$$

Sendo $g(y) = \log \left(y - \alpha + \sqrt{(y - \alpha)^2 + c} \right)$, uma solução da equação anterior, na qual $c = \frac{\sigma_\varepsilon^2}{\sigma_\eta^2}$. Obtém-se desta forma que, $AV(y^*) = \text{Var}(e^\eta)$, na qual $y^* = g'(y)$ é o dado transformado.

3.5 Estimação dos parâmetros do modelo

Para aplicar o modelo apresentado na Seção 3.1 é preciso estimar os parâmetros do mesmo. Note porém, que há vários casos possíveis de desenho de experimento para um determinado conjunto de dados de *Microarray*. Primeiramente, é importante definir se no conjunto de dados analisado foram usadas lâminas de vidro ou membranas de nylon, para as quais temos dois e um canal, respectivamente. Outro elemento a considerar é o fato de existirem ou não espaços em branco, também conhecido como controle negativo, isto é, *spots* nos quais foram fixados genes que não existem nas amostras de mRNAs testadas, como por exemplo: genes de outro organismo. Finalmente, é necessário saber se dentro de cada lâmina temos ou não réplicas, ou seja, mais de uma cópia para cada gene. Note que estes três parâmetros são independentes uns dos outros e para cada um deles temos duas possibilidades, o que gera oito configurações possíveis e, como na maioria dos casos, o desenho da lâmina é improvisado e não planejado, em uma análise podemos estar na presença de qualquer uma das oito configurações. A seguir apresentamos como o modelo é adaptado a cada uma delas.

3.5.1 Estimação do *background* usando espaços em branco

A idéia proposta neste caso é deixar espaços em branco na lâmina, também conhecido como controle negativo, isto é, *spots* nos quais foram fixados genes que não existem nas amostras de mRNAs testadas, como por exemplo: genes de outro organismo. As duas medições para cada um dos

3.5. ESTIMAÇÃO DOS PARÂMETROS DO MODELO

espaços em branco serão usadas para estimar α e σ_ϵ . Embora este método pareça uma boa opção, ele exige planejamento no desenho da lâmina. É sabido que muitas vezes as lâminas são preenchidas de modo a maximizar a quantidade de genes testados e/ou a quantidade de réplicas de cada um deles; neste caso, o método não poderá ser usado. Por outro lado, para que as estimativas sejam confiáveis é necessário usar um número suficientemente grande de espaços em branco, eliminando uma porcentagem considerável de *spots* em cada uma das lâminas. As estimativas dos parâmetros, usando somente os espaços em branco, são calculadas da seguinte forma,

$$\hat{\alpha} = \bar{x}, \text{ e } \hat{\sigma}_\epsilon = S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

na qual n é o número de espaços em branco.

3.5.2 Estimação do *background* com réplicas

Para estimar o *background* podem ser usados os genes com sinal menos intenso, ou seja, aqueles com baixo nível de expressão. A seguir é apresentado um algoritmo iterativo [61] que permite achá-los,

- (i) É escolhido um subconjunto de genes, formado pelos 10% com sinal menos intenso;
- (ii) Estima-se α e σ_ϵ , média e desvio padrão, usando somente o subconjunto de genes selecionado no passo anterior pelas expressões,

$$\hat{\alpha} = \bar{x}, \text{ e } \hat{\sigma}_\epsilon = S = \sqrt{\frac{1}{n-m} \sum_{i=1}^m (n_i - 1) S_i^2},$$

na qual: m é 10% da quantidade total de genes fixados na lâmina, n_i é o número de réplicas do i -ésimo gene, $n = \sum_{i=1}^m n_i$ e $S_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$;

- (iii) O subconjunto de genes com sinal menos intenso é então redefinido, constando apenas os genes que possuíam intensidade no intervalo $[\hat{\alpha} - 2\hat{\sigma}_\epsilon, \hat{\alpha} + 2\hat{\sigma}_\epsilon]$;
- (iv) Repetir (ii) e (iii) até que o subconjunto de genes selecionado fique constante.

3.5.3 Estimação do *background* sem réplicas

Quando não temos réplicas, a idéia da estimação é similar à apresentada na Seção 3.5.2, porém mais simples. A seguir é apresentado o algoritmo iterativo [61] para a busca dos genes com sinal menos intenso na ausência de réplicas,

- (i) É escolhido um subconjunto de genes, formado pelos 10% com sinal menos intenso;
- (ii) Estima-se α e σ_ε , média e desvio padrão, usando somente o subconjunto dos genes com sinal menos intenso,

$$\hat{\alpha} = \bar{x}, \text{ e } \hat{\sigma}_\varepsilon = S = \sqrt{\frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2};$$

- (iii) O subconjunto de genes com sinal menos intenso é então redefinido, constando apenas os genes que possuíam intensidade no intervalo $[\hat{\alpha} - 2\hat{\sigma}_\varepsilon, \hat{\alpha} + 2\hat{\sigma}_\varepsilon]$;
- (iv) Repetir (ii) e (iii) até que o subconjunto de genes selecionado fique constante.

3.5.4 Previsão do erro multiplicativo

A proposta para prever η consiste em usar os genes super-expressos da mesma forma que foram usados os sub-expressos na estimação do *background*. Na Seção 3.1 foi visto que para genes super-expressos, $y \approx \mu e^\eta$, desta forma tem-se $\log(y) \approx \log(\mu) + \eta$. É por isto que a expressão para calcular a estimativa do desvio padrão de η é

$$\hat{\sigma}_\eta = S = \sqrt{\frac{1}{n-m} \sum_{i=1}^m (n_i - 1) S_i^2},$$

para o caso onde tem-se réplicas dos genes na lâmina e

$$\hat{\sigma}_\eta = S = \sqrt{\frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2},$$

para quando não temos réplicas dos genes na lâmina. Os genes super-expressos são encontrados de

3.5. ESTIMAÇÃO DOS PARÂMETROS DO MODELO

forma equivalente aos sub-expressos, como explicado nas Seções 3.5.2 e 3.5.3.

Capítulo 4

Técnicas de agrupamento

A idéia básica de agrupamento (*clustering*) [35, 48] começa quando temos vários objetos e queremos agrupá-los pelas suas propriedades que podem ser quantitativas ou qualitativas. Os grupos são formados de modo que as distâncias entre os elementos de um grupo sejam mínimas e as distâncias entre os grupos sejam máximas. Muitas vezes a análise de agrupamento é uma das principais técnicas usadas para olhar as questões específicas de interesse no caso, em particular, dos conjuntos de dados de expressão gênica. Além disso, é utilizado para achar padrões de comportamento e fazer associações de genes com base em diferentes características ou associações de amostras de pacientes em estados similares de diagnóstico.

Neste processo existem três parâmetros importantes: (i) a função de distância usada para medir a proximidade (similaridade) entre dois pontos (objetos) no espaço p -dimensional (p é a quantidade de propriedades sendo estudadas); (ii) escolha dos pontos que são usados para medir a distância entre dois grupos (método de união); (iii) quantidade de grupos que serão formados. Existe toda uma abordagem na qual é discutido o número ótimo de grupos que serão formados a partir de um determinado conjunto de dados [74].

Em nosso objeto de estudo o mais comum é que os objetos sejam amostras de pacientes diferentes, enquanto que as propriedades desses objetos serão as características de expressão dos genes que

estão sendo analisados. Mas também pode ser o contrário, quando estivermos preocupados em avaliar a similaridade existente entre os genes, que neste caso seriam os objetos, e suas características as amostras de pacientes.

4.1 Distância entre dois pontos

A distância entre dois pontos é utilizada para medir a similaridade que existe entre dois objetos. Sejam $x = (x_1, \dots, x_p)$ e $y = (y_1, \dots, y_p)$ duas observações (p -dimensionais) e $d(x, y)$ a distância entre elas. A seguir são apresentadas algumas medidas bastante utilizadas para cálculo da distância.

a) Euclideana (“linha reta”)

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

sendo esta a mais utilizada na análise de agrupamento.

b) Métrica de *Minkowski*

$$d(x, y) = \sqrt[m]{\sum_{i=1}^p |x_i - y_i|^m},$$

para $(m = 1)$ temos *city-block*, *Manhattan* e para $(m = 2)$ temos a Euclideana. Em geral, variando m , mudamos os pesos atribuídos a pequenas e grandes diferenças.

c) Métrica de Canberra (só para variáveis não negativas)

$$d(x, y) = \sum_{i=1}^p \frac{|x_i - y_i|}{(x_i + y_i)}.$$

d) Coeficiente de Czkanowski (só para variáveis não negativas)

$$d(x, y) = 1 - \frac{2 \sum_{i=1}^p \min_{1 \leq i \leq p} (x_i, y_i)}{\sum_{i=1}^p (x_i + y_i)}.$$

e) Máximo

$$d(x, y) = \max_{1 \leq i \leq p} \{|x_i - y_i|\}.$$

f) Mahalanobis

$$d(x, y) = \sqrt{(x - y)' \hat{\Sigma}^{-1} (x - y)},$$

na qual Σ é uma estimativa para a matriz de variância-covariância das p variáveis estudadas.

4.2 Métodos de união

Para avaliar a similaridade entre dois grupos, ou um grupo e um elemento, precisamos obter um ponto que represente o grupo, o qual vai ser usado no cálculo de $d(x_i, y_i)$. A seguir são apresentados alguns destes métodos.

a) Vizinho mais próximo, *single*

Escolhe de cada grupo um ponto de tal forma que $d(x_i, y_i)$ seja a menor possível.

b) Vizinho mais distante, *complete*

Escolhe de cada grupo um ponto de forma que $d(x_i, y_i)$ seja a maior possível.

c) Centroide, *average*.

$d(r, s) = d(\bar{x}_r, \bar{x}_s)$, na qual: r e s são os dois grupos analisados, $\bar{x}_r$ e $\bar{x}_s$ os respectivos pontos centróides. Ou seja, para cada grupo, calcula-se seu ponto médio, achando para cada uma das p -variáveis o valor médio determinado pelo valor que ela tem nos n_r e n_s objetos respectivamente.

d) Divisão, ward

$$d^*(r, s) = \frac{n_r n_s}{n_r + n_s} d^2(r, s),$$

na qual: n_r e n_s são o número de elementos nos grupos r e s respectivamente.

e) Usando a correlação como medida de similaridade

Na análise de agrupamento, assim como na dos classificadores, muitas vezes a medida de similaridade usada é a Correlação, obtendo-se bons resultados [20]. Quando estamos no espaço p -dimensional, definido pelo número de genes analisados para cada uma das amostras em estudo, podemos escolher para medir a similaridade uma função de distância, sendo a Euclideana a mais comumente usada. A seguir verifica-se que:

$$d(x', y') \equiv \rho(x, y),$$

na qual: x e y são pontos p -dimensionais, contendo os valores originais das amostras sendo comparadas, p é o número de genes, $d(., .)$ é a distância Euclideana, e $\rho(., .)$ é o coeficiente de correlação. Os pontos x' e y' referem-se a x e y padronizados, respectivamente. Sendo $x' = (x - \bar{x})/S_x$, de forma que $\bar{x'} = 0$ e $S_{x'}^2 = 1$, do mesmo modo para y' .

$$\begin{aligned} d(x', y') &= \sqrt{\sum_{i=1}^p (x'_i - y'_i)^2} = \sqrt{\sum_{i=1}^p \left(\frac{x_i - \bar{x}}{S_x} - \frac{y_i - \bar{y}}{S_y} \right)^2} \\ &= \sqrt{\frac{\sum_{i=1}^p (S_y(x_i - \bar{x}) - S_x(y_i - \bar{y}))^2}{S_x^2 S_y^2}} = \sqrt{\frac{2p S_x^2 S_y^2 - 2 S_x S_y \sum_{i=1}^p (x_i - \bar{x})(y_i - \bar{y})}{S_x^2 S_y^2}} \\ &= \sqrt{2 \left(p - \frac{\rho(x, y)}{1/p} \right)} = \sqrt{2p(1 - \rho(x, y))}. \end{aligned}$$

Note que $\rho(., .) \in (-1, 1)$ e $p > 1$, então temos que para valores grandes da correlação obtemos valores pequenos de distância, enquanto que com a diminuição da correlação aumenta a distância,

desta forma consegue-se que a medida de similaridade seja a correlação, sendo mais similares aquelas mais correlacionadas.

$$d(x', y') \equiv \rho(x, y),$$

ou seja, embora os valores numéricos sejam diferentes, usar a distância Euclideana nos dados padronizados e usar a Correlação nos dados originais é equivalente, pois gera os mesmos resultados.

4.3 Agrupamento Hierárquico

Neste método de agrupamento a quantidade de grupos a serem formados não é um parâmetro, pelo fato de que na realidade o resultado final é só um grupo com todos os objetos. A idéia é colocar as uniões numa hierarquia decrescente de similaridade. Primeiro são agrupados dois objetos com maior similaridade, depois este grupo de dois objetos é unido com o terceiro objeto mais similar e assim sucessivamente até se unir o grupo com $n - 1$ objetos com o n -ésimo objeto.

4.4 Agrupamento K -Médias

Neste método é realizada uma divisão dos objetos em k partições, número de grupos a serem formados [74]. A partir do método de união escolhido [Seção 4.2], para cada grupo é determinado seu ponto de referência, e para cada objeto em particular é determinada a sua similaridade com cada um dos grupos, usando para isto os pontos de referência e a função de distância definida [Seção 4.1]. Se algum dos objetos for mais similar com outro grupo que não seja o dele, então este é realocado. O algoritmo é iterado, recalculando os pontos de referência, procurando aqueles que ainda não estejam no grupo mais apropriado e realocando os objetos novamente. É realocado um objeto em cada iteração, mas só precisam ser recalculados dois pontos de referência, aqueles que referem-se aos dois grupos modificados. O processo de realocação termina com o posicionamento de cada objeto no grupo onde ele melhor se enquadra (ou seja, naquele grupo de objetos mais similar com ele), ou quando o algoritmo atinge um número máximo de iterações indicado pelo usuário.

Note que o agrupamento final vai depender de parâmetros do algoritmo, tais como número de grupos, função de distância, método de união, assim como do agrupamento inicial, o qual pode ser aleatório ou proposto por um resultado prévio, empregando outro método de agrupamento ou não. Podem existir casos em que o agrupamento final não fica estável, embora o número de iterações seja definido o suficientemente grande. Também é possível que, uma vez fixados os parâmetros e depois de rodar o algoritmo várias vezes, o resultado final nem sempre seja o mesmo.

4.5 Agrupamento usando SOM (*Self-Organizing Maps*)

Um método que tem sido bastante utilizado em pesquisas recentes de agrupamento é o SOM [43], o qual pode ser usado no reconhecimento de padrões, como também para projetar e visualizar objetos, originalmente num espaço p -dimensional, em duas dimensões, assumindo-se uma rede neural como uma matriz regular de uma ou duas dimensões.

Assumo que o espaço do conjunto de dados analisados é $\mathbb{R}^p$, no qual p é o número de genes sendo estudados. Seja $m_i = (\mu_{i1}, \dots, \mu_{ip})' \in \mathbb{R}^p$ o vetor de referência associado ao i -ésimo nó da matriz de duas dimensões. O tipo de topologia pode ser assumido como retangular ou hexagonal. Este método de redes neurais não supervisionado, consiste num conjunto de neurônios em uma grade regular de baixa dimensão. Cada um dos neurônios está conectado aos adjacentes por uma relação de vizinhança a qual impõe a topologia ou estrutura do mapa. O SOM é treinado iterativamente e em cada iteração uma amostra $x = (x_1, \dots, x_p)'$ do conjunto de dados é escolhida ao acaso e é determinada a sua similaridade com cada um dos neurônios (vetor de referência). Desta forma obtém-se o BMU (*Best Matching Unit*), ou seja, o nó mais similar da amostra escolhida.

Seja $x \in \mathbb{R}^p$ um vetor de dados estocásticos. O SOM pode ser interpretado como uma projeção não linear da função de densidade de probabilidade do vetor x p -dimensional em duas dimensões. Na comparação do vetor x com os m_i pode ser usada qualquer medida de similaridade, sendo a distância Euclideana a mais usada. O nó mais similar a x é definido como, $c = \operatorname{argmin}_i \{d(x, m_i)\}$,

na qual $d(\cdot, \cdot)$ é a função de distância considerada.

Depois de achar o BMU, os vetores de referência são atualizados, de modo que o BMU é movido no sentido da amostra escolhida; a topologia dos vizinhos do BMU é tratada similarmente. Esta adaptação do procedimento reajusta o BMU e a topologia dos vizinhos na direção da amostra escolhida. No aprendizado, o processo no qual a projeção não linear é formada, os nós que são próximos seguindo a topologia empregada na matriz de duas dimensões são ativados por cada um dos outros no aprendizado do mesmo vetor x . O resultado disto é um relaxamento ou efeito de suavizamento local nos m_i na vizinhança do c -ésimo nó, o qual em aprendizado contínuo leva à ordenação global.

Os valores iniciais $m_i(0)$ são definidos aleatoriamente, e para cada uma das iterações temos que:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] ,$$

na qual t é o número da iteração do algoritmo. Note que $h_{ci}(t)$, conhecida como função de vizinhança, tem um papel importante no processo de suavizamento, pois define a região de influência que a amostra escolhida tem no SOM. É necessário que $h_{ci}(t) \xrightarrow{t \rightarrow \infty} 0$ para garantir a convergência do algoritmo. Geralmente, $h_{ci}(t) = h(d(r_c, r_i), t)$, na qual: $r_c, r_i \in \mathbb{R}^2$ são os vetores de locação dos nós c e i , respectivamente, de tal forma que com o crescimento de $d(r_c, r_i)$, $h_{ci}(t) \xrightarrow{t \rightarrow \infty} 0$. A largura e forma de h_{ci} definem a “rigidez” da “superfície elástica” (grade bidimensional) a ser ajustada. A seguir são apresentados dois exemplos para $h_{ci}(t)$, mais comumente usados:

a) pelo conjunto de vizinhos,

$$h_{ci}(t) = \begin{cases} \alpha(t), & \text{se } i \in N_c(t) \\ 0, & \text{se } i \notin N_c(t) \end{cases} ,$$

na qual: $\alpha(t) \in (0, 1)$, é o fator da taxa de aprendizado e $N_c(t)$ é o conjunto de índices dos pontos ao redor do nó c na rede de duas dimensões. Sendo os dois monotonicamente decrescentes em função de t .

b) pela distribuição Normal padrão, $N(0, 1)$,

$$h_{ci}(t) = \alpha(t) \exp \left(-\frac{d^2(r_c, r_i)}{2\sigma^2(t)} \right),$$

na qual $\sigma(t)$ define uma medida da largura do núcleo, sendo também, monotonicamente decrescente em função de t .

A taxa de aprendizado pode ser linear, exponencial ou inversamente proporcional a t . Por exemplo, uma boa escolha pode ser, $\alpha(t) = .9(1 - t/1000)$ [43]. A ordenação dos m_i ocorre durante a primeira etapa, enquanto que os passos restantes são necessários para o ajuste do mapa, o qual é assumido como contínuo ao longo da grade bidimensional reajustada. Quando o algoritmo é executado um número suficientemente grande de passos e os valores de m_i , no processo definido anteriormente, convergem para determinados valores fixos, então considera-se o mapa como ajustado. Por fim, cada uma das amostras do conjunto de dados é comparada com todos os neurônios e encontra-se o m_i mais próximo dela, gerando assim o agrupamento desejado.

Capítulo 5

Discriminação - Classificação

Em classificação estamos interessados em achar genes ou grupos de genes que, segundo determinados critérios, permitam distinguir bem amostras de tipos diferentes, como por exemplo: células ou tecidos normais de tumorais. Inicialmente precisa-se de um acúmulo grande de amostras classificadas, as quais serão usadas na definição dos critérios, de tal forma que os classificadores distingam as classes existentes [20, 52].

Seja X uma matriz contendo a informação referente a p genes e n amostras de mRNAs, na qual o elemento x_{ij} representa o nível de expressão do j -ésimo gene (variável) na i -ésima amostra (observação). Para cada amostra temos $x_i = (x_{i1}, \dots, x_{ip})$ e y_i , perfil de expressão e tipo respectivamente. Assumimos que o conjunto de dados tem t tipos de amostras, $y_i \in \{1, \dots, t\}$ e denotamos com n_k a quantidade de amostras do k -ésimo tipo no conjunto de dados, $k \in \{1, \dots, t\}$.

Seja $C(., \mathcal{L})$ um classificador treinado a partir do LS (*Learning Set*), $\mathcal{L} = \{(x_1, y_1), \dots, (x_{nL}, y_{nL})\}$, na qual nL é o número de amostras no LS. Este pode ser usado para classificar novas amostras, denominadas de TS (*Test Set*), $\mathcal{T} = \{x_1, \dots, x_{nT}\}$, no qual nT é o número de amostras no TS. Além disso, também podemos estar interessados em estimar a taxa de erro do classificador, e isto pode ser feito comparando $C(x_i, \mathcal{L})$ (classificação atribuída a x_i) com y_i (classificação real), para $i \in \{1, \dots, nL\}$.

5.1 Classificação usando K -Vizinhos

Este método de classificação é, quanto à sua metodologia, muito lógico e intuitivo. Inicialmente precisa ser definida a função de distância que vai ser usada para medir a similaridade entre as amostras, como abordado na Seção 4.1. Para classificar uma nova amostra x , são encontradas as k observações mais próximas $x_1, \dots, x_k$ das que formam o LS. A classificação vai ser determinada por $C(x, \mathcal{L}) = \text{Moda}(y_1, \dots, y_k)$. A determinação do valor ótimo para o parâmetro $k \in \{1, \dots, nL - 1\}$ é realizada a partir do LS, para o qual são avaliados todos seus valores possíveis e escolhido aquele para o qual obtemos o melhor resultado. Ele é definido por validação cruzada, ou seja, fixando um determinado valor para k , cada uma das amostras no LS é retirada e classificada usando este método, sendo isto realizado para todo possível valor de k . Finalmente determina-se aquele que gerou a maior quantidade de acertos. Note que não podemos usar $k = nL$, pois quando retiramos uma amostra para classificá-la o LS usado para treinar o classificador é reduzido a $nL - 1$ amostras. Além disto, note que se fosse considerado $k = nL$ para classificar x uma amostra realmente nova ($x \notin \text{LS}$) temos que $C(x, \mathcal{L}) = \text{Moda}(y_1, \dots, y_{nL}) = \text{cte } \forall x$, neste caso não é preciso definir função de distância nem treinar o classificador.

Na Figura 5.1 são apresentados vários gráficos de k vs o número de acertos, gerados a partir do conjunto de dados da Seção 6.8. Inicialmente pensamos em como escolher adequadamente a medida de proximidade (similaridade) entre as amostras, por isso testamos a distância Euclideana e a Correlação, sendo que a Correlação apresentou um maior número de acertos, para quase todo valor de k . Além da medida de similaridade, também foram observadas outras questões de interesse neste conjunto de dados em particular. Podemos falar, por exemplo, da pouca diferença que existe entre algumas das condições biológicas, como é o caso de Normal e Gastrite. No gráfico observa-se que quando assumimos Normal e Gastrite, como um tipo só, os resultados melhoram. Outro ponto importante é a escolha dos genes a serem analisados. Não parece ser uma boa proposta usar o total de 370 genes, pois acredita-se que alguns deles não estão envolvidos com a diferenciação entre os tipos. Este elemento foi avaliado no gráfico usando aqueles genes mais DE em cada uma das comparações. Também testou-se os resultados, assumindo só dois tipos de amostras, Tumoriais e

não Tumoraes, mostrando certa melhoria em comparação de quando se usam os quatro tipos originaes. De modo geral, o valor ótimo de k para cada uma das comparações fica no intervalo $[7, 13]$, muito embora, o padrão de comportamento varia para cada uma das curvas.

Figura 5.1: Comparação da eficiência do algoritmo K -Vizinhos em função de k , empregando validação cruzada. Na curva (a) foi usada a distância Euclideana enquanto que em (b), (c), (d) e (e) a medida de similaridade utilizada foi a Correlação. Em (c) e (d) amostras Normais e Gastrite foram consideradas do mesmo tipo. Em (d) e (e) só foram considerados os 18 genes mais DE nos pares de comparações. Já em (e) só foram levados em conta dois tipos de amostras, as Tumoraes e as não Tumoraes.

5.2 Discriminante linear de Fisher

Este método também é não paramétrico, pois ele não assume distribuições condicionais para cada um dos tipos de amostras. A idéia é achar a função linear $a'x$ que maximiza R_{ed} , razão entre a soma de quadrados entre t grupos (cada um representado por um tipo de amostra) e a soma de

quadrados dentro dos grupos.

Seja $y = Xa = (X_1a, \dots, X_ta)' = (y_1, \dots, y_t)'$, na qual:

X é a matriz $(n \times p)$ de dados, n é o número de amostras e p é o número de genes;

X_i é a sub-matriz $(n_i \times p)$ com as amostras do tipo i e n_i é o número de amostras deste tipo;

t é a quantidade de tipos de amostras;

a é um vetor $(p \times 1)$ tal que a função linear $a'x$ maximiza R_{ed} , definida a seguir.

Sejam:

$y'H y = a'X'H X a = a'T a$ a soma de quadrados total;

$\sum y_i'H_i y_i = \sum a'X_i'H_i X_i a = a'W a$ a soma de quadrados dentro dos grupos;

$\sum n_i(\bar{y}_i - \bar{y})^2 = \sum n_i(a'(\bar{x}_i - \bar{x}))^2 = a'B a$ a soma de quadrados entre os grupos,

nas quais $H_i = \mathbf{I} - \frac{1}{n_i}\mathbf{1}\mathbf{1}'$ é a matriz $(n_i \times n_i)$ centralizadora.

Segue que a razão $R_{ed} = \frac{a'B a}{a'W a}$, é maximizada quando a é o autovetor de $(W^{-1}B)$ correspondente ao maior autovalor, o que pode-se verificar no Teorema 11.5.1 [48]. E uma nova amostra será classificada da seguinte forma, $C(x, \mathcal{L}) = \operatorname{argmin}_i \{ | a'x - a'\bar{x}_i | \}$.

5.3 Árvores de classificação

A necessidade de usar árvores de classificação aparece quando a estrutura observada nos dados não nos permite distinguir entre os tipos usando regras simples. Neste caso, surge a idéia de criar uma estrutura de rede formada por nós. Inicialmente precisam ser definidas as divisões, regras que permitem o deslocamento na rede. Também é necessário definir cada um dos nós como folha ou não, assim como atribuir cada nó folha a um tipo. A classificação de uma amostra seguindo esta estrutura é um processo que implica se deslocar na rede, onde o caminho vai ser definido pelas regras (funções), criadas a partir do LS, aplicadas na amostra sob análise, até se chegar em

um nó folha, onde vai ser obtido o tipo da amostra. Existem aplicativos que têm este algoritmo implementado, como o CART [9] e o S-Plus [45].

5.4 Classificação usando Máxima Verossimilhança

Este método é paramétrico e assume o conhecimento das distribuições de densidade condicionais dado cada um dos tipos de amostras. Os parâmetros podem ser assumidos como conhecidos ou não, sendo que, no segundo caso, eles são estimados a partir do LS. A Normal Multivariada, pelas propriedades que tem, muitas vezes é a densidade considerada. A idéia do método é o cálculo da verossimilhança da nova amostra, dado cada um dos tipos, e a classificação é obtida por aquela classe, para a qual determina-se a maior verossimilhança, ou seja, $C(x, \mathcal{L}) = \operatorname{argmax}_i \{ \mathbb{P}(x \mid y = i) \}$.

A seguir vamos obter a função de classificação no caso da Normal Multivariada e a forma que esta adquire em alguns casos particulares, assumindo determinadas restrições.

Seja $(x \mid y = i) \sim N_p(\mu_i, \Sigma_i)$, $i \in \{1, \dots, t\}$, na qual: t é o número de tipos de amostras, μ_i e Σ_i são os parâmetros do i -ésimo tipo de amostras, um vetor $p \times 1$ de médias e uma matriz $p \times p$ de variâncias-covariâncias, respectivamente, e p é o número de genes sendo analisados. A seguir é apresentada a densidade condicional $x \mid y$, assim como a transformação $\log()$ da mesma.

$$\begin{aligned} f(x \mid y = i) &= |2\pi\Sigma_i|^{-1/2} \exp \left\{ -\frac{1}{2}(x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) \right\} \\ &= (2\pi)^{-p/2} |\Sigma_i|^{-1/2} \exp \left\{ -\frac{1}{2}(x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) \right\}, \\ \log(f(x \mid y = i)) &= -\frac{1}{2}(x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) - \frac{1}{2} \log((2\pi)^p |\Sigma_i|) \\ &= -\frac{1}{2} ((x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) + \log(|\Sigma_i|)) - \frac{p}{2} \log(2\pi). \end{aligned}$$

Note que $\frac{p}{2} \log(2\pi)$ é constante em função de i , então:

$$\begin{aligned} C(x, \mathcal{L}) &= \operatorname{argmax}_i \{ \mathbb{P}(x \mid y = i) \} = \operatorname{argmax}_i \{ f(x \mid y = i) \} \\ &= \operatorname{argmax}_i \{ \log(f(x \mid y = i)) \} = \operatorname{argmin}_i \{ -\log(f(x \mid y = i)) \} \end{aligned}$$

Segue que,

$$C(x, \mathcal{L}) = \operatorname{argmin}_i \{ (x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) + \log(| \Sigma_i |) \}.$$

Caso 1, $\Sigma_i = \Sigma, \forall i \in \{1, \dots, t\}$:

$$\begin{aligned} C(x, \mathcal{L}) &= \operatorname{argmin}_i \{ (x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) + \log(| \Sigma_i |) \} \\ &= \operatorname{argmin}_i \{ (x - \mu_i)' \Sigma^{-1} (x - \mu_i) + \log(| \Sigma |) \}. \end{aligned}$$

Neste caso $\log(| \Sigma |)$ é constante em função de i , então:

$$C(x, \mathcal{L}) = \operatorname{argmin}_i \{ (x - \mu_i)' \Sigma^{-1} (x - \mu_i) \}.$$

Caso 2, $\Sigma_i = \Delta_i = \operatorname{diag}(\sigma_{i1}^2, \dots, \sigma_{ip}^2), \forall i \in \{1, \dots, t\}$:

$$\begin{aligned} C(x, \mathcal{L}) &= \operatorname{argmin}_i \{ (x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) + \log(| \Sigma_i |) \} \\ &= \operatorname{argmin}_i \left\{ \sum_{j=1}^p \frac{(x_j - \mu_{ij})^2}{\sigma_{ij}^2} + \log(\sigma_{ij}^2) \right\}. \end{aligned}$$

Caso 3, $\Sigma_i = \Delta = \operatorname{diag}(\sigma_1^2, \dots, \sigma_p^2), \forall i \in \{1, \dots, t\}$:

$$\begin{aligned} C(x, \mathcal{L}) &= \operatorname{argmin}_i \left\{ (x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) + \log(|\Sigma_i|) \right\} \\ &= \operatorname{argmin}_i \left\{ (x - \mu_i)' \Sigma^{-1} (x - \mu_i) + \log(|\Sigma|) \right\}. \end{aligned}$$

Neste caso, que contém as restrições consideradas nos Casos 1 e 2, $\log(\sigma_{ij}^2)$ é constante em função de i , então,

$$C(x, \mathcal{L}) = \operatorname{argmin}_i \left\{ \sum_{j=1}^p \frac{(x_j - \mu_{ij})^2}{\sigma_j^2} \right\}.$$

Geralmente os parâmetros são desconhecidos e o LS é usado para calculá-los pelos estimadores de máxima verossimilhança, após a correção do vício.

$$\hat{\mu}_i = \bar{X}_i;$$

$$\hat{\Sigma}_i = S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_i^j - \bar{X}_i)(X_i^j - \bar{X}_i)',$$

na qual: X_i^j é a j -ésima amostra do tipo i .

Nos casos que assumimos $\Sigma_i = \Sigma, \forall i \in \{1, \dots, t\}$ temos que:

$$\hat{\Sigma} = \frac{1}{n-g} \sum_{j=1}^t (n_j - 1) S_j^2.$$

Para o Caso 3, com $t = 2$, temos que uma amostra $x = (x_1, \dots, x_p)$ é classificada do tipo 1 se:

$$\begin{aligned} \sum_{j=1}^p \frac{(x_j - \bar{x}_{2j})^2}{S_j^2} &\geq \sum_{j=1}^p \frac{(x_j - \bar{x}_{1j})^2}{S_j^2}, \\ \sum_{j=1}^p \frac{(x_j - \bar{x}_{2j})^2 - (x_j - \bar{x}_{1j})^2}{S_j^2} &\geq 0, \\ \sum_{j=1}^p \frac{((x_j - \bar{x}_{2j}) + (x_j - \bar{x}_{1j}))((x_j - \bar{x}_{2j}) - (x_j - \bar{x}_{1j}))}{S_j^2} &\geq 0, \end{aligned}$$

obtendo desta forma a função de discriminação para este caso em particular,

$$\sum_{j=1}^p \frac{(\bar{x}_{1j} - \bar{x}_{2j})}{S_j^2} \left(x_j - \frac{(\bar{x}_{1j} + \bar{x}_{2j})}{2} \right) \geq 0,$$

a qual também pode ser expressa na forma, $\sum_{j=1}^p v_j \geq 0$, onde: $v_j = a_j(x_j - b_j)$, $a_j = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{S_j^2}$ e $b_j = \frac{\bar{x}_{1j} + \bar{x}_{2j}}{2}$. Podemos pensar ainda em mudanças na função anterior, como por exemplo, trocar a_j por $a_j^* = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{S_{1j} + S_{2j}}$, [31].

Pode ser definida uma medida de PS (*Prediction Strength*), da seguinte forma:

$$PS = \frac{\max(V1, V2) - \min(V1, V2)}{\max(V1, V2) + \min(V1, V2)},$$

na qual: $V1 = \sum_{j=1}^p \max(v_j, 0)$ e $V2 = \sum_{j=1}^p \max(-v_j, 0)$. Note que $V1$ e $V2$ representam o somatório dos v_j positivos e negativos, respectivamente, sendo ambos não negativos. Desta forma $PS \in (0, 1)$.

5.4.1 Estimadores viciados ou não viciados?

Na Seção 5.4 empregamos os estimadores não viciados, mas podemos verificar que dentro de determinadas classes existem estimadores viciados “melhores” que o estimador não viciado.

Seja $X_1, X_2, \dots, X_n$ uma amostra aleatória (iid) da $X \sim N(\mu, \sigma^2)$,

$\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+$.

Seja $T_c = \varphi(X_1, X_2, \dots, X_n) = \frac{1}{c} \sum_{i=1}^n (X_i - \bar{X})^2$, na qual $c > 0$, uma classe de estimadores para σ^2 .

Sabe-se que: $(\hat{\sigma}_{mv}^2 = T_n)$ é o estimador de máxima verossimilhança para σ^2 .

Dado que

$$\frac{T_1}{\sigma^2} \sim \chi_{n-1}^2,$$

segue que

$$\mathbb{E}(T_n) = \frac{\sigma^2}{n} \mathbb{E} \left[\frac{T_1}{\sigma^2} \right] = \frac{n-1}{n} \sigma^2.$$

$\Rightarrow T_n$ é viciado para σ^2 ,

$\Rightarrow T_{n-1}$ é não viciado para σ^2 .

Para T_n temos que

$$\begin{aligned}\text{Var}(T_n) &= \frac{\sigma^4}{n^2} \text{Var} \left[\frac{T_1}{\sigma^2} \right] = \frac{\sigma^4}{n^2} 2(n-1) = \frac{2(n-1)\sigma^4}{n^2}; \\ b(T_n) &= \mathbb{E}(T_n) - \sigma^2 = \frac{n-1}{n} \sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}; \\ \text{EQM}(T_n) &= \text{Var}(T_n) + b^2(T_n) = \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}.\end{aligned}$$

Para T_{n-1} temos que

$$\begin{aligned}\text{Var}(T_{n-1}) &= \frac{\sigma^4}{(n-1)^2} \text{Var} \left[\frac{T_1}{\sigma^2} \right] = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{n-1}; \\ b(T_{n-1}) &= \mathbb{E}(T_{n-1}) - \sigma^2 = 0; \\ \text{EQM}(T_{n-1}) &= \text{Var}(T_{n-1}) = \frac{2\sigma^4}{n-1}.\end{aligned}$$

Para $n > 1$, temos que $\text{EQM}(T_{n-1}) > \text{EQM}(T_n)$, pois verificou-se que $\forall n > 1$, T_n (estimador viciado) é melhor em termos de EQM que T_{n-1} (estimador não viciado) dentro da classe T_c , para $c > 0$. Agora vamos achar o “melhor” estimador para σ^2 dentro desta classe. Primeiro precisamos achar a expressão geral do EQM de T_c , em função de c e depois minimizar essa expressão.

Para T_c temos que

$$\begin{aligned}\text{Var}(T_c) &= \frac{\sigma^4}{c^2} \text{Var} \left[\frac{T_1}{\sigma^2} \right] = \frac{\sigma^4}{c^2} 2(n-1) = \frac{2(n-1)\sigma^4}{c^2}; \\ b(T_c) &= \mathbb{E}(T_c) - \sigma^2 = \frac{\sigma^2}{c} \mathbb{E} \left[\frac{T_1}{\sigma^2} \right] - \sigma^2 = \frac{\sigma^2}{c} (n-1) - \sigma^2 = \left(\frac{n-1}{c} - 1 \right) \sigma^2; \\ \text{EQM}(T_c) &= \text{Var}(T_c) + b^2(T_c) = \frac{2(n-1)\sigma^4}{c^2} + \left(\frac{n-1}{c} - 1 \right)^2 \sigma^4 \\ &= \sigma^4 \left[\frac{2(n-1)}{c^2} + \left(\frac{n-1}{c} - 1 \right)^2 \right].\end{aligned}$$

Agora temos a expressão geral do EQM de T_c , e podemos minimizá-la,

$$\frac{\partial \text{EQM}(T_c)}{\partial c} = \sigma^4 \left[\frac{-4(n-1)}{c^3} + 2 \left(\frac{n-1}{c} - 1 \right) \left(-\frac{n-1}{c^2} \right) \right],$$

$$\frac{\partial \text{EQM}(T_c)}{\partial c} = 0,$$

$$\Rightarrow \frac{-4(n-1)}{c^3} - \frac{2(n-1)^2}{c^3} + \frac{2(n-1)}{c^2} = 0.$$

assumindo ($n > 1$) e multiplicando por $\frac{c^2}{2(n-1)}$ temos que,

$$\frac{2}{c} + \frac{n-1}{c} = 1 \quad \Rightarrow \quad \frac{n+1}{c} = 1,$$

$$\Rightarrow c = n + 1,$$

$\therefore T_{n+1}$ é o “melhor” estimador para σ^2 na classe T_c ($c > 0$).

5.5 Classificação usando *Aggregating*

A proposta deste método é a criação de um grande número de classificadores, cada um deles a partir de um LS perturbado. A geração dos LS perturbados é realizada a partir do LS original (conjunto de amostras de pacientes estudados) usando *bootstrap*, podendo ser este paramétrico [Seção 5.5.2] ou não [Seção 5.5.1]. O *bootstrap* não paramétrico não assume distribuição nos dados observados, e a geração de novos subconjuntos de amostras é realizada usando reamostragem com reposição das amostras originais, onde cada uma delas é equiprovável, enquanto o *bootstrap* paramétrico assume como conhecidas as distribuições condicionais para cada um dos tipos de amostras existentes. Neste caso, as amostras originais são usadas na estimação dos parâmetros das densidades assumidas e a geração de novos subconjuntos de amostras é feita por meio das distribuições estimadas. É comum estimar as probabilidades de mistura a partir das proporções de cada um dos tipos de amostra no LS original. Nos dois casos é gerado um grande número de subconjuntos de amostras, sendo cada um deles do mesmo tamanho que o LS original, e a nossa estimativa é calculada para cada um deles. A estimativa final é a média de todas. É necessário ter a consideração de que em cada um dos subconjuntos gerados precisamos ter pelo menos uma

amostra de cada tipo.

Seja $C(\cdot, \mathcal{L}_b)$ o classificador criado com o b -ésimo LS perturbado e w_b o peso atribuído na classificação feita por ele. Os classificadores são treinados com seus respectivos conjuntos de dados e a classificação de uma nova amostra é realizada da seguinte forma,

$$C(x, \mathcal{L}) = \operatorname{argmax}_i \left\{ \sum_b w_b \mathbf{1}(C(x, \mathcal{L}_b) = i) \right\}.$$

Seja $PV(x)$ os votos da predição de $C(x, \mathcal{L})$,

$$PV(x) = \frac{\max_i \sum_b w_b \mathbf{1}(C(x, \mathcal{L}_b) = i)}{\sum_b w_b}.$$

Uma possível interpretação para $PV(x)$ é, por exemplo, quando temos $w_b = 1 \forall b$, neste caso representa a proporção de votos do tipo escolhido. Em geral, assume-se $w_b \in (0, 1) \forall b$. Note que para usar este método é preciso definir um tipo de classificador, como os apresentados anteriormente neste capítulo.

Depois de ter gerado os LS perturbados e criado os classificadores podemos pensar em estimar os w_b de maneira adequada. Cada um dos classificadores, treinado com seu respectivo LS perturbado, pode ser avaliado usando alguma medida de qualidade. Por exemplo, podemos pensar em classificar todas as amostras do LS original, e w_b pode ser estimado da seguinte forma,

$$\hat{w}_b = \frac{\sum_{i=1}^{nL} \mathbf{1}(C(x_i, \mathcal{L}_b) = y_i)}{nL},$$

na qual: x_i e y_i são a i -ésima amostra do LS original e seu verdadeiro tipo respectivamente, e nL é a quantidade de amostras no LS original.

5.5.1 Geração dos LS perturbados usando *Bagging* não paramétrico

Neste caso é usada a idéia do *bootstrap* não paramétrico na geração dos LS perturbados. Por exemplo, em [20] acredita-se que 37% das amostras originais não fazem parte de um determinado LS perturbado, e estas podem ser usadas para avaliar o classificador criado, inclusive na estimação de w_b . As não escolhidas formam um TS para $C(., \mathcal{L}_b)$, que pode ser usado na estimação de ce (proporção de classificações erradas para $C(., \mathcal{L})$),

$$\hat{ce} = \frac{1}{nL} \sum_i^{nL} \mathbb{1} \left(y_i \neq \operatorname{argmax}_k \left\{ \sum_{b: x_i \notin \mathcal{L}_b} \mathbb{1}(C(x_i, \mathcal{L}_b) = k) \right\} \right).$$

Para cada amostra do LS original são encontrados aqueles LS perturbados que não a contenham, e são usados para classificá-la. O tipo atribuído a ela, $\operatorname{Moda}()$ das classificações realizadas, é comparado com o verdadeiro. Isto é feito para cada amostra e tem-se a proporção de erros.

5.5.2 Geração dos LS perturbados usando *Bagging* paramétrico

Neste caso é usada a idéia do *bootstrap* paramétrico na geração dos LS perturbados. Assumindo como conhecidas as distribuições condicionais para cada um dos tipos e estimando os parâmetros a partir do LS original. As probabilidades de mistura são estimadas a partir das proporções dos tipos de amostra no LS original.

5.5.3 Geração dos LS perturbados usando *Boosting*

No *boosting* a geração dos LS perturbados também é realizada por reamostragem, similar ao *Bagging* não paramétrico [Seção 5.5.1], mas neste caso as amostras não são equiprováveis, sendo usado um vetor de probabilidades de reamostragem $p = (p_1, \dots, p_{nL})'$. Inicialmente assume-se um p uniforme, ou seja, $p_i = 1/nL, \forall i \in (1, \dots, nL)$. A seguir, é apresentado o algoritmo de geração do b -ésimo LS perturbado [20]:

- i) Gera-se $\mathcal{L}_b$ usando o p atual;
- ii) Cria-se o classificador $C(., \mathcal{L}_b)$;

- iii) Calculam-se, $d_i = \mathbf{1}(y_i \neq C(x_i, \mathcal{L}_b))$, $\epsilon_b = \sum_{i=1}^{nL} p_i d_i$ e $\beta_b = (1 - \epsilon_b)/\epsilon_b$;
- iv) Recalcula-se $p, p_i = \begin{cases} p_i \beta_b^{d_i} / \sum_{i=1}^{nL} p_i \beta_b^{d_i}, & \text{se } \epsilon_b \in (0, 1/2) \\ p_i, & \text{se } \epsilon_b = 0 \text{ ou } \epsilon_b \geq 1/2 \end{cases}$;
- v) Calcula-se, $w_b = -\log(\beta_b)$.

5.6 Alerta no Diagnóstico

Poderíamos pensar, no caso, que realmente o objetivo não seja o de classificar as amostras, pelo fato de que isto já esteja feito. Cabe salientar que, nesta área do estudo do câncer, que tem sido o centro da análise dos conjuntos de dados estudados, os patologistas conseguem identificar o estado da doença com uma precisão muito grande. Então, a idéia de classificação pode ser adaptada, não para determinar o tipo de uma amostra nova, o que pode ser feito pelo patologista, mas possa ser interpretada como um possível alerta no diagnóstico. A proposta, que foi testada no conjunto de dados da Seção 6.8, é apresentada na Figura 5.2.

5.6. ALERTA NO DIAGNÓSTICO

Figura 5.2: Proposta para alerta no diagnóstico. A primeira coluna representa a descrição das 99 amostras no conjunto de dados da Seção 6.8 organizadas pelos tipos, as outras cinco colunas foram usadas para representar os resultados obtidos referente às comparações. Cada uma das amostras foi classificada usando os melhores trios de genes de cada uma das cinco comparações, sendo o número de trios 100, 17, 20, 52 e 4 da comparação NT, GT, MT, NM e GM respectivamente [Tabela 6.3]. Ao lado da figura é apresentada a escala de cor empregada, a qual representa a frequência com que as amostras são classificadas, do tipo da primeira condição biológica do par, pelos trios de genes de cada comparação (100% = verde e 0% = vermelho).

Na Figura 5.2 a primeira coluna tem a descrição para cada uma das amostras estudadas no con-

junto de dados da Seção 6.8, ordenadas pelo tipo e a quantidade para cada um deles. As próximas cinco colunas representam os pares de condições biológicas: Normal-Tumor (NT), Gastrite-Tumor (GT), Metaplasia-Tumor (MT), Normal-Metaplasia (NM) e Gastrite-Metaplasia (GM). A cor atribuída à i -ésima amostra na j -ésima comparação, na qual: $i \in \{1, \dots, 99\}$ e $j \in \{NT, GT, MT, NM, GM\}$, representa a concordância ou não dos melhores classificadores (trios de genes) dessa comparação no tipo da amostra analisada. Onde a cor verde representa a concordância na classificação como tipo “Condição1”, enquanto que a cor vermelha indica o tipo “Condição2”. Se olharmos por exemplo a comparação Normal-Tumor (NT), então os tipos Condição1 e Condição2 são Normal e Tumor, respectivamente. Uma cor escura indicaria dúvida, ou seja, aproximadamente metade dos classificadores indicam um tipo e a outra metade o outro tipo.

Primeiro, foram encontrados os melhores classificadores para cada par de condições biológicas, depois de ter definido um determinado critério de qualidade. Estes classificadores, formados neste caso por trios de genes, foram treinados nas amostras pertencentes ao par de condições biológicas que ele deve distinguir. Depois foram usados na classificação do total de amostras. Note que, por exemplo, se pensarmos na comparação Normal-Tumor, estaremos tentando atribuir um destes dois tipos a amostras para as quais é conhecido que pertencem a outro tipo, como é o caso das amostras de tipo Gastrite e Metaplasia. Depois de feito isto, observou-se a concordância entre os classificadores, da seguinte forma,

$$C^*(x, \mathcal{L}, cp) = \text{Moda}(C_1(x, \mathcal{L}, cp), \dots, C_{n_{cp}}(x, \mathcal{L}, cp)),$$

na qual: cp é uma determinada comparação (envolvendo dois tipos de amostras), $C_i(x, \mathcal{L}, cp)$ é a classificação atribuída para x pelo i -ésimo melhor classificador da comparação cp , $i \in \{1, \dots, n_{cp}\}$, sendo n_{cp} a quantidade de classificadores “bons” escolhido da comparação cp .

No caso de x , uma amostra Normal ou Tumor, continuando o exemplo, é interessante observar que tipo de amostra ela “parece ser” ($C^*(x, \mathcal{L}, NT)$), dado o restante de amostras desse par de condições. Por outro lado, no caso da amostra ser de um outro tipo, diferente dos envolvidos no par

de condições, pode ser muito interessante. Imaginemos, por exemplo, o caso do tipo Metaplasia, que pode ser considerado como um estágio “perto” do Tumor, apesar de ser possível nunca chegar nele. Neste caso, a concordância dos classificadores Normal-Tumor em que uma determinada Metaplasia é um Tumor pode ser considerado um alerta para o posterior acompanhamento do paciente. Um exemplo disto é a amostra “Metaplasia-GH828”, classificada como Tumor por 99 dos melhores 100 classificadores (trios de genes) Normal-Tumor. Além disto, pode ser usado na determinação de amostras raras, isto com respeito às amostras restantes no conjunto de dados. Por exemplo, a amostra “Normal-GF63”, classificada como Tumor por 93 dos melhores 100 classificadores (trios de genes) Normal-Tumor.

Vale ressaltar que algumas das amostras classificadas como raras, utilizando esse método, foram avaliadas novamente pelos patologistas. Exemplo disso é a amostra “Tumor-GH965”, da qual os patologistas confirmaram que era realmente um tumor, mas que tinha características diferentes das demais amostras tumorais, pelo fato de tratar-se de um tumor precoce. Ao tumor que acomete apenas as camadas mucosa e submucosa da parede gástrica, dá-se o nome de tumor precoce ou inicial e está associado a excelente prognóstico [15, 14].

5.7 Método BE (Busca e Escolha) de clique de genes

Uma questão interessante é o número de genes envolvidos num determinado classificador. Já foi discutido o fato de que quando temos um grande conjunto de genes e estamos estudando um problema biológico, só alguns deles devem estar realmente vinculados. Isto foi testado no conjunto de dados da Seção 6.8, para o qual mostramos que usando o total de genes os resultados obtidos não são muito bons, mas melhora quando selecionamos alguns genes por determinado critério, como por exemplo os DE. Poderíamos estar falando de uma quantidade de genes ótima, para a qual obtém-se o melhor resultado usando um tipo de classificador.

Neste trabalho fizemos uma procura exaustiva até para o caso de trios de genes, busca que, com-

putacionalmente, fica muito pesada (seis computadores por três semanas no conjunto de dados da Seção 6.8), inclusive num conjunto de genes relativamente pequeno. A proposta até chegar nos melhores cliques de genes (grupos de genes) é a seguinte: primeiro achar os genes que “distinguem” melhor entre as duas condições biológicas sendo analisadas, isto por algum critério, o proposto é o de genes DE [Seção 6.2]. Depois podemos pensar num algoritmo no qual para cada um dos melhores genes é achado o outro, com o qual ele faz o melhor par, aqui estamos aumentando a dimensão do clique de gene de um para dois. Para isto precisamos usar um determinado tipo de classificador, como por exemplo K -Vizinhos, Fisher, Máxima Verossimilhança, ou outro, e precisamos definir uma medida de qualidade para estar criando os grupos. O processo anteriormente definido para genes e pares pode ser extrapolado para trios, quadras, etc.

5.8 Classificação usando RL (Regressão Local)

A idéia desta proposta é usar Loess [Seção 2.5.1] na classificação. A seguir, o método é explicado sobre um exemplo prático. Primeiro são atribuídos valores zero ou um para cada uma das amostras Tumorais e não Tumorais, respectivamente. Depois é realizado um ajuste usando Loess, para determinados genes, escolhidos por algum critério, e, a partir dos seus níveis de expressão explicam-se os tipos (zero ou um). Desta forma, empregando a superfície de classificação por regressão local, podemos pensar na soma de quadrados dos resíduos para medir a qualidade de um determinado grupo de genes. Também usamos este método da mesma forma que os demais, comparando-os quanto ao número de acertos por validação cruzada, na qual para cada uma das amostras: primeiro ela é retirada, depois treina-se o classificador e finalmente a amostra inicialmente excluída é classificada e este resultado é comparado com o real, obtido pelos patologistas.

Figura 5.3: Exemplo de duas populações quaisquer 1 e 2 [35], nas quais uma discriminação linear não é apropriada. Sejam f_1 e f_2 as densidade empíricas das duas populações e R_1 , R_2 as referentes regiões de classificação. Também são apresentadas as curvas de contorno, assim como as curvas de separação por classificação linear e a ideal.

Tal classificador pode ser uma proposta muito adequada para determinados casos como, por exemplo, quando as nuvens referentes aos tipos de amostras não podem ser muito bem separadas por um plano p -dimensional, como visto na Figura 5.3. Em casos como este, o fato de se colocar mais uma dimensão, variável qualitativa, atribuída a cada amostra de acordo com seu tipo real e realizar o ajuste usando Loess, pode conduzir a uma melhor separação. O resultado obtido por este método e os outros anteriormente apresentados são comparados na Tabela 5.1.

5.9 Comparação dos métodos de classificação

Na Tabela 5.1 são apresentados os resultados obtidos para cada um dos algoritmos de classificação descritos anteriormente. Para cada um deles foi calculado o número de acertos por validação cruzada nas 99 amostras do conjunto de dados da Seção 6.8. Só estão sendo considerados os 18 genes mais DE nos pares de comparações, e só duas condições biológicas: Tumoriais e não Tumoriais. O critério usado na seleção dos genes DE foi o pv do teste t , isto considerando somente

5.9. COMPARAÇÃO DOS MÉTODOS DE CLASSIFICAÇÃO

cinco pares de condições biológicas, não levando em conta a comparação Normal-Gastrite, pela similaridade verificada nos dois estados biológicos. No caso do K -Vizinhos [Seção 5.1] a medida de similaridade empregada foi a Correlação. No caso da Regressão Local só foram usados quatro genes, os mais DE de cada um dos cinco pares de comparações. Já no *Aggregating* o método de classificação usado foi o K -Vizinhos com $k = 5$, veja Tabela 5.1.

Método	Acertos
<i>Aggregating</i> empregando <i>Bagging</i> paramétrico	99
K -Vizinhos com $k = (1, 5, 7)$	94
<i>Aggregating</i> empregando <i>Bagging</i> não paramétrico	94
Discriminante linear de Fisher	92
Máxima Verossimilhança	89
Regressão Local	77

Tabela 5.1: Avaliação dos métodos de classificação testados no conjunto de dados da Seção 6.8 usando validação cruzada. Foram considerados os 18 genes mais DE nos pares de comparações, e só duas condições biológicas: Tumoraís e não Tumoraís. No caso do K -Vizinhos a medida de similaridade empregada foi a Correlação. No caso da Regressão Local só foram usados quatro genes, os mais DE de cada um dos pares de comparações. Já no *Aggregating* não paramétrico o método de classificação usado foi o K -Vizinhos com $k = 5$, enquanto que no *Aggregating* paramétrico foi assumida a Normal Multivariada como a densidade das distribuições condicionais, da mesma forma que no método de Máxima Verossimilhança.

A seguir, na Tabela 5.9 são apresentados os resultados obtidos na procura de genes, pares, trios e quadras no conjunto de dados da Seção 6.8 usando o método BE. Os cliques de genes encontrados e a qualidade destes são comparados com os encontrados na busca exaustiva. Além disso foram estimados os pv no caso do BE, os quais também são apresentados na Tabela 5.9, na qual só foi considerada a comparação Normal-Tumor. A estimação dos pv foi realizada da seguinte forma: os tipos das amostras são permutados aleatoriamente, e o processo anteriormente descrito é realizado. Isto é feito um grande número de vezes e é computado a proporção de vezes que se observa um resultado melhor que quando usamos cada amostra com seu tipo original.

5.9. COMPARAÇÃO DOS MÉTODOS DE CLASSIFICAÇÃO

[H]	Método	Clique	Gene	Par	Trio	Quadra
	Busca Exaustiva	Grupo	COL4A1	COL4A1,XBP1	COL4A1,XBP1,RPL14	-
		Qualidade	11.03361	12.73118	14.94010	-
	BE com Fisher	Grupo	COL4A1	COL4A1,XBP1	COL4A1,XBP1,RPL14	COL4A1,CTSB,TOP2B,HS.326445
		Qualidade	11.03361	12.73118	14.94010	17.40290
		pv	0	0	0	0
	BE com RL	Grupo	COL4A1	CHGA,HS.327751	COL1A2,ATM,SPP1	COL1A2,NME1,HS.7243,PRO1051
		Qualidade	11.03361	0.4099842	0.7272739	1.6839779
		pv	0	0	0	0

Comparação dos melhores genes clique achados pelo método BE [Seção 5.7] usando Fisher [Seção 5.2] e Regressão Local [Seção 5.8] com os obtidos na busca exaustiva Tabela 6.3. No processo iterativo de *bootstrap* foi usado $N = 1000$. A medida de qualidade para genes foi o tv . Já para pares, trios e quadras foi o SVD e a soma de quadrado dos resíduos para o caso de Fisher e Regressão Local, respectivamente.

Capítulo 6

Análise de Dados de Experimentos de Expressão Gênica

No desenvolvimento deste trabalho tivemos a oportunidade de trabalhar com diferentes problemas reais, tanto de desenho experimental como de análise dos resultados obtidos dos experimentos. A parte relacionada ao acompanhamento de experimentos foi realizada no Instituto Ludwig, Instituto de Pesquisas de Câncer em São Paulo (<http://www.ludwig.org.br/>). Com relação à análise dos resultados, vários conjuntos de dados foram testados, levando à aplicação de diferentes técnicas estatísticas: análise preliminar, detecção de erros, normalização [40, 79], determinação dos genes mais DE [4, 22, 23, 47], testes paramétricos e não paramétricos, assim como questões de classificação e agrupamento.

6.1 Acompanhamento de experimentos de *Microarray*

Parte do trabalho de pesquisa incluiu a colaboração nos experimentos planejados dentro de nosso grupo MAIGES e realizados por Gustavo H. Esteves [27]. A participação envolveu todas as etapas de um experimento de *Microarray*: planejamento do desenho das lâminas, amplificação usando PCR, colocação nas lâminas pelo robô, fixação por laser ou forno, hibridização, digitalização e quantificação pelos aplicativos.

O fato de ter participado desde o começo do desenho do experimento [39, 37, 38, 78] foi muito interessante, pois possibilitou um planejamento cuidadoso de todos seus detalhes. A primeira etapa foi escolher a distribuição dos genes nos *spots*, que foi feita com a intenção de procurar padrões de dependência entre a expressão detectada pelo *scanner*, quantificada pelo aplicativo e a posição do *spot* na lâmina.

Um dos objetivos centrais da pesquisa com estes experimentos é gerar estimadores melhores, que meçam a quantidade de material colocado inicialmente. Por isso trabalhou-se com diferentes diluições para cada um dos seis cDNAs sintéticos usados: fago λ , *trpC*, *lysA*, *thrB*, *dapB* e *pheB* (bactéria *B. Subtilis*). Estas diluições foram feitas sob diferentes concentrações (%): 100, 50, 25, 12.5 e 6.25.

Outro objetivo central é avaliar alguns dos aplicativos de quantificação mais usados como Scanalyze [51], Quantarray [59] e Spot [34], e fazer comparações entre eles. Também entrou no estudo e nas comparações um aplicativo com este mesmo fim desenvolvido no BIOINFO [3, 2, 32]. Por exemplo no Scanalyze, há ferramentas para simplificar e semi-automatizar este processo. Em geral os aplicativos disponíveis têm uma parte manual, na qual normalmente se dá a introdução de erros sistemáticos. O aplicativo do BIOINFO usa princípios de morfologia matemática e tenta ser completamente automatizado, a partir do ponto em que são definidas as propriedades específicas do experimento.

Entre os tamanhos dos seis cDNAs existem diferenças significativas, que serão usadas para avaliar a influência que tem o tamanho no nível de expressão. Como tínhamos amostras muito pequenas, o primeiro passo consistiu na amplificação por PCR. Também buscamos “verificar” se a razão entre os sinais obtidos correspondem às razões das diluições que foram aplicadas. Em geral, temos três tipos de lâminas: *probe* e *target* nas mesmas concentrações, *probe* em concentração maior que o *target*, e *target* em concentração maior que o *probe*. Além das questões acima mencionadas, foi usada a inversão de corante, pois é conhecido que pode ser usada na “eliminação” dos

erros sistemáticos introduzidos por este reagente.

6.2 Genes DE (Diferencialmente Expressos)

Quando temos duas condições biológicas, por exemplo, Normal e Tumor e estamos estudando o perfil de expressão de um determinado conjunto de genes, podemos estar interessados em achar genes com “comportamento” mais diferenciado, o que é conhecido na literatura como genes DE [4, 22, 23, 47, 20, 79]. De fato, esta questão é uma das mais importantes quando se realiza um experimento para medir os níveis de expressão gênica. Esta busca é realizada da seguinte forma: primeiro temos que definir o critério que vai ser usado para ordenar os genes. Depois, para cada um dos genes, são tomadas as amostras que se tem para cada uma das duas condições biológicas e é calculado o critério estabelecido. Vale ressaltar que quando temos um conjunto de genes suficientemente grande, espera-se que a maioria deles seja praticamente constante nas duas condições, ou seja, quase sem variação, enquanto somente uma pequena proporção de genes sejam DE.

Neste trabalho têm sido usados os seguintes critérios: *Fold-Change*, t_v , assim como o pv tanto para o teste t, como para o teste de *Mann-Whitney*, ambos testes não pareado. O *Fold-Change*, $fc = \bar{X}_N - \bar{X}_T$, na qual $\bar{X}_N$ e $\bar{X}_T$ são as médias das amostras Normais e Tumoraes, respectivamente, faz uma comparação simples e não leva em conta a variabilidade do gene em cada uma das duas condições biológicas. Por outro lado, o t_v leva em conta as variâncias, $t_v = fc / \hat{\sigma}_{\bar{X}_N - \bar{X}_T}$, onde tem-se que:

$$\hat{\sigma}_{\bar{X}_N - \bar{X}_T} = \begin{cases} \sqrt{\frac{S_N^2}{n_N} + \frac{S_T^2}{n_T}}, & \text{se assumimos variâncias diferentes} \\ \sqrt{\frac{(n_N-1)S_N^2 + (n_T-1)S_T^2}{n_N + n_T - 2}}, & \text{se assumimos variâncias iguais} \end{cases}$$

Sabe-se que $t_v \sim t_{(n_N+n_T-2)}$ sob H_0 (e normalidade), na qual os graus de liberdade da estatística vão depender do número de amostras em cada uma das condições biológicas. É por isso que o pv parece ser o critério mais adequado, sobretudo quando queremos comparar, quanto à significância, genes indicados como DE em diferentes análises ou pares de condições biológicas. O

6.2. GENES DE (DIFERENCIALMENTE EXPRESSOS)

pv , $pv = P(Z \geq z \mid H_0)$, pode ser interpretado como a probabilidade de observar o valor registrado em um experimento assumindo H_0 verdadeira. Além do teste t, que assume normalidade nas duas amostras, existe o teste de *Mann-Whitney*, não paramétrico, que não faz considerações sobre as distribuições das amostras e é similar a um teste de medianas. A hipótese nula do teste de *Mann-Whitney* é que as medianas são iguais, enquanto que o teste t é com as médias iguais. Por seu caráter não paramétrico é que neste trabalho escolhemos o pv do teste de *Mann-Whitney* como o critério mais adequado na seleção dos genes DE.

Figura 6.1: Genes DE no conjunto de dados da Seção 6.8 nos seis pares de comparações, considerando as quatro condições biológicas: Normal (N), Gastrite (G), Metaplasia (M) e Tumor (T), sendo 28, 21, 22 e 28 as respectivas quantidades de amostras de cada um dos tipos. Os pontos vermelhos e verdes representam os 42 genes com $| -\log(pv) * \text{sinal}(\text{fold_change}) | > 7$. A cor vermelha foi usada para genes com expressão baixa na primeira condição do par e alta na segunda, enquanto que a cor verde foi usada para genes com expressão alta na primeira condição do par e baixa na segunda.

Na Figura 6.1 são apresentados os genes mais DE em cada um dos pares de condições biológicas. Foi usado o critério do *p*_v, e o *cut-off* foi definido em 7. Nota-se que a quantidade de genes DE encontrados em comparações que envolvem Tumor é relativamente alta em comparação com as demais, enquanto que, para a comparação Normal-Gastrite, não foi encontrado nenhum. À esquerda ficam os genes conhecidos como sub-expressos (mais no Tumor que no Normal, no caso da comparação Normal-Tumor), enquanto que à direita ficam os mais sobre-expressos (mais no Normal que no Tumor, no caso da comparação Normal-Tumor).

6.3 Análise do conjunto de dados de câncer de mama disponível na web

O câncer de mama [41, 29] é o câncer mais freqüente nas mulheres, sendo por isso alvo de muitas pesquisas nesta área. Um dos grupos que vem alcançando muitos resultados é o grupo de Perou C et. al. [56, 57, 62, 69], da Universidade de Stanford. Eles começaram fazendo coleta de amostras cirúrgicas de diferentes indivíduos com a doença. Este conjunto de dados se encontra disponível na *web* (http://genome-www.stanford.edu/breast_cancer/).

Um dos objetivos deste grupo foi fazer uma classificação dos carcinomas (câncer de origem epitelial) [71] de mama usando a técnica de *clustering*, tendo um total de 85 cDNAs *Microarrays* (78 câncer, 3 fibroadenoma e 4 normal) [69]. Outro aspecto interessante desses experimentos foi que eles conseguiram amostrar duas vezes, a segunda após 16 semanas, 20 dos tumores, com o objetivo de avaliar o efeito de um tratamento de quimioterapia [57].

Em nossas análise centramos a atenção no comportamento da expressão gênica associado à quimioterapia. Consideramos apenas parte do conjunto de dados referente à 40 lâminas, divididas em pares, com uma lâmina antes da quimioterapia e outra 16 semanas depois do tratamento para cada um dos 20 pacientes. São estudados um total de 9216 genes. Na Figura 6.2 são apresentados os dados normalizados referentes às 40 lâminas analisadas.

Figura 6.2: Distribuição empírica do $\log_2(T)$ normalizado referente às 40 lâminas.

Neste caso, não temos duas condições biológicas diferentes, mas a mesma, só que em instantes de tempo diferentes. O critério usado para achar os genes DE foi o tv . A Figura 6.3 compara a distribuição empírica dos tv com a distribuição normal padrão. Usando 3.5 como tv mínimo, obtivemos 22 genes com variação significativa, sendo que para 21 destes genes a expressão diminui. A Tabela 6.1 lista os 5 genes mais DE. Korn E et. al. [44] analisaram este conjunto de dados e identificaram 4 dos 5 genes listados na Tabela 6.1 como os mais importantes. Também verificamos que a expressão da maioria dos genes mais DE diminui. Isto mostra a concordância entre os resultados publicados e a nossa análise, atestando a corretude da mesma.

Figura 6.3: Comparação da distribuição empírica do tv do total de genes com a $N(0, 1)$.

Código, <i>GenBank Accession Num</i>	Nome	tv
AA040944	v-fos FBJ murine osteosarcoma viral oncogene homolog	-7
R12840	v-fos FBJ murine osteosarcoma viral oncogene homolog	-6.1
W96134	v-jun sarcoma virus 17 oncogene homolog (avian)	-5.6
AA485377	v-fos FBJ murine osteosarcoma viral oncogene homolog	-5.2
AA044993	connective tissue growth factor	-5.1

Tabela 6.1: Informação dos genes mais DE, verificada no NCBI [54] e SOURCE [70].

6.4 Análise do conjunto de dados de testes de *Microarrays* comerciais

Neste conjunto de dados gerado no INCOR (Instituto do Coração do Hospital das Clínicas) da USP e fornecido pela pesquisadora Silvana Chiavegatto, temos apenas duas lâminas, cada uma com 19200 *spots* onde foram fixados um total de 9366 genes. Para 9140 genes foram fixadas duas cópias, outros 218 genes foram fixados em quatro cópias cada e mais seis cópias foram feitas para

6.4. ANÁLISE DO CONJUNTO DE DADOS DE TESTES DE *MICROARRAYS* COMERCIAIS

os oito genes restantes. A idéia neste experimento, no qual temos poucos dados para cada gene, é avaliar a utilidade do *rank* (posicionamento), como medida para escolha dos genes mais DE, sem realizar o processo de normalização. Note que o trabalho com o *rank* é muito interessante pois dá a possibilidade de avaliar os dados “brutos”. No experimento temos dois mRNAs, que representam duas situações biológicas diferentes.

Começamos a análise do *rank* dos valores do sinal com a correção do *background* para os genes que têm apenas duas cópias em cada lâmina. Avaliamos a variação do *rank* para cada um dos genes, ordenando-os por nível de expressão. Como temos duas lâminas, cada uma com dois canais, e duas cópias por gene, teremos oito *rank*s para cada gene. Considerando cada um dos canais, definem-se:

x_{ij}^k o sinal no k -ésimo canal do j -ésimo gene na i -ésima lâmina;

$y_{ij} = | \text{rank}(x_{ij}^1) - \text{rank}(x_{ij}^2) |$, variação do *rank* do j -ésimo gene na i -ésima lâmina;

$z_j = | \text{rank}(y_{1j}) - \text{rank}(y_{2j}) |$, variação do *rank* do j -ésimo gene entre as duas lâminas.

Efeitos globais da lâmina não alteram o *rank*, pois alteram todas as medições da mesma forma, mantendo constante o posicionamento relativo. Desta forma, podemos identificar os genes mais DE nas duas condições. Esta técnica, usando o *rank* como medida de DE, é muito simples e gera um resultado rápido. Primeiro determina-se o quanto varia o *rank* dentro de cada lâmina, em cada um dos dois canais, (y_{ij}); depois determina-se a variação de y_{ij} entre as duas lâminas (z_j). Na Figura 6.4 a assimetria à direita da distribuição empírica mostra que a variação da maioria dos genes é menor que 1000 posições.

Figura 6.4: Histograma da distribuição empírica para z_j .

Foi feito o teste de *Mann-Whitney* para cada um dos genes nas duas situações biológicas e os resultados dos pv obtidos são apresentados na Figura 6.5.

Figura 6.5: Histograma dos pv do teste *Mann-Whitney*.

6.5 Análise do conjunto de dados de câncer de mama gerado no Instituto Ludwig

Neste conjunto de dados, gerado no Instituto Ludwig e fornecido pela pesquisadora Helena Brentani, temos onze lâminas sendo que, para cada uma delas, foi feita a inversão de corante, ou seja, temos no total, vinte e duas lâminas. O estudo envolveu onze pacientes, mulheres após a menopausa, antes e depois da quimioterapia, aplicada para tratar um tumor em estado avançado grau 3. A idéia é avaliar se há resposta ou não ao tratamento, indicada pela presença da substância cErbs. Além disso, tem-se para cada paciente, se o tratamento foi ou não efetivo. A amostra de mRNA de referência usada em todas as lâminas foi HB4a. A aplicação de quimioterapia, neste caso, visa diminuir o tamanho do tumor, a princípio até desaparecer, muito embora isto não tenha sido observado na prática. Depois da diminuição do tumor o próximo passo é a remoção do mesmo. Note-se a importância da quimioterapia nestes casos, nos quais a operação não seria possível com o tumor em estado avançado. Assim, estamos interessados em saber a chance da quimioterapia ter ou não efeito antes de sua aplicação.

Além disto, o estudo envolve dois tipos de cDNAs obtidos por Orestes [13] e a partir de fragmentos clonados, para os quais obtivemos resultados similares, apesar das diferenças entre os métodos de obtenção dos cDNAs. Na Figura 6.6 são apresentados os genes mais DE, onde comparamos o perfil de expressão gênica com a amostra de referência. Inicialmente analisamos a média do $\log_2()$ da razão das intensidades dentro das lâminas para cada gene, depois a média da diferença de intensidade entre pares formados por uma lâmina e sua cópia (com inversão de corante). Finalmente, determinamos o pv para um teste t com uma amostra, $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. Seleccionamos, então os genes com menor pv , apresentados na Figura 6.6, onde para cada gene é visualizado o *boxplot* da distribuição empírica do $\log_2(T)$ normalizado, considerando os onze pares de lâminas.

Figura 6.6: Genes mais DE, no eixo x temos a descrição dos genes e entre parênteses é indicado C (clone) ou F (fragmento). Para cada um dos genes é apresentado o *boxplot* da distribuição empírica do $\log_2(T)$ normalizado, considerando os onze pares de lâminas.

Note que, para os genes escolhidos, existe certa tendência dos genes no Tumor serem menos expressos que o HB4a, da mesma forma que também foram achados mais Clones que Fragmentos.

6.6 Análise do conjunto de dados de testes de protocolo bioquímico

Este foi um experimento realizado no Instituto Ludwig, no qual a concentração de material fluotante aplicado para hibridização em cada um dos canais foi controlada. Foram usadas três lâminas com proporções 1:1, 1:2 e 2:1 de mRNAs com fluoróforo verde e vermelho, respectivamente. O objetivo deste experimento foi avaliar o efeito destas proporções nas estimativas das intensidades obtidas para os dois canais nas três lâminas. Desta forma, estamos avaliando o efeito da quan-

6.6. ANÁLISE DO CONJUNTO DE DADOS DE TESTES DE PROTOCOLO BIOQUÍMICO

tidade de material flutuante em cada um dos canais sobre as estimativas de intensidade obtidas. Note que para analisar este efeito não é possível realizar o processo de normalização, pois este assume que ambos os canais apresentam a mesma intensidade em toda a lâmina. Se realizássemos a normalização, estaríamos eliminando o efeito conseguido pelo fato de controlar as proporções de material flutuante usadas nos dois canais para uma mesma lâmina.

No tipo de lâmina empregado o número de *spots* é 2592. Na Figura 6.7 é apresentado o resultado experimental obtido para os sinais nos dois canais nas três lâminas, sobre os quais foi aplicada apenas a correção de *background*.

Figura 6.7: Sinais medidos nas três lâminas, apenas com a correção de *background*.

Note a tendência do sinal ser mais forte no canal Verde, um padrão observado nos conjuntos de dados analisados neste trabalho. Na Tabela 6.2 são apresentados os resultados obtidos fazendo o

6.6. ANÁLISE DO CONJUNTO DE DADOS DE TESTES DE PROTOCOLO BIOQUÍMICO

ajuste de um modelo linear para cada uma das lâminas.

lâmina	Intersecção	erro-padrão	tv	pv	Inclinação	erro-padrão	tv	pv	R^2 (%)
1:1	23.750	2.835	8.377	<0.001	0.323	0.001	307.195	<0.001	97.33
1:2	18.651	4.278	4.360	<0.001	0.525	0.002	184.240	<0.001	92.91
2:1	52.690	1.510	34.880	<0.001	0.119	<0.001	234.450	<0.001	95.50

Tabela 6.2: Resultado obtido do ajuste do modelo linear, explicando a intensidade em um canal pela intensidade no outro, isto nas três lâminas.

Nota-se que os valores de intersecção obtidos ao considerar o intervalo onde variam as intensidades estão muito próximos de zero. Por outro lado, observa-se que o valor da inclinação estimada é menor do que 0.5, com exceção da lâmina 1:2, mas mesmo nesse caso ficou muito perto de 0.5. Isto demonstra a existência do fator de um efeito do canal na intensidade. Na Figura 6.8 são apresentadas as distribuições empíricas de M , observadas nas lâminas, a partir dos dados brutos.

Figura 6.8: Distribuições empíricas de M , observadas nos dados brutos (sem normalizar) nas lâminas 1:1, 1:2 e 2:1.

Se for feita alguma normalização, empregando um dos métodos apresentados neste trabalho, as linhas horizontais representando o valor médio desta distribuição é deslocada para zero. Assim os dados obtidos das três lâminas poderão ser comparados, desprezando as diferentes proporções aplicadas nos dois canais.

6.7 Análise do conjunto de dados “Amplificado vs não Amplificado”

Em pesquisas de expressão gênica que envolvem experimentos com *Microarray*, são necessárias amostras de cDNAs inicialmente obtidas em pequenas quantidades. Técnicas de amplificação por PCR (Reação de Polimerização em Cadeia) são usadas para aumentar a concentração de cDNAs nas amostras disponíveis.

O objetivo deste experimento é testar se os resultados obtidos variam significativamente quando as amostras de cDNAs são submetidas a amplificação. Visa também determinar a qualidade e a reprodutibilidade dos dados obtidos a partir de amostras amplificadas. O experimento foi desenhado com inversão de corante, usando 6 lâminas de vidro de 2592 *spots*, fixando nelas um total de 368 genes em três ou seis réplicas para cada um deles. Foram analisados os coeficientes de correlação, que indicam se a fidelidade do processo de amplificação é adequada e mostramos que esta técnica não provoca um erro sistemático, confirmando desta forma que as variações introduzidas pela amplificação são da mesma ordem das obtidas em função da escolha do método de normalização ou da escolha do método usado para definir a área de *background*. Os resultados obtidos [Apêndice A] indicam que o cDNA amplificado pode ser usado sem impacto negativo na qualidade dos dados gerados.

6.8 Análise do conjunto de dados de câncer de estômago

O conjunto de dados, gerado no Instituto Ludwig, pela pesquisadora Sibele Inacio Meireles, sob orientação do Prof. Luiz F. Reis, consiste em amostras de tecido estomacal colhidas de 99 pacientes, que foram classificadas como Normal, Gastrite, Metaplasia e Tumor (Difuso ou Intestinal). O tamanho de cada um dos grupos de amostras é 28, 21, 22 e 28, respectivamente e a amostragem foi realizada em três etapas, a primeira com 65 pacientes, a segunda com 12 e a última com 23. O substrato utilizado foi membrana de nylon. Na análise deste conjunto de dados [Apêndice B]

trabalhou-se com o total de 99 pacientes, citados acima. O estudo foi realizado com tecido de estômago, o que se justifica por, em mortes relacionadas com câncer [56, 57, 41, 62, 69, 29, 71] é precisamente o câncer de estômago a maior causa. Neste experimento, concentramos a atenção em um total de 604 clones representando 370 genes. Em geral, foram colocados na lâmina três cópias de cada clone na mesma sub-matriz. E, na maioria dos casos foram fixados um ou três clones por gene, embora existam alguns genes para os quais temos 39, 24 e 12 cópias.

O objetivo da análise foi a geração de classificadores que permitissem criar regras de decisão para o diagnóstico de novos pacientes. Pretendíamos, dessa forma, identificar padrões de expressão que permitissem diagnosticar tumores em formação ou já existentes, ajudando na escolha da estratégia terapêutica a ser usada.

A análise foi baseada no discriminante linear de Fisher, apresentado na Seção 5.2. A análise numérica foi implementada usando o *cluster* de 16 nós, *AMD* de 1.2 *GigaHz*, *RAM* de 726 e 30 *Gigabytes* de *HD* do BIOINFO/USP. Inicialmente, fizeram-se estimativas do tempo de cálculo necessário, que resultou da ordem de meses para o caso da busca exaustiva de trios de clones. Foi então realizada uma busca exaustiva, porém apenas para pares de genes e clones e, no caso de trios, apenas para genes. Tal análise, usando seis computadores, exigiu três semanas. A procura exaustiva foi feita sobre o primeiro subconjunto de amostras de 65 pacientes, e os classificadores obtidos foram treinados e testados na segunda e terceira amostras. Também, neste conjunto de dados foram usadas várias opções de normalização pela energia total, apresentadas na Seção 2.1.1, o que aumentou a robustez, embora os resultados apresentados foram baseados na normalização pela energia total pela média.

Uma questão interessante neste conjunto de dados é a busca de genes para os quais temos estimativas de expressão com comportamento monotônico (crescente ou decrescente) nas condições biológicas: Normal, Gastrite, Metaplasia e Tumor. Note que Normal refere-se a uma amostra sem nenhuma lesão, Gastrite significa uma lesão muito leve. Já a Metaplasia é uma lesão avançada,

6.8. ANÁLISE DO CONJUNTO DE DADOS DE CÂNCER DE ESTÔMAGO

que não necessariamente evolui para Tumor, muito embora as duas sejam similares, sendo Tumor a forma mais grave e avançada da lesão. Nas Figuras 6.9 e 6.10 são apresentados os sete genes com melhor comportamento monotonicamente crescente e decrescente, respectivamente nas patologias.

Figura 6.9: Genes com melhor comportamento monotonicamente crescente nas patologias. Ao lado de cada gene tem-se o pv do teste t referente à comparação Normal-Tumor (só as tumorais tipo Intestinal), e abaixo o pv do teste de *Mann-Whitney* da mesma comparação. Acima de cada degrau tem-se o pv do teste t na comparação que ele representa. A altura de um degrau é $-\log(pv)$, na qual o pv é o referente ao teste t .

Figura 6.10: Genes com melhor comportamento monotonicamente decrescente nas patologias. Ao lado de cada gene tem-se o pv do teste t referente à comparação Normal-Tumor (só as tumorais tipo Intestinal), e abaixo o pv do teste de *Mann-Whitney* da mesma comparação. Acima de cada degrau tem-se o pv do teste t na comparação que ele representa. A altura de um degrau é $-\log(pv)$, na qual o pv é o referente ao teste t .

A busca exaustiva foi realizada sobre genes, clones, pares de genes, pares de clones e trios de genes, não sendo viável para trios de clones dado o tempo de processamento computacional necessário. Concentramos nossa atenção nos grupos de genes. Os resultados obtidos na busca exaustiva para trios de genes são apresentados na Tabela 6.3. Estudando detalhadamente os pares considerados bons percebemos que os genes bons estavam repetidos várias vezes. O mesmo foi observado para os trios bons, nos quais repetem-se os pares bons, numa determinada comparação.

Comparação	0	1	2	3
Normal-Tumor	1044	12278	45636	66428
Normal-Metaplasia	0	52	354	10810
Gastrite-Tumor	17	512	4741	27548
Normal-Gastrite	0	0	0	2
Gastrite-Metaplasia	0	4	110	1223
Metaplasia-Tumor	1	19	808	24648

Tabela 6.3: Melhores trios de genes achados na busca exaustiva. O valor x na posição (i, j) da tabela indica que na comparação i do total de 8,510,740 classificadores, exatamente x deles classificaram, erroneamente, segundo a validação cruzada, j amostras das 99. Por exemplo, na comparação Normal-Tumor, 1044 dos 8,510,740 não erraram em nenhuma.

Além de medir a qualidade dos classificadores pela validação cruzada, foi usada a razão proposta por Fisher, R_{ed} , como qualidade para ordená-los. Na Figura 6.11 é apresentado o “melhor” trio de genes na comparação Normal-Tumor. Graficamente nota-se a criação de dois grupos, um com a maioria das Tumoraes e outro com a maioria das não Tumoraes (Normais, Gastrite e Metaplasias).

Figura 6.11: Separação realizada nas amostras pelo melhor trio de genes achado na busca exaustiva.

Neste conjunto de dados foram testados os métodos de agrupamento apresentados neste trabalho. Inicialmente foram feitos alguns testes para o caso do agrupamento hierárquico; na Figura 6.12 é apresentado o dendograma resultante usando os mesmos 18 genes usados na Seção 5.9 e usando a correlação como medida de similaridade.

Figura 6.12: Dendrograma para os 18 genes mais DE usando a Correlação como medida de similaridade e Divisão [Seção 4.2], como método de união.

Note-se que são criados dois grupos bem marcados, um com a maioria das Tumorais e outro com a maioria das não Tumorais (Normais, Gastrite e Metaplasias), como observado na Figura 6.11. Outro caso interessante é, por exemplo, quando fazemos o dendrograma usando somente os dois genes mais DE da comparação Normal-Tumor, o qual é apresentado na Figura 6.13.

Figura 6.13: Dendrograma para os 2 genes mais DE no Normal-Tumor usando a Correlação como medida de similaridade e Divisão [Seção 4.2], como método de união.

Os melhores resultados foram obtidos usando K -Médias, onde também foram usados os 18 genes e a Correlação como medida de similaridade. Na Figura 6.14 são apresentados os agrupamentos produzidos por este método para o caso de quatro e três grupos. Sendo quatro o número de grupos na primeira tentativa pelo fato de que temos quatro tipos de amostras. Para quatro grupos obtivemos um grupo que contém a maioria das Normais e Gastrite, outro com a maioria das Metaplasias, outro com a maioria das tumorais, e um pequeno grupo com seis amostras, contendo uma ou duas de cada um dos tipos. É interessante observar que as seis amostras do grupo menor sempre foram indicadas pela análise como amostras com comportamento anômalo em relação ao conjunto de dados. No caso de três grupos é criado um grupo com a maioria das Normais e Gastrite, um com a maioria das metaplasias e outro com a maioria das Tumorais. Vale destacar que de quatro para três não se observaram mudanças significativas, uma vez que os grupos ficaram muito similares, sendo as seis amostras no grupo menor incorporadas ao grupo das Tumorais.

Figura 6.14: *K*-Médias para (a) quatro e (b) três grupos respectivamente. São consideradas as 99 amostras, os 18 genes mais DE nos pares de comparações e é usada a Correlação como medida de similaridade.

6.8. ANÁLISE DO CONJUNTO DE DADOS DE CÂNCER DE ESTÔMAGO

Outro método de agrupamento testado foi o SOM, para o qual usamos diferentes configurações e topologias. Os resultados obtidos com este método seguem apresentados nas Figuras 6.16, 6.17 e 6.18.

Figura 6.15: Disposição das amostras nas topologias do SOM. São consideradas as 99 amostras e os 18 genes mais DE nos pares de comparações.

Figura 6.16: SOM para topologia (a) Retangular 1x2 e (b) Retangular 1x3 respectivamente. São consideradas as 99 amostras e os 18 genes mais DE nos pares de comparações.

Figura 6.17: SOM para topologia Retangular 1x4. São consideradas as 99 amostras e os 18 genes mais DE nos pares de comparações.

Figura 6.18: SOM para topologia (a) Retangular 2x2 e (b) Hexagonal 2x2 respectivamente. São consideradas as 99 amostras, os 18 genes mais DE nos pares de comparações.

Conclusões

Neste trabalho foram estudados diferentes conjuntos de dados de *Microarray*, considerando a complexidade envolvida na análise e interpretação de tais dados. Foram apresentadas e discutidas questões estatísticas específicas desses tipos de dados, como aquelas referentes a normalização, por exemplo. Também foram tratadas técnicas multivariadas (Agrupamento e Classificação), considerando as características dos experimentos. Todas as técnicas foram introduzidas teoricamente, testadas e comparadas em dados experimentais tanto de colaboradores quanto dados disponíveis na *web*.

Os diferentes métodos de normalização discutidos são direcionados a compensar diversas fontes de variação sistemáticas, assumindo diferentes hipóteses de modelagem. Assim, não está muito claro como compará-los. No conjunto de dados da Seção 6.8 foi empregada a normalização pela energia total usando a média, a mediana e a média aparada e verificou-se que os classificadores, assim como a qualidade destes não variou muito em função do método empregado para normalização. Também consideramos a normalização empregando Loess (a qual contém, como caso particular, a correção feita pela energia total) pelo fato dela compensar o efeito da média das intensidades nos dois canais (A) na razão das intensidades (M , medida de interesse).

Nas técnicas de Agrupamento, testadas nas amostras do conjunto de dados da Seção 6.8 segundo os níveis de expressão dos genes verificou-se que empregar o total de genes não é a melhor solução, embora também não o seja empregar uma quantidade muito pequena. Escolhemos pequenos grupos de genes que se mostrassem “suficientes” para distinguir amostras em condições

biológicas diferentes. Entre as diferentes idéias testadas para escolha de genes diferencialmente expressos, o *pv* do teste de *Mann-Whitney*, nos parece ser muito adequada pelas suas propriedades. Já no caso de pares, trios, quadras, etc, a medida de qualidade vai depender do tipo de classificador empregado. O agrupamento Hierárquico, o qual tem como finalidade a geração de só um grupo, no qual aparecem as agrupações ordenadas numa hierarquia por similaridade, nos permitiu uma primeira visão dos dados. Por outro lado o *K*-Médias e SOM proporcionam uma divisão num número de grupos, previamente especificado; os resultados obtidos para estes dos métodos foram muito similares. Exploramos o SOM detalhando nas diferentes topologias para este método de redes neurais, variando a configuração da rede bidimensional, usada no processo de separação. No conjunto de dados da Seção 6.8, os melhores resultados foram obtidos com o *K*-Médias considerando as características e as especificações do tipo para cada amostra, bem como o agrupamento das mais similares.

Em Classificação foram comparados diversos métodos e verificou-se que cada um deles pode ajustar-se melhor ou pior, dependendo das características próprias de cada conjunto de dados. Os resultados obtidos apontam que os métodos que envolvem *bootstrap* podem gerar bons resultados. No conjunto de dados da Seção 6.8 verificou-se que para o *K*-Vizinhos os resultados também foram muito bons. Outras questões, tais como a Classificação por Regressão Local e a proposta para alerta no diagnóstico, parecem ser tópicos muito interessantes para estudo.

Verificou-se que a eficiência do método *K*-Vizinhos, considerando o número de acertos por validação cruzada, melhora consideravelmente para quase todo *k* (número de vizinhos envolvidos no classificador) quando:

1. Usamos a Correlação como medida de similaridade;
2. Assumimos as classes Normal e Gastrite como uma classe só;
3. Só usamos os 18 genes mais DE;
4. Temos só duas classes: Tumerais e não Tumerais.

CONCLUSÕES

O método BE proposto consegue encontrar de forma eficiente e rápida os melhores clique de genes, sem ser necessário realizar uma busca exaustiva, a qual pode levar meses dependendo da quantidade de genes estudados.

Verificou-se a maior eficiência, considerando o número de acertos por validação cruzada, para o método *Aggregating* empregando *Bagging* paramétrico, para o qual obtivemos 100% de acertos no conjunto de dados da Seção 6.8.

Trabalhos futuros

A partir do ponto atingido pela pesquisa neste trabalho, as questões teóricas envolvidas ficaram bem definidas. Em sua continuação, poderia-se dedicar maior esforço maior às aplicações dos métodos (normalização, classificação e agrupamento) em diversos conjuntos de dados e comparar os resultados obtidos dentro e entre conjuntos de dados. Também poderia-se ajustar o modelo (misto não linear, Capítulo 3), assim como a estimação dos seus parâmetros. Quanto ao diagnóstico do modelo poderia-se avaliar a forma mais adequada de validação a partir da bibliografia existente.

Referências Bibliográficas

- [1] Armelin H et. al. **“CAGE - Cooperation for Analysis of Gene Expression”**. Project Proposal proj FAPESP n 99/07390-0, 1999.
- [2] Armelin H, Barrera J, Hirata R and Neves E. **“Microarray image analysis”**. Technical Report, August 2000.
- [3] Barrera J, Cesar R, Ferreira J, Gubitoso M, Hirata N, Hirata R and Neves E. **“An integrated environment for storage and analysis of genetic data”**. Relatório Técnico BIOINFO 001/2000, Setemember 2000.
- [4] Bassett D, Eisen M and Boguski M. **“Gene expression informatics - it’s all in your mine”**. Nature Genetics Vol 21 pg 51-55, 1999.
- [5] **“BIOINFO/USP”**. <http://www.bioinfo.usp.br>
- [6] Bittner M. **“Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling”**. Nature Vol 406 pg 536-540, August 2000.
- [7] Bolfarini H and Sandoval M. **“Introdução a Inferência Estatística”**. SBM (Sociedade Brasileira de Matemática), 2001.
- [8] Brazma A, Hingamp P, Quackenbush J et. al. **“Minimum information about a microarray experiment (MIAME)-toward standards for microarray data”**. Nature Genetics Vol 29, December 2001.
- [9] Breiman L, Friedman J, Olshen R and Stone C. **“Classification and Regression Trees”**. Chapman & Hall, New York 1984.
- [10] Brown M, Grundy W, Lin D, Cristianini N, Sugnet C, Ares M and Haussler D. **“Support Vector Machine classification of Microarray Gene Expression Data”**. Technical Report UCSC-CRL-99-09, University of California, June 1999.
- [11] Bussab W. **“Análise de Variância e Regressão”**. Editora Atual s/a, 1999.
- [12] **“CAGE - Cooperation for Analysis of Gene Expression”**.
<http://www.vision.ime.usp.br/çage/>
- [13] Camargo A, Samaia H, Neto E et. al. **“From the Cover: The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome”**. PNAS Vol 98 pg 12103-12108, 2001.
- [14] **“Câncer de Estômago”**. <http://www.hcanc.org.br/dmeds/abdo/abdo3.html>
- [15] **“Centro de Tratamento e Pesquisa, Hospital do Câncer, A. C. Camargo”**. São Paulo, Brasil.
<http://www.hcanc.org.br/>

- [16] Cleveland W. **“Robust Locally Weighted Regression and Smoothing Scatterplots”**. American Statistical Association Vol 74 # 368, December 1979.
- [17] **“Computational System Biology Reviews”**.
http://lansbury.bwh.harvard.edu/computational_system_biology_review_2003.htm. 34 references, 2003.
- [18] Davison A and Hinkley D. **“Bootstrap Methods and their Application”**. Cambridge University Press, 1997.
- [19] Debouck C and Goodfellow P. **“DNA Microarrays in Drug Discovery and Development”**. Nature Genetics Supplement Vol 21 pg 48-50, January 1999.
- [20] Dudoit S, Fridlyand J and Speed T. **“Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data”**. Technical Report # 576, June 2000.
- [21] Dudoit S, Shaffer J and Boldrick J. **“Multiple Hypothesis Testing in Microarray Experiments”**. U.C. Berkeley Division of Biostatistics Working Paper Series, Paper 110, 2002.
- [22] Dudoit S, Yang Y, Callow M and Speed T. **“Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments”**. Technical Report # 578, August 2000.
- [23] Duggan D, Bittner M, Chen Y, Meltzer P and Trent J. **“Expression profiling using cDNA microarrays”**. Nature Genetics Vol 21 pg 10-14, 1999.
- [24] Edwards D. **“Non-linear normalization and background correction in one-channel cDNA microarray studies”**. Bioinformatics Vol 19, 2003.
- [25] Eisen M, Spellman P, Brown P and Botstein D. **“Cluster analysis and display of genome-wide expression patterns”**. Proc Natl Acad Sci USA Vol 95 pg 14863-14868, 1998.
- [26] Efron B and Tibshirani R. **“An Introduction to the Bootstrap”**. CRC Press LLC, 1998.
- [27] Esteves G. **“Validação de procedimentos para medida de expressão gênica a partir de imagens de cDNA microarray”**. Master’s thesis, Instituto Ludwig de Pesquisas contra o Câncer, 2002.
- [28] Fuhrman S, Cunningham M, Wen X, Zweiger G, Seilhamer J and Somogyi R. **“The application of Shannon entropy in the identification of putative drug targets”**. Biosystems Vol 55 pg 5-14, February 2000.
- [29] Furey T, Cristianini N, Duffy N, Bednarski D, Schummer M and Haussler D. **“Support vector machine classification and validation of cancer tissue samples using microarray expression data”**. Bioinformatics Vol 16, pg 906-914, 2000.
- [30] **“GenePix Pro 5.0 Microarray Image Analysis”**. http://www.axon.com/gn_GenePixSoftware.html
- [31] Golub T et. al. **“Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring”**. Science Vol 286 pg 531-537, October 1999.
- [32] Hirata R, Barrera J, Hashimoto R, Dantas D and Esteves G. **“Segmentation of Microarray Images by Mathematical Morphology”**. Real-Time Imaging Vol 8 pg 491-505, December 2002.
- [33] Jain A, Duin R and Mao J. **“Statistical Pattern Recognition: A Review”**. IEEE Transactions on Pattern Analysis and Machine Intelligence Vol 22, January 2000.

- [34] Jain A, Tokuyasu T, Snijders A, Segraves R, Albertson D and Pinkel D. **“Fully Automatic Quantification of Microarray Image Data”**. Genome Research Vol 12 pg 325-332, 2002.
- [35] Johnson R and Wichern D. **“Applied Multivariate Statistical Analysis”**. New Jersey: Prentice-Hall, third edition, 1995.
- [36] Kerr M and Churchill G. **“Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments”**. PNAS Vol 98 pg 8961-8965, July 2001.
- [37] Kerr M and Churchill G. **“Design and Analysis of Gene Expression Microarray Experiments”**. http://www.jax.org/staff/gary_churchill.html
- [38] Kerr M and Churchill G. **“Experimental design for gene expression microarrays”**. Biostatistics Vol 2 pg 183-201, 2001.
- [39] Kerr M, Leiter E, Picard L and Churchill G. **“Analysis of a designed microarray experiment”**. Proceedings of the IEEE-Eurasip Nonlinear Signal and Image Processing Workshop, June 3-6 2001.
- [40] Kerr M, Martin M and Churchill G. **“Analysis of variance for gene expression microarray data”**. Comput Biol Vol 7 pg 819-37, 2000.
- [41] Khan J et. al. **“Expression profiling in cancer using cDNA microarrays”**. Electrophoresis Vol 20 pg 223-229, 1999.
- [42] Khan J, Bittner M, Chen Y, Meltzer P and Trent J. **“DNA microarray technology: the anticipated impact on the study of human disease”**. BBA (Biochimica et Biophysica Acta) - Reviews on Cancer, Vol 1423 pg 17-28, March 1999.
- [43] Kohonen T. **“Self-Organizing Maps”**. Springer, second edition, 1997.
- [44] Korn E, McShane L, Troendle J, Rosenwald A and Simon R. **“Identifying pre-post chemotherapy differences in gene expression in breast tumours: a statistical method appropriate for this aim”**. British Journal of Cancer, January 2002.
- [45] Krause A and Olson M. **“The Basic of S and S-Plus”**. Statistics and Computing Springer, 1997.
- [46] Laan M and Bryan J. **“Gene Expression Analysis with the Parametric Bootstrap”**. Biostatistics Vol 2 pg 1-17, 2001.
- [47] Lazzeroni L and Owen A. **“Plaid Models for Gene Expression Data”**. Stanford University, March 2000.
- [48] Mardia K, Kent J and Bibby J. **“Multivariate Analysis”**. New York: Academic Press, 1979.
- [49] **“MATLAB The Language of Technical Computing”**. Using Matlab, version 6. Math Works, Inc 2000.
- [50] Meireles S, Carvalho A, Hirata R, Montagnini A, Martins W, Runza F, Stolf B, Termini L, Neto C, Silva R, Soares F, Neves E and Reis L. **“Differentially expressed genes in gastric tumors identified by cDNA array”**. Cancer Letters Vol 190 pg 199-211, 2003.
- [51] **“Microarrays Image Analysis, IMAGE QUANTITATION using ScanAlyze”**. http://sib-dea.unil.ch/dqnguyen/module_8/ExoMicroarraysImageAnalysis/ExoImageAnalysis.html

- [52] Mukherjee S, Tamayo P, Slonin D, Verri A, Golub T, Mesirov J and Poggio T. **“Support Vector Machine Classification of Microarray Data”**. Technical Report MIT, 1998.
- [53] Nature Genetics Supplement Vol 21 no. 1, January 1999.
- [54] **“NCBI (National Center for Biotechnology Information)”**. <http://www.ncbi.nlm.nih.gov/>
- [55] Nurse P. **“Regulation of the Eukaryotic Cell Cycle”**. European Journal of Cancer Vol 33 pg 1002-1004, 1997.
- [56] Perou C et. al. **“Distinctive gene expression patterns in human mammary epithelial cells and breast cancers”**. PNAS Vol 96 pg 9212-9217, August 1999.
- [57] Perou C et. al. **“Molecular portraits of human breast tumours”**. Nature Genetics Vol 406 pg 747-752, August 2002.
- [58] Quackenbush J. **“Microarray data normalization and transformation”**. Nature Genetics Vol 32 pg 496-501, December 2002.
- [59] **“QuantArray Microarray Analysis Software Operating Manual Version 2.0.”**. GSI LUMONICS.
- [60] **“R”**. <http://www.r-project.org/>
- [61] Rocke D and Durbin B. **“A Model for Measurement Error for Gene Expression Arrays”**. Computational Biology Vol 8 pg 557-569, 2001.
- [62] Ross D and Perou C. **“A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines ”**. Disease Markers Vol 17 pg 99-109, 2001.
- [63] Saal L, Troein C and Christersson J. **“BioArray Software Enviroment (BASE): a platform for comprehensive management and analysis of microarray data”**. Software Report, July 2002.
- [64] Schena M. **“Microarray Analysis”**. John Wiley & Sons, 2002.
- [65] Sebastiani P, Gussoni E, Kohane I and Ramoni M. **“Statistical Challenges in Functional Genomics”**. Statistical Science, December 2002.
- [66] Shaffer J. **“Multiple hypothesis testing”**. Annu Rev Psychol Vol 46 pg 561-584, 1995.
- [67] Smolen P, Baxter D and Byrne J. **“Modeling Transcriptional Control in Gene Networks-Methods, Recent Results, and Future Directions”**. Bulletin of Mathematical Biology, Vol 62 pg 247-292, 2000.
- [68] Smyth G, Yang Y and Speed T. **“Statistical Issues in cDNA Microarray Data Analysis”**. Research Report, May 2002.
- [69] Sorlie T, Perou C, Tibshirani R et. al. **“Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications”**. PNAS Vol 98 pg 10869-10874, September 2001.
- [70] **“SOURCE (Stanford Online Universal Resource for cDNA Clones & ESTs)”**. <http://genome-www.stanford.edu/source/>
- [71] Szallasi Z and Liang S. **“Modeling the Normal and Neoplastic Cell Cycle With ‘Realistic Boolean Genetic Networks’: Their Application for Understanding Carcinogenesis and Assessing Therapeutic Strategies”**. In PSB Proceedings Vol 3 pg 66-76, 1998.

- [72] Tamayo S, Mesirov J, Golub T and Lander E. **“Class prediction and Discovery Using Gene Expression Data”**. Technical report Whitehead/MIT Center for Genome Research, 1999.
- [73] Tibshirani R, Hastie T, Eisen M, Ross D, Botstein D and Brown P. **“Clustering methods for the analysis of DNAmicroarray data”**. Stanford University, October 1999.
- [74] Tibshirani R, Walther G and Hastie T. **“Estimating the number of clusters in a dataset via the Gap statistic”**. JRSSB, March 2000.
- [75] Venables W and Ripley B. **“Modern Applied Statistics with S-PLUS”**. Springer. ISBN 0-387-98825-4, third edition, 1999.
- [76] Vingron M. **“Computational analysis of expression data”**. Workshop on computation of biochemical pathways and genetic networks, Heidelberg 1999.
- [77] Westfall P and Young S. **“Resampling-based multiple testing: examples and methods for p-value adjustment”**. Wiley series in probability and mathematical statistics. Wiley, 1993.
- [78] Winer B, Brown D and Michels K. **“Statistical Principles in Experimental Design”**. McGraw-Hill Inc, 1991.
- [79] Yang Y, Dudoit S, Luu P and Speed T. **“Normalization for cDNA microarray data”**. Proceedings of Photonics West, Biomedical Optics Symposium (BiOS), The International Society of Optical Engineering (SPIE), San Jose, California, 2001.

Capítulo 7

Apêndices: Artigos associados à presente dissertação

Apêndice A

Comparative analysis of amplified and nonamplified RNA for hybridization in cDNA microarray

Apêndice B

MOLECULAR CLASSIFIERS FOR GASTRIC CANCER AND PRE-MALIGNANT DISEASES OF THE GASTRIC MUCOSA (in print)