Estudo de coeficientes de correlação para
medidas de proximidade em dados de
expressão gênica

Pablo Andretta Jaskowiak

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Estudo de coeficientes de correlação para medidas de proximidade em dados de expressão gênica¹

Pablo Andretta Jaskowiak

Orientador: *Prof. Dr. Ricardo José Gabrielli Barreto Campello*

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA.*

USP – São Carlos
Abril/2011

¹ Trabalho realizado com auxílio financeiro do CNPq.

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

A39e Andretta Jaskowiak, Pablo
 Estudo de Coeficientes de Correlação para Medidas
de Proximidade em Dados de Expressão Gênica / Pablo
Andretta Jaskowiak; orientador Ricardo José
Gabrielli Barreto Campello -- São Carlos, 2011.
140 p.

 Dissertação (Mestrado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2011.

 1. Coeficientes de Correlação. 2. Análise de Dados
de Expressão Gênica. 3. Seleção de Atributos. 4.
Classificação. 5. Agrupamento. I. José Gabrielli
Barreto Campello, Ricardo, orient. II. Título.

Aos meus pais.

Agradecimentos

A toda minha família, em especial aos meus pais, Lidete e Paulo, por todo o apoio, amor e carinho. Sem seu incentivo e auxílio eu certamente não teria chegado até aqui.

Ao Professor Dr. Ricardo José Gabrielli Barreto Campello, meu orientador, pela oportunidade de trabalhar neste projeto e por toda confiança e empenho em mim depositados. Sua orientação além de possibilitar a realização e conclusão deste trabalho teve grande impacto positivo tanto em meu crescimento pessoal quanto profissional.

Ao Professor Dr. Ivan Gesteira Costa Filho, pelas discussões, sugestões e conselhos relativos ao meu trabalho.

Aos colegas e amigos de laboratório, Bruno, Danilo, Débora, Jonathan, Lucas, Luiz, Marcelo, Márcio, Murilo, Ricardo, Rosane e Thiago. Em especial, agradeço ao Thiago pelas discussões, sugestões e conselhos pertinentes ao meu trabalho.

Aos amigos e amigas de São Carlos-SP, sem vocês meu mestrado teria acabado antes, ou talvez não teria acabado. Por vários momentos vocês foram minha válvula de escape e me deram a tranquilidade necessária para seguir em frente. Ao Adriano, Anderson, André Domingues, André Vicente, Cássio, Cyntia, David, Heline, Jean, Lucas, Gabriel, Rafael, Renato, Roberto e Tácito.

A todos meus amigos de Cascavel-PR, pelos curtos, porém inesquecíveis momentos de paz e tranquilidade que pude ter quando retornei para casa. Em especial agradeço ao grande amigo e irmão Thiago Schuh e ao amigo Carlos Diógenes.

Aos professores e professoras do ICMC-USP que de alguma forma participaram em minha formação. Em especial, agradeço ao Professor Dr. Eduardo Raul Hruschka.

Aos Professores do curso de Informática da Universidade Estadual do Oeste do Paraná. Todos vocês foram de fundamental importância em minha formação. Em especial, agradeço ao Professor Josué Pereira de Castro e à Professora Adriana Postal.

A todos os funcionários do ICMC-USP, em especial as funcionárias da seção de pós-graduação, pela paciência, gentileza e prontidão em seu atendimento. Aos guardas da portaria do bloco 6, pelas conversas e pelo café sempre quente nas madrugadas também quentes de São Carlos.

Finalmente, agradeço ao CNPq pelo suporte financeiro para a realização do trabalho.

Resumo

O desenvolvimento da tecnologia de *microarray* tornou possível a medição dos níveis de expressão de centenas ou até mesmo milhares de genes simultaneamente para diversas condições experimentais. A grande quantidade de dados disponível gerou a demanda por métodos computacionais que permitam sua análise de forma eficiente e automatizada. Em muitos dos métodos computacionais empregados durante a análise de dados de expressão gênica é necessária a escolha de uma medida de proximidade apropriada entre genes ou amostras. Dentre as medidas de proximidade disponíveis, coeficientes de correlação têm sido amplamente empregados, em virtude da sua capacidade em capturar similaridades entre tendências das sequências numéricas comparadas (genes ou amostras). O presente trabalho possui como objetivo comparar diferentes medidas de correlação para as três principais tarefas envolvidas na análise de dados de expressão gênica: agrupamento, seleção de atributos e classificação. Dessa forma, é apresentada nesta dissertação uma visão geral da análise de dados de expressão gênica e das diferentes medidas de correlação consideradas para tal comparação. São apresentados também resultados empíricos obtidos a partir da comparação dos coeficientes de correlação para agrupamento de genes, agrupamento de amostras, seleção de genes para o problema de classificação de amostras e classificação de amostras.

Abstract

The development of microarray technology made possible the expression level measurement of hundreds or even thousands of genes simultaneously for various experimental conditions. The huge amount of available data generated the need for computational methods that allow its analysis in an efficient and automated way. In many of the computational methods employed during gene expression data analysis the choice of a proximity measure is necessary. Among the proximity measures available, correlation coefficients have been widely employed because of their ability to capture similarity trends among the compared numeric sequences (genes or samples). The present work has as objective to compare different correlation measures for the three major tasks involved in the analysis of gene expression data: clustering, feature selection and classification. To this extent, in this dissertation an overview of gene expression data analysis and the different correlation measures considered for this comparison are presented. In the present work are also presented empirical results obtained from the comparison of correlation coefficients for gene clustering, sample clustering, gene selection for sample classification and sample classification.

Sumário

Lista de Abreviaturas	xix
Lista de Figuras	xxi
Lista de Tabelas	xxiii
Lista de Algoritmos	xxvii
1 Introdução	1
1.1 Motivação	2
1.2 Objetivo	3
1.3 Organização do Texto	3
2 Fundamentos Biológicos	5
2.1 Ácidos Nucléicos	6
2.1.1 Ácido Desoxirribonucleico	6
2.1.2 Ácido Ribonucleico	7
2.2 Proteínas	7
2.3 Expressão Gênica	8
2.4 Considerações Finais	12
3 Análise de Dados de Expressão Gênica	13
3.1 Tecnologias de <i>Microarray</i>	14
3.1.1 <i>Microarrays</i> de cDNA	15
3.1.2 <i>Microarrays</i> de Oligonucleotídeos	17
3.2 Obtenção dos Dados	20
3.3 Pré-Processamento	23
3.4 Tarefas Comuns e Aplicações	24
3.4.1 Agrupamento	25
3.4.2 Seleção de Atributos	28
3.4.3 Classificação	30

3.5	Medidas de Proximidade em Dados de Expressão Gênica	32
3.6	Considerações Finais	36
4	Medidas de Proximidade	37
4.1	Coefficientes de Correlação	37
4.1.1	Pearson	38
4.1.2	Spearman	39
4.1.3	Kendall	40
4.1.4	Goodman-Kruskal	40
4.1.5	Goodman-Kruskal Ponderado	41
4.1.6	Rank-Magnitude	42
4.1.7	Jackknife	42
4.2	Medidas Baseadas em Coeficientes de Correlação	43
4.2.1	Son e Baek	43
4.3	Dissimilaridade <i>Short Time-Series</i>	45
4.4	Considerações Finais	46
5	Experimentos em Agrupamento	47
5.1	Algoritmos Utilizados	47
5.1.1	<i>k-medoids</i>	48
5.1.2	Hierárquicos Aglomerativos	49
5.2	Experimentos com Amostras	49
5.2.1	Metodologia	50
5.2.2	Bases de dados	52
5.2.3	Resultados	52
5.2.4	Compilação dos Resultados	62
5.3	Experimentos com Genes	64
5.3.1	Metodologia	64
5.3.2	Bases de Dados	67
5.3.3	Resultados	68
5.3.4	Compilação dos Resultados	74
5.4	Considerações Finais	75
6	Experimentos em Seleção de Atributos e Classificação	77
6.1	Algoritmos Utilizados	77
6.1.1	<i>Simplified Silhouette Filter</i>	78
6.1.2	<i>k-Nearest Neighbors</i>	79
6.2	Metodologia	80
6.3	Experimentos com Amostras	83
6.3.1	Bases de Dados	84
6.3.2	Seleção de Atributos	84
6.3.3	Classificação	90

6.4	Considerações Finais	93
7	Conclusões	95
	Referências Bibliográficas	97
A	Tabelas dos Experimentos de Agrupamento de Amostras	109
A.1	Número de Grupos Fixo	110
A.2	Número de Grupos Variável	114
A.3	Número de Grupos Estimado	122
B	Tabelas dos Experimentos de Seleção de Atributos	131

Lista de Abreviaturas

ARI *Adjusted Rand Index*

CAST *Cluster Affinity Search Technique*

cDNA *Complementary Deoxyribonucleic Acid*

DNA *Deoxyribonucleic Acid*

FMG *Finite Mixture of Gaussians*

KNN *k-Nearest Neighbors*

MM *MissMatch*

MPSS *Massive Parallel Signature Sequencing*

MRMR *Minimum Redundancy - Maximum Relevance*

mRNA *Messenger Ribonucleic Acid*

NB *Naïve Bayes*

PCR *Polymerase Chain Reaction*

PM *Perfect Match*

RNA *Ribonucleic Acid*

rRNA *Ribosomal Ribonucleic Acid*

SAGE *Serial Analysis of Gene Expression*

SS *Silhueta Simplificada*

SSF *Simplified Silhouette Filter*

SVMs *Support Vector Machines*

TIFF *Tagged Image File Format*

tRNA *Transfer Ribonucleic Acid*

Lista de Figuras

2.1	Exemplo de uma dupla hélice de molécula de DNA (adaptado de Human Genome Program (1992)).	6
2.2	Estrutura básica de um aminoácido.	8
2.3	Dogma central da biologia molecular. Adaptado de Setubal e Meidanis (1997)	9
2.4	Replicação de DNA (adaptado de Human Genome Program (1992)).	9
2.5	Processo de expressão gênica (adaptado de Human Genome Program (1992)).	11
3.1	<i>Microarray</i> de cDNA (Adaptado de Harrington et al. (2000)).	16
3.2	<i>Microarray</i> de cDNA (Lashkari et al., 1997).	18
3.3	<i>Microarray</i> de oligonucleotídeos (Adaptado de Harrington et al. (2000)).	19
3.4	<i>Microarray</i> de oligonucleotídeos (GeneChip) (Wilson et al., 2004).	20
3.5	Matriz de dados de expressão gênica (adaptado de Zhang (2006)).	22
3.6	Exemplo de dendrograma. O corte indicado pela linha tracejada produz como resultado uma partição de 3 grupos.	26
3.7	Duas sequências numéricas com tendências similares.	33
5.1	Valores médios de Rand Ajustado para as bases de cDNA, com valor k fixo.	54
5.2	Valores médios de Rand Ajustado para as bases Affymetrix, com valor k fixo.	55
5.3	Valores médios de Rand Ajustado para as bases cDNA, com valor k variável.	58
5.4	Valores médios de Rand Ajustado para as bases Affymetrix, com k variável.	59
5.5	Valores médios de Rand Ajustado para bases cDNA. As partições comparadas com o critério externo e seu número de grupos foram estimados com a Silhueta.	61
5.6	Valores médios de Rand Ajustado para bases Affymetrix. As partições comparadas com o critério externo e seu número de grupos foram estimados com a Silhueta.	62

5.7	Comparação entre os enriquecimentos obtidos para as duas medidas com melhores resultados para cada um dos quatro algoritmos de agrupamento considerados.	71
5.8	Comparação entre os enriquecimentos obtidos para o <i>Average-Linkage</i> . . .	72
5.9	Comparação entre os enriquecimentos obtidos para o <i>Complete-Linkage</i> . . .	73
5.10	Comparação entre os enriquecimentos obtidos para o <i>k-medoids</i>	74
6.1	Procedimento para avaliação dos métodos de seleção de atributos.	81
6.2	Erros médios de classificação em bases de cDNA.	85
6.3	Erros médios de classificação em bases Affymetrix.	87
6.4	Erros médios do classificador 1NN em bases cDNA.	90
6.5	Erros médios do classificador 1NN em bases Affymetrix	92

Lista de Tabelas

2.1	Mapeamento entre <i>códons</i> e aminoácidos (Lodish et al., 2003).	11
4.1	Sensibilidade e complexidade computacional dos coeficientes de correlação discutidos. Tabela adaptada de Campello e Hruschka (2009)	38
5.1	Bases de dados utilizadas, adaptada de de Souto et al. (2008)	53
5.2	Valores de Rand Ajustado para o algoritmo <i>k-medoids</i> com <i>k</i> fixo.	56
5.3	Valores de Rand Ajustado para o algoritmo <i>k-medoids</i> com <i>k</i> variável.	60
5.4	Resultados de Rand Ajustado para <i>k-medoids</i> . O número de grupos foi estimado com o uso da Silhueta.	63
5.5	Conjunto de bases de dado não rotuladas utilizado nos experimentos.	68
5.6	Número de vitórias, empates e derrotas entre os algoritmos de agrupamento.	70
5.7	Vitórias, empates e derrotas para o algoritmo <i>Average-Linkage</i>	72
5.8	Vitórias, empates e derrotas para o algoritmo <i>Complete-Linkage</i>	73
5.9	Vitórias, empates e derrotas para o algoritmo <i>k-medoids</i>	74
6.1	Porcentagens médias e números médios de atributos selecionados em bases de cDNA. Em ambos casos valores de desvio padrão são apresentados entre parênteses.	86
6.2	Vitórias, empates e derrotas para o classificador NB em bases cDNA.	86
6.3	Vitórias, empates e derrotas para o classificador 1NN em bases cDNA.	87
6.4	Porcentagens médias e números médios de atributos selecionados em bases Affymetrix. Em ambos casos valores de desvio padrão são apresentados entre parênteses.	88
6.5	Vitórias, empates e derrotas para o classificador NB em bases Affymetrix.	89
6.6	Vitórias, empates e derrotas para o classificador 1NN em bases Affymetrix.	89
6.7	Vitórias, empates e derrotas do classificador 1NN em bases cDNA.	91
6.8	Vitorias, empates e derrotas do classificador 1NN em bases Affymetrix	92
6.9	Erros do classificador 1NN para os sete coeficientes de correlação avaliados.	94

A.1	Valores de Rand Ajustado para o algoritmo <i>Single-Linkage</i> com k fixo. . .	110
A.2	Valores de Rand Ajustado para o algoritmo <i>Average-Linkage</i> com k fixo. .	111
A.3	Valores de Rand Ajustado para o algoritmo <i>Complete-Linkage</i> com k fixo.	112
A.4	Valores de Rand Ajustado para o algoritmo <i>k-medoids</i> com k fixo.	113
A.5	Valores de Rand Ajustado para o algoritmo <i>Single-Linkage</i> com k variável.	114
A.6	Valores de Rand Ajustado para o algoritmo <i>Average-Linkage</i> com k variável.	115
A.7	Valores de Rand Ajustado para o algoritmo <i>Complete-Linkage</i> com k variável.	116
A.8	Valores de Rand Ajustado para o algoritmo <i>k-medoids</i> com k variável. . . .	117
A.9	Número de grupos obtido com o algoritmo <i>Single-Linkage</i>	118
A.10	Número de grupos obtido com o algoritmo <i>Average-Linkage</i>	119
A.11	Número de grupos obtido com o algoritmo <i>Complete-Linkage</i>	120
A.12	Número de grupos obtido com o algoritmo <i>k-medoids</i>	121
A.13	Resultados de Rand Ajustado para <i>Single-Linkage</i> . O número de grupos foi estimado com o uso da Silhueta.	122
A.14	Resultados de Rand Ajustado para <i>Average-Linkage</i> . O número de grupos foi estimado com o uso da Silhueta.	123
A.15	Resultados de Rand Ajustado para <i>Complete-Linkage</i> . O número de grupos foi estimado com o uso da Silhueta.	124
A.16	Resultados de Rand Ajustado para <i>k-medoids</i> . O número de grupos foi estimado com o uso da Silhueta.	125
A.17	Número de grupos obtido com o algoritmo <i>Single-Linkage</i>	126
A.18	Número de grupos obtido com o algoritmo <i>Average-Linkage</i>	127
A.19	Número de grupos obtido com o algoritmo <i>Complete-Linkage</i>	128
A.20	Número de grupos obtido com o algoritmo <i>k-medoids</i>	129
B.1	Erros de classificação obtidos para a base de dados <i>alizadeh-2000-v1</i>	132
B.2	Erros de classificação obtidos para a base de dados <i>alizadeh-2000-v2</i>	132
B.3	Erros de classificação obtidos para a base de dados <i>alizadeh-2000-v3</i>	132
B.4	Erros de classificação obtidos para a base de dados <i>armstrong-2002-v1</i> . . .	132
B.5	Erros de classificação obtidos para a base de dados <i>armstrong-2002-v2</i> . . .	133
B.6	Erros de classificação obtidos para a base de dados <i>bhattacharjee-2001</i> . . .	133
B.7	Erros de classificação obtidos para a base de dados <i>bittner-2000</i>	133
B.8	Erros de classificação obtidos para a base de dados <i>bredel-2005</i>	133
B.9	Erros de classificação obtidos para a base de dados <i>chen-2002</i>	134
B.10	Erros de classificação obtidos para a base de dados <i>chowdary-2006</i>	134
B.11	Erros de classificação obtidos para a base de dados <i>dyrskjot-2003</i>	134
B.12	Erros de classificação obtidos para a base de dados <i>garber-2001</i>	134
B.13	Erros de classificação obtidos para a base de dados <i>golub-1999-v1</i>	135
B.14	Erros de classificação obtidos para a base de dados <i>golub-1999-v2</i>	135
B.15	Erros de classificação obtidos para a base de dados <i>gordon-2002</i>	135
B.16	Erros de classificação obtidos para a base de dados <i>khan-2001</i>	135

B.17	Erros de classificação obtidos para a base de dados <i>laiho-2007</i>	136
B.18	Erros de classificação obtidos para a base de dados <i>lapointe-2004-v1</i>	136
B.19	Erros de classificação obtidos para a base de dados <i>lapointe-2004-v2</i>	136
B.20	Erros de classificação obtidos para a base de dados <i>liang-2005</i>	136
B.21	Erros de classificação obtidos para a base de dados <i>nutt-2003-v1</i>	137
B.22	Erros de classificação obtidos para a base de dados <i>nutt-2003-v2</i>	137
B.23	Erros de classificação obtidos para a base de dados <i>nutt-2003-v3</i>	137
B.24	Erros de classificação obtidos para a base de dados <i>pomeroy-2002-v1</i>	137
B.25	Erros de classificação obtidos para a base de dados <i>pomeroy-2002-v2</i>	138
B.26	Erros de classificação obtidos para a base de dados <i>ramaswamy-2001</i>	138
B.27	Erros de classificação obtidos para a base de dados <i>risinger-2003</i>	138
B.28	Erros de classificação obtidos para a base de dados <i>shipp-2002-v1</i>	138
B.29	Erros de classificação obtidos para a base de dados <i>singh-2002</i>	139
B.30	Erros de classificação obtidos para a base de dados <i>su-2001</i>	139
B.31	Erros de classificação obtidos para a base de dados <i>tomlins-2006</i>	139
B.32	Erros de classificação obtidos para a base de dados <i>tomlins-2006-v2</i>	139
B.33	Erros de classificação obtidos para a base de dados <i>west-2001</i>	140
B.34	Erros de classificação obtidos para a base de dados <i>yeoh-2002-v1</i>	140
B.35	Erros de classificação obtidos para a base de dados <i>yeoh-2002-v2</i>	140

Lista de Algoritmos

1	<i>k-medoids</i>	48
2	<i>Algoritmo Hierárquico Aglomerativo Genérico</i>	49
3	<i>Simplified Silhouette Filter</i>	80

Introdução

O sequenciamento do genoma de diversos organismos, em especial dos seres humanos, constituiu um grande passo para a Biologia Molecular. Agora, na chamada era pós-genômica ([Zhang, 2006](#)), os esforços de pesquisa antes concentrados principalmente no sequenciamento de genomas voltam-se para a análise das sequências obtidas, com o objetivo de compreender seu funcionamento ([Brazma e Vilo, 2000](#)). Para que tal compreensão seja alcançada, é fundamental que tanto as funções de genes quanto suas interações sejam identificadas, possibilitando a inferência de suas responsabilidades nos mais diversos processos biológicos. Desta forma, genes possivelmente responsáveis pelo surgimento de doenças podem ser determinados, permitindo a criação de métodos eficientes para a sua detecção ([Bhaskar et al., 2006](#)). Além disso, as respostas de genes a diferentes tipos de tratamentos podem ser observadas, auxiliando na criação de medicamentos mais eficazes ([Brazma e Vilo, 2000](#)).

Dada a grande quantidade de genes presentes em um organismo e a complexidade das interações existentes entre os mesmos ([Ji et al., 2003](#)), são necessários métodos eficientes que permitam *medir* e *analisar* os níveis de expressão de uma grande quantidade de genes em paralelo. Tais níveis de expressão fornecem indícios acerca do nível de atividade de cada um dos genes monitorados para uma amostra em particular (tecido celular ou conjunto de células). No que diz respeito à *medição*, a tecnologia de *microarray* tem se mostrado promissora, com um grande aumento em sua utilização nos últimos anos ([Bhaskar et al., 2006](#)). A tecnologia em si possibilita a *medição* do nível de expressão gênica de milhares de genes simultaneamente, permitindo que pesquisadores determinem quais genes se expressam em um dado tipo de célula sob diferentes condições. Por outro lado, em virtude da grande quantidade de dados gerados e de suas características peculiares, sua *análise* tem se mostrado desafiadora e até mesmo o gargalo para utilização da tecnologia, como apontado por [Brazma e Vilo \(2000\)](#) e [Sherlock \(2001\)](#).

A necessidade de interpretar e *analisar* a grande quantidade de dados biológicos disponíveis motivou e proporcionou o desenvolvimento do campo interdisciplinar de pesquisa denominado bioinformática (Bertone e Gerstein, 2001; Zhang, 2006), no qual se encontra inserida a análise de dados de *microarray*, com suas diferentes motivações e objetivos. A análise de bases de dados de expressão gênica permite identificar genes potencialmente associados a determinadas doenças, atribuir funções a genes com função ainda desconhecida, dentre outras possibilidades. Em função da grande quantidade de dados produzidos e de suas características especiais, percebe-se na literatura de análise de dados de expressão gênica a utilização de métodos e metodologias provenientes da estatística, do aprendizado de máquina e da mineração de dados durante a análise desse tipo de dados.

1.1 Motivação

Os dados obtidos a partir de experimentos de *microarray* constituem, em essência, valores reais dispostos em uma matriz, na qual cada uma das linhas está associada a um gene específico e cada uma das colunas representa uma condição experimental distinta (diferentes instantes de tempo ou diferentes tipos de tecidos, por exemplo), também chamada amostra. Dessa forma, cada um dos valores reais contidos na matriz de dados, representa o nível de expressão de um gene específico para uma determinada amostra.

Após a realização dos experimentos de *microarray* propriamente ditos, a matriz de dados obtida é então fornecida como entrada a métodos computacionais, iniciando assim o processo de análise dos dados. Em muitos dos métodos ou algoritmos empregados, é necessária a escolha de uma medida de proximidade¹, seja entre pares de genes ou de amostras (*e.g.* tecido saudável e canceroso), os quais, em última análise, podem ser interpretados simplesmente como sequências numéricas provenientes de uma linha ou coluna da matriz de dados anteriormente mencionada.

Uma característica dos dados de expressão gênica é a de que, durante o processo de comparação entre duas sequências numéricas quaisquer, as tendências dos valores observados possuem fundamental importância. Isso significa que duas sequências devem ser consideradas similares se possuem forma similar (tendência similar em seus valores), ainda que possuam diferenças em seus valores. Tal característica se deve ao fato de que formas similares representam indícios de genes co-expressos (Heyer et al., 1999). Além disso, tendências similares observadas entre amostras não rotuladas e amostras já identificadas de doenças podem levar à obtenção de bons classificadores (Dudoit et al., 2002).

A fim de capturar este tipo de informação, coeficientes de correlação têm sido amplamente utilizados (Hardin et al., 2007), tal como a bem conhecida correlação de Pearson (Zhang, 2006). Para dados já padronizados (normalização *score-z* (Zhang, 2006)), nota-se ainda, na literatura, o uso da distância Euclidiana, a qual produz resultados consistentes

¹O termo proximidade é utilizado nesta dissertação como uma forma genérica de quantificar o quão parecidos são dois objetos e pode se referir tanto a uma medida de similaridade quanto a uma medida de dissimilaridade (Jain e Dubes, 1988).

e até mesmo equivalentes aos obtidos com o uso da correlação de Pearson ([Jiang et al., 2004](#)). Além destas duas medidas, outros coeficientes de correlação encontram-se descritos na literatura (vide por exemplo, [Campello e Hruschka \(2009\)](#)), oferecendo alternativas às medidas já mencionadas. Algumas destas medidas têm sido utilizadas para dados de expressão gênica (*e.g.* [Kotlyar et al. \(2002\)](#) e [Fujita et al. \(2009\)](#)), ou foram propostas especificamente para esse tipo de dados (*e.g.* [Heyer et al. \(1999\)](#), [Balasubramaniyan et al. \(2005\)](#) e [Son e Baek \(2008\)](#)).

Apesar da diversidade de medidas disponíveis que são capazes de capturar a tendência entre os valores sob consideração, não existem na literatura estudos abrangentes envolvendo uma grande quantidade de tarefas, algoritmos, medidas e bases de dados, com o intuito de identificar cenários que possam favorecer uma ou mais medidas em dados de *microarray*. Em alguns trabalhos, como nos de [Costa et al. \(2002\)](#) e [Kim et al. \(2007\)](#), algumas dessas medidas são comparadas para a tarefa de agrupamento de genes, porém a quantidade de bases de dados utilizadas é relativamente pequena e tanto a metodologia quanto as medidas comparadas diferem nos dois trabalhos. Recentemente no trabalho de [de Souto et al. \(2008\)](#) algoritmos de agrupamento são comparados para a tarefa de agrupamento de amostras. Durante a comparação, algumas medidas de correlação são consideradas, mas este não é o foco principal do trabalho. Desta forma, a motivação para o presente trabalho surge das alternativas existentes quanto às medidas de proximidade para dados de expressão gênica e das diferenças em suas formulações, sendo que tais diferenças podem vir a tornar algumas destas medidas mais adequadas em cenários específicos de aplicação.

1.2 Objetivo

O objetivo do presente trabalho é avaliar de maneira experimental diferentes coeficientes de correlação como medida de proximidade para dados de expressão gênica. A partir desta avaliação, cenários específicos para os quais uma medida seja mais apropriada podem vir a ser identificados. As medidas consideradas são avaliadas em diferentes tarefas comumente relacionadas à análise de dados de *microarray*, a saber:

- Agrupamento
- Seleção de atributos
- Classificação

1.3 Organização do Texto

Este documento está organizado da seguinte maneira. No Capítulo 2 são apresentados conceitos biológicos básicos que se encontram relacionados à análise de dados de expressão gênica. Uma visão geral da tecnologia de *microarray* é fornecida no Capítulo 3. Neste

capítulo ainda são apresentados conceitos relacionados à análise de dados de expressão gênica referentes às três tarefas consideradas neste trabalho, bem como aspectos relativos a medidas de proximidade em dados de expressão gênica. No Capítulo 4, os coeficientes de correlação e outras medidas que foram consideradas para análise no referido trabalho são apresentadas e discutidas quanto às suas principais propriedades. No Capítulo 5, são apresentados e discutidos experimentos comparativos envolvendo coeficientes de correlação na tarefa de agrupamento de dados. No Capítulo 6, são apresentados e discutidos experimentos comparativos envolvendo coeficientes de correlação nas tarefas supervisionadas de seleção de atributos e classificação. Finalmente, no Capítulo 7, são apresentadas as considerações finais do presente trabalho.

Fundamentos Biológicos

O aparecimento de organismos vivos remonta a um passado relativamente recente em termos geológicos, sendo os primeiros organismos existentes muito simples ([Setubal e Meidanis, 1997](#); [Zhang, 2006](#)). Ao decorrer do tempo alguns destes organismos evoluíram e deram origem a formas de vida mais complexas, como as que encontramos atualmente. Nesse processo de manutenção e evolução da vida, duas moléculas são de fundamental importância: as proteínas e os ácidos nucleicos. Em uma interpretação simples, as proteínas são responsáveis por definir o que somos e fazemos, enquanto que os ácidos nucleicos são responsáveis por codificar características dos organismos e informações necessárias para a síntese de proteínas. Além disso, são os ácidos nucleicos que possibilitam a transmissão de informações de geração em geração, processo este chamado de hereditariedade.

Neste capítulo, é fornecida uma visão geral de conceitos básicos de Biologia Molecular a fim de familiarizar o leitor com o vocabulário utilizado no restante desta dissertação. Como apontado por [Setubal e Meidanis \(1997\)](#), para cada uma das regras da biologia geralmente existem também diversas exceções. Sendo assim, este capítulo não cobre de maneira ampla ou profunda os fundamentos da biologia molecular. O leitor interessado em mais detalhes pode consultar os trabalhos de [Setubal e Meidanis \(1997\)](#); [Lodish et al. \(2003\)](#); [Mount \(2004\)](#); [Lesk \(2008\)](#) e referências neles contidas. O capítulo encontra-se organizado da seguinte forma. Na Seção 2.1 os conceitos básicos referentes aos ácidos nucleicos são apresentados. Na Seção 2.2 são discutidas as proteínas e sua composição. Por fim, na Seção 2.3, o processo de expressão gênica é descrito, relacionando os conceitos vistos nas seções anteriores.

2.1 Ácidos Nucléicos

Os ácidos nucleicos são responsáveis pela codificação, armazenamento e transmissão de material genético. É a informação armazenada nestas moléculas que possibilita o processo de hereditariedade, no qual informações genéticas são transmitidas de uma geração a outra. Além disso, é a informação codificada nos ácidos nucleicos que permite a síntese de proteínas, moléculas importantes e necessárias para a manutenção da vida. Existem dois tipos básicos de ácidos nucleicos: ácido desoxirribonucleico (DNA) e ácido ribonucleico (RNA) (Setubal e Meidanis, 1997). Ambos são brevemente detalhados a seguir.

2.1.1 Ácido Desoxirribonucleico

O DNA contém a informação necessária para a síntese de proteínas e pode ser considerado uma fórmula que é passada de geração em geração (Setubal e Meidanis, 1997). Sua estrutura, comumente referenciada como estrutura de dupla hélice (Lodish et al., 2003) é formada por ligações entre bases chamadas adenina (A), timina (T), citosina (C) e guanina (G) (Human Genome Program, 1992), as quais obedecem a duas regras básicas: adenina (A) só se liga a timina (T) e citosina (C) só se liga a guanina (G). O conjunto de diversas dessas ligações formam uma molécula de DNA (Figura 2.1), sendo seu comprimento mensurado pelo número de pares de bases (bp, *base pairs*) que a compõe. O genoma humano, por exemplo, possui por volta de 3 bilhões de bp (Human Genome Program, 1992). São as diferenças nas sequências de DNA que possibilitam a existência de diversas espécies de organismos vivos e que diferentes organismos de uma mesma espécie possuam características peculiares, tornando-os únicos (Human Genome Program, 1992).

Figura 2.1: Exemplo de uma dupla hélice de molécula de DNA (adaptado de Human Genome Program (1992)).

Cada uma das células de um organismo possui em seu interior uma cópia do DNA

do organismo, também referenciado como seu material genético. Em células procariontes, nas quais não existe uma carioteca¹, o material genético encontra-se no citoplasma da célula, condensado em uma única molécula maior, geralmente circular, denominada cromossomo (Setubal e Meidanis, 1997). Já em células eucariontes, dotadas de uma carioteca, o material genético encontra-se armazenado internamente ao núcleo celular em moléculas também chamadas cromossomos. Na grande maioria das células eucariontes, os cromossomos são encontrados em pares, sendo cada um deles herdado de um dos genitores do organismo em questão. O conjunto de todos os cromossomos do organismo é denominado de genoma do organismo (Human Genome Program, 1992).

Uma característica importante das moléculas de DNA é que existem regiões específicas que são responsáveis pela codificação de proteínas e moléculas de RNA. Tais regiões são geralmente únicas, ou seja, para cada molécula de proteína ou RNA de um organismo, existe geralmente somente uma sequência de DNA codificadora (Setubal e Meidanis, 1997). Essas regiões codificadoras correspondem a apenas uma parcela do comprimento total do DNA de um organismo e são chamadas de genes (Human Genome Program, 1992), sendo as regiões que as antecedem responsáveis por determinar a quantidade de proteínas ou RNA a ser produzida.

2.1.2 Ácido Ribonucléico

As moléculas de RNA possuem semelhanças com as moléculas de DNA anteriormente discutidas, porém, pelo menos duas diferenças estruturais são encontradas entre as mesmas e merecem destaque. A primeira delas é que moléculas de RNA geralmente não possuem estrutura de dupla hélice como as moléculas de DNA. A segunda diz respeito às quatro bases encontradas no RNA: adenina (A), uracila (U), citosina (C) e guanina (G), sendo a única alteração a presença da uracila (U) no lugar da base timina (T). As propriedades das ligações mantêm-se inalteradas, salvo que para moléculas de RNA, a base de adenina (A) se liga à uracila (U).

Feitas tais considerações, é importante destacar que existem três tipos diferentes de ácidos ribonucléicos (Lodish et al., 2003), a saber: RNA mensageiro (mRNA), RNA ribossômico (rRNA) e RNA transportador (tRNA). Estes três tipos de RNA serão discutidos quanto a suas funções durante a apresentação do processo de expressão gênica na Seção 2.3.

2.2 Proteínas

As proteínas são moléculas presentes em grande abundância nos organismos vivos, sendo sua produção essencial para a manutenção da vida (Zhang, 2006). Existem diversos tipos de proteínas, cada qual com uma função específica, de forma que proteínas com funções similares são agrupadas em famílias de proteínas (Lodish et al., 2003). Proteínas

¹Membrana que define um núcleo celular.

chamadas de catalisadoras, por exemplo, possuem como função acelerar reações químicas (Setubal e Meidanis, 1997). Já proteínas regulatórias controlam o processo de expressão gênica, bem como a atividade de outras proteínas (Zhang, 2006).

Cada molécula de proteína é composta de uma cadeia de moléculas mais simples, chamadas aminoácidos. Os aminoácidos, por sua vez, possuem uma estrutura básica em comum. Nessa estrutura está presente um átomo central de carbono (C_α), ao qual estão ligados um grupo amina (H_2N), um grupo carboxila ($COOH$), um átomo de hidrogênio (H) e uma cadeia lateral, sendo tal cadeia lateral responsável pela determinação de qual é o aminoácido em questão. A estrutura básica de um aminoácido é ilustrada pela Figura 2.2.

Figura 2.2: Estrutura básica de um aminoácido.

Todas as proteínas são constituídas de combinações entre os 20 diferentes aminoácidos existentes na natureza (Zhang, 2006). Durante a formação de uma cadeia de aminoácidos (também chamada de cadeia polipeptídica), moléculas de água são liberadas em virtude das ligações químicas decorrentes. Dessa forma, em última instância, uma proteína é formada por sequências de resíduos de aminoácidos que possuem comprimento variável (Setubal e Meidanis, 1997). É comum que proteínas sejam formadas por cadeias possuindo de 100 a 1000 resíduos de aminoácidos de comprimento (Lodish et al., 2003).

2.3 Expressão Gênica

Como argumentado anteriormente, são nas moléculas de DNA que se encontram as informações necessárias para a produção de moléculas de RNA e proteínas. Essas informações encontram-se armazenadas em regiões específicas do DNA chamadas genes. É por meio do processo denominado de expressão gênica que genes tornam-se ativos ou expressos. Durante este processo, as informações contidas nos genes são transcritas em moléculas de RNA que posteriormente podem vir a ser traduzidas em proteínas, as quais são seu produto final. O processo de expressão gênica, por sua vez, é explicado pelo dogma central da biologia molecular, que estabelece como ocorre a transferência de informação genética internamente nas células (Setubal e Meidanis, 1997). Tal processo, na sua forma original, pode ser dividido em três subprocessos, a saber: replicação, transcrição e tradução. Estes

três processos são ilustrados na Figura 2.3 e são detalhados a seguir.

Figura 2.3: Dogma central da biologia molecular. Adaptado de [Setubal e Meidanis \(1997\)](#).

O processo de replicação possibilita a cópia de informação genética contida em uma molécula de DNA. Por meio desse processo, uma molécula de DNA é duplicada, dando origem a duas novas moléculas idênticas de DNA. Este fenômeno é possível graças às propriedades de ligações entre bases vistas na Seção 2.1.1. De forma simplificada, a molécula de DNA a ser duplicada tem sua estrutura de dupla hélice desenrolada e aberta com o auxílio de proteínas especiais. Após tal abertura inicia-se o processo de síntese das duas novas moléculas de DNA, tal qual ilustrado pela Figura 2.4. A cada uma das bases que compõem as tiras formadas com a abertura da molécula de DNA pai fixam-se bases complementares, dando origem a duas novas estruturas de dupla hélice, ou seja, duas novas moléculas de DNA. Todo o processo de replicação ocorre com o auxílio de diversas proteínas. Existem, por exemplo, proteínas responsáveis por auxiliar na abertura da dupla hélice, bem como proteínas responsáveis por mantê-las abertas durante todo o processo.

Figura 2.4: Replicação de DNA (adaptado de [Human Genome Program \(1992\)](#)).

A transcrição é o processo pelo qual informações codificadas em um gene originam uma molécula de RNA, que posteriormente pode vir a ser utilizada durante o processo de tradução. De forma semelhante ao que ocorre no processo de replicação, o primeiro passo para o início da transcrição é a abertura da molécula de DNA. O local de abertura corresponde ao início da região codificadora de um gene e é identificado por sequências especiais de bases (Zhang, 2006). Após tal identificação e abertura, o processo de cópia tem início. A molécula de RNA é formada a partir do complemento das bases que compõem o gene e são lidas da molécula de DNA. O processo de cópia pára quando regiões terminadoras, que indicam o fim do código do gene em questão, são encontradas. Ao final do processo de transcrição, obtém-se uma molécula de RNA, a qual pode ser de três diferentes tipos (Lodish et al., 2003), como citado na Seção 2.1.2, a saber: rRNA, tRNA ou mRNA. As moléculas de rRNA são encontradas como parte da estrutura dos ribossomos, organelas responsáveis pela síntese de proteínas (Lodish et al., 2003). Moléculas de tRNA são transportadoras de aminoácidos utilizados durante a síntese de proteínas. Por fim, as moléculas de mRNA são as responsáveis por codificar a informação necessária para a síntese de proteínas, que ocorre por meio do processo de tradução.

Antes de explicar o processo de tradução, é interessante elucidar como os aminoácidos encontram-se codificados no mRNA. Essa codificação é feita utilizando os chamados *códons*, que são formados por sequências de três bases (uracila (U), citosina (C), adenina (A) ou guanina (G)), possibilitando 64 diferentes combinações entre estas. Todas as 64 possíveis combinações são mostradas na Tabela 2.1, que estabelece o mapeamento entre cada um dos *códons* e os 20 aminoácidos existentes. Considerando o *códon* formado pela sequência AUG, por exemplo, seu aminoácido correspondente pode ser obtido da seguinte forma. Na primeira posição da Tabela 2.1 é escolhida a base A, restringindo assim a tabela a apenas quatro linhas (9 à 12). Na segunda posição é escolhida a base U que restringe a tabela apenas à primeira coluna das linhas 9 à 12. Finalmente, na terceira posição é escolhida a base G que restringe a tabela à décima segunda linha e primeira coluna, posição que corresponde ao aminoácido Metionina, representado na Tabela 2.1 por seu símbolo, Met. Pode-se notar ainda, que alguns dos *códons* não codificam aminoácidos, mas sim sinais de parada (denotados por *Stop*) que indicam o fim do código do gene em questão, como é o exemplo do *códon* UAA. É interessante notar também que em alguns casos mais de um *códon* codifica um mesmo aminoácido.

Feitas tais considerações, o processo de tradução ocorre como segue. A molécula de mRNA produzida durante o processo de transcrição sai do núcleo celular (no caso dos eucariontes) e migra até o citoplasma da célula. Como discutido anteriormente, essa molécula de mRNA é a cópia de um gene específico e contém a informação necessária para a síntese de uma proteína.

A síntese de proteínas ocorre em organelas conhecidas como ribossomos. Quando a molécula de mRNA chega ao ribossomo, este inicia o processo de leitura de seus *códons*. Quando um *códon* é lido da molécula de mRNA internamente ao ribossomo, uma molécula

Tabela 2.1: Mapeamento entre *códons* e aminoácidos (Lodish et al., 2003).

Posição 1	Posição 2				Posição 3
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

de tRNA complementar se fixa a um aminoácido livre presente no citoplasma celular. A molécula de tRNA transporta então o aminoácido correspondente até o ribossomo, onde os diversos aminoácidos codificados pelo mRNA são unidos por ligações químicas.

A união dos diversos aminoácidos forma uma cadeia crescente de resíduos de aminoácidos. Quando um *códon* terminador é encontrado na molécula de mRNA não há uma molécula de tRNA associada ao mesmo, de forma que o processo termina (Zhang, 2006), dando origem a uma proteína. O processo completo, que tem início com uma molécula de DNA e termina com a síntese de uma molécula de proteína é ilustrado pela Figura 2.5.

Figura 2.5: Processo de expressão gênica (adaptado de Human Genome Program (1992)).

2.4 *Considerações Finais*

Neste capítulo foram apresentados conceitos básicos de Biologia Molecular necessários para a compreensão das tecnologias de *microarray*. Os ácidos nucléicos e as proteínas, moléculas fundamentais para evolução e manutenção da vida foram discutidas e relacionadas pelo processo de expressão gênica. Por meio deste processo, a informação codificada nas moléculas de DNA, em regiões denominadas genes, é transportada e interpretada em organelas chamadas ribossomos para a síntese de proteínas. Como será visto a seguir, a existência de diferentes tipos de células está associada a diferenças em seus níveis de expressão gênica, níveis estes, que podem ser mensurados e quantificados pelas tecnologias de *microarray*. Tais tecnologias são apresentadas e discutidas no próximo capítulo.

Análise de Dados de Expressão Gênica

Praticamente todas as células de um ser vivo contêm uma cópia idêntica de seu material genético, seu genoma¹ (Belk e Borden, 2003). Embora contenham o mesmo material genético, células diferentes, do ponto de vista funcional, são encontradas nos mais diversos organismos. A existência de células especializadas ocorre graças as diferenças nos níveis de expressão de seus genes, de forma que, genes distintos são expressos em diferentes quantidades, de acordo com a funcionalidade da célula em questão. Neste sentido, a análise dos níveis de expressão gênica pode fornecer indícios valiosos acerca do funcionamento do código genômico, auxiliando na determinação de funções de genes e relações existentes entre os mesmos. Não obstante, a análise dos níveis de expressão gênica pode auxiliar na detecção de padrões de expressão gênica relacionados com doenças, como câncer (Golub et al., 1999), na identificação de genes alvo para o uso de terapias genéticas (Eisen et al., 1998) e no desenvolvimento de novos medicamentos (Kuo et al., 2004).

Atualmente, diferentes tecnologias, métodos ou técnicas possibilitam a realização de análise a nível molecular, quantificando os níveis de expressão gênica para uma ou mais amostras² de interesse. De forma geral, estas tecnologias podem ser divididas em dois grandes grupos (Kuo et al., 2004), de acordo com a quantidade de genes passível de análise: tecnologias de *low throughput* e tecnologias de *high throughput*. As tecnologias de *low throughput*, como PCR (*Polymerase Chain Reaction*) (VanGuilder et al., 2008) e *Northern Blot* (Lodish et al., 2003), permitem a medição do nível de expressão de uma quantidade relativamente pequena de genes (dezenas) com um alto grau de precisão. Já tecnologias de *high throughput*, como SAGE (*Serial Analysis of Gene Expression*) (Velculescu et al., 1995), MPSS (*Massive Parallel Signature Sequencing*) (Brenner et al., 2000) e *microarray* (Harrington et al., 2000) permitem a medição do nível de expressão

¹Com exceção de células reprodutivas.

²O termo amostra se refere a uma amostra biológica, proveniente de um tecido ou célula.

de uma grande quantidade de genes (milhares), ao preço de um menor grau de precisão.

Mesmo sendo menos precisas do que tecnologias de *low throughput*, consideradas tradicionais (Kuo et al., 2004), tecnologias de *high throughput* têm recebido crescente atenção, com destaque para as tecnologias de *microarray* (Bhaskar et al., 2006). Isso se deve, em parte, ao fato de que as informações globais que são obtidas a partir da medição da expressão de milhares de genes simultaneamente são consideradas significativas, principalmente em fases exploratórias de pesquisa (Kuo et al., 2004). É importante ressaltar, porém, que métodos de *low throughput* ainda são utilizados. Como apontado por Kuo et al. (2004) e VanGuilder et al. (2008), estas técnicas têm sido empregadas em fases posteriores à realização de experimentos de *microarray*, a fim de validar os resultados obtidos.

Em virtude de sua ampla utilização e da grande quantidade de bases de dados disponíveis, o foco desta dissertação encontra-se na análise de dados de expressão gênica provenientes de experimentos de *microarray*, análise esta que é discutida neste capítulo. O capítulo encontra-se organizado da seguinte maneira. Na Seção 3.1 as tecnologias de *microarray* são descritas de forma breve. Na Seção 3.2 o processo de obtenção dos dados a partir dos *microarrays* é elucidado. Na Seção 3.3 são apresentados conceitos de pré-processamento aplicados a dados de expressão gênica. Na Seção 3.4 são discutidos conceitos envolvendo as três principais tarefas comumente empregadas na análise de dados de *microarray* (Piatetsky-Shapiro e Tamayo, 2003). Algoritmos e métodos pertencentes a essas tarefas serão considerados durante a avaliação dos coeficientes de correlação, avaliação esta que constitui o principal interesse do presente trabalho. Por fim, o uso de medidas de proximidade no contexto de dados de expressão gênica é discutido na Seção 3.5.

3.1 Tecnologias de Microarray

A criação das tecnologias de *microarray* foi possível graças às propriedades de complementaridade entre pares de bases de DNA e RNA (Tarca et al., 2006) vistas no Capítulo 2. Nestas tecnologias, de forma geral, os níveis de expressão gênica são quantificados com base nas quantidades de mRNAs encontradas em um tecido ou célula sob estudo. O princípio básico das tecnologias é relativamente simples. Cada *array*³ é constituído de uma pastilha, geralmente feita de vidro. Nesta pastilha, regiões fixas correspondem a genes específicos e contém milhares de cópias complementares de seu material genético, ou seja, milhares de cópias da sequência ou de partes da sequência (formada pelas bases A, T, C e G) que constitui o gene em questão. Para a realização de um experimento, o material genético extraído de uma amostra é preparado com substâncias fluorescentes ou radioativas que possibilitem sua detecção posterior, como será visto adiante, e é então depositado sobre a pastilha. Tal material genético é composto pelas moléculas de mRNA que são produzidas durante o processo de expressão gênica, descrito na Seção 2.3. Dessa

³Os termos *array* e *microarray* serão usados com o mesmo significado ao longo do capítulo.

forma, a ausência de mRNA correspondente a um gene específico indica que o mesmo não está expresso na amostra em questão. Por outro lado, para um gene expresso, a quantidade de mRNA presente na amostra está relacionada e é proporcional ao nível de expressão daquele gene. Após a deposição, o material genético dos genes expressos na amostra (seu mRNA) hibridiza com as cópias de seu material genético em sua região correspondente no *microarray*. Em outras palavras a hibridização é o processo pelo qual duas sequências complementares se ligam uma à outra, graças as propriedades de complementaridade vistas no Capítulo 2. Ao fim do experimento, a quantidade de material hibridizado em cada região é mensurada, permitindo assim a quantificação dos níveis de expressão de cada um dos genes analisados.

Embora possuam os mesmos princípios básicos em comum, diferentes tecnologias de *microarray* encontram-se disponíveis comercialmente, cada qual com suas peculiaridades. Quanto à sua fabricação, duas abordagens distintas podem ser identificadas (Zhang, 2006; Harrington et al., 2000): deposição e síntese *in situ*. Na abordagem de fabricação por deposição, sequências de genes são fixadas roboticamente nas pastilhas. Já na abordagem de fabricação *in situ*, as sequências são sintetizadas nos próprios *arrays*. Dentre todas as tecnologias de *microarrays* atualmente existentes, duas se destacam (Zhang, 2006): a de cDNA (DNA complementar), representante do processo de fabricação por deposição, e de oligonucleotídeos (muitas vezes referida como Affymetrix, nome da empresa fabricante), representante do processo de fabricação *in situ*. Ambas são brevemente descritas a seguir.

3.1.1 *Microarrays de cDNA*

O processo de produção de um *microarray* de cDNA (Schena et al., 1995) tem início com a seleção dos genes que estarão representados no *microarray* (Duggan et al., 1999). A partir desta seleção, as sequências dos genes de interesse, que possuem centenas ou até mesmo milhares de bases de comprimento (Tarca et al., 2006), podem ser obtidas e submetidas a um processo de amplificação por PCR (*Polymerase Chain Reaction*). Pelo processo de PCR, milhares de cópias idênticas de cada uma das sequências de interesse são obtidas para posterior fixação no *array*. Tais sequências são então purificadas, a fim de eliminar resíduos do processo de PCR que possam vir a contaminar o experimento de *microarray*.

Uma vez obtidas milhares de sequências idênticas para cada um dos genes de interesse, inicia-se sua deposição. Em uma pastilha, geralmente feita de vidro ou *nylon* (Duggan et al., 1999), o processo consiste basicamente da fixação robótica das sequências obtidas em pontos pré-determinados do *microarray* chamados *probes* (Duggan et al., 1999). Cada *probe* contém milhares de sequências idênticas, geradas anteriormente por PCR e corresponde a um único gene. Após o processo de fixação para cada um dos *probes*, a pastilha ainda passa por um tratamento químico e térmico a fim de evitar o desprendimento das sequências depositadas. O processo descrito é ilustrado pela Figura 3.1(a).

Com o término do processo fabricação, o *microarray* de cDNA encontra-se pronto para

Figura 3.1: *Microarray* de cDNA (Adaptado de Harrington et al. (2000)).

utilização. Uma característica peculiar da tecnologia de cDNA é que nos experimentos realizados, os níveis de expressão dos genes de duas amostras são avaliados simultaneamente⁴. Neste sentido, geralmente, uma das amostras é proveniente de um tecido de interesse enquanto a outra é uma amostra de controle (Freeman et al., 2000).

O experimento propriamente dito tem início com a extração de moléculas de mRNA das duas amostras: interesse e controle. Como as moléculas de mRNA se degradam rapidamente e são relativamente instáveis, moléculas de cDNA (DNA complementar), mais estáveis, são geralmente obtidas a partir das moléculas de mRNA para subsequente utilização nos experimentos (Kuo et al., 2004). As moléculas de cDNA obtidas são então submetidas a um processo de marcação que permite sua posterior detecção no *microarray*. O cDNA proveniente da amostra de interesse é marcado com corante fluorescente vermelho (Cy5), enquanto que o cDNA proveniente da amostra de controle é marcado com corante fluorescente verde (Cy3). Uma vez marcadas, as moléculas de ambas amostras são então depositadas sobre o mesmo *microarray*.

Com a deposição do material genético das duas amostras (cDNA obtido a partir das moléculas de mRNA presentes em cada uma das amostras), o processo de hibridização tem início. Moléculas de cDNA de genes específicos se hibridizam, se ligam, ao seu

⁴Em virtude desta característica, *microarrays* de cDNA são também conhecidos como *double channel microarrays* (Tarca et al., 2006).

DNA correspondente fixado em *probes* do *microarray*⁵. Como o material genético das duas amostras (interesse e controle) é misturado e depositado simultaneamente, ambos participam do processo de hibridização simultaneamente. Após um período de tempo determinado, o *array* é lavado a fim de remover moléculas de cDNA que não hibridizaram, ou seja, que não se fixaram a nenhum *probe* ou que se fixaram apenas parcialmente a um *probe* que não corresponde ao seu gene. Dessa forma, o processo de lavagem possui por finalidade limpar o *array*, deixando neste somente as moléculas de interesse que hibridizaram ao *probe* correto.

Uma vez lavado, o *microarray* é fornecido como entrada a um *scanner* que emite feixes de luz com comprimento de onda específico para cada um dos corantes utilizados na preparação das amostras. Como resultado, duas imagens são obtidas: uma para cada amostra analisada. Estas imagens são então combinadas via *software* e resultam em uma única imagem, que possui três cores (verde, vermelho e amarelo) em diferentes níveis de intensidade, tal qual mostrado na Figura 3.2. São as cores da imagem final e suas intensidades que quantificam os níveis de expressão dos genes analisados nas duas amostras submetidas ao experimento. A cor de cada *probe* determina em qual amostra o gene em questão é predominantemente expresso, enquanto sua intensidade revela o nível de expressão relativo do gene para as duas amostras. Tal coloração se deve aos corantes fluorescentes introduzidos inicialmente nas amostras biológicas: vermelho na amostra de interesse e verde na amostra de controle. Como exemplo, considere a Figura 3.1(b) que resume a realização de um experimento de cDNA. A presença da cor vermelha no *probe* do gene X indica que aquele gene esteve mais expresso na amostra de interesse, ou seja, uma maior quantidade de material genético da amostra de interesse hibridizou ao *probe* do gene em questão. A presença da cor verde no *probe* do gene Y indica que aquele gene esteve mais expresso na amostra de controle, ou seja, uma maior quantidade de material genético da amostra de controle hibridizou ao *probe* em questão. Já a presença da cor amarela no *probe* do gene Z indica que aquele gene esteve igualmente expresso tanto na amostra de interesse quanto na amostra de controle⁶, ou seja, material genético de ambas as amostras (interesse e controle) hibridizaram em quantidades similares ao *probe* em questão. Por fim, pontos com ausência de coloração correspondem a genes não expressos em nenhuma das amostras, ou seja, o gene do *probe* em questão não esteve expresso tanto na amostra de interesse quanto na amostra de controle.

3.1.2 *Microarrays de Oligonucleotídeos*

Dentre as tecnologias de *microarray* de oligonucleotídeos, destaca-se a desenvolvida pela empresa Affymetrix (Lockhart et al., 1996; Tarca et al., 2007), sob o nome de

⁵A amostra depositada no *array* é composta por material genético de uma única tira (como as moléculas de RNA vistas no Capítulo 2). A hibridização é possível, uma vez que o *array* é composto por tiras complementares que permitem a ligação do material genético depositado, conforme visto no Capítulo 2.

⁶Note que a cor amarela indica que o gene esteve igualmente expresso nas amostras de interesse e controle. Tal cor é obtida com a combinação das cores vermelha (interesse) e verde (controle).

Figura 3.2: *Microarray* de cDNA (Lashkari et al., 1997).

GeneChip (Dalma-Weiszhausz et al., 2006). Além dos próprios *microarrays*, a empresa comercializa uma plataforma que auxilia na realização de experimentos controlados (Dalma-Weiszhausz et al., 2006). Uma vez que os *microarrays* e toda plataforma experimental são padronizados, a flexibilidade dos experimentos é menor que na tecnologia de cDNA. Nesta tecnologia, diferentemente da tecnologia de cDNA vista anteriormente, as sequências dos genes de interesse são impressas (sintetizadas) diretamente no *microarray* e não mais fixadas roboticamente. Desta forma, não há a necessidade de produção e purificação de milhares de sequências para a construção do *array* (Lockhart et al., 1996). Outra diferença entre as tecnologias diz respeito ao tamanho das sequências que compõem os *microarrays*. Na tecnologia GeneChip, os *arrays* são compostos por sequências muito mais curtas que na tecnologia de cDNA chamadas oligonucleotídeos. Cada oligonucleotídeo, por sua vez, possui em torno de 25 bases de comprimento (Dalma-Weiszhausz et al., 2006).

A construção de um GeneChip se dá por processos similares aos empregados na construção de circuitos integrados de larga escala (Dalma-Weiszhausz et al., 2006; Zhang, 2006). De forma simplificada, o processo tem início com a preparação da superfície do *microarray* por meio da aplicação de reagentes. Feita tal preparação, uma fonte luminosa é combinada com uma máscara, que permite a passagem de luz apenas por locais específicos, bloqueando os demais. Após a passagem da luz pela máscara, bases específicas (A, C, T ou G) são então sintetizadas no *array*, apenas nos locais anteriormente iluminados. Dessa forma, com a repetição do processo e combinação de diversas máscaras, as sequências de oligonucleotídeos desejadas podem ser sintetizadas em posições específicas do *array*. O processo termina quando milhares de cópias das sequências de interesse tiverem sido sintetizadas, cada qual com 25 bases de comprimento.

Em um *microarray* de oligonucleotídeos já pronto, regiões específicas, geralmente chamadas *probes* correspondem aos genes (Dalma-Weiszhausz et al., 2006). Cada *probe* corresponde a um único gene e é formado por milhares de oligonucleotídeos. É importante notar que um único gene é representado por mais de um *probe*, cada qual contendo oligonucleotídeos correspondentes a apenas fragmentos de sua sequência total. Uma característica peculiar desta tecnologia é a de que para cada *probe* que representa um gene, chamado *Perfect Match* (PM), existe um outro *probe* associado, chamado de *Mismatch* (MM). A função do PM é a mesma dos *probes* presentes na tecnologia de cDNA: permitir que o material genético daquele gene específico presente na amostra hibridize e possa ser posteriormente quantificado. Os *probes* MM por sua vez, são formados por sequências idênticas as que formam os *probes* PM, exceto por diferenças em uma única base na sua posição central. Sua principal função é a identificação de fenômenos indesejados de hibridização cruzada, que ocorrem quando sequências apenas parcialmente complementares hibridizam nos *microarrays* (Lockhart et al., 1996). A Figura 3.3(a) ilustra um *microarray* de oligonucleotídeos e as posições dos *probes* PM e MM. Note que cada *probe* corresponde a um único gene, mas que um gene é representado por diversos *probes*.

Figura 3.3: *Microarray* de oligonucleotídeos (Adaptado de Harrington et al. (2000)).

Feitas tais considerações sobre o *array* e sua fabricação, pode-se falar então de sua utilização, que é muito semelhante a de um *microarray* de cDNA, salvo que apenas uma amostra é analisada por vez. Ou seja, para cada *array*, os níveis de expressão de apenas

uma amostra de interesse são quantificados. O mRNA é extraído da amostra de interesse e marcado com biotina para possibilitar sua detecção posterior. Após a marcação do material genético, este é então depositado sobre o *microarray* para que possa hibridizar aos *probes* do mesmo. Com o término do processo de hibridização o *microarray* é lavado, a fim de remover as moléculas de mRNA que não hibridizaram a nenhum *probe* específico. Às moléculas de mRNA que permaneceram no *array* é adicionado um composto corante que se liga à biotina. Com a adição do corante, o *array* é finalmente submetido a um *scanner*, para que uma imagem de resultado possa ser gerada. O processo experimental, descrito aqui de forma sucinta, é ilustrado pela Figura 3.3(b). A Figura 3.4 apresenta uma imagem obtida ao final de um experimento. É importante notar que para a análise de duas amostras distintas, dois experimentos de *microarray* de oligonucleotídeos devem ser conduzidos separadamente, conforme mostra a Figura 3.3(b).

Figura 3.4: *Microarray* de oligonucleotídeos (GeneChip) (Wilson et al., 2004).

3.2 Obtenção dos Dados

Com a realização de experimentos de *microarray*, os níveis de expressão de uma grande quantidade de genes podem ser medidos para os mais diversos tipos de amostras, ou condições experimentais, as quais podem estar relacionadas a diferentes tipos de tecidos (*e.g.* saudáveis ou cancerosos), diferentes sub-tipos de câncer, diferentes instantes de tempo durante a ocorrência de um processo celular, ou ainda diferentes processos celulares, por exemplo. O resultado de um experimento, para ambas tecnologias anteriormente descritas, é uma imagem do *microarray* em questão (Figuras 3.2 e 3.4), geralmente no formato de dados TIFF (*Tagged Image File Format*). As imagens obtidas com a tecnologia

de cDNA são coloridas, enquanto que as imagens obtidas com a tecnologia GeneChip são monocromáticas (Zhang, 2006). Mesmo com certas diferenças quanto à cor e localização das informações, as imagens provenientes de ambas tecnologias devem ser transformadas em dados numéricos para que sua análise seja possível.

Para a tecnologia GeneChip, existem pacotes de *software* pertencentes à plataforma experimental que realizam todo o processamento das imagens obtidas e produzem como resultado dados numéricos. De forma básica, cada um dos *probes* do *array* são identificados pelo *software* de análise de imagem. Os dados resultantes são então obtidos a partir dos *probes* PM ou de combinações das intensidades entre os *probes* PM e MM (Dalma-Weiszhausz et al., 2006). Para mais informações a respeito do processamento das imagens, Dalma-Weiszhausz et al. (2006) e referências contidas podem ser consultados.

Para imagens provenientes de experimentos de cDNA não há uma plataforma comum de análise. De forma geral, técnicas de processamento de imagens são aplicadas a fim de converter as imagens em valores numéricos. Segundo Zhang (2006), quatro passos são tomados para obtenção dos resultados em forma numérica: (i) identificação dos *probes*, (ii) segmentação da imagem, (iii) quantificação dos *probes* e (iv) avaliação de qualidade dos *probes*. Mais informações quanto à aquisição de dados a partir de imagens de *arrays* de cDNA podem ser obtidas em Zhang (2006), Esteves (2007) e referências contidas.

Com o processamento de imagens de diversos experimentos relacionados, conduzidos com uma mesma tecnologia, uma quantidade considerável de dados pode ser obtida e condensada na forma de uma única matriz de dados. Antes da união dos resultados obtidos, estes são geralmente submetidos a etapas de pré-processamento que possuem, por exemplo, a função de normalizar os dados entre os diferentes *microarrays* considerados (Smyth e Speed, 2003). Esta etapa, de fundamental importância, permite a comparação entre níveis de expressão de diferentes *arrays*, com a remoção de variações sistemáticas introduzidas durante a realização dos experimentos propriamente ditos (Sherlock, 2001), fazendo com que diferentes *arrays* (experimentos) possuam as mesmas distribuições. As matrizes de dados, obtidas ao final desta etapa de pré-processamento, encontram-se em um formato adequado para análise de dados por algoritmos de mineração de dados e aprendizado de máquina, e possuem a forma apresentada na Figura 3.5. Nesta matriz, cada uma das linhas está associada a um gene e cada uma das colunas refere-se a uma amostra distinta, de forma que, cada uma das posições da matriz representa o nível de expressão de um gene para uma amostra específica. Vale ressaltar que cada uma das colunas da matriz corresponde a um experimento de *microarray* distinto.

Ainda com respeito aos níveis de expressão presentes na matriz de dados, cabem duas observações. Para dados provenientes de *microarrays* de oligonucleotídeos, os níveis de expressão são absolutos, ou seja, relativos a um único gene em uma única amostra. Tais níveis refletem a quantidade de cópias de mRNA de cada um dos genes em questão para a amostra em particular. Quanto à tecnologia cDNA, é importante notar que os níveis de expressão de cada um dos genes analisados são dados pela Equação (3.1), onde $Cy5$ e $Cy3$

Figura 3.5: Matriz de dados de expressão gênica (adaptado de Zhang (2006)).

correspondem aos níveis de expressão em cada uma das duas amostras analisadas no *array* em questão. Essa representação é adotada uma vez que *microarrays* de cDNA medem o nível relativo de expressão de genes entre duas amostras e não seus níveis absolutos.

$$e = \log_2 \left(\frac{Cy5}{Cy3} \right) \quad (3.1)$$

Com a obtenção dos resultados na forma de matriz de dados, tem início o processo de análise dos dados. Como apontado por Brazma e Vilo (2000) e Sherlock (2001), a análise dos dados de expressão gênica gerados tem se mostrado o gargalo para a utilização da tecnologia de *microarray*, visto que a quantidade de dados disponíveis tem se tornado cada vez maior. Além disso, algumas características peculiares presentes nesse tipo de dados tornam sua análise desafiadora (Kuo et al., 2004).

Uma destas características diz respeito às dimensões da matriz de dados que deve ser analisada. É comum que dados de expressão gênica possuam uma grande quantidade de genes e poucas amostras, ou seja, a matriz resultante possui uma grande quantidade de linhas (na ordem de milhares) e poucas colunas (na ordem de dezenas), o que pode ser um problema para certos algoritmos, como será visto mais adiante. Tal característica se deve ao fato de que cada uma das colunas da matriz de dados é proveniente de um experimento e possui um custo associado, custo este que limita o número de amostras na matriz. Outras características indesejadas, porém comuns neste tipo de dados, são a presença de ruídos, valores ausentes, *outliers* e variações sistemáticas inerentes aos experimentos (Heyer et al., 1999; Jiang et al., 2004; Zhang, 2006; Yang et al., 2009).

Em função das características supramencionadas, fica evidente a necessidade de métodos computacionais eficientes que possibilitem a análise automática de tais dados ou que forneçam auxílio aos especialistas de domínio durante esse processo. Dentre os métodos computacionais que vêm sendo aplicados para essa finalidade, encontram-se métodos da estatística (Zhang, 2006), do aprendizado de máquina (Tarca et al., 2007) e da mineração de dados (Bertone e Gerstein, 2001). Além dos métodos empregados durante a própria fase de análise, ainda são necessários métodos de pré-processamento a fim de minimizar o efeito negativo causado pela presença de ruídos, *outliers* e atributos não informativos nos

algoritmos utilizados. Tal pré-processamento é discutido a seguir.

3.3 *Pré-Processamento*

Seja com o objetivo de melhorar a qualidade dos dados disponíveis ou de torná-los mais adequados para um algoritmo em particular, a etapa de pré-processamento de dados é de fundamental importância e pode influenciar os resultados obtidos durante a análise que será posteriormente realizada ([Tan et al., 2005](#)). Nesta seção são brevemente apresentadas algumas das técnicas de pré-processamento comumente adotadas para dados de *microarray*. O tópico não será abordado em profundidade, uma vez que alguns dos problemas envolvidos na fase de pré-processamento desse tipo de dados constituem tópicos de pesquisa por si sós ([Jiang et al., 2004](#)).

É importante ressaltar que antes mesmo da obtenção da matriz de dados na forma apropriada para análise já apresentada, os dados brutos provenientes de experimentos são geralmente “preparados”. Os procedimentos de preparação dos dados adotados antes da obtenção da matriz em sua forma final estão relacionados a questões e protocolos experimentais da própria tecnologia e não serão aqui abordados. Informações sobre tais procedimentos podem ser encontradas, por exemplo, em [Sherlock \(2001\)](#), [Smyth e Speed \(2003\)](#), [Hua et al. \(2008\)](#) e em referências contidas. Os procedimentos de pré-processamento aqui discutidos são aplicados após a obtenção da matriz de dados na forma da Figura 3.5.

Com a obtenção da matriz de dados, uma das primeiras questões a ser tratada diz respeito aos valores ausentes. Valores ausentes são comumente encontrados em dados obtidos com a tecnologia de cDNA ([Tuikkala et al., 2008](#)), a qual, possui menor padronização, se comparada a tecnologia de oligonucleotídeos (Affymetrix). Dessa forma, a discussão sobre valores ausentes se limita ao escopo de dados obtidos com a tecnologia de cDNA. A presença de valores ausentes pode ser atribuída a diversos fatores relacionados à realização dos experimentos de *microarrays* de cDNA, como apontado por [Troyanskaya et al. \(2001\)](#). Neste sentido, riscos ou poeira presentes na pastilha de vidro ou deformidades nos próprios *probes* podem levar a erros de leitura e consequentemente valores ausentes ([Troyanskaya et al., 2001](#)). Uma solução simples para o problema é a repetição do experimento que deu origem aos dados, o que nem sempre é possível e possui um custo associado. Uma vez que muitos dos algoritmos utilizados na análise da matriz de dados não toleram ou são afetados pela presença de valores ausentes, medidas devem ser tomadas para contornar ou de certa forma amenizar este problema.

Basicamente, três alternativas têm sido empregadas na literatura para o tratamento de valores ausentes em dados de expressão gênica. A primeira delas é descartar os objetos (genes ou amostras, dependendo do contexto da aplicação) para os quais existem valores faltantes. Essa alternativa pode ser considerada quando o número de objetos com valores ausentes é pequeno, porém não é considerada uma alternativa muito interessante, uma vez que informações contidas nos demais atributos do objeto que será descartado tam-

bém serão perdidas (Sherlock, 2001). Uma segunda abordagem que pode ser tomada é a substituição de todos os valores ausentes por um único valor constante, tipicamente zero. Essa abordagem tem sido empregada em dados de expressão gênica, porém não é muito indicada, uma vez que não leva em consideração a correlação de estrutura existente nos dados (Troyanskaya et al., 2001). A terceira alternativa é a estimação dos valores ausentes, também chamada de imputação de dados. Muitos dos métodos de imputação de dados provenientes do aprendizado de máquina têm sido utilizados na imputação de valores ausentes em dados de expressão gênica (Tuikkala et al., 2008). De acordo com Tuikkala et al. (2008), mesmo não havendo um consenso sobre qual método de imputação produz melhores resultados para dados de expressão gênica, estimar os valores ausentes sempre produz melhores resultados do que as outras duas alternativas já mencionadas. Troyanskaya et al. (2001) realizaram uma comparação de três diferentes métodos de imputação de valores ausentes para bases de dados de expressão gênica e sugerem que o método de imputação baseado no método *k-Nearest Neighbors* (kNNImpute) é uma escolha razoável para a tarefa. Para mais informações sobre o tópico de imputação de valores em dados de expressão gênica, os trabalhos de Troyanskaya et al. (2001) e Tuikkala et al. (2008) podem ser consultados.

Uma vez tratado o problema de valores ausentes, duas abordagens ainda são comumente empregadas antes da análise propriamente dita dos dados. Em se tratando das tarefas de agrupamento e classificação, genes que não apresentam variações para diferentes amostras, ou apresentam apenas uma pequena variação, são tipicamente removidos. Isso se deve ao fato de que agrupar esses genes não revela nenhuma informação interessante (Heyer et al., 1999). Além disso, em problemas de classificação tais genes não auxiliam na discriminação entre diferentes classes de amostras. É ainda comum, antes da aplicação de algoritmos de análise, que as linhas da matriz de dados (genes) sejam padronizadas para possuírem média zero e desvio padrão um, tanto em problemas de agrupamento (Heyer et al., 1999) como de classificação (Lee et al., 2005), a fim de que as similaridades em tendência ou forma entre genes sejam evidenciadas.

É importante notar que as medidas de pré-processamento adotadas podem variar, não existindo uma metodologia padrão. Isso se deve ao fato de que os métodos de pré-processamento são empregados de acordo com características da base de dados e dos algoritmos que serão posteriormente empregados em sua análise. Mais informações sobre métodos de pré-processamento aplicados a dados de expressão gênica podem ser encontradas, por exemplo, em Kuo et al. (2004) e referências contidas.

3.4 Tarefas Comuns e Aplicações

Assumindo que a base de dados encontra-se devidamente pré-processada, três diferentes tipos de tarefas de análise de dados são necessárias, segundo Piatetsky-Shapiro e Tamayo (2003), para a compreensão de dados de expressão gênica, a saber: agrupamento,

seleção de atributos e classificação, sendo a avaliação do desempenho de coeficientes de correlação nestas três tarefas o principal objetivo do presente trabalho. Nesta seção estas tarefas são apresentadas brevemente do ponto de vista do aprendizado de máquina e relacionadas a seus objetivos biológicos.

3.4.1 Agrupamento

A análise de agrupamento de dados pode ser compreendida como o processo de organizar n objetos dispostos em um espaço de m dimensões (atributos) em k grupos que possuam significado no contexto de um problema em particular (Jain e Dubes, 1988). Mesmo não havendo uma definição precisa e globalmente aceita para o termo grupo, diversos métodos de agrupamento têm sido criados e utilizados para revelar padrões intrínsecos aos dados em diversas áreas (Everitt et al., 2001). Uma vez que os grupos aos quais os objetos devem ser atribuídos não são conhecidos previamente, a tarefa de agrupamento de dados é considerada parte do aprendizado de máquina não-supervisionado. De acordo com Jain e Dubes (1988), seus métodos podem ser divididos em duas grandes categorias, as quais compreendem os chamados métodos particionais e métodos hierárquicos.

Algoritmos ou métodos de agrupamento particionais produzem como saída uma única partição do conjunto de dados recebido como entrada. De acordo com Jain e Dubes (1988), o problema de particionar dados pode ser definido da seguinte forma: a partir de um conjunto de n objetos representados por m atributos, determine uma partição dos objetos em k grupos, de forma que objetos em um grupo sejam mais similares uns aos outros do que a objetos em grupos diferentes. Os métodos particionais podem ainda ser subdivididos em métodos exclusivos e não exclusivos. Neste trabalho serão considerados apenas métodos particionais exclusivos, os quais produzem partições nas quais cada objeto pertence precisamente a um único grupo.

Diferentemente dos métodos particionais, que produzem como saída uma única partição dos dados, métodos de agrupamento hierárquicos produzem como resultado um conjunto de partições aninhadas e podem ainda ser divididos em duas classes: aglomerativos e divisivos. Métodos aglomerativos iniciam atribuindo cada objeto a um grupo distinto, de forma que o número de grupos seja igual ao número de objetos. A cada passo subsequente os dois grupos mais similares são então unidos e formam um novo grupo. Tal passo é então repetido até que todos os objetos pertençam a um único grupo. Nos métodos divisivos o processo ocorre de forma inversa, com todos os objetos iniciando em um único grupo. Durante a aplicação de um método divisivo os grupos são divididos até que ao final sejam obtidos n grupos, cada um formado por um único objeto. O resultado obtido com a aplicação de métodos hierárquicos é geralmente representado na forma de um dendrograma, tal qual ilustrado na Figura 3.6. Diferentes partições com diferentes números de grupos podem ser obtidas com a escolha de pontos de corte distintos no dendrograma gerado por um algoritmo hierárquico. A obtenção ou não de uma partição com k grupos é de escolha do usuário do algoritmo, assim como o valor de k desejado.

Figura 3.6: Exemplo de dendrograma. O corte indicado pela linha tracejada produz como resultado uma partição de 3 grupos.

Em dados de expressão gênica, a tarefa de agrupamento desempenha um papel de fundamental importância durante a análise exploratória dos dados. Suas duas principais aplicações são o agrupamento de genes ([Kerr et al., 2008](#)) e o agrupamento de amostras ([Golub et al., 1999](#)). Na primeira forma de aplicação, genes são os objetos e amostras os atributos do problema em questão. A descoberta de grupos de genes similares (chamados genes co-expressos) pode auxiliar na identificação da função de genes para os quais ainda não há informações disponíveis ([Jiang et al., 2004](#)). Além disso, grupos de genes podem fornecer informações a respeito dos sistemas que regulam tais genes, uma vez que genes co-expressos podem estar associados a um único sistema regulatório ([Heyer et al., 1999](#)). É importante ressaltar que neste contexto de aplicação as amostras geralmente estão associadas a diferentes tempos de um mesmo processo (*e.g.* ciclo celular, experimento de estresse celular, dentre outros), de forma que cada gene pode ser visto como uma série temporal. Já no segundo tipo de aplicação, amostras são objetos enquanto genes são atributos do problema. Estudos conduzidos neste cenário possuem como objetivo a identificação de grupos ou sub-grupos ainda desconhecidos relacionados a doenças, geralmente câncer ([Zhang, 2006](#)).

Diversos métodos de agrupamento tradicionais baseados em proximidades têm sido aplicados para dados provenientes de experimentos de *microarray*, tanto no contexto de agrupamento de genes quanto de amostras. O bem conhecido algoritmo particional *k-means* ([MacQueen, 1967](#); [Steinley, 2006](#)) foi aplicado e comparado por ([Yeung et al., 2001](#)) com outros cinco algoritmos de agrupamento em quatro bases de dados envolvidas no problema de agrupamento de genes. O algoritmo apresentou os melhores resultados juntamente com o algoritmo de agrupamento CAST (*Cluster Affinity Search Technique*) ([Ben-Dor e Yakhini, 1999](#)), o qual foi proposto especificamente para dados de expressão gênica. Nos resultados obtidos por [Costa \(2003\)](#), quanto à comparação de métodos de agrupamento

também para agrupamento de genes, o algoritmo *k-means* novamente produziu resultados satisfatórios. Métodos de agrupamento hierárquicos, tanto divisivos (Datta e Datta, 2003) como aglomerativos (Gibbons e Roth, 2002), também têm sido aplicados para dados de expressão gênica. Os resultados obtidos por Yeung et al. (2001) sugerem que os algoritmos hierárquicos aglomerativos *Complete-Linkage* e *Average-Linkage* devem ser utilizados em detrimento do algoritmo *Single-Linkage*, que, no geral, produz os piores resultados. No contexto de agrupamento de amostras, de Souto et al. (2008) apresentam um estudo comparativo considerando 35 bases de dados e sete algoritmos. No estudo realizado, os autores destacam os algoritmos *Finite Mixture of Gaussians* (FMG) e *k-means* em virtude de seus bons resultados nas bases de dados consideradas. Novamente, o algoritmo *Single-Linkage* produziu os piores resultados dentre os métodos hierárquicos.

Além da aplicação de métodos tradicionais, algoritmos de agrupamento têm sido desenvolvidos especificamente para o cenário de dados de expressão gênica. Heyer et al. (1999) propõem um método de agrupamento particional chamado QT-Clust, o qual, segundo seus autores, elimina a necessidade da escolha do número de grupos desejados, além de automaticamente eliminar genes que não devem pertencer a nenhum grupo. Porém, a eliminação do parâmetro que configura o número de grupos desejados vem ao custo da introdução de dois novos parâmetros, os quais devem ser ajustados pelo usuário. Mais tarde, De Smet et al. (2002) apresentam o algoritmo de agrupamento particional chamado Adap-Cluster, o qual, segundo os autores, possui como idéia básica o mesmo algoritmo proposto por Heyer et al. (1999), porém, com a capacidade de induzir grupos de tamanhos diferentes. O algoritmo proposto não necessita a definição do número de grupos desejados, mas também apresenta dois parâmetros que devem ser configurados pelo usuário. Eisen et al. (1998) propõem a aplicação de um algoritmo hierárquico aglomerativo *Average-Linkage* juntamente com a adoção de um procedimento de ordenação do dendrograma obtido a fim de maximizar a similaridade de objetos adjacentes. O resultado produzido é então representado graficamente, auxiliando a análise da hierarquia obtida. A metodologia proposta tem sido amplamente utilizada para dados de expressão gênica (Jiang et al., 2004), porém depende basicamente da análise visual de especialistas de domínio.

Além da escolha do algoritmo de agrupamento a ser aplicado, diversos algoritmos de agrupamento demandam a escolha de uma medida de proximidade entre pares de objetos, sendo de fundamental importância a escolha de uma medida que permita ao algoritmo extrair grupos significativos dos dados. Essa importância tem sido enfatizada tanto no contexto mais geral de agrupamento de dados (Jain e Dubes, 1988) como no cenário mais específico de agrupamento de dados de expressão gênica (Costa et al., 2002), (Kerr et al., 2008). Como será visto mais adiante (Seção 3.5), as medidas de proximidade comumente empregadas durante o agrupamento de dados de expressão gênica são coeficientes de correlação, sendo a investigação do desempenho de algoritmos de agrupamento quando equipados com diferentes coeficientes de correlação parte integrante do presente trabalho.

A literatura de agrupamento de dados é vasta e diversas referências sobre o tema

podem ser encontradas. Ao leitor interessado em mais informações no contexto geral de agrupamento de dados, os livros tradicionais de [Jain e Dubes \(1988\)](#) e [Everitt et al. \(2001\)](#) podem ser consultados. Para uma revisão de métodos de agrupamento de dados no contexto de dados de expressão gênica, [Jiang et al. \(2004\)](#), [D'haeseleer \(2005\)](#) e [Zhang \(2006\)](#) podem ser consultados. A comparação de diversos algoritmos de agrupamento de dados no contexto de dados de expressão gênica é realizada por [Yeung et al. \(2001\)](#), [Datta e Datta \(2003\)](#), [Costa \(2003\)](#) e [de Souto et al. \(2008\)](#).

3.4.2 Seleção de Atributos

A seleção de atributos possui por objetivo encontrar, dentre vários subconjuntos possíveis de atributos, aquele que maximiza um determinado critério de avaliação, eliminando atributos irrelevantes e redundantes, além de ruído contido nos dados ([Liu e Yu, 2005](#)). Segundo [Guyon e Elisseeff \(2003\)](#), a redução do número de atributos é desejável por apresentar diversos benefícios em potencial, dentre os quais destacam-se maior facilidade de visualização e compreensão dos dados e diminuição nos tempos de treinamento e utilização de algoritmos. Considerando os efeitos inerentes ao problema da maldição da dimensionalidade ([Mitchell, 1997](#)), a seleção de atributos pode ainda melhorar a acurácia de predição dos modelos obtidos⁷, evitando o problema de superajuste aos dados de treinamento (*overfitting*).

O superajuste aos dados de treinamento ocorre tipicamente quando o número de objetos disponíveis para a construção de um classificador é pequeno, porém descritos por uma grande quantidade de atributos (dimensões). Nesses casos, encontrar uma função que separe perfeitamente os dados se torna mais fácil (em virtude da grande quantidade de atributos disponíveis), porém tal função pode não se comportar bem durante a classificação de objetos não vistos durante a fase de construção do modelo. Uma vez que enumerar todos os subconjuntos possíveis de atributos torna-se impraticável quando o número de atributos cresce (para m atributos existem 2^m subconjuntos possíveis), diversos métodos heurísticos de seleção de atributos têm sido desenvolvidos e empregados na prática.

Os métodos de seleção de atributos existentes podem ser divididos em três categorias ([Guyon e Elisseeff, 2003](#)), levando em conta a maneira com que o algoritmo de aprendizado de máquina, que será posteriormente utilizado, se relaciona com o método de seleção para a avaliação dos subconjuntos de atributos. *Wrappers* avaliam os subconjuntos de atributos com a aplicação do próprio método de aprendizado de máquina que será posteriormente utilizado com os atributos selecionados para resolver o problema em questão, podendo este método ser visto como uma sub-rotina do processo de seleção de atributos ([Blum e Langley, 1997](#)). Por empregarem como critério de avaliação o próprio algoritmo que será posteriormente utilizado durante a fase de análise, métodos pertencentes a essa categoria tipicamente possuem um maior custo computacional associado

⁷A seleção de atributos será considerada somente no contexto de classificação durante o trabalho.

(Liu e Yu, 2005). Diferentemente dos *wrappers*, filtros selecionam atributos sem nenhuma participação do algoritmo de aprendizado de máquina, de forma que a avaliação dos subconjuntos de atributos é feita exclusivamente com base em características inerentes aos dados (Liu et al., 2006). Devido à sua baixa complexidade computacional (se comparado à maioria dos métodos classificados como *wrappers*), filtros têm ganhado grande atenção e têm sido a principal escolha para seleção de atributos em dados de *microarray*, como apontado por Saeys et al. (2007). Por fim, métodos de seleção de atributos embarcados (*embedded*) são aqueles nos quais o subconjunto de atributos é selecionado como parte e durante a construção do modelo de aprendizado de máquina, estando os critérios de seleção intimamente relacionados com o algoritmo utilizado. Para uma revisão e descrição mais detalhada sobre métodos de seleção de atributos, podem ser consultadas (Langley, 1994), (Blum e Langley, 1997), (Guyon e Elisseeff, 2003) e (Liu e Yu, 2005).

Transportando os conceitos de seleção de atributos apresentados para dados de expressão gênica, a principal aplicação de seleção de atributos é encontrada previamente à tarefa de classificação de diferentes condições experimentais, a chamada seleção de genes. O processo de seleção de genes possui por objetivo encontrar genes informativos que expliquem ou permitam discriminar, de forma mais eficiente, diferentes condições de amostras (*e.g.* células saudáveis e cancerosas). Nesse sentido, a seleção de genes busca por genes que apresentem níveis de expressão mais dissimilares sob diferentes condições, juntamente com a identificação e remoção de genes irrelevantes e redundantes entre si, de forma a permitir melhor discriminação entre diferentes tipos de amostras.

Filtros de atributos baseados na ordenação de genes têm sido aplicados para a seleção de genes informativos (Golub et al., 1999). Métodos baseados nesse conceito criam uma ordenação de genes de acordo com algum critério, como, por exemplo, sua capacidade individual de discriminar entre classes, e possuem um baixo custo computacional associado. Mesmo selecionando os genes mais informativos, essa abordagem não considera correlações entre genes, o que pode levar à seleção de subconjuntos com número elevado de atributos, uma vez que as possíveis redundâncias entre estes não são consideradas durante o processo de seleção. Como apontado por Li et al. (2004), grande parte dos métodos aplicados durante a seleção de genes ignora suas possíveis correlações. Dessa forma, a utilização de métodos que considerem tais correlações pode auxiliar na seleção de melhores subconjuntos de genes (Dudoit et al., 2002).

Métodos de seleção de atributos que consideram as redundâncias existentes entre genes também têm sido aplicados em bases de dados de expressão gênica. Guyon et al. (2002) propõem um *wrapper* de seleção de atributos combinando *Support Vector Machines* (SVMs) com eliminação recursiva de atributos e compara os resultados com o método de ordenação de genes utilizado por (Golub et al., 1999), produzindo melhores resultados. Ding e Peng (2005) propõem o filtro chamado *Minimum Redundancy - Maximum Relevance* (MRMR), o qual possui como objetivos minimizar a redundância (correlação) entre genes e ao mesmo tempo maximizar sua relevância para a distinção entre diferentes

classes. Os autores comparam os resultados de seu método com os resultados de métodos de ordenação comumente empregados para a seleção de genes, quanto à acurácia de classificação de diferentes métodos, e observam que, ao considerar a correlação entre genes, melhores acurácias nos classificadores construídos foram obtidas, o que indica também um melhor subconjunto de genes. Mesmo considerando a correlação entre atributos, o método de seleção proposto por [Ding e Peng \(2005\)](#) (assim como os demais métodos baseados em ordenação de atributos) demandam a escolha, por parte do usuário, do número de atributos que serão posteriormente utilizados na construção do modelo de classificação, o que não é uma tarefa trivial. Uma comparação entre métodos comumente empregados de seleção de genes por ordenação é realizada por [Li et al. \(2004\)](#) no contexto de classificação de tecidos, porém nenhum método é identificado como melhor para a tarefa.

Mesmo sendo parte integrante da etapa de pré-processamento de dados, a qual não é o foco principal do presente trabalho, a seleção de atributos recebe atenção neste trabalho por dois motivos. O primeiro deles é que, em dados de expressão gênica, a quantidade de genes analisados é elevada. Em problemas de classificação de amostras, a grande quantidade de genes (atributos) pode produzir efeitos negativos sob a acurácia do algoritmo de classificação utilizado, tornando a seleção de atributos um passo tipicamente necessário antes da aplicação de modelos de classificação. O segundo motivo é que alguns dos algoritmos de seleção de atributos existentes na literatura utilizam medidas de correlação entre atributos, o que se insere diretamente no escopo do presente trabalho. Além do uso de uma medida de correlação entre atributos, alguns destes algoritmos estimam de forma automática o tamanho do subconjunto de atributos selecionados (ver, por exemplo, [\(Mitra et al., 2002\)](#), [\(Yu e Liu, 2003\)](#) e [\(Covões et al., 2009\)](#)), diferentemente dos métodos de [Guyon et al. \(2002\)](#) e [Golub et al. \(1999\)](#), por exemplo. Uma vez que não há a necessidade da escolha de um limiar para a definição do número de atributos selecionados estes algoritmos tornam-se uma alternativa atrativa aos demais métodos já citados.

3.4.3 Classificação

Segundo [Tan et al. \(2005\)](#), classificação é o processo pelo qual objetos são atribuídos a uma classe dentre um conjunto de classes previamente determinadas. Nesse sentido, o processo de classificação difere do agrupamento de dados, uma vez que se conhece *a priori* tanto o número de classes como quais são as classes, inserindo assim os métodos e algoritmos de classificação no chamado aprendizado supervisionado. Dessa forma, o objetivo do processo de classificação é, a partir de dados já rotulados, também chamados dados de treinamento, induzir um modelo de classificação que permita a posterior classificação de dados ainda não vistos durante o processo de indução.

Os métodos ou algoritmos empregados para realização da tarefa de classificação podem ser vistos como possuindo duas fases distintas, as quais são chamadas fases de indução e de classificação. Durante a fase de indução, também chamada de fase de treinamento, dados já rotulados são fornecidos a um algoritmo para a construção de um modelo de

classificação. Os dados já rotulados formam o conjunto de treinamento, no qual cada um dos objetos possui, além de seus atributos descritivos, um atributo especial, chamado atributo meta, o qual associa cada um dos objetos a uma classe específica que deve ser aprendida pelo modelo de classificação. Após a fase de indução espera-se que o modelo criado tenha se ajustado aos dados de treinamento, bem como adquirido a capacidade de generalizar, ou seja, atribuir rótulos de classes corretamente a dados ainda não vistos durante o processo de indução do modelo (Tan et al., 2005). O processo de atribuir a uma classe cada um dos novos objetos apresentados ao modelo constitui a fase de classificação.

Em dados de expressão gênica, a principal aplicação de classificadores é encontrada na discriminação de diferentes tipos de amostras (Lee et al., 2005). Esse tipo de aplicação tem como objetivo a construção de modelos que permitam discriminar conjuntos de amostras ainda não rotuladas a partir de dados de expressão gênica já rotulados. Diferentes tipos de amostras podem representar, por exemplo, tecidos saudáveis e doentes ou ainda diferentes classes de câncer já conhecidas. De acordo com Lu e Han (2003), a classificação de amostras cancerosas está relacionada com a descoberta de novos medicamentos para seu tratamento, além de poder proporcionar a indicação de tratamentos mais adequados a cada um dos tipos de câncer identificados, minimizando assim os efeitos toxicológicos sofridos pelos pacientes. Além da classificação de amostras, há ainda a possibilidade da utilização de classificadores na discriminação de genes, associando genes com função ainda não conhecida a classes funcionais já estabelecidas (Kuramochi e Karypis, 2005). Tal aplicação, porém, é raramente encontrada na literatura e não será abordada neste trabalho.

Um passo geralmente tomado antes da aplicação de um algoritmo de classificação em dados de expressão gênica é a seleção de atributos, conforme discutido na Seção 3.4.2. Técnicas de seleção de atributos têm sido principalmente aplicadas como um passo de pré-processamento para a tarefa de classificação de amostras. Isso se deve ao grande número de genes e ao pequeno número de amostras encontrados em dados de expressão gênica, os quais podem influenciar negativamente o desempenho de algoritmos de classificação (Liu e Yu, 2005).

Diversos algoritmos têm sido empregados durante a tarefa de classificação de amostras. Os classificadores vão desde modelos mais simples (Dudoit et al., 2002), tais como Naïve Bayes e kNN (*k-Nearest Neighbors*), até modelos mais complexos, como SVMs (*Support Vector Machines*) e redes neurais artificiais (Lee et al., 2005). Dentre os métodos de classificação que vêm sendo utilizados para dados de *microarray*, é de principal interesse neste trabalho o método kNN. Tal interesse se dá uma vez que o método faz uso de uma medida de proximidade durante a classificação de objetos. Em dados de expressão gênica, tal medida tem sido implementada através do coeficiente de correlação de Pearson (Ben-Dor et al., 2000; Lu e Han, 2003), tornando assim interessante a investigação do comportamento de outras medidas. Além disso, estudos como os de Dudoit et al. (2002)

e [Lee et al. \(2005\)](#) sugerem que o método kNN produz bons resultados⁸ para dados de expressão gênica, ainda que comparado a métodos mais sofisticados de classificação, possuindo assim um bom compromisso entre simplicidade e acurácia obtida nos resultados. Mais informações sobre a tarefa de classificação no contexto de dados de expressão gênica podem ser obtidas, por exemplo, em [Dudoit et al. \(2002\)](#) e [Lee et al. \(2005\)](#).

3.5 Medidas de Proximidade em Dados de Expressão Gênica

Durante a apresentação das três tarefas comumente aplicadas a dados de expressão gênica fica evidente a necessidade da escolha de uma medida de proximidade (similaridade ou dissimilaridade) adequada, uma vez que diversos algoritmos utilizam internamente uma medida de proximidade. Neste contexto, as medidas podem ser aplicadas a fim de comparar pares de genes ou amostras, dependendo do cenário no qual o algoritmo é utilizado. O problema de calcular proximidades entre pares de genes ou amostras pode ser simplificado se ambos os objetos a serem comparados forem compreendidos como sequências numéricas reais de comprimento fixo, para as quais uma medida de proximidade pode ser diretamente aplicada.

Em dados de *microarray* as formas das sequências numéricas sob comparação possuem fundamental importância, enquanto as diferenças absolutas entre pares de valores das sequências não são de interesse na maioria dos problemas. Isso significa que, ao comparar duas sequências numéricas, as tendências de seus valores possuem papel fundamental, mesmo que seus valores para cada um de seus atributos sejam diferentes. Esta característica possui motivação biológica, podendo ser explicada pelo fato de que proteínas trabalham em diversos complexos. Em alguns casos, tais complexos requerem mais de uma cópia da mesma proteína, o que acarreta em um maior nível de expressão para o gene que origina tal proteína ([Krause et al., 2004](#)). Neste sentido, espera-se que dois genes que participam da codificação de um mesmo complexo possuam níveis de expressão similares ([Heyer et al., 1999](#)). Além disso, similaridades em tendência podem trazer informações relevantes a respeito das regiões que regulam tais genes ([Heyer et al., 1999](#)).

Para tornar mais claro esse conceito, considere uma matriz de dados de expressão gênica $\mathbf{M}$. A fim de comparar, por exemplo, os níveis de expressão obtidos para dois genes $\mathbf{g}_1$ e $\mathbf{g}_2$ nas m diferentes condições experimentais apresentadas na matriz $\mathbf{M}$, podemos representar graficamente cada um desses níveis como pares ordenados em um gráfico, no qual o eixo x está associado aos m experimentos e o eixo y está associado aos níveis de expressão obtidos para os mesmos, tal qual apresentado na Figura 3.7. Cada uma das linhas apresentadas na figura representa o nível de expressão para um gene, sendo o nível obtido em cada experimento (atributo) marcado pelo símbolo * contido nestas linhas. Analisando as sequências atributo a atributo é fácil perceber que as grandezas de seus valores diferem. Porém, se o par de genes considerado for comparado quanto às suas

⁸É importante ressaltar que, em geral, os melhores resultados de classificação são obtidos com SVMs.

formas ou tendências, sua similaridade fica evidente.

Figura 3.7: Duas sequências numéricas com tendências similares.

Em virtude da necessidade de comparar sequências numéricas dando ênfase às suas similaridades em tendências e não à similaridade entre valores, coeficientes de correlação têm sido as medidas de proximidade tipicamente aplicadas a dados de expressão gênica. Uma das medidas comumente utilizadas neste tipo de dados é a correlação de Pearson (Zhang, 2006). Essa medida tem sido adaptada como medida de distância tanto para métodos envolvidos na tarefa de classificação de amostras (Dudoit et al., 2002) como para as tarefas de agrupamento (Eisen et al., 1998) e seleção de genes (Ding e Peng, 2005). Quando os dados obtidos são padronizados (cada sequência com média zero e desvio padrão igual a um), a correlação de Pearson pode ser substituída pela distância Euclidiana, a qual produz resultados similares aos obtidos com o uso da correlação de Pearson. De fato, pode-se provar que as duas medidas são intimamente relacionadas quando comparam sequências padronizadas, cada qual com média zero e desvio padrão igual a um (Jiang et al., 2004).

Mesmo sendo amplamente utilizada em dados de expressão gênica, a correlação de Pearson possui algumas características indesejadas. Dentre estas destacam-se a sensibilidade a ruído e *outliers*, comumente encontrados neste tipo de dados. A presença de *outliers* pode fazer, por exemplo, com que genes ou amostras similares/dissimilares sejam consideradas equivocadamente dissimilares/similares. Tal tipo de equivoco pode afetar negativamente qualquer uma das três tarefas anteriormente apresentadas. Durante o processo de agrupamento tais erros podem fazer, por exemplo, com que dois objetos altamente dissimilares sejam atribuídos a um mesmo grupo. Na seleção de atributos, por exemplo, correlações não existentes entre atributos podem ser identificadas, causando a eliminação de um dos atributos, o qual talvez ajudasse durante o processo posterior de indução de um classificador qualquer. Finalmente, na classificação de dados, erros desse tipo podem levar a classificação errônea de objetos.

Por estes motivos, outras medidas de proximidade têm sido aplicadas para dados de expressão gênica. Segundo Kerr et al. (2008), quando se sabe previamente da presença

de ruído e *outliers* nos dados as correlações de Spearman e Kendall (vide Seções 4.1.2 e 4.1.3) podem ser mais apropriadas. A correlação de Spearman, por exemplo, foi utilizada por [Kotlyar et al. \(2002\)](#) para a tarefa de agrupamento de genes. A medida foi aplicada juntamente com o algoritmo hierárquico *Complete-Linkage* e permitiu a geração de grupos nos quais genes com funções já conhecidas e previamente associadas foram encontrados. Além das medidas de correlação já citadas, novas medidas têm sido propostas recentemente, sendo que algumas dessas foram propostas especificamente para o cenário de análise de dados de expressão gênica.

[Balasubramaniyan et al. \(2005\)](#) apresentam uma medida de similaridade baseada na correlação de Spearman para a detecção de padrões locais em dados de séries temporais de expressão gênica. A medida proposta possibilita também a detecção de padrões locais com deslocamentos em função do tempo (*time-shifts*). A similaridade é construída de forma a atribuir valores altos a pares de objetos que possuem subsequências altamente similares, sendo, segundo seus autores, menos sensível a presença de ruídos nos dados. Além da medida, os autores propõem também um novo algoritmo de agrupamento e utilizam ambos para o agrupamento de genes, encontrando alguns grupos para os quais genes de função similar foram identificados. Uma característica do trabalho é que os resultados de agrupamento obtidos são comparados aos de [Tavazoie et al. \(1999\)](#), porém, nenhum critério de validação de agrupamento ([Jain e Dubes, 1988](#)) é utilizado durante tal comparação. Embora interessante, tal medida necessita do uso de simulações de Monte Carlo, o que pode tornar sua aplicação impraticável em alguns casos. Tal fato, somado ao número de bases de dados utilizadas durante a avaliação realizada no presente trabalho, inviabilizou o uso de tal medida⁹.

[Heyer et al. \(1999\)](#) propõem uma medida de correlação chamada Jackknife (discutida na Seção 4.1.7), a qual, segundo seus autores, é mais robusta que a correlação de Pearson à presença de *outliers*. Além da medida de correlação, os autores propõem um novo algoritmo de agrupamento. Utilizando o algoritmo proposto equipado com a correlação Jackknife os autores agrupam uma única base de dados de expressão gênica e validam os resultados obtidos somente por inspeção visual (como em muitos dos trabalhos da área); porém, uma comparação com outras medidas não é considerada no trabalho. Uma vez que tal comparação não é realizada, não fica evidente se sua utilização pode trazer benefícios práticos, tendo em vista que a medida possui também um maior custo computacional se comparada a outras medidas utilizadas em dados de expressão gênica.

Outras medidas especificamente propostas para dados de expressão gênica são ainda apresentadas por [Möller-Levet et al. \(2005\)](#), [Kim et al. \(2007\)](#), [Hardin et al. \(2007\)](#) e [Son e Baek \(2008\)](#). As cinco medidas (duas no trabalho de [Son e Baek, 2008](#)) são propostas com ênfase em seus potenciais benefícios para a tarefa de agrupamento de dados; porém,

⁹Além disso, o código fonte fornecido pela autora apresentou incompatibilidade com a descrição do artigo original. Em comunicação pessoal realizada com a autora, pontos fundamentais para a utilização da medida (que não são detalhados no artigo) não foram esclarecidos, fato que reforçou a decisão de não incluir esta medida nas análises realizadas.

em nenhum dos trabalhos uma avaliação comparativa exaustiva das medidas é realizada. Quanto às medidas propostas por [Hardin et al. \(2007\)](#) e [Kim et al. \(2007\)](#) cabem algumas observações. Mesmo apresentando bons resultados em alguns cenários de agrupamento de genes, as duas medidas propostas possuem duas limitações em potencial. A primeira, como destacado pelos próprios autores, é referente a seu alto custo computacional, custo este que pode limitar seu uso quando um grande número de genes é considerado. A segunda é a questão de convergência dos algoritmos utilizados durante o cálculo das medidas, que pode não ocorrer. Em virtude destas características as duas medidas não foram consideradas para comparação neste trabalho. [Möller-Levet et al. \(2005\)](#) propõem uma medida que leva em conta os tamanhos dos intervalos de amostragem em experimentos de séries temporais. Tal medida é discutida na Seção 4.3. No trabalho de [Son e Baek \(2008\)](#) duas medidas baseadas nas correlações de Pearson e Spearman são apresentadas. Estas medidas são discutidas na Seção 4.2.1.

Além das medidas de proximidade desenvolvidas especificamente para aplicações de expressão gênica, duas novas medidas de correlação foram propostas recentemente por [Campello e Hruschka \(2009\)](#). As medidas propostas são variações de medidas já existentes, porém, com diferentes sensibilidades às características das sequências numéricas sob consideração. As medidas propostas são sensíveis tanto aos *ranks* quanto às magnitudes dos valores numéricos de duas sequências, sendo que tais sensibilidades podem tornar as medidas adequadas em cenários específicos da análise de dados de expressão gênica. Essas medidas são discutidas nas Seções 4.1.5 e 4.1.6.

A diversidade de medidas propostas na literatura, em especial as medidas recentemente propostas para dados de expressão gênica, indica que em alguns cenários específicos de aplicação, a medida de correlação de Pearson, comumente empregada, pode produzir resultados insatisfatórios. Porém, mesmo com o aumento da quantidade de medidas disponíveis, pouco esforço tem sido empregado em sua comparação empírica no que diz respeito às três principais tarefas de análise de dados aplicadas a dados de *microarray* discutidas na Seção 3.4. Neste sentido, destaca-se o trabalho de [Costa et al. \(2002\)](#) que tem por principal objetivo comparar tais medidas no cenário de dados de expressão gênica. Os autores realizam uma comparação de 3 diferentes medidas em 3 bases dados de expressão gênica que envolvem a tarefa de agrupamento de séries temporais de dados de *microarray*. Além do trabalho de [Costa et al. \(2002\)](#), o trabalho de [de Souto et al. \(2008\)](#) considera algumas das medidas mencionadas, porém tem seu principal foco na comparação de algoritmos de agrupamento e não de medidas de proximidade.

Uma das características dos trabalhos encontrados nos quais uma análise comparativa entre medidas é realizada ou uma nova medida é proposta é a falta de uma metodologia rigorosa que permita avaliar o desempenho empírico das medidas em bases de dados reais. Mesmo nos trabalhos de [Costa et al. \(2002\)](#) e [de Souto et al. \(2008\)](#), nos quais tal comparação é realizada com o uso de critérios de validação rigorosos, a quantidade de medidas consideradas é pequena se comparada às diversas medidas atualmente disponíveis.

3.6 *Considerações Finais*

Neste capítulo foram apresentadas inicialmente as duas principais tecnologias de *microarray* atualmente utilizadas, a saber: cDNA e Affymetrix. A partir de sua apresentação, o processo de obtenção de dados em forma adequada para análise computacional foi elucidado. Após tal apresentação, os procedimentos de pré-processamento comumente aplicados a este tipo de dados foram discutidos, juntamente com as três tarefas comumente empregadas para sua análise: agrupamento, seleção de atributos e classificação. Por fim, foi realizada uma discussão sobre medidas de proximidade no contexto da análise de dados de expressão gênica. A partir da discussão realizada torna-se evidente que coeficientes de correlação são as medidas comumente adotadas na análise deste tipo de dados, uma vez que as tendências entre sequências comparadas são de fundamental importância. Tais medidas serão discutidas em maiores detalhes no próximo capítulo.

Medidas de Proximidade

No Capítulo 3 as três principais tarefas empregadas na análise de dados de expressão gênica foram discutidas. Dentre os métodos e algoritmos empregados durante a fase de análise de dados, são abordados neste trabalho aqueles que empregam de alguma forma uma medida de proximidade entre pares de genes ou amostras, os quais podem ser interpretados como duas sequências numéricas $\mathbf{a}$ e $\mathbf{b}$, ambas com p valores reais na forma $\mathbf{a} = (a_1, a_2, \dots, a_p)$ e $\mathbf{b} = (b_1, b_2, \dots, b_p)$. Como visto anteriormente, neste cenário são de principal interesse aquelas medidas capazes de capturar similaridade em forma ou tendência entre sequências, ficando evidente que coeficientes de correlação são as medidas de proximidade predominantemente empregadas. Neste capítulo são apresentados e discutidos de forma mais detalhada os coeficientes de correlação e demais medidas de proximidade que foram objeto de estudo e comparação no presente trabalho. Na Seção 4.1 são apresentados os coeficientes de correlação considerados durante a avaliação, os quais constituem o principal foco desta dissertação. Na Seção 4.2 são apresentadas medidas baseadas em coeficientes de correlação introduzidas especificamente para o contexto de agrupamento de genes. Finalmente, na Seção 4.3 é apresentada uma medida de dissimilaridade introduzida também para o cenário específico de agrupamento de genes.

4.1 Coeficientes de Correlação

Considerando duas sequências numéricas $\mathbf{a}$ e $\mathbf{b}$, um coeficiente de correlação qualquer pode ser aplicado a fim de mensurar o grau e tipo de relacionamento existente entre as duas sequências, produzindo como resultado um único valor real entre -1 e 1. A magnitude do valor produzido está associada ao grau ou força da relação entre as sequências consideradas, de forma que valores próximos a 1, em módulo, indicam um forte relacionamento

(correlação) entre as duas sequências. Quanto menor o valor absoluto produzido, mais fraco é o relacionamento entre as sequências consideradas, com valores absolutos próximos a 0 indicando a ausência de correlação entre as sequências, ou seja, uma distribuição aleatória dos seus valores (Chinchilli e Gruemer, 1995). Já o tipo do relacionamento existente entre as duas sequências é representado pelo sinal do valor da correlação. Valores de correlação para os quais o sinal positivo é obtido indicam uma correlação positiva, de forma que as duas sequências crescem ou decrescem juntas. Valores de correlação para os quais o sinal negativo é observado indicam correlação negativa entre as sequências, de forma que quando os valores de uma sequência crescem/decrescem os da outra decrescem/crescem.

Diferentes medidas de correlação podem ser utilizadas a fim de mensurar o grau e tipo de relacionamento entre duas sequências, cada qual com um tipo de sensibilidade às características presentes nas sequências consideradas. As medidas que serão aqui discutidas podem ser categorizadas como sensíveis a três tipos de características, a saber: somente magnitudes, somente *ranks* e tanto magnitudes quanto *ranks*. Medidas sensíveis às magnitudes das sequências levam em conta as próprias magnitudes dos valores de ambas sequências durante o cálculo de sua correlação. Mesmo sendo sensíveis às magnitudes das sequências, tais medidas capturam similaridades referentes à suas tendências e não somente entre a diferença de seus valores. Medidas baseadas ou sensíveis somente a seus *ranks* desprezam qualquer informação contida nas magnitudes das sequências, observando somente a ordem relativa de cada um dos seus n valores. Por fim, medidas sensíveis a magnitudes e *ranks* consideram ambas informações durante o cálculo de seu valor para as sequências consideradas. Um resumo quanto às características de sensibilidade e complexidade computacional dos coeficientes de correlação que serão descritos a seguir é apresentado pela Tabela 4.1.

Tabela 4.1: Sensibilidade e complexidade computacional dos coeficientes de correlação discutidos. Tabela adaptada de Campello e Hruschka (2009).

Correlação	Sensibilidade	Complexidade
Pearson (ρ)	Magnitudes	$O(p)$
Spearman ($\hat{\rho}$)	<i>Ranks</i>	$O(p \log p)$
Kendall (τ)	<i>Ranks</i>	$O(p \log p)$
Goodman-Kruskal (γ)	<i>Ranks</i>	$O(p \log p)$
Goodman-Kruskal Ponderado ($\hat{\gamma}$)	Magnitudes e <i>ranks</i> de ambas as sequências	$O(p^2)$
Rank-Magnitude (r)	<i>Ranks</i> de a e magnitudes de b	$O(p \log p)$
Jackknife (ϱ)	Magnitudes	$O(p^2)$

4.1.1 Pearson

O coeficiente de correlação de Pearson (Pearson, 1895; Zhang, 2006) é considerado uma medida de proximidade paramétrica, uma vez que permite a identificação de correlações

do tipo linear entre duas sequências sob a premissa de que suas distribuições são aproximadamente normais (Estelberger e Reibnegger, 1995). A correlação de Pearson é sensível somente às magnitudes dos valores das sequências numéricas consideradas e é descrita pela Equação (4.1), na qual $\bar{a}$ e $\bar{b}$ correspondem às médias das respectivas sequências.

$$\rho(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^p (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^p (a_i - \bar{a})^2} \sqrt{\sum_{i=1}^p (b_i - \bar{b})^2}} \quad (4.1)$$

Uma consideração importante que deve ser feita a respeito deste coeficiente de correlação é a sua elevada sensibilidade à presença de *outliers*, bem como de distribuições não-normais das sequências consideradas (Zhang, 2006), as quais podem ser verificadas experimentalmente, conforme demonstrado por Fujita et al. (2009). Mesmo possuindo tais características indesejáveis, a correlação de Pearson tem sido amplamente utilizada em dados de expressão gênica (Zhang, 2006). Para casos nos quais se sabe previamente da existência de *outliers* ou de distribuições não normais das sequências, medidas de proximidade alternativas podem então ser preferidas.

Outro ponto que deve ser considerado é a tendência da correlação de Pearson em produzir falsos positivos (Zhang, 2006). Falsos positivos, neste contexto, podem ser definidos como pares de sequências pouco similares entre si, mas que recebem um alto valor de similaridade pela medida utilizada (Heyer et al., 1999). A produção de falsos positivos se deve em grande parte à sensibilidade da correlação quanto à presença de *outliers* nas sequências numéricas. Tal sensibilidade da correlação de Pearson motivou a criação da correlação Jackknife (Heyer et al., 1999), que será apresentada na Seção 4.1.7.

4.1.2 Spearman

O coeficiente de correlação de Spearman (Spearman, 1904; Kendall, 1970) pode ser interpretado como um caso particular da correlação de Pearson, no qual somente os *ranks* de ambas as sequências numéricas são considerados. Por considerar somente os *ranks* das sequências, desprezando a informação contida na magnitude de seus valores, a correlação de Spearman é mais robusta à presença de ruído e *outliers* se comparada à correlação de Pearson (Estelberger e Reibnegger, 1995) e tem sido utilizada para dados de expressão gênica nos quais se sabe que tais características estão presentes (Kerr et al., 2008). Em função da utilização dos *ranks* dos valores, a correlação de Spearman é capaz de identificar correlações não lineares entre duas sequências, desde que estas possuam seus valores crescendo ou decrescendo de forma monotônica (Estelberger e Reibnegger, 1995; Fujita et al., 2009). Outra potencial vantagem da correlação de Spearman é a de que uma distribuição aproximadamente normal das sequências não é assumida *a priori*, ao contrário do que ocorre com a correlação de Pearson. Porém, segundo Zhang (2006), de uma maneira geral a correlação de Spearman produz resultados inferiores à correlação de Pearson em dados de expressão gênica, em função da perda de informação que ocorre quando os valores são transformados para *ranks*. Apesar disso, esta correlação vem sendo

utilizada para dados de expressão gênica pelas razões citadas (Kotlyar et al., 2002; Fujita et al., 2009).

O valor da correlação de Spearman para duas sequências numéricas $\mathbf{a}$ e $\mathbf{b}$ pode ser obtido com a aplicação da Equação (4.1), dado que ambas as sequências sejam constituídas somente de *ranks*. Para sequências constituídas de valores reais, como usualmente ocorre em dados de expressão gênica, cada um de seus p valores devem ser substituídos pelos seus respectivos *ranks* nas sequências, de forma que o cálculo das médias e desvios padrão seja realizado em função desses. No caso de empates durante a geração dos *ranks* (dois ou mais valores iguais, que devem receber o mesmo *rank*), todos os valores iguais recebem como valor de *rank* a soma dos *ranks* dividida pela quantidade de valores. No caso de uma sequência $\mathbf{s}$ com os valores $\mathbf{s} = (5; 1; 5; 8)$, por exemplo, os *ranks* $\mathbf{s}_r = (2,5; 1; 2,5; 4)$ são obtidos.

4.1.3 Kendall

Levando em consideração somente os *ranks* das sequências sob análise, a correlação de Kendall (Kendall, 1970) é definida em função do número de pares concordantes (S_+) e o número de pares discordantes (S_-) entre as duas sequências $\mathbf{a}$ e $\mathbf{b}$. Pares concordantes são aqueles para os quais uma mesma relação de ordem é aplicada, ou seja, $a_i < a_j$ e $b_i < b_j$ ou $a_i > a_j$ e $b_i > b_j$. De forma análoga, pares discordantes são aqueles para os quais $a_i < a_j$ e $b_i > b_j$ ou $a_i > a_j$ e $b_i < b_j$. Todos os demais tipos de pares possíveis que não se enquadram na definição de concordantes ou discordantes são chamados pares neutros. Feitas tais considerações, a correlação de Kendall pode ser definida pela Equação (4.2), para a qual o termo de normalização $p(p-1)/2$ corresponde ao número possível de pares existentes para as sequências $\mathbf{a}$ e $\mathbf{b}$.

$$\tau(\mathbf{a}, \mathbf{b}) = \frac{S_+ - S_-}{p(p-1)/2} \quad (4.2)$$

É fácil observar que a correlação só atinge seus valores extremos caso não existam pares neutros presentes nas sequências e todos os $p(p-1)/2$ pares sejam concordantes ou discordantes, fazendo com que um dos termos (S_+ ou S_-) do numerador da Equação (4.2) torne-se igual a zero. Uma forma mais eficiente para o cálculo da correlação de Kendall foi proposta recentemente por Campello e Hruschka (2009), reduzindo seu custo computacional de $O(p^2)$ para $O(p \log p)$.

4.1.4 Goodman-Kruskal

A correlação de Goodman-Kruskal (Goodman e Kruskal, 1954) também toma por base a idéia de pares concordantes, discordantes e neutros apresentada para a correlação de Kendall, mantendo a forma de obtenção dos valores S_+ e S_- inalterada. A diferença existente em relação à correlação de Kendall se deve ao fato da correlação de Goodman-Kruskal não levar em conta os pares neutros em sua normalização. Esta alteração faz com

que o coeficiente consiga atingir seus valores máximos mesmo na presença de pares neutros nas sequências, por considerar somente o número de pares concordantes e discordantes em seu termo de normalização. A correlação de Goodman-Kruskal é apresentada pela Equação (4.3), para a qual S_+ representa o número de pares concordantes e S_- o número de pares discordantes.

$$\gamma(\mathbf{a}, \mathbf{b}) = \frac{S_+ - S_-}{S_+ + S_-} \quad (4.3)$$

4.1.5 Goodman-Kruskal Ponderado

A correlação de Goodman-Kruskal Ponderado (Campello e Hruschka, 2009) utiliza em seu cálculo tanto informações dos *ranks* quanto das magnitudes das sequências, sendo dessa forma sensível a ambas características para as duas sequências. O coeficiente é definido pela Equação (4.4) para duas sequências numéricas reais $\mathbf{a}$ e $\mathbf{b}$. De forma geral, a medida proposta expande a correlação de Goodman-Kruskal por considerar que dois pares podem ser concordantes ou discordantes em diferentes graus e não somente de forma binária, como considera a correlação de Goodman-Kruskal original.

$$\hat{\gamma}(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^{p-1} \sum_{j=i+1}^p \hat{w}_{ij}}{\sum_{i=1}^{p-1} \sum_{j=i+1}^p |w_{ij}|} \quad (4.4)$$

O termo $\hat{w}_{ij}$ é definido pela Equação (4.5). Ele é positivo para pares concordantes, negativo caso os pares sejam discordantes e zero para neutros. Os termos $\hat{w}_{ij}^{\mathbf{a}}$ e $\hat{w}_{ij}^{\mathbf{b}}$ da Equação (4.5) são fornecidos pelas Equações (4.6) e (4.7) respectivamente. Cada um desses termos representa a diferença percentual (com sinal) entre o i -ésimo e o j -ésimo elementos das sequências correspondentes. Os valores a_{max} e a_{min} representam o maior e menor valor dentre todos os valores da sequência $\mathbf{a}$, enquanto b_{max} e b_{min} são o maior e menor valores de $\mathbf{b}$, respectivamente.

$$\hat{w}_{ij} = \begin{cases} \min\{\hat{w}_{ij}^{\mathbf{a}}/\hat{w}_{ij}^{\mathbf{b}}, \hat{w}_{ij}^{\mathbf{b}}/\hat{w}_{ij}^{\mathbf{a}}\} & \hat{w}_{ij}^{\mathbf{a}} \hat{w}_{ij}^{\mathbf{b}} > 0 \\ \max\{\hat{w}_{ij}^{\mathbf{a}}/\hat{w}_{ij}^{\mathbf{b}}, \hat{w}_{ij}^{\mathbf{b}}/\hat{w}_{ij}^{\mathbf{a}}\} & \hat{w}_{ij}^{\mathbf{a}} \hat{w}_{ij}^{\mathbf{b}} < 0 \\ 1 & \hat{w}_{ij}^{\mathbf{a}} = \hat{w}_{ij}^{\mathbf{b}} = 0 \\ 0 & \text{demais casos} \end{cases} \quad (4.5)$$

$$\hat{w}_{ij}^{\mathbf{a}} = \begin{cases} \frac{a_i - a_j}{a_{max} - a_{min}} & \text{se } a_{max} \neq a_{min} \\ 0 & \text{outro caso} \end{cases} \quad (4.6)$$

$$\hat{w}_{ij}^{\mathbf{b}} = \begin{cases} \frac{b_i - b_j}{b_{max} - b_{min}} & \text{se } b_{max} \neq b_{min} \\ 0 & \text{outro caso} \end{cases} \quad (4.7)$$

O termo w_{ij} da Equação (4.4) é definido pela Equação (4.8), com os termos $w_{ij}^{\mathbf{a}} = \text{sign}(a_i - a_j)$ e $w_{ij}^{\mathbf{b}} = \text{sign}(b_i - b_j)$ sendo iguais aos sinais das operações $a_i - a_j$ e $b_i - b_j$

respectivamente. Em outras palavras, a Equação (4.8) simplesmente indica se um par é concordante (+1), discordante (-1) ou neutro (0), o que faz com que o denominador em (4.4) corresponda à mesma normalização realizada na correlação de Goodman-Kruskal original.

$$w_{ij} = \begin{cases} w_{ij}^{\mathbf{a}}/w_{ij}^{\mathbf{b}} & \text{se } w_{ij}^{\mathbf{b}} \neq 0 \\ 1 & \text{se } w_{ij}^{\mathbf{a}} = 0 \text{ e } w_{ij}^{\mathbf{b}} = 0 \\ 0 & \text{demais casos} \end{cases} \quad (4.8)$$

4.1.6 Rank-Magnitude

Quando somente os *rank*s de uma das duas sequências são de interesse ou quando uma das sequências é composta de valores qualitativos (*e.g.* valores ordinais), o coeficiente de correlação Rank-Magnitude pode ser utilizado (Campello e Hruschka, 2009). O coeficiente é definido pela Equação (4.9), com $R(a_i)$ representando o *rank* da posição i para a sequência $\mathbf{a}$.

$$\hat{r}(\mathbf{a}, \mathbf{b}) = \frac{2 \sum_{i=1}^p R(a_i)b_i - r^{max} - r^{min}}{r^{max} - r^{min}} \quad (4.9)$$

Os valores de r_m^{min} e r_m^{max} são fornecidos pelas Equações (4.10) e (4.11) respectivamente, nas quais o valor de $\bar{b}_i$ é obtido rearranjando simultaneamente (de forma pareada) as sequências $\mathbf{a}$ e $\mathbf{b}$ de forma que $\mathbf{b}$ fique ordenada.

$$r^{min}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^p (n+1-i)\bar{b}_i \quad (4.10)$$

$$r^{max} = \sum_{i=1}^p i\bar{b}_i \quad (4.11)$$

Uma das características da medida é sua assimetria. De forma a ser utilizada para dados nos quais as duas sequências consideradas são constituídas de valores numéricos, a falta de simetria da medida pode ser corrigida (Clarkson, 2006). A nova medida obtida, simétrica, é definida pela Equação (4.12). Qualquer menção à correlação Rank-Magnitude no decorrer desta dissertação se refere apenas à versão simétrica da correlação, apresentada pela Equação (4.12).

$$r(\mathbf{a}, \mathbf{b}) = \frac{\hat{r}(\mathbf{a}, \mathbf{b}) + \hat{r}(\mathbf{b}, \mathbf{a})}{2} \quad (4.12)$$

4.1.7 Jackknife

A correlação Jackknife foi introduzida por Heyer et al. (1999) a fim de minimizar o número de falsos positivos encontrados durante o processo de agrupamento de genes.

Falsos positivos são pares de objetos pouco similares que recebem um alto valor de similaridade pela medida utilizada. Como observado por Heyer et al. (1999), uma grande parte dos falsos positivos encontrados em dados de expressão gênica se deve à presença de *outliers*.

Tomando como base o procedimento estatístico de estimativa de *bias* conhecido como Jackknife (Efron, 1982), Heyer et al. (1999) propõem a correlação Jackknife, tal qual descrita pela Equação (4.13). Os valores $\rho^i(\mathbf{a}, \mathbf{b})$ indicam o valor de correlação de Pearson entre as sequências $\mathbf{a}$ e $\mathbf{b}$, com seu i -ésimo valor removido, enquanto $\rho(\mathbf{a}, \mathbf{b})$ representa a correlação de Pearson entre as sequências $\mathbf{a}$ e $\mathbf{b}$ considerando todos seus valores.

$$\varrho(\mathbf{a}, \mathbf{b}) = \min\{\rho^1(\mathbf{a}, \mathbf{b}), \rho^2(\mathbf{a}, \mathbf{b}), \rho^3(\mathbf{a}, \mathbf{b}), \dots, \rho^p(\mathbf{a}, \mathbf{b}), \rho(\mathbf{a}, \mathbf{b})\} \quad (4.13)$$

A idéia subjacente à correlação Jackknife é a de que, se as sequências não possuem *outliers* dominantes, então a remoção ou não remoção de qualquer par de valores presente nas sequências não irá afetar significativamente o valor obtido com o calculo da correlação. Porém, no caso em que as sequências possuem *outliers*, sua remoção diminuirá de forma significativa o valor da correlação, o que neste caso indica que as sequências eram correlacionadas devido em grande parte ou totalmente à presença de um único *outlier*.

Caso existam indícios de que as sequências sob consideração possuam mais de um *outlier*, a correlação Jackknife pode ser alterada para eliminar um maior número de valores a cada passo do cálculo da correlação, porém a um maior custo computacional associado. Como observado por Zhang (2006), esta variante que considera diferentes valores para remoção é raramente utilizada em função de seu custo computacional. Além da remoção de uma maior quantidade de valores, a correlação Jackknife pode ser generalizada para considerar outras medidas de correlação internamente, mesmo tendo sido proposta originalmente com a correlação de Pearson.

4.2 Medidas Baseadas em Coeficientes de Correlação

Nesta seção são apresentadas duas medidas baseadas em coeficientes de correlação propostas por Son e Baek (2008).

4.2.1 Son e Baek

Considerando o cenário de agrupamento de séries temporais de dados de expressão gênica, no qual os níveis de expressão de diversos genes são mensurados em diferentes experimentos (amostras) ao longo do tempo, Son e Baek (2008) propõem duas diferentes medidas de similaridade que combinam o resultado obtido com a aplicação de um coeficiente de correlação e outras informações extraídas das sequências sob comparação. Os autores argumentam que o uso de coeficientes de correlação neste cenário não captura de forma completa a informação temporal contida nas sequências. Ainda segundo

os autores, isso se deve ao fato de que os valores de correlação entre duas sequências $\mathbf{a} = (a_1, a_2, \dots, a_p)$ e $\mathbf{b} = (b_1, b_2, \dots, b_p)$ são idênticos aos valores obtidos com a aplicação da correlação a sequências $\mathbf{a}'$ e $\mathbf{b}'$, para as quais foram aplicadas as mesmas permutações entre valores.

A fim de considerar a ordem relativa entre os valores de duas sequências, estas são interpretadas da seguinte forma. Dada uma sequência $\mathbf{x}$ com p valores, na forma já apresentada, esta pode ser interpretada como $p - 1$ funções lineares, que interligam os seus p valores, de forma que a inclinação de cada uma dessas funções pode ser definida de acordo com a Equação (4.14), na qual t_i e t_{i+1} representam os valores de tempo para os quais os valores x_i e x_{i+1} foram mensurados.

$$\text{inclinação}(\mathbf{x}, i) = \frac{x_{i+1} - x_i}{t_{i+1} - t_i} \quad (4.14)$$

A partir da informação de inclinação para um par de valores de uma sequência, três casos distintos de inclinação podem ser identificados: inclinação positiva, inclinação negativa ou ausência de inclinação, como expressa a Equação (4.15).

$$L(\mathbf{x}, i) = \begin{cases} 1, & \text{se } \text{inclinação}(\mathbf{x}, i) > 0 \\ -1, & \text{se } \text{inclinação}(\mathbf{x}, i) < 0 \\ 0, & \text{se } \text{inclinação}(\mathbf{x}, i) = 0 \end{cases} \quad (4.15)$$

Com a definição da Equação (4.15), um índice de concordância entre duas sequências é então sugerido, a fim de mensurar se as sequências possuem comportamento similar em tendência nos seus $p - 1$ intervalos considerados. Tal índice é descrito pela Equação (4.16), que indica a proporção de intervalos em que há concordância de inclinações entre as sequências. Nesta equação, a função I retorna 1 no caso de igualdade e 0 nos demais.

$$A(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{p-1} \frac{I(L(\mathbf{a}, i) = L(\mathbf{b}, i))}{p - 1} \quad (4.16)$$

Outra informação considerada pelos autores diz respeito aos valores máximos e mínimos das duas sequências sob comparação. Neste sentido, o índice expresso na Equação (4.17) é proposto. Os termos $t_{\mathbf{a}}^{\min}$ e $t_{\mathbf{b}}^{\max}$ indicam valores de tempo para os quais a sequência $\mathbf{a}$ obteve seus valores mínimo e máximo, com a mesma interpretação para a sequência $\mathbf{b}$. Ainda quanto à Equação (4.17), seu valor é 1 caso os valores máximos e mínimos das duas sequências ocorrem no mesmo instante de tempo. Caso só um destes valores coincida (máximo ou mínimo), a função retorna o valor 0.5. Por fim, quando nenhum dos valores coincide o valor obtido é igual a 0.

$$M(\mathbf{a}, \mathbf{b}) = \begin{cases} 1 & \text{se } t_{\mathbf{a}}^{\min} = t_{\mathbf{b}}^{\min} \text{ e } t_{\mathbf{a}}^{\max} = t_{\mathbf{b}}^{\max} \\ 0.5 & \text{se } t_{\mathbf{a}}^{\min} = t_{\mathbf{b}}^{\min} \text{ ou } t_{\mathbf{a}}^{\max} = t_{\mathbf{b}}^{\max} \\ 0 & \text{se } t_{\mathbf{a}}^{\min} \neq t_{\mathbf{b}}^{\min} \text{ e } t_{\mathbf{a}}^{\max} \neq t_{\mathbf{b}}^{\max} \end{cases} \quad (4.17)$$

Os dois índices propostos pelas Equações (4.16) e (4.17) são então combinados com dois diferentes coeficientes de correlação, dando origem a duas medidas de similaridade, apresentadas nas Equações (4.18) e (4.19). Na medida *YR1*, a correlação de Pearson é utilizada, com a função $R(\mathbf{a}, \mathbf{b}) = (\rho(\mathbf{a}, \mathbf{b}) + 1)/2$. Já na medida *YS1* a correlação de Spearman é utilizada, com $S(\mathbf{a}, \mathbf{b}) = (\hat{\rho}(\mathbf{a}, \mathbf{b}) + 1)/2$.

$$YR1(\mathbf{a}, \mathbf{b}) = \omega_1 R(\mathbf{a}, \mathbf{b}) + \omega_2 A(\mathbf{a}, \mathbf{b}) + \omega_3 M(\mathbf{a}, \mathbf{b}) \quad (4.18)$$

$$YS1(\mathbf{a}, \mathbf{b}) = \omega_1 S(\mathbf{a}, \mathbf{b}) + \omega_2 A(\mathbf{a}, \mathbf{b}) + \omega_3 M(\mathbf{a}, \mathbf{b}) \quad (4.19)$$

Nas duas equações das medidas de similaridade propostas existem ainda três pesos, definidos por ω_1 , ω_2 e ω_3 , os quais devem satisfazer $\sum_{i=1}^3 \omega_i = 1$. Os autores sugerem uma forma de estimar os pesos ideais para cada base de dados utilizada, porém o custo computacional associado é elevado, o que torna tal abordagem impraticável durante a realização dos diversos experimentos aqui realizados. A fim de comparar as medidas, os pesos foram fixados da seguinte forma: $\omega_1 = 0.5$, $\omega_2 = 0.25$ e $\omega_3 = 0.25$, valores utilizados pelos autores em experimentos realizados (Son e Baek, 2008). Por fim, a complexidade computacional das medidas *YR1* e *YS1* são $O(p)$ e $O(p \log p)$, respectivamente. Estas medidas só foram consideradas durante a realização dos experimentos de agrupamento de genes (discutidos na Seção 5.3), cenário para o qual foram especificamente propostas.

4.3 Dissimilaridade *Short Time-Series*

A medida de dissimilaridade *Short Time-Series* (STS), foi proposta por Möller-Levet et al. (2005) para o cenário específico de agrupamento de séries temporais de expressão gênica. Os autores apontam que uma vez que as séries temporais neste tipo de experimentos são curtas, métodos tradicionais de análise de séries temporais não são adequados para tratar o problema. Considerando cada gene como uma sequência numérica com p valores, os autores argumentam que tal sequência pode ser compreendida como uma composição de $p - 1$ funções lineares que interligam os seus p valores. Neste sentido, a dissimilaridade entre duas sequências numéricas pode ser mensurada de acordo com as diferenças entre as inclinações de suas $p - 1$ funções lineares, capturando assim as dissimilaridades entre as tendências para cada intervalo. A fim de que as inclinações reais das funções lineares sejam mensuradas e que a medida considere os diferentes tamanhos de intervalos existentes entre dois pontos, os autores adicionam à medida de dissimilaridade informações a respeito

do tamanho dos intervalos. A medida final proposta é apresentada pela Equação (4.20). Os valores t_i e t_{i+1} representam os valores de tempo para os quais os valores de ambas sequências foram mensurados. Note que o tamanho dos intervalos definidos por $t_{i+1} - t_i$ podem variar, mas possuem o mesmo tamanho para as duas sequências sob comparação.

$$STS(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{p-1} \left(\frac{b_{i+1} - b_i}{t_{i+1} - t_i} - \frac{a_{i+1} - a_i}{t_{i+1} - t_i} \right)^2} \quad (4.20)$$

Dada tal formulação, os autores (Möller-Levet et al., 2005) argumentam que a inclusão da informação do tamanho dos intervalos na medida de dissimilaridade atua como uma ponderação. Quanto maior o tamanho do intervalo, menor o peso das diferenças entre os valores na dissimilaridade final entre as duas sequências sob consideração. Por fim, é fácil observar que a complexidade computacional da medida proposta é $O(p)$ na qual p , novamente, é o tamanho das duas sequências avaliadas. Esta medida só foi considerada durante a realização dos experimentos de agrupamento de genes (discutidos na Seção 5.3), cenário para o qual foi especificamente proposta.

4.4 Considerações Finais

Neste capítulo foram apresentados e discutidos sete coeficientes de correlação quanto às suas principais características e propriedades. Além dos coeficientes de correlação propriamente ditos, foram apresentadas três medidas recentemente propostas para o cenário de análise de séries temporais de dados de expressão gênica, sendo duas destas baseadas em medidas de correlação. A comparação empírica destas medidas em tarefas de análise de expressão gênica constitui o principal foco desta dissertação. Nos próximos capítulos são apresentados experimentos e resultados envolvendo a avaliação das medidas aqui descritas nos cenários de agrupamento, seleção de atributos e classificação.

Experimentos em Agrupamento

Neste capítulo são apresentados e discutidos os experimentos realizados para a avaliação de medidas de similaridade em agrupamento de dados de expressão gênica. Para uma melhor discussão dos resultados obtidos, os mesmos foram divididos em duas grandes partes: agrupamento de amostras (Seção 5.2) e agrupamento de genes (Seção 5.3). Tal separação é motivada por diferenças existentes nas características das bases de dados utilizadas, bem como nos processos de validação que foram empregados em cada uma dessas subtarefas. Além disso, algumas das medidas discutidas no Capítulo 4 são investigadas somente no cenário de agrupamento de genes, em virtude das características especiais das bases de dados deste cenário. Antes da apresentação dos resultados propriamente ditos, os algoritmos de agrupamento para os quais as medidas foram avaliadas serão apresentados.

5.1 Algoritmos Utilizados

Para avaliar as medidas de similaridade na tarefa de agrupamento de dados de expressão gênica quatro algoritmos de agrupamento foram considerados: *k-medoids*, *Single-Linkage*, *Average-Linkage* e *Complete-Linkage*. É importante neste ponto fazer uma consideração a respeito da escolha do algoritmo *k-medoids* em detrimento do mais comumente utilizado algoritmo *k-means*. Tal escolha tem relação com o cálculo de centróide realizado pelo algoritmo *k-means* durante o agrupamento. Para a distância Euclidiana, o centróide de um grupo é obtido pela média aritmética dos objetos que pertencem ao grupo. Porém, até mesmo para outras métricas da norma L_p (da qual faz parte a distância Euclidiana) (Jain e Dubes, 1988) o cálculo do centróide deve ser alterado para que as propriedades de otimização e convergência do algoritmo sejam mantidas (Steinley, 2006). Quando a distância de Manhattan é utilizada, por exemplo, o protótipo do grupo deve

ser substituído pela mediana de seus objetos (Steinley, 2006). Uma vez que as formas corretas de cálculo do centróide para as diversas medidas de correlação analisadas neste trabalho¹ ainda são desconhecidas (até onde se sabe), o algoritmo *k-medoids* foi preferido, em virtude da substituição dos centróides (objetos artificiais) pelos *medoids* (objetos reais da base de dados), o que garante a convergência do algoritmo independente da dissimilaridade utilizada.

Feitas tais considerações, os algoritmos considerados são descritos a seguir. As descrições dos algoritmos são realizadas a fim de que o trabalho permaneça auto contido. O leitor já familiarizado com os algoritmos considerados pode prosseguir diretamente para as Seções 5.2 e 5.3, sem nenhum prejuízo para o entendimento dos resultados e da discussão que será realizada ao longo deste capítulo.

5.1.1 *k-medoids*

O algoritmo de agrupamento particional *k-medoids* (Bishop, 2007) é semelhante ao algoritmo *k-means* (Wu et al., 2008), exceto pela substituição do centróide do grupo (objeto artificial) pelo *medoid* (objeto real mais representativo). Ainda quanto ao algoritmo *k-medoids* empregado neste trabalho, é importante ressaltar que um coeficiente de correlação é adaptado como medida de distância, como será visto adiante.

Feitas tais considerações, o algoritmo *k-medoids*, que recebe como entrada um conjunto de n objetos, pode ser descrito conforme o Algoritmo 1, juntamente com a Equação (5.1), que define como deve ser realizada a escolha de um novo *medoid* η_r para um grupo G_r . O valor k , fornecido ao algoritmo como parâmetro de entrada, define o número de grupos que devem ser formados para os n objetos disponíveis. Ainda quanto a Equação (5.1), $Dist(\mathbf{x}^i, \mathbf{x}^j) = 1 - Correlação(\mathbf{x}^i, \mathbf{x}^j)$, sendo a função correlação um dos coeficientes de correlação definidos no Capítulo 4.

Algoritmo 1: *k-medoids*.

- 1: **Entradas:** $k \geq 2$, base de dados
 - 2: Selecione k objetos distintos como k *medoids* iniciais
 - 3: **repita**
 - 4: **para cada** objeto $\mathbf{x}^i$ da base de dados, com $1 \leq i \leq n$ **faça**
 - 5: Adicione $\mathbf{x}^i$ ao grupo G_r se
 - 6: $Dist(\mathbf{x}^i, \eta_r) \leq Dist(\mathbf{x}^i, \eta_s), \forall s \in \{1, \dots, k\} - \{r\}$
 - 7: **fim para**
 - 8: Redefinir os *medoids* de acordo com a Equação (5.1)
 - 9: **até** *medoids* permaneçam inalterados
-

$$\eta_r = \arg \min_{\mathbf{x}^i \in G_r} \left\{ \sum_{\mathbf{x}^j \in G_r} Dist(\mathbf{x}^i, \mathbf{x}^j) \right\} \quad (5.1)$$

¹Note que para que as propriedades de convergência do algoritmo sejam mantidas, o cálculo do centróide deve ser realizado e modificado de acordo com a medida de proximidade utilizada. No trabalho de Steinley (2006), por exemplo, são apresentados os diferentes centróides para métricas da norma L_p .

5.1.2 Hierárquicos Aglomerativos

Algoritmos de agrupamento hierárquicos produzem como resultado uma hierarquia de partições, obtida a partir de uma matriz de proximidades entre pares de objetos. Métodos aglomerativos iniciam atribuindo cada objeto a um grupo distinto, de forma que o número de grupos inicial é igual ao número de objetos. A cada passo subsequente, os dois grupos mais similares são então unidos para formarem um novo grupo. Tal passo é então repetido até que todos os objetos pertençam a um único grupo, como descreve o Algoritmo 2.

Algoritmo 2: *Algoritmo Hierárquico Aglomerativo Genérico.*

- 1: **Entrada:** Matriz de proximidade entre objetos
 - 2: **repita**
 - 3: Encontre e unifique os dois grupos mais similares
 - 4: Atualize a matriz de proximidade entre grupos
 - 5: **até** apenas um grupo seja obtido
-

Note que no algoritmo apresentado o terceiro passo se refere à unificação dos dois grupos mais similares. A partir da definição de diferentes medidas de similaridade entre pares de grupos, diferentes algoritmos hierárquicos podem ser obtidos, mesmo que sigam o mesmo esquema geral. Neste trabalho, três diferentes medidas de similaridade entre pares de grupos são consideradas, resultando em três diferentes algoritmos hierárquicos:

- **Single-Linkage:** A distância entre dois grupos é definida como a menor distância entre todos os pares de objetos pertencentes aos dois grupos. Para um par específico, cada um dos dois objetos pertence a um grupo distinto.
- **Average-Linkage:** A distância entre dois grupos é definida como a distância média entre todos os pares de objetos pertencentes aos dois grupos. Para um par específico, cada um dos dois objetos pertence a um grupo distinto.
- **Complete-Linkage:** A distância entre dois grupos é definida como a maior distância entre todos os pares de objetos pertencentes aos dois grupos. Para um par específico, cada um dos dois objetos pertence a um grupo distinto.

5.2 Experimentos com Amostras

Nesta seção são apresentados os experimentos realizados no contexto de agrupamento de amostras. É importante ressaltar que, neste cenário, as amostras ou condições experimentais são consideradas objetos enquanto os genes são seus atributos. Na Seção 5.2.1 a metodologia adotada para a realização dos experimentos é discutida. Na Seção 5.2.2 as bases de dados utilizadas durante os experimentos são apresentadas. Finalmente, na Seção 5.2.3 os resultados obtidos são introduzidos e avaliados.

5.2.1 Metodologia

Para a avaliação dos coeficientes de correlação os quatro algoritmos de agrupamento anteriormente descritos foram considerados. Nos quatro algoritmos os coeficientes de correlação foram adaptados como medidas de distância, na forma descrita pela Equação (5.2), na qual $\mathbf{x}^i$ e $\mathbf{x}^j$ se referem a duas amostras distintas.

$$Dist(\mathbf{x}^i, \mathbf{x}^j) = 1 - Correlação(\mathbf{x}^i, \mathbf{x}^j) \quad (5.2)$$

Uma vez que as bases de dados utilizadas para a realização dos experimentos de agrupamento de amostras possuem rotulação disponível, ou seja, uma partição de referência já existe *a priori*, um índice externo de validação pode ser utilizado para comparar as conformidades entre as partições obtidas a partir da execução de cada um dos algoritmos e a partição de referência existente. Segundo Jain e Dubes (1988), dada uma partição $\mathcal{U}$, com c grupos ($c \geq 2$), obtida a partir da aplicação de um algoritmo de agrupamento qualquer, e uma partição de referência $\mathcal{V}$, com r grupos ($r \geq 2$), índices externos de comparação entre partições podem ser expressos com base em quatro diferentes variáveis:

- a: Número de pares de objetos que estão no mesmo grupo em $\mathcal{U}$ e no mesmo grupo em $\mathcal{V}$.
- b: Número de pares de objetos que estão no mesmo grupo em $\mathcal{U}$ e em grupos diferentes em $\mathcal{V}$.
- c: Número de pares de objetos que estão em grupos diferentes em $\mathcal{U}$ e no mesmo grupo em $\mathcal{V}$.
- d: Número de pares de objetos que estão em grupos diferentes em $\mathcal{U}$ e em grupos diferentes em $\mathcal{V}$.

A partir dessas quatro variáveis, índices externos como Jaccard, Rand e Rand Ajustado podem ser definidos (Jain e Dubes, 1988). Em virtude da correção que leva em conta conformidades entre partições encontradas ao acaso (Milligan e Cooper, 1986), além de sua ampla utilização na literatura de dados de expressão gênica, a avaliação dos resultados se deu com a aplicação do índice Rand Ajustado (Hubert e Arabie, 1985; Jain e Dubes, 1988), o qual é definido pela Equação (5.3). Quanto maior o valor encontrado pelo índice, maior é a concordância entre as partições sob comparação, sendo seu valor máximo igual a 1. Segundo Milligan e Cooper (1986), o limite inferior para o índice Rand Ajustado depende do particionamento dos dados e é geralmente um valor negativo.

$$ARI = \frac{a - \frac{(a+c)(a+b)}{(a+b+c+d)}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)+(a+b)}{(a+b+c+d)}} \quad (5.3)$$

Em virtude da necessidade de realizar comparações entre partições sem rotulação pré-definida em parte dos experimentos realizados, um índice relativo de comparação entre partições foi também utilizado. O índice escolhido foi o da Silhueta (Kaufman e Rousseeuw, 1990), em virtude de seus resultados superiores ou comparáveis aos obtidos com outros índices relativos, conforme apresentado em (Vendramin et al., 2009, 2010). Além disso, Bolshakova e Azuaje (2003) e Yang et al. (2006) reportam bons resultados para o índice também no cenário de dados de expressão gênica. O índice da Silhueta é definido pela Equação (5.4) para uma partição de dados com n objetos e k grupos. Para cada um dos n objetos, $u(i)$ representa o valor da distância média entre o objeto $\mathbf{x}^i$ e os demais objetos de seu grupo. O valor de $v(i)$ é obtido da seguinte forma: calcula-se a distância média entre o objeto $\mathbf{x}^i$ e todos os objetos de um único grupo. Tal passo é executado para todos os $k - 1$ grupos (exclui-se do cálculo o grupo ao qual $\mathbf{x}^i$ pertence), de forma que $k - 1$ distâncias médias entre $\mathbf{x}^i$ e cada um dos $k - 1$ grupos sejam obtidas. Ao final deste processo, atribui-se a $v(i)$ o menor valor de distância média encontrado. Ou seja, $v(i)$ representa a distância média entre o objeto $\mathbf{x}^i$ e o grupo mais próximo. É simples verificar que o valor da Silhueta varia no intervalo de -1 a 1 e que quanto maior seu valor, melhor é a partição obtida, de acordo com o critério.

$$S(\mathcal{P}) = \frac{1}{n} \sum_{i=1}^n \frac{v(i) - u(i)}{\max\{u(i), v(i)\}} \quad (5.4)$$

Por fim, para a avaliação estatística dos resultados obtidos, o procedimento proposto por Demšar (2006) que consiste na aplicação dos testes não-paramétricos de Friedman e Nemenyi foi empregado. Tais testes foram utilizados uma vez que permitem a comparação de diversos coeficientes de correlação aplicados a diferentes bases de dados. Sua aplicação se deu da seguinte forma. Como será visto na próxima seção, dois grandes conjuntos de bases de dados foram considerados durante a realização dos experimentos: um conjunto de bases de cDNA e um conjunto de bases Affymetrix. Em cada cenário de experimentos considerado (os três cenários considerados serão descritos nas próximas seções), as medidas foram avaliadas separadamente para cada um dos dois tipos de bases de dados e cada um dos algoritmos considerados. Para exemplificar a aplicação dos testes estatísticos, considere apenas um cenário de aplicação, para o qual existem dois diferentes conjuntos de bases de dados, cDNA e Affymetrix. Para cada conjunto de bases, quatro algoritmos foram aplicados utilizando as sete medidas avaliadas. A fim de comparar as medidas quando utilizadas com um mesmo algoritmo em um mesmo tipo de base de dados, ao final dos experimentos, oito testes estatísticos foram aplicados. Por exemplo, um teste foi aplicado para comparar as diferentes medidas quanto ao seu desempenho em bases de cDNA considerando o algoritmo *Single-Linkage*, outro teste foi aplicado a fim de comparar as medidas em bases Affymetrix também com o algoritmo *Single-Linkage* e assim sucessivamente, cobrindo dessa forma todas combinações possíveis entre tipos de bases de dados e algoritmos de agrupamento considerados.

5.2.2 Bases de dados

Para a realização dos experimentos de agrupamento de amostras o conjunto de 35 bases de dados de *benchmark* introduzido por de Souto et al. (2008) foi utilizado. As bases do conjunto se referem a dados provenientes de estudos sobre diversos tipos de câncer², como, por exemplo, de próstata, pulmão e pele. Destas 35 bases, 21 foram obtidas a partir do uso da tecnologia Affymetrix, enquanto as demais 14 são provenientes de experimentos envolvendo a tecnologia cDNA. A Tabela 5.1 apresenta um sumário com as características das bases de dados utilizadas. É importante ressaltar que as bases de dados desse conjunto já se encontram pré-processadas, pré-processamento este realizado por de Souto et al. (2008). Para bases de dados Affymetrix, valores de expressão abaixo de 10 ou acima de 16000 foram setados para 10 ou 16000, respectivamente, uma vez que valores extremos fora deste intervalo geralmente não são confiáveis. Ainda para o conjunto de bases de dados Affymetrix, para cada base foram selecionados por volta de 10% do número de genes da base original. Tal redução foi realizada a fim de remover genes não informativos ou com baixas diferenças em seus níveis de expressão para diferentes amostras. A obtenção deste subconjunto de genes se deu com a aplicação do procedimento de *fold change* (Faceli et al., 2004), o qual tem por objetivo selecionar genes que possuam uma variação de pelo menos *l*-folds no valor de sua razão em, no mínimo, *c* amostras. Para bases de dados de cDNA, o primeiro passo tomado foi a eliminação de genes com mais de 10% de valores ausentes. Genes com uma menor porcentagem de valores ausentes tiveram seus valores imputados com a média de seus demais valores. Ainda quanto as bases de cDNA, finalmente, o procedimento de *fold change* foi aplicado a fim de manter também 10% do número de genes da base original. É importante ressaltar, que tanto para bases Affymetrix quanto cDNA os valores de *l* e *c* foram ajustados a fim de obter por volta de 10% do número de genes. Tais valores podem ser obtidos a partir do material suplementar fornecido pelos autores em <http://algorithmics.molgen.mpg.de/Supplements/CompCancer/>.

5.2.3 Resultados

Uma vez definidos os algoritmos de agrupamento, as bases de dados e a metodologia de avaliação, os experimentos realizados podem então ser descritos e examinados. Três cenários experimentais foram considerados de forma a avaliar as medidas de correlação. Os dois primeiros são baseados no trabalho de de Souto et al. (2008), no qual são comparados diferentes algoritmos de agrupamento para a tarefa de agrupamento de amostras. Já o terceiro cenário foi criado visando a simulação de uma aplicação real de agrupamento de amostras. Durante a apresentação dos resultados apenas algumas das tabelas produzidas são direta e explicitamente utilizadas. As demais tabelas, não apresentadas durante a discussão, podem ser encontradas no Apêndice A.

²Informações acerca dos tipos de tecidos que originaram as bases de dados, assim como referências aos trabalhos nos quais foram originalmente publicadas podem ser obtidas em de Souto et al. (2008).

Tabela 5.1: Bases de dados utilizadas, adaptada de [de Souto et al. \(2008\)](#).

	Nome	Amostras	Grupos	Distribuição dos Grupos	Genes
cDNA	alizadeh-v1	42	2	21,21	1095
	alizadeh-v2	62	3	42,9,11	2093
	alizadeh-v3	62	4	21,21,9,11	2093
	bittner	38	2	19, 19	2201
	bredel	50	3	31,14,5	1739
	chen	180	2	104,76	85
	garber	66	4	17,40,4,5	4553
	khan	83	4	29,11,18,25	1069
	lapointe-v1	69	3	11,39,19	1625
	lapointe-v2	110	4	11,39,19,41	2496
	liang	37	3	28,6,3	1411
	risinger	42	4	13,3,19,7	1771
	tomlins-v1	104	5	27,20,32,13,12	2315
	tomlins-v2	92	4	27,20,32,13	1288
Affymetrix	armstrong-v1	72	2	24,48	1081
	armstrong-v2	72	3	24,20,28	2194
	bhattacharjee	203	5	139,17,6,21,20	1543
	chowdary	104	2	62,42	182
	dyrskjot	40	3	9,20,11	1203
	golub-v1	72	2	47,25	1877
	golub-v2	72	3	38,9,25	1877
	gordon	181	2	31,150	1626
	laiho	37	2	8,29	2202
	nuttt-v1	50	4	14,7,14,15	1377
	nuttt-v2	28	2	14,14	1070
	nuttt-v3	22	2	7,15	1152
	pomeroy-v1	34	2	25,9	857
	pomeroy-v2	42	5	10,10,10,4,8	1379
	ramaswamy	190	14	11,10,11,11,22,10,11,10,30,11,11,11,11,20	1363
	shipp	77	2	58,19	798
	singh	102	2	58,19	339
	su	174	10	26,8,26,23,12,11,7,27,6,28	1571
	west	49	2	25,24	1198
	yeoh-v1	248	2	43,205	2526
	yeoh-v2	248	6	15,27,64,20,79,43	2526
	Nome	Amostras	Grupos	Distribuição dos Grupos	Genes

Número de Grupos Fixo

No primeiro cenário de aplicação considerado a informação disponível do número de grupos foi utilizada pelos algoritmos de agrupamento, no sentido de que somente partições com o número de grupos da partição de referência foram consideradas durante a avaliação. Seguindo o mesmo procedimento experimental tomado por [de Souto et al. \(2008\)](#), para cada uma das 35 bases de dados consideradas foram geradas partições contendo apenas o número de grupos definido na partição de referência fornecida. O intuito deste cenário é avaliar a capacidade da combinação algoritmo de agrupamento e medida de correlação em recuperar partições em conformidade com a estrutura definida *a priori* para as bases dados. Para cada um dos pares algoritmo hierárquico e coeficiente de correlação considerados, somente uma execução em cada base de dados foi realizada, uma vez que os algoritmos são determinísticos, ou seja, para uma entrada fixa produzem sempre os mesmos resultados. Ao final das execuções dos algoritmos, em cada base de dados, para cada

par algoritmo hierárquico e coeficiente de correlação uma única partição com o número de grupos da partição de referência foi obtida com um corte no dendrograma gerado. Já para o algoritmo *k-medoids*, 30 execuções (inicializações aleatórias) foram realizadas para cada coeficiente de correlação em cada uma das bases de dados consideradas. Em cada uma das bases de dados, ao final das 30 inicializações, para cada um dos coeficientes de correlação considerados, somente a partição com melhor Rand Ajustado foi selecionada para posterior comparação. É importante destacar, que com esta abordagem, o algoritmo *k-medoids* possui um viés que não é observado nos algoritmos de agrupamento hierárquicos. Isto se deve ao fato de que os algoritmos hierárquicos são determinísticos e executados somente uma vez para cada coeficiente de correlação considerado. Desta forma, espera-se que o algoritmo *k-medoids* produza valores de Rand Ajustado superiores aos observados com os algoritmos hierárquicos em virtude de suas múltiplas execuções.

Na Figura 5.1 são mostrados os resultados médios de Rand Ajustado obtidos para cada par algoritmo e coeficiente de correlação nas 14 bases de cDNA. De forma geral, os métodos hierárquicos apresentaram os piores resultados independentemente da medida de correlação utilizada. Considerando o algoritmo *Single-Linkage*, medidas que consideram apenas as magnitudes dos valores (Pearson e Jackknife) obtiveram os piores resultados em média. Uma possível explicação para tal fato pode ser a grande sensibilidade do algoritmo *Single-Linkage* à presença de *outliers* e valores ruidosos, sensibilidade esta que pode vir a ser atenuada com o uso de medidas baseadas em *ranks*. Considerando o algoritmo *Average-Linkage*, a correlação de Pearson apresentou desempenho inferior às demais. Para este algoritmo a correlação de Goodman-Kruskal Ponderado obteve os piores resultados em média, enquanto as correlações Jackknife e Rank-Magnitude apresentaram os melhores resultados. Considerando o algoritmo *Complete-Linkage*, as correlações de Pearson e Jackknife produziram os melhores resultados. Levando em conta o algoritmo *k-medoids*, para o qual os melhores resultados médios foram obtidos, três medidas se destacaram: Jackknife, Pearson e Rank-Magnitude.

Figura 5.1: Valores médios de Rand Ajustado para as bases de cDNA, com valor k fixo.

Os resultados médios de Rand Ajustado obtidos para as bases Affymetrix são ap-

resentados pela Figura 5.2. Novamente, os algoritmos hierárquicos obtiveram os piores resultados em média. Considerando o algoritmo *Single-Linkage*, as diferenças entre correlações já observadas nas bases de cDNA ficaram mais evidentes, com as correlações de Pearson e Jackknife produzindo resultados médios de Rand Ajustado muito próximos a zero. Considerando o algoritmo *Average-Linkage*, praticamente todas as medidas produziram resultados médios de Rand Ajustado similares, salvo pequenas diferenças. Para o algoritmo *Complete-Linkage*, a correlação Rank-Magnitude apresentou os melhores resultados médios de Rand Ajustado, seguida pelas medidas de Jackknife e Pearson. Estas mesmas três medidas se destacam quando o algoritmo *k-medoids* é considerado. Para este algoritmo as três medidas produziram valores próximos em média, porém consideravelmente superiores aos obtidos com as demais medidas de correlação avaliadas. É interessante notar ainda que a correlação de Kendall, baseada em *ranks* e não tão difundida na literatura de análise de expressão gênica, obteve resultados médios superiores à correlação de Spearman, que por sua vez, também é baseada em *ranks* e mais comumente utilizada.

Figura 5.2: Valores médios de Rand Ajustado para as bases Affymetrix, com valor k fixo.

É importante ressaltar porém que a apresentação dos valores médios apenas indica o comportamento geral das medidas para cada um dos algoritmos. Considerando os valores de Rand Ajustado obtidos para cada uma das medidas, em cada uma das bases de dados consideradas, alguns casos interessantes podem ser observados. A fim de exemplificar alguns cenários interessantes encontrados, são apresentados a seguir, na Tabela 5.2, os resultados de Rand Ajustado obtidos para cada uma das bases de dados consideradas com a aplicação das medidas em conjunto com o algoritmo *k-medoids*. Para facilitar a visualização dos resultados, as células da tabela foram coloridas de acordo com o desempenho das medidas. Desta forma, quanto mais forte o tom de cinza do fundo da célula, melhor o desempenho da medida em questão na base de dados considerada.

Nas bases de dados *singh-2002* e *pomeroy-2002-v1*, por exemplo, a correlação de Pearson (ρ), que obteve bons resultados em média, falhou em identificar as estruturas de grupos contidas nas bases de dados. Nestas duas bases, a utilização de medidas como

Kendall (τ) e Goodman-Kruskal Ponderado ($\hat{\gamma}$) produziram resultados de Rand Ajustado superiores. É ainda interessante notar, que na base de dados *pomeroy-2002-v1*, a correlação Jackknife (ϱ) apresentou uma melhora significativa em relação à correlação de Pearson (ρ). Considerando a base de dados *gordon-2002*, para a qual as correlações de Pearson (ρ) e Jackknife (ϱ) produziram bons valores de Rand Ajustado, a utilização da correlação de Rank-Magnitude (r) proporcionou melhoras aos resultados obtidos. Tais exemplos são aqui ressaltados a fim de demonstrar que, em alguns casos específicos, as melhores medidas em média podem não se comportar de maneira satisfatória. Outro caso que merece destaque é o da base de dados *dyrskjot-2003*, para a qual a correlação de Goodman-Kruskal Ponderado ($\hat{\gamma}$), que apresentou os piores resultados em média, proporcionou o melhor valor de Rand Ajustado (0.07 melhor que a segunda colocada).

Tabela 5.2: Valores de Rand Ajustado para o algoritmo *k-medoids* com *k* fixo.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.50	0.43	0.43	0.37	0.37	0.43	0.37
	alizadeh-2000-v2	0.79	0.79	0.80	0.80	0.80	0.84	0.80
	alizadeh-2000-v3	0.56	0.62	0.62	0.62	0.51	0.62	0.53
	bittner-2000	0.53	0.45	0.38	0.38	0.38	0.45	0.08
	bredel-2005	0.45	0.45	0.46	0.42	0.42	0.42	0.44
	chen-2002	0.60	0.60	0.62	0.62	0.62	0.64	0.69
	garber-2001	0.33	0.38	0.33	0.33	0.36	0.33	0.35
	khan-2001	0.77	0.89	0.40	0.33	0.29	0.65	0.40
	lapointe-2004-v1	0.24	0.16	0.22	0.21	0.22	0.22	0.19
	lapointe-2004-v2	0.29	0.31	0.30	0.30	0.30	0.32	0.34
	liang-2005	0.20	0.16	0.16	0.16	0.16	0.16	0.16
	risinger-2003	0.25	0.25	0.20	0.20	0.47	0.35	0.25
	tomlins-2006	0.37	0.44	0.38	0.34	0.37	0.35	0.34
	tomlins-2006-v2	0.37	0.43	0.32	0.27	0.30	0.33	0.30
Affymetrix	armstrong-2002-v1	0.64	0.30	0.52	0.33	0.52	0.56	0.60
	armstrong-2002-v2	0.81	0.81	0.89	0.89	0.88	0.88	0.70
	bhattacharjee-2001	0.56	0.54	0.40	0.58	0.39	0.57	0.37
	chowdary-2006	0.85	0.81	0.78	0.78	0.78	0.78	0.78
	dyrskjot-2003	0.53	0.62	0.51	0.59	0.52	0.40	0.69
	golub-1999-v1	0.89	0.69	0.30	0.48	0.30	0.69	0.48
	golub-1999-v2	0.89	0.81	0.61	0.55	0.63	0.87	0.48
	gordon-2002	0.82	0.82	0.83	0.85	0.85	0.92	0.83
	laiho-2007	0.38	0.32	0.36	0.51	0.42	0.24	0.24
	nut-2003-v1	0.20	0.19	0.22	0.19	0.17	0.17	0.14
	nut-2003-v2	-0.01	0.00	0.05	0.05	0.05	0.05	-0.02
	nut-2003-v3	0.06	0.03	0.09	0.27	0.09	0.17	0.27
	pomeroy-2002-v1	0.05	0.32	0.33	0.33	0.33	0.39	0.40
	pomeroy-2002-v2	0.62	0.58	0.58	0.56	0.56	0.52	0.59
	ramaswamy-2001	0.56	0.68	0.61	0.66	0.63	0.66	0.01
	shipp-2002-v1	0.39	0.58	0.24	0.21	0.21	0.39	0.02
	singh-2002	0.07	0.06	0.06	0.55	0.06	0.16	0.55
	su-2001	0.60	0.57	0.64	0.60	0.63	0.72	0.66
	west-2001	0.50	0.44	0.44	0.29	0.39	0.56	0.50
	yeoh-2002-v1	0.53	0.59	0.39	0.19	0.22	0.44	0.14
	yeoh-2002-v2	0.46	0.45	0.11	0.11	0.12	0.36	0.03
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Por fim, os testes estatísticos de Friedman e Nemenyi foram aplicados a fim de detectar se uma ou mais medidas de correlação apresentaram diferenças estatisticamente signifi-

cantes entre os valores de Rand Ajustado obtidos. Para a aplicação dos testes, as medidas foram comparadas somente quando da aplicação em um mesmo algoritmo, ou seja, os resultados obtidos para cada uma das medidas não foram comparados entre algoritmos. Considerando valores de $\alpha = 10\%$, não houve diferenças estatisticamente significativas entre as medidas de correlação para nenhum dos algoritmos de agrupamento considerados, tanto em bases de cDNA quanto em bases Affymetrix.

Número de Grupos Variável

Neste segundo cenário de aplicação, em cada uma das bases de dados consideradas, para cada par algoritmo e correlação, foi escolhida a partição com melhor valor de Rand Ajustado considerando números de grupos no intervalo $[2, \lceil \sqrt{n} \rceil]$, faixa comumente utilizada na literatura de agrupamento de dados (Pal e Bezdek, 1995; Pakhira et al., 2005). Para o algoritmo *k-medoids* 30 inicializações aleatórias (execuções) foram realizadas para cada valor de k considerado, porém, ao final de todas as execuções, somente uma partição (aquela com melhor valor de Rand Ajustado) foi selecionada para comparação para cada medida em cada uma das bases de dados. Para os algoritmos hierárquicos, partições com números de grupos entre $[2, \lceil \sqrt{n} \rceil]$ foram geradas por meio de cortes nos dendrogramas obtidos. Novamente, somente uma partição foi selecionada para posterior comparação para cada par algoritmo e correlação em cada uma das bases de dados. O intuito deste cenário é o de avaliar as medidas quanto aos melhores valores de Rand Ajustado gerados, de forma independente do número de grupos. Tal cenário é considerado, uma vez que partições com um número de grupos diferente da original podem, em certos casos, apresentar grupos mais naturais que partições com o número de grupos correto, definido pela partição de referência (Vendramin et al., 2010). Durante a apresentação dos resultados, as medidas serão comparadas somente quanto aos valores de Rand Ajustado obtidos. Ao leitor interessado, tabelas contendo os números de grupos das melhores partições, para cada par algoritmo e correlação em cada uma das bases de dados consideradas podem ser encontradas no Apêndice A. Novamente, é importante destacar, que com esta abordagem, o algoritmo *k-medoids* possui um viés que não é observado nos algoritmos de agrupamento hierárquicos. Isto se deve ao fato de que os algoritmos hierárquicos são determinísticos e executados somente uma vez para cada coeficiente de correlação considerado. Desta forma, espera-se que o algoritmo *k-medoids* produza valores de Rand Ajustado superiores aos observados com os algoritmos hierárquicos em virtude de suas múltiplas execuções.

Feitas tais considerações, na Figura 5.3 são apresentados os resultados médios de Rand Ajustado obtidos para as 14 bases de dados provenientes de experimentos de cDNA. Em comparação com os resultados anteriormente apresentados (para valores de k fixo), no geral, houve uma melhora nos valores médios de Rand Ajustado obtidos para todos os pares algoritmo-correlação considerados. Mais uma vez, independentemente da medida utilizada, o algoritmo *Single-Linkage* apresentou os piores resultados em média. Para este algoritmo, todas as medidas de correlação apresentaram resultados médios de Rand

Ajustado similares. Considerando o algoritmo *Average-Linkage*, os piores resultados, em média, foram obtidos com a correlação de Goodman-Kruskal Ponderado, enquanto que os melhores resultados foram obtidos com a correlação Rank-Magnitude, seguida por Jackknife e Pearson. Para o algoritmo *Complete-Linkage*, as correlações Jackknife e Pearson apresentaram os melhores resultados em média. Novamente, o algoritmo particional *k-medoids* apresentou os melhores resultados médios, independentemente da medida de correlação utilizada. É importante notar, porém, que a escolha da medida de correlação neste algoritmo proporcionou diferenças consideráveis nos resultados. Observando a Figura 5.3 é perceptível que três coeficientes de correlação se destacaram em relação aos demais: Jackknife, Rank-Magnitude e Pearson. Estas medidas proporcionaram resultados médios melhores que correlações baseadas somente em *ranks* (Goodman-Kruskal, Kendall e Spearman) e que a correlação de Goodman-Kruskal Ponderado, para a qual foram obtidos os piores resultados médios de Rand Ajustado.

Figura 5.3: Valores médios de Rand Ajustado para as bases cDNA, com valor k variável.

Na Figura 5.4 podem ser observados os resultados obtidos para as bases provenientes de experimentos da tecnologia Affymetrix. Ao comparar os valores médios de Rand Ajustado obtidos para número de grupos variável com aqueles obtidos para número de grupos fixo é possível perceber uma melhora quando o número de grupos é variável. Considerando os três algoritmos que produziram os melhores resultados em média (*Average-Linkage*, *Complete-Linkage* e *k-medoids*) é evidente que três coeficientes de correlação produziram resultados superiores em média: Jackknife, Pearson e Rank-Magnitude. É interessante notar que, novamente, a medida Rank-Magnitude apresentou resultados médios competitivos com as correlações Jackknife e Pearson, comumente utilizadas na literatura de dados de expressão gênica. Outro fato que merece atenção é o de que a correlação de Spearman, geralmente utilizada na literatura de dados de expressão gênica, apresentou resultados similares ou inferiores em média às outras duas correlações baseadas em *ranks* menos utilizadas ou até mesmo desconsideradas na literatura de expressão gênica: as correlações de Kendall e de Goodman-Kruskal. Por fim, de modo geral, a correlação de Goodman-Kruskal Ponderado apresentou os piores resultados médios de Rand Ajustado, com exceção do algoritmo *Single-Linkage*.

Figura 5.4: Valores médios de Rand Ajustado para as bases Affymetrix, com k variável.

Na Tabela 5.3 são apresentados os resultados detalhados de Rand Ajustado para cada uma das medidas considerando apenas o algoritmo *k-medoids*. A fim de facilitar a compreensão dos resultados, células com tons de cinza mais escuro indicam melhores valores de Rand Ajustado para a base de dados em questão. É interessante notar que em alguns casos específicos em que as correlações de Pearson (ρ) e Jackknife (ϱ) produzem valores baixos de Rand Ajustado é possível obter melhores resultados com a aplicação de correlações baseadas em *ranks*. Tal fato pode ser observado, por exemplo, nas bases: *risinger-2003*, *singh-2002* e *su-2001*. Vale ressaltar ainda que em casos específicos, a escolha de uma medida em particular pode proporcionar melhoras significativas nos resultados obtidos, como ocorre, por exemplo, para as correlações de Pearson (ρ), Jackknife (ϱ) e Rank-Magnitude (r) na base de dados *khan-2001* ou com a correlação Rank-Magnitude (r) para a base de dados *gordon-2002*. Na base *golub-1999-v1* o uso da correlação de Pearson (ρ) proporciona os melhores valores de Rand Ajustado. É interessante notar ainda, que na base de dados *khan-2001*, por exemplo, o uso da correlação Jackknife (ϱ) trouxe melhoras significativas em relação à correlação de Pearson (ρ).

Novamente os testes de Friedman e Nemenyi foram aplicados a fim de determinar a existência ou não de diferenças estatisticamente significativas entre os coeficientes de correlação. Com $\alpha = 10\%$ foi identificada diferença estatística favorável à medida Rank-Magnitude em relação à medida de Goodman-Kruskal Ponderado quando o algoritmo de agrupamento *k-medoids* é utilizado em bases de dados do tipo cDNA. Nos demais casos não foram observadas diferenças estatisticamente significativas.

Número de Grupos Estimado

No último dos três cenários considerados, as informações referentes aos números de grupos presentes nas bases de dados não foram utilizadas durante o agrupamento. Este cenário representa uma aplicação real em que o usuário não possui nenhuma informação *a priori* de quantos ou quais grupos deseja encontrar nos dados que possui em mãos. Neste sentido, durante a realização dos experimentos foram considerados números de

Tabela 5.3: Valores de Rand Ajustado para o algoritmo *k-medoids* com *k* variável.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.50	0.43	0.43	0.37	0.37	0.43	0.37
	alizadeh-2000-v2	0.79	0.79	0.80	0.80	0.80	0.84	0.80
	alizadeh-2000-v3	0.58	0.62	0.62	0.62	0.62	0.62	0.61
	bittner-2000	0.53	0.45	0.38	0.38	0.38	0.45	0.21
	breidel-2005	0.45	0.45	0.46	0.43	0.43	0.43	0.44
	chen-2002	0.60	0.60	0.62	0.62	0.62	0.64	0.69
	garber-2001	0.37	0.38	0.33	0.33	0.36	0.37	0.36
	khan-2001	0.77	0.89	0.45	0.39	0.43	0.77	0.44
	lapointe-2004-v1	0.24	0.28	0.24	0.27	0.25	0.26	0.25
	lapointe-2004-v2	0.31	0.31	0.33	0.38	0.32	0.33	0.34
	liang-2005	0.36	0.36	0.36	0.36	0.36	0.36	0.36
	risinger-2003	0.29	0.35	0.37	0.39	0.47	0.43	0.31
	tomlins-2006	0.41	0.44	0.38	0.38	0.37	0.43	0.37
	tomlins-2006-v2	0.37	0.43	0.32	0.33	0.32	0.37	0.32
Affymetrix	armstrong-2002-v1	0.64	0.57	0.55	0.56	0.58	0.56	0.60
	armstrong-2002-v2	0.82	0.81	0.89	0.89	0.88	0.88	0.71
	bhattacharjee-2001	0.56	0.54	0.42	0.60	0.56	0.62	0.41
	chowdary-2006	0.87	0.82	0.84	0.84	0.83	0.81	0.79
	dyrskjot-2003	0.58	0.64	0.58	0.61	0.61	0.40	0.69
	golub-1999-v1	0.89	0.69	0.48	0.52	0.57	0.69	0.48
	golub-1999-v2	0.89	0.81	0.61	0.60	0.69	0.87	0.48
	gordon-2002	0.84	0.82	0.85	0.87	0.87	0.92	0.83
	laiho-2007	0.38	0.32	0.36	0.51	0.44	0.24	0.33
	nut-2003-v1	0.22	0.22	0.22	0.19	0.18	0.20	0.17
	nut-2003-v2	0.02	0.07	0.08	0.05	0.07	0.05	0.16
	nut-2003-v3	0.15	0.11	0.18	0.27	0.17	0.24	0.27
	pomeroy-2002-v1	0.22	0.32	0.34	0.33	0.34	0.43	0.40
	pomeroy-2002-v2	0.67	0.61	0.58	0.57	0.62	0.58	0.63
	ramaswamy-2001	0.60	0.68	0.61	0.66	0.64	0.68	0.01
	shipp-2002-v1	0.39	0.58	0.24	0.21	0.27	0.39	0.12
	singh-2002	0.28	0.29	0.36	0.55	0.37	0.35	0.55
	su-2001	0.60	0.60	0.70	0.67	0.70	0.78	0.67
	west-2001	0.50	0.47	0.44	0.36	0.39	0.56	0.50
	yeoh-2002-v1	0.54	0.59	0.39	0.20	0.22	0.44	0.19
	yeoh-2002-v2	0.54	0.49	0.15	0.17	0.15	0.39	0.06
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

grupos k variando entre os valores 2 e $\lceil \sqrt{n} \rceil$, faixa comumente utilizada na literatura de agrupamento de dados (Pal e Bezdek, 1995; Pakhira et al., 2005). Com relação ao algoritmo *k-medoids*, para cada número de grupos k este foi inicializado aleatoriamente 30 vezes. Uma vez que a informação do número de grupos não foi utilizada durante os experimentos, o critério da Silhueta (descrito na Seção 5.2.1) foi empregado a fim de escolher a melhor partição dentre todas as geradas no intervalo de interesse considerado, estimando o número de grupos. Ao final dos experimentos, em cada base de dados, para cada par algoritmo-correlação a melhor partição (de acordo com a Silhueta) foi tomada como resultado e utilizada para comparação com a partição de referência, utilizando o critério externo de Rand Ajustado.

Na Figura 5.5 são apresentados os resultados médios de Rand Ajustado obtidos quanto à comparação das melhores partições escolhidas pelo critério da Silhueta em bases cDNA com as partições de referência. De forma similar aos cenários anteriores, o algoritmo

Single-Linkage produziu os piores resultados médios independentemente da medida de correlação utilizada. As diferenças antes notadas entre os algoritmos hierárquicos e o algoritmo particional *k-medoids* são aqui minimizadas, de forma que os melhores resultados são obtidos com o uso das correlações de Pearson e Jackknife juntamente com o algoritmo *Average-Linkage*. De forma geral, estas duas medidas se destacam, seguidas pela correlação Rank-Magnitude que obteve resultados inferiores às duas primeiras na média. Esta última destaca-se das demais no presente cenário somente quando aplicada em conjunto com o algoritmo *k-medoids*. Outro fato interessante aqui observado é o de que a correlação Goodman-Kruskal Ponderado produziu melhores resultados (se comparada às demais correlações) que nos demais cenários anteriormente considerados, para os quais havia obtido geralmente os piores resultados.

Figura 5.5: Valores médios de Rand Ajustado para bases cDNA. As partições comparadas com o critério externo e seu número de grupos foram estimados com a Silhueta.

Os resultados médios de Rand Ajustado obtidos para bases Affymetrix são apresentados na Figura 5.6. Como esperado, os piores desempenhos foram obtidos com o algoritmo de agrupamento *Single-Linkage* para todos os coeficientes de correlação considerados. Levando em conta os outros três algoritmos, três correlações se destacaram em média: Jackknife, Pearson e Rank-Magnitude. Para os três algoritmos, estas correlações apresentaram resultados similares em média e visivelmente superiores aos das demais medidas comparadas. Ainda considerando os três melhores algoritmos (*Average-Linkage*, *Complete-Linkage* e *k-medoids*), a correlação de Goodman-Kruskal Ponderado apresentou em média os piores resultados de Rand Ajustado. Considerando somente o algoritmo *Complete-Linkage* as correlações baseadas em *ranks* de Goodman-Kruskal e Kendall superaram em média a correlação de Spearman. Estes resultados não se repetem, porém, quando os algoritmos *Average-Linkage* e *k-medoids* são considerados.

Como já discutido anteriormente, os valores médios representam apenas uma tendência observada. A fim de proporcionar uma visão geral dos resultados obtidos no presente cenário, são apresentados na Tabela 5.4 os resultados de Rand Ajustado obtidos com a aplicação do algoritmo *k-medoids* juntamente com as medidas avaliadas tanto em bases de cDNA quanto Affymetrix. Como nas demais tabelas apresentadas, quanto mais escuro o

Figura 5.6: Valores médios de Rand Ajustado para bases Affymetrix. As partições comparadas com o critério externo e seu número de grupos foram estimados com a Silhueta.

tom de cinza de uma determinada célula, melhor o valor de Rand Ajustado da correlação em questão em relação às demais para a base de dados considerada. Para as bases de dados de cDNA, mais especificamente na base de dados *khan-2001* é interessante notar que o uso das correlações de Jackknife (ϱ) e Pearson (ρ) proporcionaram uma melhora significativa nos valores de Rand Ajustado se comparadas às demais medidas. Comportamento similar ocorre para a base de dados *pomeroy-2002-v2*. Outras bases que merecem destaque são *west-2001* e *yeoh-2002-v2*, para as quais as correlações de Pearson (ρ), Jackknife (ϱ) e Rank-Magnitude (r) produziram melhores resultados de Rand Ajustado se comparadas às demais correlações.

A fim de verificar diferenças estatisticamente significativas os testes de Friedman e Nemenyi foram aplicados aos resultados de Rand Ajustado obtidos. Com valores de $\alpha = 10\%$, não houve diferenças significativas entre nenhuma das medidas de correlação avaliadas neste cenário, tanto em bases cDNA quanto Affymetrix.

5.2.4 Compilação dos Resultados

As comparações realizadas em três diferentes cenários permitiram observar o comportamento de sete diferentes medidas de correlação quando aplicadas em conjunto com quatro diferentes algoritmos de agrupamento de dados, tanto em bases cDNA quanto Affymetrix. Uma primeira observação que merece ser feita não diz respeito às medidas comparadas em si, mas sim ao algoritmo *Single-Linkage*. Como já observado no trabalho de [de Souto et al. \(2008\)](#), este algoritmo apresenta os piores desempenhos dentre os algoritmos hierárquicos tradicionalmente utilizados. Este fato foi mais uma vez confirmado com os experimentos aqui realizados e é reforçado, no sentido de que diferentes medidas de correlação antes não consideradas foram aqui avaliadas e não proporcionaram melhora nos resultados produzidos pelo algoritmo em questão. Ainda quanto aos algoritmos considerados durante a avaliação dos coeficientes de correlação, é interessante notar que o algoritmo *k-medoids* apresentou, em média, resultados no mínimo comparáveis aos obti-

Tabela 5.4: Resultados de Rand Ajustado para *k-medoids*. O número de grupos foi estimado com o uso da Silhueta.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.25	0.12	0.20	0.06	0.23	0.28	0.16
	alizadeh-2000-v2	0.71	0.71	0.71	0.71	0.71	0.77	0.71
	alizadeh-2000-v3	0.38	0.38	0.38	0.38	0.38	0.41	0.38
	bittner-2000	-0.03	0.18	-0.02	-0.00	-0.00	-0.02	0.01
	bredel-2005	0.43	0.43	0.43	0.43	0.43	0.43	0.37
	chen-2002	0.59	0.56	0.57	0.57	0.59	0.61	0.55
	garber-2001	0.24	0.24	0.16	0.16	0.16	0.21	0.16
	khan-2001	0.51	0.52	-0.01	-0.01	-0.01	0.16	-0.01
	lapointe-2004-v1	0.04	0.04	0.04	0.04	0.04	0.04	0.04
	lapointe-2004-v2	0.01	0.02	0.02	0.02	0.02	0.02	0.02
	liang-2005	0.16	0.16	0.16	0.16	0.16	0.16	0.16
	risinger-2003	0.10	0.10	0.04	0.04	0.02	0.07	0.04
	tomlins-2006	0.15	0.15	0.18	0.18	0.18	0.16	0.19
	tomlins-2006-v2	0.18	0.18	0.16	0.16	0.16	0.19	0.16
Affymetrix	armstrong-2002-v1	0.18	0.27	0.40	0.33	0.33	0.27	0.33
	armstrong-2002-v2	0.56	0.51	0.58	0.59	0.59	0.50	0.58
	bhattacharjee-2001	0.31	0.31	0.39	0.37	0.38	0.35	0.35
	chowdary-2006	0.85	0.81	0.78	0.78	0.78	0.78	0.78
	dyrskjot-2003	0.18	0.47	0.20	0.20	0.21	0.03	0.40
	golub-1999-v1	0.47	0.54	0.42	0.43	0.57	0.62	0.24
	golub-1999-v2	0.70	0.76	0.43	0.53	0.50	0.87	0.31
	gordon-2002	0.77	0.82	0.83	0.85	0.85	0.92	0.83
	laiho-2007	-0.04	-0.12	-0.11	-0.07	-0.11	-0.04	-0.03
	nutt-2003-v1	-0.00	-0.00	0.06	0.06	0.06	0.07	0.13
	nutt-2003-v2	-0.03	-0.03	-0.02	-0.02	-0.02	0.00	-0.02
	nutt-2003-v3	0.01	-0.05	-0.06	-0.06	-0.06	-0.08	-0.06
	pomeroy-2002-v1	0.08	0.04	0.11	0.11	0.11	0.05	0.05
	pomeroy-2002-v2	0.53	0.51	0.17	0.14	0.15	0.21	0.01
	ramaswamy-2001	0.56	0.68	0.18	0.24	0.24	0.18	-0.00
	shipp-2002-v1	0.08	0.58	0.19	0.21	0.21	0.39	0.00
	singh-2002	0.05	0.05	0.05	0.05	0.05	0.04	0.04
	su-2001	0.49	0.51	0.67	0.60	0.61	0.68	0.60
	west-2001	0.50	0.44	-0.01	-0.01	-0.01	0.56	-0.01
	yeoh-2002-v1	0.12	0.07	-0.01	-0.00	-0.00	0.13	-0.01
	yeoh-2002-v2	0.40	0.44	0.00	0.04	0.04	0.27	0.00
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

dos com o uso dos algoritmos *Average-Linkage* e *Complete-Linkage*, independentemente do coeficiente de correlação considerado nestes algoritmos.

Considerando os resultados gerais obtidos, mesmo não havendo diferenças estatisticamente significativas entre as medidas comparadas na maioria dos casos considerados, algumas tendências gerais podem ser observadas. A primeira delas é quanto ao desempenho das correlações de Pearson e Jackknife, que em média e em diferentes cenários sempre produziram resultados comparáveis ou superiores aos obtidos com a aplicação de outros coeficientes de correlação. Outro ponto que merece ser considerado diz respeito aos desempenhos médios obtidos pela correlação Rank-Magnitude. De forma geral, a medida produziu resultados competitivos aos obtidos com a utilização das correlações de Pearson e Jackknife, comumente empregadas na literatura de expressão gênica. Dessa forma, os presentes experimentos podem indicar que tal medida seja adequada às características dos

dados de expressão gênica e deva vir a ser considerada como uma nova alternativa às medidas de similaridade já empregadas no contexto de agrupamento de amostras. Quanto às medidas baseadas somente em *ranks* (Spearman, Kendall e Goodman-Kruskal), no geral e em média, estas apresentaram resultados inferiores às correlações de Pearson, Jackknife e Rank-Magnitude. Ainda quanto às correlações baseadas somente em *ranks* é interessante notar que as correlações de Kendall e Goodman-Kruskal produziram resultados comparáveis e em alguns casos melhores aos obtidos com a correlação de Spearman, mais comumente utilizada na literatura de dados de expressão gênica.

5.3 Experimentos com Genes

Nesta seção são apresentados os experimentos realizados no contexto de agrupamento de genes. Nestes experimentos, os genes são os objetos do problema enquanto as amostras ou condições experimentais são seus atributos. É importante destacar que cada amostra de uma base de dados se refere a um instante distinto de tempo e que, dessa forma, cada gene pode ser visto como uma série temporal, geralmente de curta duração (Ernst et al., 2005). Além dos sete coeficientes de correlação já comparados durante os experimentos envolvendo amostras, outras três medidas são também consideradas³. Tais medidas são avaliadas somente nestes experimentos, uma vez que foram propostas especificamente para o agrupamento de séries temporais de genes. Feitas tais considerações, a presente seção encontra-se organizada da seguinte forma. Na Seção 5.3.1 é discutida a metodologia de avaliação empregada para a realização dos experimentos do presente cenário. Na Seção 5.3.2 as bases de dados utilizadas durante os experimentos são apresentadas. Finalmente, na Seção 5.3.3, os resultados obtidos são introduzidos, avaliados e discutidos.

5.3.1 Metodologia

Para a avaliação das medidas consideradas na tarefa de agrupamento de genes, os algoritmos de agrupamento *k-medoids*, *Single-Linkage*, *Average-Linkage* e *Complete-Linkage*, descritos na Seção 5.1, foram empregados. De forma similar aos experimentos de agrupamento de amostras, os coeficientes de correlação foram adaptados como medidas de distância, na forma descrita pela Equação (5.5). Quanto às demais medidas, estas também foram adaptadas como medidas de dissimilaridade, quando necessário.

$$Dist(\mathbf{x}^i, \mathbf{x}^j) = 1 - Correlação(\mathbf{x}^i, \mathbf{x}^j) \quad (5.5)$$

O cenário de agrupamento de genes possui características peculiares, as quais levaram a adoção de uma metodologia diferenciada de avaliação. A primeira destas características é a de que na literatura de dados de expressão gênica a quantidade de bases de dados rotuladas para a tarefa de agrupamento de genes é relativamente pequena. A segunda diz

³Todas as dez medidas avaliadas encontram-se descritas no Capítulo 4.

respeito às dimensões das bases rotuladas encontradas. Nestas bases, o número de objetos (genes) é consideravelmente reduzido se comparado ao de bases de dados para as quais não há nenhuma rotulação disponível. Dessa forma, ainda que permitam a avaliação com um critério externo, tal qual o de Rand Ajustado, tais bases de dados não representam, de fato, cenários reais de aplicação, nos quais milhares de genes são encontrados. Em virtude de tais características, optou-se por utilizar bases de dados sem prévia rotulação. Nestas bases de dados não há uma partição de referência conhecida, ou seja, um índice de validação externo não pode ser empregado a fim de avaliar e comparar cada uma das partições obtidas.

Uma vez que o número de grupos das bases de dados não é conhecido *a priori*, o critério da Silhueta (Kaufman e Rousseeuw, 1990), definido na Seção 5.2.1, foi empregado para estimar o número de grupos para cada par algoritmo de agrupamento e medida de dissimilaridade em uma base de dados específica. O índice da Silhueta foi escolhido em virtude dos resultados superiores ou comparáveis aos obtidos com outros índices relativos, conforme apresentado em (Vendramin et al., 2009, 2010). Além disso, os trabalhos de Bolshakova e Azuaje (2003) e Yang et al. (2006) reportam bons resultados para o índice da Silhueta também no cenário de agrupamento de dados de expressão gênica.

É importante ressaltar que a Silhueta não foi utilizada com o intuito de comparar as medidas de proximidade uma vez que os valores obtidos com o uso de diferentes medidas não são comensuráveis. Dessa maneira, a fim de comparar os resultados produzidos com cada uma das diferentes medidas avaliadas, uma heurística de avaliação similar à utilizada por Ernst et al. (2005) e Costa et al. (2008) foi empregada. Nesta heurística de avaliação, uma partição obtida com a aplicação de um método em particular é submetida a uma análise de enriquecimento de genes (*gene enrichment analysis*) (Beißbarth e Speed, 2004), a qual, de forma simplificada, atribui significado biológico aos grupos de genes da partição. Dessa forma, duas ou mais partições podem ser comparadas de forma indireta, com base em seus níveis de enriquecimento, obtidos a partir das análises individuais realizadas. Para tornar mais claro o procedimento de validação empregado, este será dividido e brevemente apresentado em duas partes: (i) a avaliação individual de uma única partição, ou análise de enriquecimento de genes, e (ii) a comparação entre duas ou mais partições, a qual permite determinar o melhor agrupamento do ponto de vista biológico, segundo a heurística de avaliação.

A primeira parte da avaliação consiste da análise de enriquecimento de cada um dos grupos de genes de uma partição gerada por um único algoritmo. Após a aplicação de um algoritmo de agrupamento com uma das medidas avaliadas, os identificadores dos genes⁴ que compõem o agrupamento resultante são avaliados utilizando a base de dados da *Gene Ontology* (Ashburner et al., 2000) com o auxílio da ferramenta Gostat (Beißbarth e Speed, 2004). A *Gene Ontology* é mantida pelo *Gene Ontology Consortium* e define um padrão

⁴Os identificadores dos genes não devem ser confundidos com os rótulos externos de uma base de dados já rotulada. Note que os rótulos externos definem grupos de genes. Já seus identificadores são únicos, ou seja, cada gene possui um identificador associado, o qual permite sua identificação de forma unívoca.

internacional para a anotação de genes, com base em suas funções moleculares, processos biológicos e componentes celulares (Beißbarth e Speed, 2004). Para cada um dos genes de um grupo da referida partição, a ferramenta GOstat determina os termos da *Gene Ontology* associados. Feito o enriquecimento para cada um dos genes do grupo, a ferramenta GOstat determina os termos da *Gene Ontology* que encontram-se estatisticamente representados para aquele grupo, ou seja, que não foram anotados ao acaso. Esta determinação é feita da seguinte forma. Conta-se o número de vezes que cada um dos termos da *Gene Ontology* foi atribuído ao grupo de genes em questão. Obtida a frequência de cada um dos termos, um teste estatístico (*Fisher's Exact Test* (Beißbarth e Speed, 2004)) é aplicado a fim de verificar as chances de que aquela anotação tenha ocorrido ao acaso, para a frequência observada. Por fim, para cada anotação estatisticamente significativa um *p-value* correspondente é atribuído. De forma resumida, ao final do processo, para cada um dos grupos de genes da partição é obtido um conjunto de termos da *Gene Ontology*. Cada um destes termos, por sua vez, possui um *p-value* associado, o qual indica a probabilidade de que tal anotação tenha ocorrido por mero acaso. É importante ressaltar que nas análises realizadas, somente termos com *p-values* ≤ 0.05 foram considerados para posterior comparação. Mais informações quanto aos testes estatísticos utilizados durante o processo e quanto às correções aplicadas em virtude dos múltiplos testes realizados podem ser obtidas em (Beißbarth e Speed, 2004).

Uma vez explicado o funcionamento do enriquecimento dos grupos de genes obtidos para uma única partição, a segunda parte da avaliação empregada, que consiste da comparação entre enriquecimentos para dois ou mais algoritmos pode ser elucidada. Por simplicidade, considere um único algoritmo de agrupamento aplicado com duas diferentes medidas de dissimilaridade quaisquer, d_1 e d_2 . A partir de sua aplicação, duas partições podem ser obtidas e submetidas ao procedimento de enriquecimento de grupos de genes já descrito, a fim de obter, para o resultado de cada um dos algoritmos, um conjunto de termos da *Gene Ontology* e seus respectivos *p-values* associados. Com tais valores em mãos, a heurística de comparação baseada nos trabalhos de (Ernst et al., 2005) e (Costa et al., 2008) pode ser empregada⁵. Dados os resultados de enriquecimento para as duas medidas d_1 e d_2 , é feita uma contagem do número de vezes em que a medida d_1 proporcionou enriquecimentos com valores de *p-value* menores que os de d_2 , ou seja, o número de vezes em que a medida d_1 proporcionou enriquecimentos mais significativos do que d_2 . O mesmo é feito para a medida d_2 , ou seja, é feita uma contagem do número de vezes que a medida d_2 proporcionou enriquecimentos mais significativos que os de d_1 . Ao final, estas duas informações são combinadas tal qual apresentado na Equação (5.6), na qual a função $\#()$ retorna o número de vezes que a condição em questão foi satisfeita. Note que valores positivos indicam que a medida d_1 obteve um maior número de enriquecimentos

⁵De forma a automatizar os processos de enriquecimento e comparação entre as partições obtidas com as diferentes medidas consideradas, *scripts* na linguagem de programação Python desenvolvidos e gentilmente cedidos pelo Prof. Dr. Ivan Gesteira Costa Filho da UFPE foram utilizados no presente trabalho.

com menores *p-values* associados quando comparada a d_2 . Já valores negativos indicam que a medida d_2 teve um maior número de enriquecimentos com menores *p-values* quando comparada com a medida d_1 . Note ainda, que a avaliação entre d_1 e d_2 ou entre d_2 e d_1 produzirá os mesmos valores, porém com sinal trocado.

$$\text{Comparação}(d_1, d_2) = \log \left(\frac{\#(p\text{-value } d_1 < p\text{-value } d_2)}{\#(p\text{-value } d_1 > p\text{-value } d_2)} \right) \quad (5.6)$$

Mesmo não permitindo uma avaliação e comparação tão precisas como as obtidas com a aplicação do critério de Rand Ajustado, por exemplo, tal heurística é empregada uma vez que possui aceitação na literatura de dados de expressão gênica. Além disso, é importante ressaltar que, em problemas de agrupamento de genes, não há uma metodologia padrão de avaliação ou comparação entre métodos aceita como definitiva na literatura. Neste sentido, alguns métodos têm sido recentemente propostos, tais como os de (Gibbons e Roth, 2002), (Datta e Datta, 2006) e (Costa et al., 2007), por exemplo.

5.3.2 Bases de Dados

Para a realização dos experimentos de agrupamento de genes três bases de dados foram utilizadas. Estas três bases são provenientes de experimentos de cDNA envolvendo o organismo *Saccharomyces cerevisiae*, comumente referido como *yeast* (levedura). Em virtude da dificuldade em encontrar bases de dados Affymetrix em formato adequado para a avaliação, somente bases de cDNA foram consideradas neste cenário. É importante destacar que duas das bases de dados contêm amostras de mais de um experimento, ou seja, diversos experimentos estão combinados em uma única base de dados. Em razão desta característica, as duas das bases de dados foram inicialmente divididas, de forma que cada base de dados resultante contivesse apenas amostras associadas a um único experimento, o que faz mais sentido para a análise realizada. A base de dados proveniente do trabalho de Spellman et al. (1998), referente a experimentos que envolvem o ciclo celular do *Saccharomyces cerevisiae*, foi dividida em quatro bases de dados. A base de dados proveniente do trabalho de Gasch et al. (2000), referente a experimentos que medem o nível de expressão dos genes deste mesmo organismo quando submetido a mudanças em seu ambiente externo, foi dividida em 12 bases de dados. Por fim, para a terceira base de dados, proveniente do trabalho de Chu et al. (1998), que investiga os níveis de expressão do organismo *Saccharomyces cerevisiae* durante a esporulação, nenhuma divisão foi realizada. Um sumário das 17 bases de dados resultantes é apresentado na Tabela 5.5. Note que na tabela são apresentadas duas colunas referentes ao número de genes das bases de dados. Na primeira destas colunas são apresentados os números de genes originais das bases, enquanto na segunda são apresentados os números de genes selecionados após aplicação de pré-processamento. O pré-processamento realizado é discutido a seguir.

Em um primeiro momento, genes que continham 10% ou mais de valores de atributos ausentes foram removidos. Uma vez que nas bases consideradas o número de atributos

Tabela 5.5: Conjunto de bases de dado não rotuladas utilizado nos experimentos.

Nome	Origem	Amostras	Genes Originais	Genes Selecionados
<i>alpha factor</i>	Spellman et al. (1998)	18	6178	1099
<i>cdc 15</i>		24	6178	1086
<i>cdc 28</i>		17	6178	1044
<i>elutriation</i>		14	6178	935
<i>1mM menadione</i>	Gasch et al. (2000)	9	6152	1050
<i>1M sorbitol</i>		7	6152	1030
<i>1.5mM diamide</i>		8	6152	1038
<i>2.5mM DTT</i>		8	6152	991
<i>constant 32nM H2O2</i>		10	6152	976
<i>diauxic shift</i>		7	6152	1016
<i>complete DTT</i>		7	6152	962
<i>heat shock 1</i>		8	6152	988
<i>heat shock 2</i>		7	6152	999
<i>nitrogen depletion</i>		10	6152	1011
<i>YPD 1</i>		12	6152	1011
<i>YPD 2</i>		10	6152	1022
<i>yeast sporulation</i>	Chu et al. (1998)	7	6118	1171

(amostras) é reduzido, a aplicação deste procedimento removeu todos os genes com valores ausentes, o que eliminou a necessidade da aplicação posterior de um método ou técnica de imputação de dados. Em um segundo momento, com o intuito de remover genes não expressos, ou expressos em níveis muito baixos, as bases de dados foram pré-processadas de forma similar à realizada em (de Souto et al., 2008). O procedimento aplicado, comumente chamado de *fold change*, possui como base o cálculo de razão entre as expressões dos genes nas duas amostras biológicas que compõem uma amostra⁶ (Faceli et al., 2004). Seu objetivo é selecionar genes que possuam uma variação de pelo menos l -folds no valor de sua razão em, no mínimo, c amostras. Para as bases de dados consideradas, o valor de c utilizado foi de 1, enquanto o valor de l foi ajustado de forma a selecionar por volta de 1000 genes na base de dados final. Tal número de genes foi escolhido com base em trabalhos realizados neste mesmo cenário, os quais, após a aplicação de um filtro de variabilidade, também selecionaram por volta de 1000 genes (Tamayo et al., 1999; Gasch et al., 2000; Qin, 2006). Por fim, é importante ressaltar que os dados foram padronizados antes da aplicação da dissimilaridade *Short Time-Series* (STS), de forma que cada gene tivesse média igual a 0 e desvio padrão igual a 1, conforme sugerem seus autores (Möller-Levet et al., 2005).

5.3.3 Resultados

Uma vez definidos os algoritmos de agrupamento, a metodologia de avaliação e as bases de dados utilizadas, os experimentos realizados podem então ser descritos e examinados.

⁶É importante lembrar que, em bases de dados de experimentos de cDNA, cada uma das amostras da matriz de dados final é composta pela razão entre duas amostras biológicas, uma amostra de interesse e uma amostra de controle, conforme visto anteriormente no Capítulo 3.

Tendo em vista que o número de grupos das bases de dados não é conhecido, a Silhueta foi utilizada de forma a estimar o número de grupos e consequentemente a partição mais adequada para cada par algoritmo de agrupamento e medida de dissimilaridade em uma base de dados específica. Neste sentido, durante a realização dos experimentos foram considerados números de grupos k variando entre os valores 2 e $\lceil \sqrt{n} \rceil$, faixa comumente utilizada na literatura de agrupamento de dados (Pal e Bezdek, 1995; Pakhira et al., 2005). Com relação ao algoritmo *k-medoids*, para cada número de grupos k este foi inicializado 50 vezes, sendo ao final das execuções a partição com maior valor de Silhueta eleita como resultado. Para cada um dos algoritmos hierárquicos, somente uma execução foi realizada para cada medida considerada, uma vez que os algoritmos são determinísticos. Após a obtenção do dendrograma para cada par algoritmo hierárquico e medida de dissimilaridade, foram obtidas partições com números de grupos k variando entre os valores 2 e $\lceil \sqrt{n} \rceil$. Novamente, para cada par algoritmo e medida de dissimilaridade, foi eleita como resultado a partição que obteve o maior valor de Silhueta.

Antes da comparação das medidas propriamente ditas é realizada uma comparação entre os algoritmos de agrupamento considerados. Após esta comparação, as medidas de proximidade serão comparadas entre si para cada um dos algoritmos de agrupamento considerados. Tal comparação preliminar é realizada uma vez que a heurística de avaliação tomada fornece apenas valores relativos de comparação entre pares de algoritmos, não um *score* absoluto tal qual o critério de Rand Ajustado. Dessa forma, a identificação de qual algoritmo ou par algoritmo de agrupamento e medida de dissimilaridade produziu os melhores resultados não é obtida de forma tão direta, como no cenário de agrupamento de amostras, por exemplo. A fim de realizar tal comparação, para cada uma das 17 bases de dados foram calculados os números de vitórias (V), empates (E) e derrotas (D) entre todos os pares algoritmos e medidas de dissimilaridade considerando os valores da heurística de avaliação utilizada. Os resultados das 17 bases foram então combinados de forma independente da medida de dissimilaridade empregada, produzindo como resultado a Tabela 5.6. Em cada uma das células desta tabela encontram-se representados 1700 pares comparados, ou seja, 10 medidas para cada um dos dois algoritmos que compõem a célula para cada uma das 17 bases de dados ($10 \times 10 \times 17 = 1700$). A tabela deve ser interpretada da seguinte forma. Dados dois métodos, um em uma linha e um em uma coluna, temos uma célula específica, com os valores $V/E/D$. Tais valores significam que o método da linha foi melhor que o da coluna em V dos pares avaliados, que os métodos empataram em E pares, e por fim, que o método da linha perdeu do método da coluna em D dos pares avaliados. Observando a Tabela 5.6 é possível notar que o algoritmo *Single-Linkage*, independentemente da medida utilizada, levou à obtenção dos piores resultados, segundo a heurística de avaliação utilizada. Este algoritmo obteve resultados de enriquecimento inferiores aos dos demais em 92% dos casos. Considerando o algoritmo *Average-Linkage*, este obteve desempenho levemente superior ao do algoritmo *Complete-Linkage* e levemente inferior ao do algoritmo *k-medoids*. O algoritmo *k-medoids*,

por sua vez, produziu resultados superiores aos do *Complete-Linkage* em 60% dos casos. De modo geral, independentemente da medida utilizada, o algoritmo *k-medoids* produziu os melhores resultados, sendo superior aos demais em 67% dos casos.

Tabela 5.6: Número de vitórias, empates e derrotas entre os algoritmos de agrupamento.

	<i>Single-Linkage</i>	<i>Average-Linkage</i>	<i>Complete-Linkage</i>	<i>k-medoids</i>
<i>Single-Linkage</i>	—	142/21/1537	135/8/1557	112/1/1587
<i>Average-Linkage</i>	1537/21/142	—	955/16/729	831/12/857
<i>Complete-Linkage</i>	1557/8/135	729/16/955	—	669/11/1020
<i>k-medoids</i>	1587/1/112	857/12/831	1020/11/669	—

Com o intuito de obter uma melhor comparação entre os algoritmos considerados, foram selecionadas, para cada um dos algoritmos, as duas medidas que proporcionaram o maior número de vitórias, considerando os resultados de avaliação individuais de cada algoritmo. Estas medidas foram então comparadas entre si em todas as bases de dados, perante os enriquecimentos de seus grupos de genes, tais quais obtidos com a aplicação da metodologia de avaliação anteriormente discutida. Os resultados de tal comparação são apresentados na forma de *heatmap* pela Figura 5.7. É importante ressaltar que o *heatmap* apresenta as médias dos valores de comparação para cada um dos pares nas 17 bases de dados utilizadas. Ainda quanto a esta figura, as cores e intensidades das células correspondem aos valores de comparação entre pares de métodos, obtidos pela Equação 5.6, ou seja, para uma linha e coluna específica, tem-se o valor médio de

$$\log \left(\frac{\#(p\text{-value linha} < p\text{-value coluna})}{\#(p\text{-value linha} > p\text{-value coluna})} \right)$$

para as 17 bases de dados consideradas. Note que células vermelhas indicam que o método da linha obteve um maior número de enriquecimentos com valores inferiores de *p-value* se comparado ao método da coluna. Já valores azuis indicam o contrário, ou seja, o método da linha obteve um menor número de enriquecimentos com valores inferiores de *p-value* se comparado ao método da coluna. Feitas tais considerações, entre as melhores medidas de dissimilaridade para os algoritmos *Complete-Linkage*, *Average-Linkage* e *k-medoids* os resultados de enriquecimento obtidos foram bastante equilibrados. Para estes três algoritmos, três correlações levaram aos melhores resultados de enriquecimento: Pearson (ρ), Jackknife (ϱ) e Rank-Magnitude (r). Ainda quanto aos resultados apresentados na Figura 5.7, é interessante notar que as duas medidas de dissimilaridade que produziram os melhores resultados para o algoritmo *Single-Linkage* produzem resultados de enriquecimento consideravelmente inferiores aos das melhores medidas para os outros três algoritmos. Estes resultados, somados aos já apresentados na Tabela 5.6, motivaram a não comparação entre os resultados das medidas de proximidade para este algoritmo individualmente. Tal decisão é justificável uma vez que mesmo os melhores resultados para este algoritmo são sempre piores caso comparados aos resultados obtidos com os

demaís algoritmos.

Figura 5.7: Comparação entre os enriquecimentos obtidos para as duas medidas com melhores resultados para cada um dos quatro algoritmos de agrupamento considerados.

Uma vez apresentada uma comparação preliminar entre os algoritmos, as correlações e demais medidas podem então ser comparadas considerando cada um dos algoritmos de agrupamento de forma individual. Inicialmente, as medidas são comparadas considerando o algoritmo *Average-Linkage*. Para este algoritmo em particular, são apresentados na Figura 5.8 os resultados médios obtidos com a heurística de avaliação entre pares de medidas nas 17 bases avaliadas. A fim de tornar a discussão mais clara, são apresentados também na Tabela 5.7 os números de vitórias (V), empates (E) e derrotas (D) entre pares de medidas. A interpretação da tabela é a mesma já descrita anteriormente. Considerando este algoritmo, a correlação Jackknife (ϱ) proporcionou os melhores resultados em média e, no geral, foi superior às demais em 67% dos casos, contra apenas 54% da correlação de Pearson (ρ). É interessante notar que a correlação Rank-Magnitude (r) produziu resultados comparáveis aos da correlação Jackknife (ϱ), sendo superior às demais em 64% dos casos. Outro fato que chama a atenção é o de que a correlação de Kendall (τ), baseada em *ranks*, foi superior à correlação de Pearson em 11 das 17 bases de dados avaliadas, fato que não se repete para as demais medidas baseadas em *ranks*, como Spearman ($\hat{\rho}$), por exemplo, comumente utilizada em dados de expressão gênica. Por fim, é de certa forma surpreendente que as três medidas introduzidas especificamente para este cenário tenham levado aos piores resultados de enriquecimento, se comparadas às demais. De fato, a pior dentre as medidas avaliadas foi a dissimilaridade *Short Time-Series* (STS).

Figura 5.8: Comparação entre os enriquecimentos obtidos para o *Average-Linkage*.

Tabela 5.7: Vitórias, empates e derrotas para o algoritmo *Average-Linkage*.

Método	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$	YR1	YS1	STS
ρ	—	9/0/8	9/0/8	6/0/11	10/1/6	6/0/11	8/1/8	8/0/9	11/0/6	17/0/0
ϱ	8/0/9	—	11/0/6	11/0/6	10/0/7	10/0/7	13/0/4	12/0/5	13/0/4	16/0/1
γ	8/0/9	6/0/11	—	7/0/10	10/0/7	7/0/10	7/0/10	9/0/8	8/0/9	17/0/0
τ	11/0/6	6/0/11	10/0/7	—	8/0/9	7/0/10	11/0/6	9/0/8	11/0/6	17/0/0
$\hat{\rho}$	6/1/10	7/0/10	7/0/10	9/0/8	—	7/1/9	10/0/7	9/0/8	12/0/5	17/0/0
r	11/0/6	7/0/10	10/0/7	10/0/7	9/1/7	—	11/0/6	12/0/5	11/0/6	17/0/0
$\hat{\gamma}$	8/1/8	4/0/13	10/0/7	6/0/11	7/0/10	6/0/11	—	8/0/9	9/0/8	17/0/0
YR1	9/0/8	5/0/12	8/0/9	8/0/9	8/0/9	5/0/12	9/0/8	—	9/1/7	16/0/1
YS1	6/0/11	4/0/13	9/0/8	6/0/11	5/0/12	6/0/11	8/0/9	7/1/9	—	17/0/0
STS	0/0/17	1/0/16	0/0/17	0/0/17	0/0/17	0/0/17	0/0/17	1/0/16	0/0/17	—

A seguir, são apresentados os resultados de comparação entre as medidas para o algoritmo de agrupamento *Complete-Linkage*. Na Figura 5.9 são novamente apresentados os resultados médios de avaliação para as 17 bases de dados, enquanto na Tabela 5.8 são apresentados os números de vitórias, empates e derrotas considerando cada uma das medidas avaliadas, também para as 17 bases de dados. Novamente, os piores resultados médios de enriquecimento foram obtidos com a utilização da medida STS. As medidas YR1 e YS1, baseadas na correlação de Pearson e Spearman, respectivamente, também produziram resultados inferiores de enriquecimento, em média, se comparadas às demais. A correlação Rank-Magnitude (r) que, em média, apresentou os melhores resultados, foi superior a todas as demais em 74% dos casos, contra apenas 59% para a correlação de Pearson (ρ). Ainda quanto à correlação Rank-Magnitude (r), esta produziu resultados de enriquecimento superiores aos obtidos com a correlação de Pearson (ρ) em 13 das 17 bases de dados (76% das bases de dados). Para este algoritmo, em particular, o uso da correlação Jackknife (ϱ) não trouxe benefícios em comparação à correlação de Pearson (ρ), se forem tomados para comparação os valores médios de enriquecimento obtidos. De fato, considerando os números de vitórias, empates e derrotas entre as duas medidas, resultados bastante similares podem ser observados em relação às demais.

Figura 5.9: Comparação entre os enriquecimentos obtidos para o *Complete-Linkage*.

Tabela 5.8: Vitórias, empates e derrotas para o algoritmo *Complete-Linkage*.

Método	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$	YR1	YS1	STS
ρ	—	9/0/8	10/1/6	10/0/7	9/0/8	3/1/13	8/3/6	13/0/4	12/1/4	17/0/0
ϱ	8/0/9	—	12/0/5	10/0/7	8/0/9	6/0/11	10/0/7	12/0/5	10/0/7	14/0/3
γ	6/1/10	5/0/12	—	8/0/9	9/0/8	6/0/11	8/0/9	11/0/6	8/0/9	17/0/0
τ	7/0/10	7/0/10	9/0/8	—	6/1/10	5/0/12	9/0/8	11/0/6	10/0/7	17/0/0
$\hat{\rho}$	8/0/9	9/0/8	8/0/9	10/1/6	—	6/0/11	8/0/9	10/0/7	9/0/8	15/0/2
r	13/1/3	11/0/6	11/0/6	12/0/5	11/0/6	—	14/0/3	13/0/4	12/0/5	17/0/0
$\hat{\gamma}$	6/3/8	7/0/10	9/0/8	8/0/9	9/0/8	3/0/14	—	11/0/6	9/0/8	16/0/1
YR1	4/0/13	5/0/12	6/0/11	6/0/11	7/0/10	4/0/13	6/0/11	—	11/0/6	14/0/3
YS1	4/1/12	7/0/10	9/0/8	7/0/10	8/0/9	5/0/12	8/0/9	6/0/11	—	15/0/2
STS	0/0/17	3/0/14	0/0/17	0/0/17	2/0/15	0/0/17	1/0/16	3/0/14	2/0/15	—

Por fim, são apresentados os resultados de comparação entre as correlações e demais medidas para o algoritmo de agrupamento *k-medoids*. Novamente, são apresentados na Figura 5.10 os valores médios de comparação obtidos com a heurística de avaliação utilizada para as 17 bases de dados. São apresentados também na Tabela 5.9 os números de vitórias, empates e derrotas entre cada uma das medidas avaliadas. A interpretação da tabela é a mesma já utilizada anteriormente, ou seja, para uma célula específica, os valores V/E/D indicam que o algoritmo da linha proporcionou enriquecimentos melhores que os da coluna em V bases de dados, enriquecimentos iguais em E bases de dados e enriquecimentos piores em D bases de dados. Feitas tais considerações, para este algoritmo, em particular, os melhores resultados, em média, foram obtidos com a correlação Jackknife (ϱ), a qual foi superior às demais medidas em 73% dos casos avaliados, contra 62% observados para a correlação de Pearson (ρ). Para este algoritmo, a correlação Rank-Magnitude (r) produziu, em média, valores mais baixos de enriquecimento se comparadas às correlações de Pearson (ρ) e Jackknife (ϱ), porém, foi ainda superior à utilização de outras medidas em 59% dos casos. É ainda interessante notar que a medida YR1 produziu, para o algoritmo *k-medoids*, resultados mais competitivos às demais, em média. Como observado para os demais algoritmos avaliados, a dissimilaridade STS apresentou o pior desempenho em média. Considerando apenas as correlações baseadas em *ranks*, Spearman ($\hat{\rho}$), Kendall (τ) e Goodman-Kruskal (γ), resultados competitivos entre si po-

dem ser observados. Por fim, a correlação de Goodman-Kruskal Ponderado ($\hat{\gamma}$) obteve resultados comparáveis e em alguns casos superiores aos das medidas baseadas em *ranks*, produzindo resultados superiores a estas em 56% dos casos.

Figura 5.10: Comparação entre os enriquecimentos obtidos para o *k-medoids*.

Tabela 5.9: Vitórias, empates e derrotas para o algoritmo *k-medoids*.

Método	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$	YR1	YS1	STS
ρ	—	6/0/11	13/0/4	11/1/5	11/0/6	11/0/6	11/0/6	10/0/7	10/1/6	12/0/5
ϱ	11/0/6	—	13/0/4	13/0/4	13/0/4	13/2/2	12/0/5	10/0/7	13/0/4	14/0/3
γ	4/0/13	4/0/13	—	6/5/6	6/1/10	3/0/14	8/0/9	6/1/10	9/2/6	12/0/5
τ	5/1/11	4/0/13	6/5/6	—	7/0/10	4/0/13	7/0/10	6/1/10	10/1/6	12/0/5
$\hat{\rho}$	6/0/11	4/0/13	10/1/6	10/0/7	—	6/0/11	7/0/10	7/0/10	9/1/7	13/0/4
r	6/0/11	2/2/13	14/0/3	13/0/4	11/0/6	—	8/0/9	11/0/6	12/0/5	12/0/5
$\hat{\gamma}$	6/0/11	5/0/12	9/0/8	10/0/7	10/0/7	9/0/8	—	8/0/9	8/0/9	11/0/6
YR1	7/0/10	7/0/10	10/1/6	10/1/6	10/0/7	6/0/11	9/0/8	—	10/1/6	13/0/4
YS1	6/1/10	4/0/13	6/2/9	6/1/10	7/1/9	5/0/12	9/0/8	6/1/10	—	11/0/6
STS	5/0/12	3/0/14	5/0/12	5/0/12	4/0/13	5/0/12	6/0/11	4/0/13	6/0/11	—

5.3.4 Compilação dos Resultados

As comparações realizadas no cenário de agrupamento de genes permitiram observar o comportamento de sete medidas de correlação e três medidas propostas especificamente para tal cenário, quando empregadas com quatro algoritmos de agrupamento comumente utilizados na literatura. A heurística de avaliação empregada permitiu a comparação das medidas em cenários reais de aplicação, nos quais milhares de genes devem ser agrupados. Neste sentido, um primeiro comentário que deve ser realizado diz respeito ao algoritmo *Single-Linkage*, o qual, independentemente da medida utilizada, produziu os piores resultados. Desta forma, os resultados aqui obtidos reforçam e vão de encontro aos resultados já reportados na literatura de dados de expressão gênica (D'haeseleer, 2005).

Quanto às medidas comparadas neste cenário específico, as correlações de Pearson, Jackknife e Rank-Magnitude apresentaram, em grande parte dos resultados obtidos, desempenhos superiores ou comparáveis aos das demais medidas avaliadas. De forma geral,

com base nos experimentos realizados, nota-se que os coeficientes de correlação baseados em *ranks* proporcionaram a obtenção de resultados inferiores aos das correlações de Pearson, Jackknife e Rank-Magnitude. Quanto às medidas baseadas em *ranks* é interessante observar que as correlações de Kendall e Goodman-Kruskal, raramente utilizadas em dados de expressão gênica, podem, em alguns casos, proporcionar resultados superiores aos obtidos com a mais difundida correlação de Spearman. Observa-se também que a correlação de Goodman-Kruskal Ponderado obteve, neste cenário, resultados competitivos aos das correlações baseadas em *ranks*. Por fim, é interessante notar que as medidas YR1, YS1 e STS, propostas especificamente para o agrupamento de genes, produziram, no geral, os piores resultados. É importante ressaltar que as medidas YR1 e YS1 são parametrizadas e foram utilizadas com seus parâmetros pré-fixados (conforme discutido na Seção 4.2.1), porém, estimá-los possui alto custo computacional, o que pode tornar sua aplicação inviável.

5.4 Considerações Finais

Neste capítulo sete coeficientes de correlação foram avaliados para as tarefas de agrupamento de amostras e agrupamento de genes. Além dos coeficientes de correlação, três medidas propostas especificamente para o agrupamento de genes foram também avaliadas nesta tarefa. Em ambas tarefas consideradas, dentre os quatro algoritmos de agrupamento utilizados, os piores resultados foram obtidos com o *Single-Linkage*, independentemente da medida empregada. Quanto às medidas avaliadas, em grande parte dos experimentos conduzidos, as medidas de Pearson, Jackknife e Rank-Magnitude apresentaram resultados competitivos entre si e superiores aos das demais. Considerando as correlações baseadas em *ranks*, sua utilização geralmente proporcionou piores resultados. Ainda quanto a estas medidas, os resultados obtidos sugerem, para ambas as tarefas, que as correlações de Kendall e Goodman-Kruskal podem, em alguns casos, proporcionar resultados superiores aos observados com a correlação de Spearman, comumente utilizada na literatura de dados de expressão gênica. Por fim, é importante ressaltar que, em virtude da quantidade de bases de dados utilizadas, os resultados foram apresentados de forma sumarizada, utilizando as médias obtidas para os conjuntos de bases de dados. Neste sentido, as médias apresentam apenas uma estimativa do comportamento geral das medidas e podem esconder comportamentos peculiares em cenários específicos. Como apontado ao longo da apresentação dos resultados, em alguns casos particulares as melhores correlações em média podem não produzir resultados satisfatórios. Neste sentido, considerar um conjunto de correlações como possíveis alternativas em um cenário de análise exploratória de dados de expressão gênica parece ser a estratégia mais apropriada.

Experimentos em Seleção de Atributos e Classificação

Neste capítulo são apresentados e avaliados os resultados obtidos com a realização de experimentos envolvendo as tarefas de seleção de atributos e classificação. Os experimentos conduzidos se restringem ao escopo de classificação de amostras e seleção de genes, este último também para o problema de classificação de amostras¹. Tal escolha se deve ao fato de que estas duas tarefas são comumente encontradas na literatura de dados de expressão gênica ([Piatetsky-Shapiro e Tamayo, 2003](#)). Por outro lado, experimentos envolvendo classificação de genes, ou seleção de amostras para problemas de classificação de genes, são raramente considerados na prática e não serão abordados neste trabalho. O capítulo encontra-se estruturado da seguinte maneira. Na Seção 6.1 são apresentados em detalhes os algoritmos de seleção de atributos e classificação utilizados para avaliação dos coeficientes de correlação. Na Seção 6.2 a metodologia de avaliação empregada é discutida. Por fim, na Seção 6.3 são discutidos separadamente os experimentos envolvendo seleção de genes para o problema de classificação de amostras e a própria classificação de amostras, sem a prévia realização de seleção de genes.

6.1 Algoritmos Utilizados

Nesta seção são apresentados os dois algoritmos utilizados para a avaliação dos coeficientes de correlação. Primeiramente é apresentado na Seção 6.1.1 o algoritmo de seleção de atributos *Simplified Silhouette Filter* (SSF). O classificador kNN é então brevemente

¹Neste capítulo, o termo amostra está sempre associado a um objeto, ou seja, as amostras da matriz de dados de *microarray* são os objetos do problema em questão, enquanto os genes são seus atributos.

discutido na Seção 6.1.2. O leitor já familiarizado com os algoritmos apresentados pode prosseguir para a Seção 6.2 sem prejuízo para o entendimento da discussão dos resultados.

6.1.1 *Simplified Silhouette Filter*

O algoritmo *Simplified Silhouette Filter* (SSF) (Covões et al., 2009) busca eliminar atributos redundantes com base em seu agrupamento. Dada uma matriz de dados $\mathbf{D}$, com n linhas representando seus objetos e m colunas representando seus respectivos atributos, sua transposta $\mathbf{D}^T$ ($m \times n$) pode ser obtida. A matriz $\mathbf{D}^T$ possui m vetores de atributos $\mathbf{x}^1, \dots, \mathbf{x}^m$, cada qual com número de objetos igual a n . Esses vetores de atributos podem ser expressos sob a forma $\mathbf{x}^i = (x_1^i, x_2^i, \dots, x_n^i)$, com $1 \leq i \leq m$. Trabalhando sobre a matriz $\mathbf{D}^T$, o SSF utiliza o algoritmo de agrupamento particional *k-medoids* (descrito na Seção 5.1.1) e um coeficiente de correlação adaptado como medida de distância, na forma descrita pela Equação (6.1), a fim de encontrar grupos de atributos correlacionados. É importante notar que, durante o processo de seleção de atributos, não há distinção entre correlações positivas e negativas entre atributos, visto que ambas indicam a presença de informações redundantes. Por este motivo, o módulo do valor da correlação é utilizado.

$$Dist(\mathbf{x}^i, \mathbf{x}^j) = 1 - |Correlação(\mathbf{x}^i, \mathbf{x}^j)| \quad (6.1)$$

Uma vez que o número de atributos selecionados pelo SSF está diretamente ligado ao número de grupos obtidos com a aplicação do algoritmo *k-medoids* e este número não é conhecido *a priori*, o SSF utiliza uma estratégia de amostragem para gerar diversas partições para diversos números de grupos. Tal estratégia é empregada tendo em vista que realizar uma busca no espaço de todas as partições possíveis é geralmente impraticável. Além disso, o *k-medoids* pode ficar preso em mínimos locais, dado um número de grupos fixo. Dessa forma, para cada número de grupos k considerado, são geradas n_p partições. Cada uma destas partições é então avaliada com o critério da Silhueta Simplificada (SS) (Hruschka et al., 2006) a fim de escolher a melhor partição dentre as geradas, segundo o critério. O critério da Silhueta Simplificada possui como base a Silhueta proposta por Kaufman e Rousseeuw (1990), descrita na Seção 5.2.1. Uma vantagem do uso da SS é a diminuição da complexidade computacional de $O(m^2)$ na Silhueta original para $O(m)$ na SS². A Silhueta Simplificada é aqui definida em função do cálculo das distâncias entre os atributos e os *medoids* dos grupos. Dado um atributo $\mathbf{x}^i$, a dissimilaridade deste ao *medoid* do grupo ao qual pertence é denotada por $a(i)$. Considerando ainda o atributo $\mathbf{x}^i$, denota-se por $b(i)$ a menor dissimilaridade entre $\mathbf{x}^i$ e os *medoids* dos demais grupos, excluindo-se o grupo ao qual $\mathbf{x}^i$ pertence. Tal valor representa a dissimilaridade entre $\mathbf{x}^i$ e seu grupo vizinho, ou seja, seu grupo mais próximo. Feitas tais considerações, para uma partição de atributos $\mathcal{P}$, a Silhueta Simplificada é definida como a média das Silhuetas individuais de cada um dos m atributos, conforme descreve a Equação (6.2).

²A complexidade está definida em função de m uma vez que os atributos estão sendo agrupados.

$$SS(\mathcal{P}) = \frac{1}{m} \sum_{i=1}^m \frac{b(i) - a(i)}{\max\{b(i), a(i)\}} \quad (6.2)$$

Após a execução da estratégia de amostragem e a obtenção da partição com o melhor valor de SS , resta a escolha de quais atributos devem ser selecionados dentre os grupos da melhor partição obtida. O SSF considera duas alternativas para realizar tal seleção. A primeira delas é a seleção do *medoid* de cada um dos grupos da partição. Esta escolha deve-se ao fato de que o *medoid* é o atributo mais representativo de seu grupo. A segunda forma de seleção proposta para o SSF é a escolha do atributo *medoid* e o atributo de fronteira de cada um dos grupos. O atributo de fronteira é definido como o mais distante do *medoid* de seu grupo, e é, portanto, o atributo menos correlacionado (em módulo) com o *medoid*. Esta forma de seleção é considerada por dois motivos. O primeiro é o de que o atributo de fronteira pode conter informações consideravelmente diferentes do atributo *medoid*, as quais podem ser relevantes. Já o segundo é tentar diminuir o viés do algoritmo em selecionar poucos atributos (Covões, 2010), viés este que foi observado também em experimentos preliminares realizados (Jaskowiak et al., 2010). Por esta razão, somente a variante que seleciona os atributos *medoid* e fronteira de cada um dos grupos foi considerada neste trabalho para avaliação dos coeficientes de correlação.

Feitas tais considerações, o algoritmo SSF pode ser finalmente descrito conforme o Algoritmo 3 (Covões et al., 2009), para o caso de seleção dos atributos *medoid* e fronteira, adotado no presente trabalho. Neste algoritmo, k_{min} e k_{max} são os limites mínimo e máximo de número de grupos que devem ser avaliados e n_p é o número de execuções (inicializações) aleatórias que devem ser realizadas para o *k-medoids*, para cada valor de k , ou seja, para cada número de grupos dentro da faixa definida por k_{min} e k_{max} .

6.1.2 k-Nearest Neighbors

O *k-Nearest Neighbors* (kNN) (Cover e Hart, 1967; Silverman e Jones, 1989; Aha e Kibler, 1991; Tan et al., 2005) é um algoritmo de classificação considerado *lazy*, visto que todo processamento realizado é adiado até o momento em que se torna necessária a classificação de um novo objeto. O processo de treinamento do algoritmo consiste simplesmente em armazenar todos os objetos de treinamento apresentados, dado que um modelo de classificação não é explicitamente derivado dos dados. Uma vez armazenados os objetos de treinamento, o algoritmo já encontra-se pronto para classificar novos objetos ainda não rotulados. O processo de classificação de um novo objeto ocorre também de forma simples. Dado um novo objeto $\mathbf{x}$, para o qual o rótulo de classe ainda não é conhecido, o algoritmo kNN calcula sua distância para todos³ os objetos armazenados durante o processo de treinamento, a fim de encontrar os k objetos mais próximos de $\mathbf{x}$, ou seja, seus k vizinhos mais próximos. Ao final deste processo, o objeto $\mathbf{x}$ em questão

³De fato, nem todas as distâncias precisam ser calculadas. Algoritmos como os propostos por (Orchard, 1991) e (Ye et al., 2009) podem reduzir o número de cálculos durante o processo de classificação.

Algoritmo 3: *Simplified Silhouette Filter.*

```

1: Entrada:  $k_{min} \geq 2$  ,  $k_{max} \leq m - 1$  ,  $n_p \geq 1$ 
2:  $MVSS \leftarrow -\infty$ ; // Melhor Valor de SS
3: para cada  $k \in \{k_{min}, \dots, k_{max}\}$  faça
4:   para cada  $i \in \{1, \dots, n_p\}$  faça
5:     Gerar partição inicial aleatória  $\mathcal{P}_{inicial}$  com  $k$  grupos;
6:     Executar  $k$ -medoids com a correlação escolhida e com  $\mathcal{P}_{inicial}$ ;
7:     Seja  $\mathcal{P}_{obtida}$  a partição obtida ao final do agrupamento;
8:      $VSS \leftarrow SS(\mathcal{P}_{obtida})$ ;
9:     se  $VSS > MVSS$  então
10:       $MVSS \leftarrow VSS$ ;
11:       $k^* \leftarrow k$ ;
12:       $\mathcal{P}^* \leftarrow \mathcal{P}_{obtida}$ ;
13:   fim se
14: fim para
15: fim para
16:  $Selecionados \leftarrow \emptyset$ ;
17: para cada grupo  $G_m \in \mathcal{P}^*$  faça
18:    $Selecionados \leftarrow Selecionados$ 
      $\cup \{\text{atributo medoid de } G_m, \text{ atributo fronteira de } G_m\}$ ;
19: fim para
20: Retorne  $Selecionados$ 

```

é classificado pelo algoritmo como pertencendo à classe majoritária dentre as classes de seus k vizinhos mais próximos. Torna-se evidente, a partir da definição do algoritmo, que o valor k é um parâmetro que deve ser escolhido ou ajustado. É importante destacar que além do parâmetro k , o algoritmo kNN necessita da definição de uma medida de proximidade entre pares de objetos, tornando-se dessa forma interessante para a avaliação realizada neste trabalho.

6.2 Metodologia

Nesta seção é discutida a metodologia adotada para a realização dos experimentos de seleção de atributos e classificação. Para a tarefa de classificação, os coeficientes de correlação foram comparados por meio das taxas de erros obtidas com a aplicação do algoritmo kNN utilizando cada uma das diferentes medidas consideradas. Neste cenário, nenhum algoritmo de seleção de atributos foi empregado e a classificação foi realizada com a utilização de todos os atributos dos conjuntos de dados. Dessa forma, nenhum viés adicional foi introduzido durante a comparação das medidas. É importante lembrar que neste cenário somente o classificador kNN foi considerado uma vez que medidas de proximidade podem ser diretamente acopladas ao mesmo. Para a tarefa de seleção de atributos, os coeficientes de correlação foram comparados a partir das taxas de erros de classificação obtidas utilizando somente os atributos selecionados e do número de atributos selecionados. Para obtenção de taxas de erros, dois classificadores comumente utilizados

na literatura foram empregados: Naïve Bayes (NB) (Tan et al., 2005) e kNN, esse último utilizando distância Euclidiana. Além de amplamente utilizados, tais classificadores foram escolhidos uma vez que podem ser particularmente sensíveis tanto ao número de atributos presente nos dados bem como a atributos redundantes e irrelevantes (Witten e Frank, 2005). Quanto ao algoritmo kNN, em ambos os casos considerados, o número de vizinhos, parâmetro k do algoritmo, foi ajustado para o valor 1, comumente utilizado na literatura. Ainda quanto a este algoritmo, a distância Euclidiana foi utilizada a fim de não introduzir nenhum viés adicional, uma vez que as correlações são neste cenário comparadas para a seleção de atributos (genes) e não para a classificação propriamente dita.

Com o intuito de obter uma estimativa imparcial dos erros dos classificadores utilizados, o procedimento de validação cruzada de 10 pastas estratificado (*stratified 10-fold cross-validation*)⁴ (Tan et al., 2005) foi empregado. Este procedimento consiste em dividir o conjunto de dados disponível em 10 pastas disjuntas, de forma que cada pasta contenha aproximadamente a mesma proporção de objetos de cada uma das classes observadas no conjunto de dados original. Após tal divisão, o classificador em questão é construído 10 vezes, de forma que, a cada uma das construções, uma pasta diferente é utilizada como pasta de teste e as nove demais são utilizadas como pasta de treinamento. Nos experimentos em que algoritmos de seleção de atributos foram empregados, o procedimento de validação cruzada de 10 pastas estratificado foi utilizado externamente ao processo de seleção de atributos. Dessa forma, a seleção de atributos foi realizada somente nas pastas de treinamento (Reunanen, 2003), desconsiderando quaisquer informações dos conjuntos de teste, o que simula uma aplicação real do processo de seleção de atributos. Para tornar mais claro o processo adotado, na Figura 6.1 é ilustrada uma única iteração da validação cruzada envolvendo a seleção de atributos. O processo de seleção é aplicado aos dados de treinamento e, posteriormente, os atributos não selecionados são removidos do conjunto de teste, possibilitando assim a construção e avaliação do classificador.

Figura 6.1: Procedimento para avaliação dos métodos de seleção de atributos.

Para verificar a possível existência de diferenças estatísticas entre os coeficientes de cor-

⁴O procedimento de validação cruzada de 10 pastas estratificado pode ser visto como uma realização de um procedimento mais genérico, onde um número $2 \leq v \leq n$ arbitrário de pastas pode ser escolhido.

relação avaliados, o procedimento sugerido por [Demšar \(2006\)](#), que consiste na aplicação dos testes não-paramétricos de Friedman e Nemenyi foi adotado. Tais testes permitem a comparação dos diferentes pares algoritmo e coeficiente de correlação, quando aplicados a um conjunto de bases de dados. Neste sentido, para a avaliação no cenário de classificação, os testes foram aplicados aos erros de classificação do 1NN utilizando as diferentes correlações. Para a avaliação no cenário de seleção de atributos, os testes foram aplicados aos erros obtidos com os classificadores 1NN e NB após a aplicação do método de seleção com cada uma das medidas avaliadas. É importante ressaltar que dois testes foram aplicados para a avaliação dos coeficientes de correlação para a tarefa de classificação, um para as bases de cDNA e outro para as bases Affymetrix. Já para a seleção de atributos, quatro testes foram aplicados aos resultados obtidos, dois para bases de cDNA (um para os erros obtidos com o classificador NB e outro para os erros do classificador 1NN) e dois para bases Affymetrix (novamente, um para os erros obtidos com o classificador NB e outro para os erros do classificador 1NN). Mais informações quanto às bases de dados utilizadas na realização dos experimentos são fornecidas na Seção 6.3.1.

Uma vez apresentada a metodologia de avaliação empregada durante a realização dos experimentos, são necessários alguns comentários adicionais no que diz respeito ao estimador de erro escolhido, o procedimento de validação cruzada de 10 pastas estratificado. Uma característica peculiar do cenário específico de classificação de amostras e seleção de genes para o problema de classificação de amostras é a pequena quantidade de amostras (neste caso objetos) disponíveis. Esta característica tem levantado na literatura a discussão de qual estratégia de estimação de erro deve ser utilizada, ou é mais adequada, para tal cenário de aplicação. De acordo com [Dougherty \(2001\)](#) o pequeno número de amostras disponíveis afeta o desempenho dos estimadores de erro e, por consequência, as próprias estimativas de erro obtidas a partir de sua utilização. A fim de comparar a qualidade das estimativas de erro geradas a partir do uso de diferentes estimadores, no cenário específico de classificação de amostras e seleção de genes em dados de *microarray*, diversos estudos têm sido realizados. Alguns destes estudos são discutidos a seguir.

No estudo realizado por [Braga-Neto e Dougherty \(2004\)](#) os autores comparam o desempenho de sete estimadores de erro em algoritmos de classificação. Os autores apontam que para estimadores de erro baseados em pastas cada objeto possui uma participação elevada no cálculo do erro, o que contribui para o aumento da variabilidade observada. Em uma base de dados com 30 objetos, por exemplo, cada objeto é responsável por mais de 3% na taxa de erro estimada. Segundo os autores, estimadores de erro como *Bootstrap* .632 e suas variantes ([Efron, 1979, 1983](#)) e o estimador cv10r (que consiste em realizar 10 repetições do procedimento de validação cruzada) não sofrem tanto deste problema, porém, possuem um alto custo computacional associado e podem se tornar impraticáveis. De fato, os autores argumentam que para experimentos onde o número de genes é elevado e muitos subconjuntos de genes devem ser avaliados durante a etapa de seleção de atributos tais estimadores de erro podem apresentar um alto custo computacional a ponto de

sua aplicação se tornar impraticável.

Molinaro et al. (2005) comparam o desempenho de sete estimadores de erro também para a tarefa de classificação de amostras. Diferentemente do trabalho de Braga-Neto e Dougherty (2004) a comparação realizada pelos autores considera ainda casos para os quais o processo de seleção de atributos foi empregado. Segundo os autores, o procedimento de validação cruzada de 10 pastas produz estimativas de erro próximas daquelas obtidas com o uso de *leave-one-out cross-validation*⁵ (LOOCV) e portanto deve ser preferido em virtude de seu menor custo computacional. Essa mesma recomendação é realizada por Ambroise e McLachlan (2002). Ainda segundo Molinaro et al. (2005), em alguns casos, quando há a aplicação de seleção de atributos, os estimadores LOOCV e *10-fold cross-validation* podem produzir estimativas melhores que as do *Bootstrap .632+* (Efron, 1983).

A breve discussão realizada possui o intuito de demonstrar que não há na literatura um consenso quanto à qual método de estimação de erro deve ser utilizado quando o número de amostras é reduzido, tal qual no caso de dados de *microarray*. Como apontado por Boulesteix et al. (2008), estudos contraditórios, tais quais os de Braga-Neto e Dougherty (2004) e Molinaro et al. (2005) têm sido publicados no que diz respeito à comparação de estimadores de erro no caso de número limitado de amostras. É importante notar que no presente trabalho tanto os métodos de classificação quanto de seleção de atributos não estão sendo comparados para a aplicação final no problema abordado. Dessa forma, uma estimativa exata do erro para as bases de dados em questão não é a principal questão envolvida. Por tal motivo acredita-se que a validação cruzada de 10 pastas estratificada seja suficiente e justa para a comparação entre diferentes medidas nos algoritmos considerados. Neste sentido, trabalhos relativamente recentes que avaliam diversos algoritmos de classificação em dados de *microarray* também têm utilizado o procedimento de validação cruzada de 10 pastas para estimação de seu erro, com a finalidade de comparar diferentes algoritmos (Romualdi et al., 2003; Statnikov et al., 2005).

Ainda quanto ao estimador de erro *Bootstrap .632+*, experimentos preliminares envolvendo seleção de atributos indicaram um alto custo computacional, o qual tornou sua utilização inviável. O leitor interessado em discussões quanto à adequabilidade de estimadores de erro no cenário de classificação de amostras pode consultar os trabalhos de (Dougherty, 2001), (Ambroise e McLachlan, 2002), (Braga-Neto e Dougherty, 2004), (Molinaro et al., 2005), (Boulesteix et al., 2008) e (Dougherty et al., 2010).

6.3 Experimentos com Amostras

Apresentados os algoritmos e a metodologia de comparação empregada durante a realização dos experimentos, os experimentos propriamente ditos podem ser apresentados e discutidos. Na Seção 6.3.1 são apresentadas as bases de dados utilizadas para a realização dos experimentos. Na Seção 6.3.2 são discutidos os resultados obtidos em seleção

⁵No LOOCV o número de pastas é igual ao número de objetos.

de genes para classificação de amostras. Finalmente, na Seção 6.3.3, são apresentados e discutidos os resultados obtidos em classificação de amostras sem a seleção de genes (atributos). Durante a apresentação e discussão dos resultados estes foram sumarizados, a fim de facilitar a compreensão da discussão realizada. Tabelas completas com resultados individuais obtidos em cada uma das bases de dados consideradas podem ser encontradas no Apêndice B.

6.3.1 Bases de Dados

Nos dois cenários considerados, seleção de genes para classificação de amostras e classificação de amostras sem seleção de atributos (genes), o conjunto de 35 bases de dados de *benchmark* introduzido por (de Souto et al., 2008) foi utilizado. Estas bases de dados são as mesma já utilizadas durante os experimentos de agrupamento de amostras. Desta forma, um sumário destas bases pode ser encontrado na Tabela 5.1, apresentada na Seção 5.2.2. Novamente, mais informações a respeito destas bases, como o número de genes originais, referências bibliográficas aos estudos que às originaram e quanto aos procedimentos de pré-processamento aplicados podem ser obtidas em (de Souto et al., 2008).

6.3.2 Seleção de Atributos

A seguir são apresentados os resultados obtidos com o uso dos diferentes coeficientes de correlação na tarefa de seleção de atributos (seleção de genes), para o problema de classificação de amostras, utilizando o algoritmo SSF, apresentado na Seção 6.1.1. Os resultados serão discutidos de forma separada para cada um dos tipos de bases de dados utilizados, a saber: cDNA e Affymetrix.

Simplified Silhouette Filter

Antes da apresentação dos resultados propriamente ditos, algumas considerações quanto ao método SSF devem ser feitas. Mesmo estimando de forma automática quais e quantos atributos devem ser selecionados, o algoritmo possui alguns parâmetros (teoricamente não críticos) que devem ser setados pelo usuário. Nos experimentos conduzidos os seguintes parâmetros (definidos no Algoritmo 3) foram utilizados: número mínimo de grupos, $k_{min} = 2$; número máximo de grupos, $k_{max} = \sqrt{m}$ (onde m é igual ao número de atributos) e número de repetições da estratégia de amostragem, $n_p = 20$ (número de inicializações aleatórias do *k-medoids* para cada valor de k avaliado). Ainda quanto ao algoritmo SSF, durante a discussão dos resultados, sua sigla é seguida do símbolo do coeficiente de correlação utilizado entre atributos durante a fase de agrupamento. Para a correlação de Pearson (ρ), por exemplo, a variante⁶ do SSF obtida é denotada por SSF- ρ . O mesmo se aplica para os seis outros coeficientes de correlação considerados (descritos

⁶Neste capítulo, o termo variante é utilizado para denotar o método SSF obtido com a aplicação de um coeficiente de correlação específico, dentre os sete avaliados. Desta forma, sete variantes são consideradas.

na Seção 4.1). Quanto ao algoritmo kNN, vale lembrar que seu parâmetro k , referente ao número de vizinhos, foi ajustado para o valor 1; dessa forma, o algoritmo será referenciado como 1NN.

Feitas tais considerações, os resultados obtidos podem ser discutidos. Inicialmente, são apresentados, na Figura 6.2, os erros médios obtidos com a aplicação dos algoritmos NB e 1NN para as 14 bases provenientes de experimentos realizados com a tecnologia de cDNA. A aplicação dos classificadores se deu após o uso do SSF, com cada uma das medidas de correlação avaliadas. Em se tratando dos resultados médios obtidos, em nenhum caso a aplicação da seleção de atributos apresentou melhoras nos erros de classificação obtidos, se os resultados forem comparados à utilização de todos os atributos. Considerando ainda os erros médios obtidos com os dois classificadores após a seleção de atributos, o uso de diferentes coeficientes de correlação não levou a grandes diferenças nos resultados observados. Além dos erros médios, são apresentadas, na Tabela 6.1, as porcentagens médias e números médios de atributos selecionados, juntamente com os desvios padrão, entre parênteses, para cada um dos métodos. Observando-se tais valores, torna-se evidente o viés do SSF em selecionar subconjuntos com número reduzido de atributos. De fato, apenas três variantes do SSF, dentre as sete consideradas, selecionaram em média mais de 1% dos atributos para as bases de dados de cDNA. Neste sentido, os subconjuntos com números de atributos muito reduzidos podem ter influenciado negativamente a performance dos classificadores nas bases de dados.

Figura 6.2: Erros médios de classificação em bases de cDNA.

São apresentados também, nas Tabelas 6.2 e 6.3, os números de vitórias (V), empates (E) e derrotas (D) relativos aos erros de classificação obtidos com a aplicação dos classificadores NB e 1NN, respectivamente. As tabelas devem ser interpretadas da seguinte forma: dados dois métodos na tabela, um em uma linha e um em uma coluna, temos uma célula específica, com os valores $V/E/D$, que representam que o método da linha foi melhor que o da coluna em V bases de dados, que os métodos empataram em E bases de dados e, por fim, que o método da linha perdeu do método da coluna em D bases de dados. Feitas tais considerações, é possível notar que para ambos classificadores a se-

Tabela 6.1: Porcentagens médias e números médios de atributos selecionados em bases de cDNA. Em ambos casos valores de desvio padrão são apresentados entre parênteses.

Método	% Atributos Selecionados	Média de Atributos Selecionados
SSF- ρ	0.64 (1.20)	5.26 (2.92)
SSF- ϱ	0.62 (1.20)	4.91 (2.20)
SSF- γ	1.26 (3.83)	4.66 (2.23)
SSF- τ	1.36 (4.21)	4.74 (2.55)
SSF- $\hat{\rho}$	0.55 (1.19)	4.07 (0.13)
SSF- r	0.97 (2.70)	4.56 (1.32)
SSF- $\hat{\gamma}$	1.72 (5.56)	5.06 (3.70)
Sem Seleção	100 (0.00)	1846.29 (1002.06)

leção de atributos, independentemente da correlação utilizada, levou à obtenção de piores resultados. Ainda é possível notar que o desempenho dos classificadores após a seleção de atributos foi bastante similar com as diferentes medidas. Porém, analisando-se os resultados obtidos para o classificador Naïve Bayes, alguns resultados interessantes podem ser observados. Considerando o método SSF empregando a correlação de Pearson (SSF- ρ) na base de dados *lapointe-2004-v1*, um erro médio de 30.24% foi obtido com apenas 4.2 atributos em média, frente a um erro médio de 30.48% obtido com a utilização de todos os 1625 atributos da base. Embora as taxas médias de erro sejam similares, a variante de seleção de atributos considerada proporcionou uma considerável redução no número de atributos da base. Na base de dados *bredel-2005*, por exemplo, as variantes que empregam as correlações de Goodman-Kruskal (SSF- γ) e Kendall (SSF- τ) proporcionaram resultados de erros médios melhores ou iguais aos obtidos com todos os atributos para o algoritmo Naïve Bayes, em ambos os casos com uma redução de 1739 atributos na base original para apenas 4 após a aplicação das variantes de seleção de atributos. Outros casos similares podem também ser observados nas bases de dados *bittner*, *garber*, *liang* e *risinger* (resultados detalhados no Apêndice B). Ainda quanto aos resultados obtidos nas bases de cDNA, não houve destaques significativos para os números de vitórias, empates e derrotas considerando o algoritmo 1NN.

Tabela 6.2: Vitórias, empates e derrotas para o classificador NB em bases cDNA.

Método	SSF- ρ	SSF- ϱ	SSF- γ	SSF- τ	SSF- $\hat{\rho}$	SSF- r	SSF- $\hat{\gamma}$	S. Seleção
SSF- ρ	—	8/1/5	7/0/7	5/0/9	8/0/6	8/0/6	8/0/6	1/0/13
SSF- ϱ	5/1/8	—	5/1/8	6/0/8	8/1/5	7/1/6	7/0/7	0/0/14
SSF- γ	7/0/7	8/1/5	—	3/4/7	9/0/5	7/0/7	8/0/6	1/2/11
SSF- τ	9/0/5	8/0/6	7/4/3	—	10/0/4	8/0/6	9/0/5	0/3/11
SSF- $\hat{\rho}$	6/0/8	5/1/8	5/0/9	4/0/10	—	5/0/9	7/1/6	1/0/13
SSF- r	6/0/8	6/1/7	7/0/7	6/0/8	9/0/5	—	7/0/7	0/0/14
SSF- $\hat{\gamma}$	6/0/8	7/0/7	6/0/8	5/0/9	6/1/7	7/0/7	—	0/0/14
S. Seleção	13/0/1	14/0/0	11/2/1	11/3/0	13/0/1	14/0/0	14/0/0	—

Na Figura 6.3 são apresentados os erros médios obtidos com os classificadores NB e 1NN após a seleção de atributos com cada uma das variantes do método SSF para as 21 bases de dados Affymetrix. De forma semelhante aos resultados obtidos para as bases de cDNA, o uso de todos os atributos proporcionou, em média, menores erros para ambos classificadores considerados, NB e 1NN. Para este conjunto de bases de dados, entretanto,

Tabela 6.3: Vitórias, empates e derrotas para o classificador 1NN em bases cDNA.

Método	SSF- ρ	SSF- ϱ	SSF- γ	SSF- τ	SSF- $\hat{\rho}$	SSF- r	SSF- $\hat{\gamma}$	S. Seleção
SSF- ρ	—	6/0/8	8/0/6	7/0/7	7/2/5	6/1/7	6/1/7	0/0/14
SSF- ϱ	8/0/6	—	7/0/7	7/0/7	6/0/8	9/0/5	8/0/6	0/0/14
SSF- γ	6/0/8	7/0/7	—	2/3/9	6/1/7	7/1/6	5/0/9	0/0/14
SSF- τ	7/0/7	7/0/7	9/3/2	—	8/0/6	9/0/5	6/1/7	0/0/14
SSF- $\hat{\rho}$	5/2/7	8/0/6	7/1/6	6/0/8	—	6/1/7	4/1/9	0/0/14
SSF- r	7/1/6	5/0/9	6/1/7	5/0/9	7/1/6	—	9/0/5	0/0/14
SSF- $\hat{\gamma}$	7/1/6	6/0/8	9/0/5	7/1/6	9/1/4	5/0/9	—	0/0/14
S. Seleção	14/0/0	14/0/0	14/0/0	14/0/0	14/0/0	14/0/0	14/0/0	—

houve diferenças visíveis entre os erros médios obtidos para ambos classificadores após a aplicação das diferentes variantes de seleção de atributos consideradas. Para ambos os classificadores empregados, a variante SSF- ϱ (Jackknife) produziu os menores resultados médios de erro. Se comparados aos resultados obtidos com a utilização de todos atributos das bases de dados, esses resultados foram, em média, 12% piores para o classificador NB e 6% piores para o classificador 1NN, . Entre as variantes consideradas, os piores resultados foram obtidos com a aplicação da correlação de Goodman-Kruskal (SSF- γ), para ambos classificadores considerados. Para o classificador NB, separadamente, o uso do SSF com a correlação de Pearson (SSF- ρ) produziu resultados próximos aos obtidos com os da correlação de Spearman (SSF- $\hat{\rho}$). Ainda considerando o classificador NB, a seleção de atributos realizada com as correlações Rank-Magnitude (SSF- r) e Goodman-Kruskal Ponderado (SSF- $\hat{\gamma}$) levou a resultados similares em média. Considerando somente o classificador 1NN, as variantes que empregaram as correlações de Kendall (SSF- τ), Spearman (SSF- $\hat{\rho}$), Rank-Magnitude (SSF- r), Goodman-Kruskal (SSF- γ) e Goodman-Kruskal Ponderado (SSF- $\hat{\gamma}$) apresentaram também resultados próximos em média.

Figura 6.3: Erros médios de classificação em bases Affymetrix.

Mesmo havendo uma maior taxa de erro média após a aplicação da seleção de atributos, tais resultados tornam-se interessantes quando as porcentagens e números médios de atributos selecionados por estes métodos são avaliadas. Estes resultados são apresentados pela Tabela 6.4. Novamente, os valores de desvios padrão são reportados entre parênteses. Note que a variante SSF- ϱ (Jackknife) selecionou em média apenas 1.7% dos atributos

nas bases de dados Affymetrix, uma redução considerável. Ainda quanto ao número de atributos selecionados, é interessante notar que com apenas 0.4% do total dos atributos, em média, a variante que emprega a correlação de Spearman (SSF- $\hat{\rho}$) obteve o segundo menor valor de erro médio, dentre as demais variantes consideradas, se os resultados do classificador NB forem considerados. Para o classificador 1NN, de forma geral, maiores médias nas porcentagens de atributos selecionados, obtidas com as variantes das correlações de Pearson (SSF- ρ), Jackknife (SSF- ϱ) e Kendall (τ), se traduziram em menores erros médios obtidos com a aplicação posterior do classificador.

Tabela 6.4: Porcentagens médias e números médios de atributos selecionados em bases Affymetrix. Em ambos casos valores de desvio padrão são apresentados entre parênteses.

Método	% Atributos Selecionados	Média de Atributos Selecionados
SSF- ρ	1.72 (1.49)	21.40 (18.60)
SSF- ϱ	1.70 (1.82)	21.35 (22.59)
SSF- γ	0.44 (0.50)	4.10 (0.21)
SSF- τ	0.68 (0.84)	10.34 (19.87)
SSF- $\hat{\rho}$	0.42 (0.45)	4.04 (0.17)
SSF- r	0.47 (0.57)	4.29 (0.62)
SSF- $\hat{\gamma}$	0.44 (0.60)	4.07 (0.46)
Sem Seleção	100 (0.00)	1425.90 (630.71)

Tendo em mente que os resultados em média podem esconder casos particulares interessantes, nas Tabelas 6.5 e 6.6 são apresentados os números de vitórias, empates e derrotas das variantes do SSF, considerando os erros obtidos para os classificadores NB e 1NN nas bases de dados Affymetrix. A interpretação das tabelas é a mesma já empregada anteriormente, ou seja, cada linha apresenta o número de vitórias (V), empates (E) e derrotas (D) do método da linha em relação ao método da coluna, na forma V/E/D. A variante que utiliza a correlação Jackknife (SSF- ϱ), que obteve melhores resultados em média, obteve melhores resultados que a utilização de todos os atributos em 14% dos casos para o classificador NB e em 28% dos casos para o classificador 1NN. Tais resultados são interessantes, uma vez que a quantidade de atributos selecionados é bastante reduzida em comparação com o número de atributos originais das bases de dados. Ainda quanto à variante SSF- ϱ (Jackknife), para o classificador NB, seus resultados foram superiores aos das demais variantes em 62% dos casos, contra apenas 45% para a correlação de Pearson (SSF- ρ). Considerando os resultados do 1NN, os subconjuntos produzidos com a utilização da correlação Jackknife (SSF- ϱ) produziram resultados superiores aos das demais variantes em 74% dos casos considerados, contra apenas 53% para Pearson (SSF- ρ). Tais resultados são evidenciados uma vez que a correlação Jackknife é baseada na correlação de Pearson e está associada a um maior custo computacional. Neste sentido, com a contrapartida de um maior custo computacional associado, sua utilização, para o presente conjunto de bases de dados, se mostrou vantajosa. Considerando os números de vitórias, empates e derrotas entre as variantes SSF- τ (Kendall) e SSF- $\hat{\rho}$ (Spearman), resultados bastante parecidos podem ser observados para o classificador NB. Ainda para o classifi-

cador NB, todas as variantes de seleção de atributos, exceto SSF- γ (Goodman-Kruskal) e SSF- $\hat{\gamma}$ (Goodman-Kruskal Ponderado), levaram a melhores resultados se comparadas à variante SSF- ρ (Pearson).

Tabela 6.5: Vitórias, empates e derrotas para o classificador NB em bases Affymetrix.

Correlação	SSF- ρ	SSF- ϱ	SSF- γ	SSF- τ	SSF- $\hat{\rho}$	SSF- r	SSF- $\hat{\gamma}$	S. Seleção
SSF- ρ	—	6/0/15	17/1/3	8/0/13	9/0/12	11/0/10	16/1/4	0/0/21
SSF- ϱ	15/0/6	—	18/0/3	11/0/10	10/1/10	17/1/3	18/0/3	3/0/18
SSF- γ	3/1/17	3/0/18	—	5/0/16	4/0/17	8/1/12	5/2/14	1/0/20
SSF- τ	13/0/8	10/0/11	16/0/5	—	10/0/11	15/0/6	16/0/5	1/0/20
SSF- $\hat{\rho}$	12/0/9	10/1/10	17/0/4	11/0/10	—	14/0/7	14/2/5	2/2/17
SSF- r	10/0/11	3/1/17	12/1/8	6/0/15	7/0/14	—	11/0/10	1/0/20
SSF- $\hat{\gamma}$	4/1/16	3/0/18	14/2/5	5/0/16	5/2/14	10/0/11	—	0/0/21
S. Seleção	21/0/0	18/0/3	20/0/1	20/0/1	17/2/2	20/0/1	21/0/0	—

Tabela 6.6: Vitórias, empates e derrotas para o classificador 1NN em bases Affymetrix.

Correlação	SSF- ρ	SSF- ϱ	SSF- γ	SSF- τ	SSF- $\hat{\rho}$	SSF- r	SSF- $\hat{\gamma}$	S. Seleção
SSF- ρ	—	4/0/17	19/0/2	11/0/10	14/0/7	14/0/7	13/0/8	3/0/18
SSF- ϱ	17/0/4	—	21/0/0	15/0/6	17/0/4	18/0/3	15/0/6	6/0/15
SSF- γ	2/0/19	0/0/21	—	4/0/17	4/0/17	2/2/17	3/1/17	1/0/20
SSF- τ	10/0/11	6/0/15	17/0/4	—	10/0/11	11/0/10	10/1/10	2/0/19
SSF- $\hat{\rho}$	7/0/14	4/0/17	17/0/4	11/0/10	—	8/1/12	9/2/10	3/0/18
SSF- r	7/0/14	3/0/18	17/2/2	10/0/11	12/1/8	—	12/1/8	1/0/20
SSF- $\hat{\gamma}$	8/0/13	6/0/15	17/1/3	10/1/10	10/2/9	8/1/12	—	2/3/16
S. Seleção	18/0/3	15/0/6	20/0/1	19/0/2	18/0/3	20/0/1	16/3/2	—

Por fim, com o intuito de verificar diferenças estatisticamente significativas entre os coeficientes de correlação avaliados, os testes de Friedman e Nemenyi foram aplicados. Neste sentido, para as bases de cDNA só foram observadas diferenças significativas favoráveis à utilização dos classificadores sem a aplicação de seleção de atributos, com $\alpha = 5\%$, tanto para o classificador 1NN quanto para o classificador Naïve Bayes. Para as bases de dados Affymetrix, as seguintes diferenças foram observadas. Com relação ao classificador Naïve Bayes, o uso de todos os atributos proporcionou resultados significativamente melhores aos produzidos com a correlação Jackknife com $\alpha = 10\%$. Considerando as demais variantes de seleção de atributos, a aplicação do classificador NB utilizando todos atributos proporcionou resultados superiores com $\alpha = 5\%$. Ainda com relação ao classificador NB, o uso da variante que emprega a correlação Jackknife (SSF- ϱ) levou a resultados superiores aos obtidos com as variantes que utilizam a correlação de Goodman-Kruskal (SSF- γ) e Goodman-Kruskal Ponderado (SSF- $\hat{\gamma}$), com $\alpha = 5\%$. Diferenças favoráveis também foram observadas para as variantes que empregam a correlação de Kendall (SSF- τ) e Spearman (SSF- $\hat{\rho}$) em detrimento da variante que utiliza a correlação de Goodman-Kruskal (SSF- γ), com $\alpha = 5\%$. Ainda considerando o conjunto de bases Affymetrix, agora para o classificador 1NN, os testes sugerem diferenças favoráveis à aplicação do algoritmo de classificação sem seleção de atributos com $\alpha = 5\%$, exceto contra a variante que emprega a correlação Jackknife (SSF- ϱ). Os testes também sugerem diferenças favoráveis frente à variante que emprega a correlação de Goodman-Kruskal (SSF- γ) para as variantes que utilizaram Pearson (SSF- ρ), Jackknife (SSF- ϱ) e Kendall (SSF- τ), com $\alpha = 5\%$ e com $\alpha = 10\%$ para Rank-Magnitude (SSF- r) e Goodman-Kruskal Ponderado (SSF- $\hat{\gamma}$). Final-

mente, a variante $SSF-\varrho$ (Jackknife) apresentou resultados estatisticamente superiores à $SSF-\hat{\rho}$ (Spearman) ($\alpha = 5\%$) e $SSF-r$ (Rank-Magnitude) ($\alpha = 10\%$).

6.3.3 Classificação

Nesta seção são abordados os experimentos conduzidos para avaliação dos coeficientes de correlação como medida de distância para o classificador kNN. Quanto ao classificador, especificamente, seu parâmetro k , correspondente ao número de vizinhos, foi setado para o valor 1, dessa forma, o classificador será daqui em diante referido como 1NN. Cada um dos coeficientes de correlação considerados foi avaliado de acordo com o erro de classificação produzido da aplicação do algoritmo 1NN quando equipado com o mesmo. Novamente, os resultados encontram-se separados de acordo com o tipo das bases de dados utilizadas.

Os resultados médios de erro de classificação obtidos para as 14 bases de dados de cDNA são apresentados na Figura 6.4. Os menores erros médios foram obtidos com a utilização da correlação Jackknife (ϱ), a qual, se comparada à correlação de Pearson (ρ), reduziu em praticamente 2% o erro médio de classificação. Com resultados muito próximos aos obtidos com a correlação de Pearson (ρ), a correlação Rank-Magnitude (r) obteve o terceiro melhor erro médio para as medidas comparadas. Dentre as sete correlações avaliadas, as baseadas em *ranks*, Goodman-Kruskal (γ), Kendall (τ) e Spearman ($\hat{\rho}$), apresentaram resultados similares entre si quanto aos erros médios observados. Por fim, os maiores erros de classificação, em média, foram obtidos com a utilização da correlação de Goodman-Kruskal Ponderado ($\hat{\gamma}$). A diferença entre as médias do melhor e pior resultados, correlação Jackknife (ϱ) e Goodman-Kruskal Ponderado ($\hat{\gamma}$), respectivamente, foi de 4%.

Figura 6.4: Erros médios do classificador 1NN em bases cDNA.

Ainda quanto aos resultados obtidos para as 14 bases de cDNA, na Tabela 6.7 são apresentados os números de vitórias, empates e derrotas obtidos com a utilização de cada uma das sete medidas avaliadas. Considerando uma linha e uma coluna específica, a célula

em questão apresenta o número de vitórias da correlação da linha sobre a da coluna, o número de empates entre as medidas e o número de derrotas sofridas pela correlação da linha em questão, todos em relação aos erros de classificação obtidos. Para estas 14 bases de dados, é possível notar que a correlação Jackknife (ϱ) produziu resultados melhores ou iguais às demais medidas em pelo menos 80% dos casos considerados. Estes resultados são próximos aos obtidos com a correlação de Pearson (ρ), que proporcionou resultados iguais ou superiores às demais correlações em 78% dos casos. Já a correlação Rank-Magnitude (r), que em média apresentou resultados muito próximos à correlação de Pearson (ρ), obteve resultados iguais ou melhores às demais medidas em 69% dos casos (9% a menos que a correlação de Pearson). Considerando as correlações de Goodman-Kruskal (γ), Kendall (γ) e Spearman ($\hat{\rho}$), baseadas em *ranks*, os resultados quanto ao número de vitórias, empates e derrotas obtidos foram bastante equilibrados. Ainda quanto a estas três medidas, é possível notar que seus resultados foram inferiores à utilização das correlações de Pearson (ρ), Jackknife (ϱ) e Rank-Magnitude (r) na maioria das bases de dados avaliadas. Por fim, considerando o número de vitórias, empates e derrotas, os piores resultados foram observados com o uso da correlação de Goodman-Kruskal Ponderado ($\hat{\gamma}$), a qual obteve resultados inferiores às demais medidas em 54% dos casos avaliados.

Tabela 6.7: Vitórias, empates e derrotas do classificador 1NN em bases cDNA.

Correlação	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
ρ	—	2/8/4	9/2/3	9/2/3	9/2/3	6/6/2	9/2/3
ϱ	4/8/2	—	9/2/3	9/2/3	9/2/3	8/4/2	9/2/3
γ	3/2/9	3/2/9	—	2/11/1	4/8/2	3/4/7	6/5/3
τ	3/2/9	3/2/9	1/11/2	—	4/7/3	3/4/7	6/5/3
$\hat{\rho}$	3/2/9	3/2/9	2/8/4	3/7/4	—	3/4/7	7/4/3
r	2/6/6	2/4/8	7/4/3	7/4/3	7/4/3	—	9/2/3
$\hat{\gamma}$	3/2/9	3/2/9	3/5/6	3/5/6	3/4/7	3/2/9	—

Considerando as 21 bases de dados Affymetrix, os erros médios de classificação para o algoritmo 1NN, quando equipado com cada uma das medidas de correlação avaliadas, são apresentados na Figura 6.5. Nestas bases de dados, a correlação Rank-Magnitude (r) obteve os melhores resultados em média, se destacando em relação às demais. Em média, a taxa de erro para esta medida foi 2% menor que a segunda melhor, a correlação de Spearman ($\hat{\rho}$). Analisando os resultados obtidos para as medidas baseadas em *ranks*, Goodman-Kruskal (γ), Kendall (τ) e Spearman ($\hat{\rho}$), é possível observar um menor erro médio em relação às correlações de Pearson (ρ) e Jackknife (ϱ). Ainda quanto a estas medidas, é possível notar que seus resultados foram bastante próximos, em média. É interessante notar que, para as bases de dados Affymetrix, o uso da correlação Jackknife (ϱ) proporcionou erros médios em torno de 1% menores aos obtidos com a correlação de Pearson (ρ). Como observado também nas bases de dados cDNA, o uso da correlação de Goodman-Kruskal Ponderado ($\hat{\gamma}$) levou à produção dos piores resultados em média.

Na Tabela 6.8 são apresentados os números de vitórias, empates e derrotas dos coeficientes de correlação avaliados para as 21 bases de dados Affymetrix. A correlação Rank-Magnitude (r), melhor em média, proporcionou resultados melhores que as demais correlações em 68% dos casos. Ainda considerando a correlação Rank-Magnitude (r), sua

Figura 6.5: Erros médios do classificador 1NN em bases Affymetrix

utilização proporcionou resultados iguais ou melhores às demais em 83% dos casos avaliados. Quando analisados os números de vitórias, empates e derrotas entre as correlações baseadas em *ranks*, Goodman-Kruskal (γ), Kendall (τ) e Spearman ($\hat{\rho}$), e a correlação Jackknife (r), não são observadas grandes diferenças, assim como nos resultados médios já apresentados. Por fim, o uso da correlação de Pearson (ρ) levou a resultados inferiores às demais em 56% dos casos, enquanto para que a correlação de Goodman-Kruskal Ponderado ($\hat{\gamma}$) resultados inferiores às demais podem ser observados em 57% dos casos. É interessante notar que, mesmo produzindo piores resultados em média, a correlação de Goodman-Kruskal Ponderado ($\hat{\gamma}$) proporcionou resultados superiores ou iguais aos obtidos com a correlação de Pearson (ρ) em 61% das bases de dados avaliadas, ou seja, 13 de 21.

Tabela 6.8: Vitórias, empates e derrotas do classificador 1NN em bases Affymetrix

Correlação	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
ρ	—	3/6/12	6/4/11	8/2/11	9/1/11	2/4/15	8/2/11
ϱ	12/6/3	—	10/1/10	10/1/10	10/2/9	4/2/15	13/1/7
γ	11/4/6	10/1/10	—	8/8/5	5/11/5	3/5/13	12/4/5
τ	11/2/8	10/1/10	5/8/8	—	3/13/5	4/3/14	12/3/6
$\hat{\rho}$	11/1/9	9/2/10	5/11/5	5/13/3	—	4/3/14	12/4/5
r	15/4/2	15/2/4	13/5/3	14/3/4	14/3/4	—	15/2/4
$\hat{\gamma}$	11/2/8	7/1/13	5/4/12	6/3/12	5/4/12	4/2/15	—

Com o intuito de proporcionar uma visão detalhada dos resultados obtidos, são apresentados na Tabela 6.9 os erros médios obtidos com a utilização de cada uma das sete medidas avaliadas. Para cada par composto por um coeficiente de correlação e uma base de dados, o valor apresentado corresponde à média obtida com o procedimento de validação cruzada de 10 pastas estratificado. Os desvios padrões obtidos com o estimador nas 10 pastas são reportados entre parênteses. Para facilitar a visualização dos resultados, cada célula da tabela recebeu um tom de cinza, seguindo o seguinte critério: para cada uma das bases de dados (linhas da tabela), tons de cinza mais claros indicam maiores erros de classificação, enquanto tons de cinza mais escuros indicam menores erros. Empates

entre correlações receberam um mesmo tom de cinza. Considerando os resultados individualmente, alguns casos interessantes podem ser observados. Para bases de cDNA como *bittner-2000* e *khan-2001*, por exemplo, é possível verificar um desempenho superior das correlações de Pearson (ρ), Jackknife (ϱ) e Rank-Magnitude (r), com as duas primeiras produzindo resultados de erro 10% menores que os obtidos para as demais medidas. É interessante notar, porém, que, em bases como *lapointe-2004-v1* e *lapointe-2004-v2*, o uso de correlações baseadas em *ranks*, Goodman-Kruskal (γ), Kendall (τ) e Spearman ($\hat{\rho}$), levou à obtenção de menores taxas de erro. Considerando as bases do tipo Affymetrix, em bases como *bhattacharjee-2001*, *laiho-2007*, *pomeroy-2002-v1* e *shipp-2002-v1* um desempenho superior das correlações baseadas em *ranks*, Goodman-Kruskal (γ), Kendall (τ) e Spearman ($\hat{\rho}$), pode ser observado. Nestas bases, diferenças de até 20% nos valores de erro podem ser observadas entre as medidas comparadas. Ainda quanto a estas bases, com exceção da base de dados *shipp-2002-v1*, o uso da correlação de Goodman-Kruskal Ponderado ($\hat{\gamma}$) também proporcionou melhoras em relação às correlações de Pearson (ρ) e Jackknife (ϱ). Já em bases de dados como *yeoh-2002-v1* e *yeoh-2002-v2*, por exemplo, o uso das correlações de Pearson (ρ) e Jackknife (ϱ) levaram a melhores resultados, superando a correlação Rank-Magnitude (r), que, em média, apresentou os melhores resultados para as bases de dados Affymetrix.

A fim de verificar diferenças estatisticamente significativas entre os coeficientes de correlação avaliados, os testes de Friedman e Nemenyi foram aplicados aos resultados de erro obtidos. Dois testes foram aplicados, um para comparar as medidas nas bases de cDNA e outro para as bases de dados Affymetrix. Para o conjunto de dados de cDNA não houve diferenças estatisticamente significativas entre as medidas comparadas. Já para as bases de dados do conjunto Affymetrix, os testes indicaram, com $\alpha = 5\%$, diferenças estatisticamente favoráveis para a correlação Rank-Magnitude (r) contra as correlações de Pearson (ρ) e Goodman-Kruskal Ponderado ($\hat{\gamma}$).

6.4 Considerações Finais

Neste capítulo foram apresentados experimentos realizados para a comparação de sete coeficientes de correlação para as tarefas de seleção de genes e classificação de amostras. Na Seção 6.1 os algoritmos de seleção de atributos e classificação utilizados durante a análise comparativa foram apresentados. A metodologia empregada durante a avaliação das medidas foi discutida na Seção 6.2. Considerando os resultados obtidos a partir da comparação dos coeficientes de correlação para a tarefa de seleção de genes, a utilização prévia da seleção de atributos levou a maiores erros de classificação, se comparada à utilização de todos os atributos. Para as bases de cDNA todas as medidas avaliadas apresentaram resultados muito similares. Já para bases Affymetrix a correlação Jackknife apresentou em média os melhores resultados. Considerando a tarefa de classificação de amostras sem a prévia seleção de atributos, para as bases de cDNA três medidas se

Tabela 6.9: Erros do classificador 1NN para os sete coeficientes de correlação avaliados.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	22.0 (24.9)	19.5 (15.7)	24.5 (26.4)	24.5 (26.4)	24.5 (26.4)	24.5 (23.6)	27.0 (25.0)
	alizadeh-2000-v2	1.4 (4.5)	1.4 (4.5)	1.4 (4.5)	1.4 (4.5)	1.4 (4.5)	1.4 (4.5)	1.4 (4.5)
	alizadeh-2000-v3	19.3 (10.1)	19.3 (10.1)	14.5 (12.3)	14.5 (12.3)	16.2 (11.2)	19.3 (12.8)	14.5 (12.3)
	bittner-2000	5.0 (15.8)	5.0 (15.8)	12.5 (21.3)	15.0 (17.5)	12.5 (17.7)	10.0 (17.5)	22.5 (21.9)
	bredel-2005	20.0 (16.3)	20.0 (16.3)	22.0 (17.5)	22.0 (17.5)	22.0 (17.5)	20.0 (18.9)	24.0 (18.4)
	chen-2002	6.1 (4.9)	6.7 (5.1)	9.0 (6.0)	9.0 (6.0)	8.4 (6.0)	6.1 (4.9)	9.0 (6.0)
	garber-2001	18.6 (10.3)	18.6 (10.3)	21.7 (13.2)	21.7 (13.2)	21.7 (13.2)	20.2 (13.2)	21.7 (13.2)
	khan-2001	0.0 (0.0)	0.0 (0.0)	10.7 (11.5)	10.8 (13.5)	10.8 (13.5)	1.1 (3.5)	11.0 (10.6)
	lapointe-2004-v1	34.8 (16.7)	27.6 (14.3)	24.5 (17.8)	24.5 (17.8)	26.0 (21.0)	33.1 (25.1)	22.9 (12.1)
	lapointe-2004-v2	39.1 (12.2)	30.9 (13.7)	25.5 (9.4)	24.6 (9.6)	25.5 (9.4)	26.4 (10.0)	28.2 (13.2)
	liang-2005	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	risinger-2003	26.0 (17.5)	23.5 (15.3)	28.5 (19.0)	28.5 (19.0)	30.5 (21.3)	26.0 (17.5)	33.5 (17.2)
	tomlins-2006	13.4 (10.8)	13.4 (10.8)	17.3 (12.9)	17.3 (12.9)	17.3 (12.9)	17.3 (12.9)	16.3 (12.2)
	tomlins-2006-v2	16.2 (8.9)	17.3 (8.9)	21.6 (13.6)	21.6 (13.6)	20.4 (12.9)	18.3 (10.8)	20.4 (10.0)
Affymetrix	armstrong-2002-v1	2.9 (6.0)	1.4 (4.5)	1.4 (4.5)	1.4 (4.5)	1.4 (4.5)	0.0 (0.0)	7.1 (10.1)
	armstrong-2002-v2	4.1 (6.6)	5.5 (9.8)	4.1 (6.6)	4.1 (6.6)	4.1 (6.6)	4.1 (6.6)	9.6 (9.5)
	bhattacharjee-2001	14.3 (4.9)	14.3 (6.8)	7.4 (4.1)	6.3 (5.1)	6.3 (5.1)	8.3 (5.0)	8.3 (4.6)
	chowdary-2006	2.8 (4.5)	2.8 (4.5)	3.8 (4.9)	5.7 (4.9)	3.8 (4.9)	2.8 (4.5)	3.8 (4.9)
	dyrskjot-2003	17.5 (16.9)	17.5 (16.9)	20.0 (19.7)	22.5 (24.9)	20.0 (19.7)	17.5 (20.6)	15.0 (17.5)
	golub-1999-v1	4.1 (6.6)	2.7 (5.7)	7.0 (10.0)	5.5 (9.8)	5.5 (9.8)	2.9 (6.0)	4.1 (6.6)
	golub-1999-v2	5.4 (6.9)	4.1 (6.6)	8.4 (9.9)	8.4 (9.9)	8.4 (9.9)	2.9 (6.0)	4.3 (6.9)
	gordon-2002	2.2 (3.9)	2.2 (2.8)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	laiho-2007	24.2 (29.8)	24.2 (29.8)	9.2 (14.9)	9.2 (14.9)	9.2 (14.9)	16.7 (14.7)	9.2 (14.9)
	nut-2003-v1	62.0 (22.0)	60.0 (24.9)	62.0 (17.5)	62.0 (17.5)	64.0 (18.4)	58.0 (22.0)	60.0 (18.9)
	nut-2003-v2	56.7 (27.4)	50.0 (32.4)	53.3 (28.1)	58.3 (26.4)	58.3 (26.4)	40.0 (37.8)	60.0 (30.6)
	nut-2003-v3	46.7 (24.6)	51.7 (29.9)	46.7 (34.1)	41.7 (28.6)	41.7 (28.6)	46.7 (24.6)	41.7 (28.6)
	pomeroy-2002-v1	35.0 (28.5)	35.8 (32.4)	15.0 (21.1)	15.0 (21.1)	15.0 (21.1)	30.8 (27.0)	16.7 (24.5)
	pomeroy-2002-v2	26.5 (14.5)	22.0 (18.4)	24.0 (23.7)	24.0 (16.8)	24.0 (16.8)	24.0 (23.7)	26.5 (18.7)
	ramaswamy-2001	24.7 (11.4)	20.5 (8.4)	24.7 (7.9)	26.3 (7.0)	25.8 (6.8)	18.4 (5.7)	90.5 (3.3)
	shipp-2002-v1	22.3 (11.3)	17.0 (10.6)	10.2 (11.5)	10.4 (10.0)	8.9 (10.4)	5.4 (9.7)	20.7 (12.1)
	singh-2002	25.6 (12.4)	23.6 (12.7)	21.6 (12.2)	22.6 (14.1)	23.6 (13.5)	19.6 (12.3)	21.7 (11.5)
	su-2001	11.4 (7.5)	10.3 (7.3)	9.8 (5.5)	8.0 (2.9)	9.8 (5.5)	5.8 (4.7)	11.5 (6.2)
	west-2001	17.0 (22.1)	15.0 (20.7)	6.5 (10.6)	8.5 (11.1)	8.5 (14.5)	6.5 (10.6)	16.5 (12.9)
	yeoh-2002-v1	0.0 (0.0)	0.0 (0.0)	10.1 (6.4)	11.7 (6.9)	10.1 (5.1)	4.0 (2.7)	15.7 (3.4)
	yeoh-2002-v2	21.8 (5.3)	19.4 (7.6)	40.3 (6.8)	37.1 (6.8)	37.1 (7.0)	26.2 (10.1)	40.3 (7.4)
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

destacam, produzindo os menores erros médios: Jackknife, Pearson e Rank-Magnitude. Quando as bases de dados Affymetrix são consideradas tal fato não se repete, uma vez que as medidas de Pearson e Jackknife produziram resultados inferiores aos observados com a utilização de correlações baseadas em *ranks*. Para este conjunto de bases de dados os melhores resultados foram observados com a utilização da correlação Rank-Magnitude.

Conclusões

Nesta dissertação foram estudados e comparados coeficientes de correlação como medidas de proximidade para dados de expressão gênica. As comparações conduzidas envolveram quatro cenários comumente encontrados na análise deste tipo de dados: agrupamento de amostras, agrupamento de genes, seleção de genes e classificação de amostras. Além dos coeficientes de correlação propriamente ditos, três medidas também foram avaliadas no cenário de agrupamento de genes, cenário para o qual foram especificamente propostas.

No Capítulo 5, comparações envolvendo a tarefa de agrupamento de dados foram apresentadas. Em ambos os cenários de avaliação, agrupamento de amostras e agrupamento de genes, os coeficientes de correlação de Pearson, Jackknife e Rank-Magnitude se mostraram, no geral, superiores aos demais. Dessa forma, a correlação Rank-Magnitude, proposta recentemente, torna-se uma alternativa interessante adicional às demais medidas geralmente utilizadas. Quanto às medidas baseadas em *ranks* é interessante notar que as correlações de Kendall e Goodman-Kruskal, raramente utilizadas na literatura de dados de expressão gênica, produziram resultados comparáveis e, em alguns casos até superiores, aos da correlação de Spearman, comumente empregada. Considerando somente o cenário de agrupamento de genes, os piores desempenhos no geral foram obtidos com a utilização das medidas propostas especificamente para este cenário. A avaliação realizada no Capítulo 5 permitiu, ainda, a obtenção de resultados referentes ao desempenho de algoritmos de agrupamento. Neste sentido, observou-se, em ambos os cenários avaliados, que o algoritmo *k-medoids* apresentou, em geral, desempenhos superiores em relação aos demais. Com relação aos algoritmos de agrupamento hierárquicos observou-se um desempenho significativamente inferior por parte do algoritmo *Single-Linkage*. Este resultado reforça os resultados já reportados na literatura de dados de expressão gênica (D'haeseleer, 2005; de Souto et al., 2008), sendo que uma maior quantidade de medidas de proximidade foram

consideradas durante a presente avaliação.

No Capítulo 6 foram apresentados os resultados obtidos a partir da comparação dos coeficientes de correlação nas tarefas de seleção de genes e classificação de amostras. Para a tarefa de seleção de genes, a aplicação do método SSF proporcionou a obtenção de maiores erros de classificação, se estes forem comparados aos obtidos com o uso dos classificadores sem a prévia seleção de genes. Acredita-se que tais resultados tenham sido observados em virtude da quantidade muito reduzida de atributos selecionados pelo método. Considerando somente os resultados obtidos com a aplicação da seleção de genes, em bases de cDNA, as diferentes medidas avaliadas produziram resultados muito similares. Já para as bases de dados provenientes de experimentos Affymetrix a correlação Jackknife apresentou, no geral, os melhores resultados. Em relação aos resultados obtidos para a tarefa de classificação de amostras, sem prévia seleção de genes, mesmo sem grandes diferenças entre os coeficientes de correlação comparados, de forma geral, para bases de dados cDNA as correlações de Pearson, Jackknife e Rank-Magnitude levaram aos menores erros de classificação. Para as bases de dados Affymetrix, os menores erros de classificação foram produzidos pela correlação Rank-Magnitude. Ainda quanto a este último tipo de bases de dados é interessante notar que as correlações baseadas em *ranks* produziram melhores resultados que as correlações de Pearson e Jackknife. Ainda considerando os experimentos de classificação, em ambos os tipos de bases de dados a utilização da correlação de Goodman-Kruskal Ponderado levou à produção dos piores resultados.

De uma maneira geral não houve um coeficiente de correlação superior aos demais em todos os casos de avaliação considerados, fato este de certa forma esperado. Neste sentido, para uma aplicação em particular a avaliação prévia de diversas medidas de correlação parece ser a estratégia mais apropriada. Quando uma avaliação preliminar mostra-se inviável, a escolha de uma dentre as correlações de Pearson, Jackknife e Rank-Magnitude aparenta ser recomendável, uma vez que estas medidas produziram resultados superiores aos das demais em grande parte dos experimentos realizados. Considerando as medidas baseadas em *ranks*, as correlações de Goodman-Kruskal e Kendall podem levar à obtenção de resultados superiores aos da correlação de Spearman, fato que sugere que tais medidas são também interessantes quando uma análise preliminar mostra-se viável.

Por fim, além da contribuição obtida com o estudo e comparação de coeficientes de correlação como medidas de proximidade em dados de expressão gênica, o presente trabalho contribui também com a revisão e apresentação de conceitos básicos para a compreensão da tecnologia de *microarray* como um todo. A revisão realizada teve início com a discussão de conceitos biológicos fundamentais apresentada no Capítulo 2. Após tal discussão, as duas tecnologias de *microarray* comumente utilizadas para a realização de experimentos de expressão gênica foram apresentadas no Capítulo 3. Ainda neste capítulo, foi levantada uma discussão envolvendo as três principais tarefas aplicadas na análise destes dados, para as quais os coeficientes de correlação foram finalmente avaliados e comparados.

Referências Bibliográficas

- Aha, D. W. e Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1):37–66. Citado na página 79.
- Ambroise, C. e McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10):6562–6566. Citado na página 83.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., e Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29. Citado na página 65.
- Balasubramaniyan, R., Hullermeier, E., Weskamp, N., e Kamper, J. (2005). Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, 21(7):1069–1077. Citado nas páginas 3 e 34.
- Beißbarth, T. e Speed, T. P. (2004). Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465. Citado nas páginas 65 e 66.
- Belk, C. e Borden, V. (2003). *Biology: Science for Life*. Prentice Hall. Citado na página 13.
- Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., e Yakhini, Z. (2000). Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7(3-4):559–583. Citado na página 31.
- Ben-Dor, A. e Yakhini, Z. (1999). Clustering gene expression patterns. In *RECOMB '99: Proceedings of the third annual international conference on Computational molecular biology*, pages 33–42, New York, NY, USA. ACM. Citado na página 26.

- Bertone, P. e Gerstein, M. (2001). Integrative data mining: The new direction in bioinformatics - machine learning for analyzing genome-wide expression profiles and proteomics data sets. *IEEE Engineering in Medicine and Biology*, 20(4):33–40. Citado nas páginas 2 e 22.
- Bhaskar, H., Hoyle, D. C., e Singh, S. (2006). Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in Biology and Medicine*, 36(10):1104 – 1125. Intelligent Technologies in Medicine and Bioinformatics. Citado nas páginas 1 e 14.
- Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer. Citado na página 48.
- Blum, A. L. e Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271. Citado nas páginas 28 e 29.
- Bolshakova, N. e Azuaje, F. (2003). Cluster validation techniques for genome expression data. *Signal Processing*, 83(4):825 – 833. Citado nas páginas 51 e 65.
- Boulesteix, A.-L., Strobl, C., Augustin, T., e Daumer, M. (2008). Evaluating microarray-based classifiers: An overview. *Cancer Informatics*, 6:77–97. Citado na página 83.
- Braga-Neto, U. M. e Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380. Citado nas páginas 82 e 83.
- Brazma, A. e Vilo, J. (2000). Gene expression data analysis. *FEBS Letters*, 480(1):17 – 24. Citado nas páginas 1 e 22.
- Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J.-i., e Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, 18:630–634. Citado na página 13.
- Campello, R. J. G. B. e Hruschka, E. R. (2009). On comparing two sequences of numbers and its applications to clustering analysis. *Information Sciences*, 179(8):1025–1039. Citado nas páginas xxiii, 3, 35, 38, 40, 41, e 42.
- Chinchilli, V. M. e Gruemer, H. D. (1995). The correlation coefficient in the interpretation of laboratory data. *Clinica Chimica Acta*, 236(1):115 – 117. Citado na página 38.
- Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., e Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, 282(5389):699–705. Citado nas páginas 67 e 68.

- Clarkson, K. L. (2006). Nearest-neighbor searching and metric space dimensions. In *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, pages 15–59. MIT Press. Citado na página 42.
- Costa, I. G. (2003). Comparative analysis of clustering methods for gene expression data. Dissertação de Mestrado, Universidade Federal de Pernambuco. Citado nas páginas 26 e 28.
- Costa, I. G., de Carvalho, F. A. T., e de Souto, M. C. P. (2002). Comparative study on proximity indices for cluster analysis of gene expression time series. *Journal of Intelligent & Fuzzy Systems*, 13(2-4):133–142. Citado nas páginas 3, 27, e 35.
- Costa, I. G., de Souto, M. C. P., e Schliep, A. (2007). Validating gene clusterings by selecting informative gene ontology terms with mutual information. In *BSB*, pages 81–92. Citado na página 67.
- Costa, I. G., Roepcke, S., Hafemeister, C., e Schliep, A. (2008). Inferring differentiation pathways from gene expression. *Bioinformatics*, 24(13):i156–i164. Citado nas páginas 65 e 66.
- Covões, T. F. (2010). Seleção de atributos via agrupamento. Dissertação de Mestrado, Universidade de São Paulo. Citado na página 79.
- Covões, T. F., Hruschka, E. R., Castro, L. N., e Santos, A. M. (2009). A cluster-based feature selection approach. In *HAIS '09: Proceedings of the 4th International Conference on Hybrid Artificial Intelligence Systems*, pages 169–176, Berlin, Heidelberg. Springer-Verlag. Citado nas páginas 30, 78, e 79.
- Cover, T. e Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21 – 27. Citado na página 79.
- Dalma-Weiszhausz, D. D., Warrington, J., Tanimoto, E. Y., e Miyada, C. G. (2006). The affymetrix genechip platform: an overview. *Methods in Enzymology*, 410:3–28. Citado nas páginas 18, 19, e 21.
- Datta, S. e Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459–466. Citado nas páginas 27 e 28.
- Datta, S. e Datta, S. (2006). Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, 7(1):397. Citado na página 67.
- De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B., e Moreau, Y. (2002). Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, 18(5):735–746. Citado na página 27.

- de Souto, M., Costa, I., de Araujo, D., Ludermir, T., e Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 9(1):497. Citado nas páginas xxiii, 3, 27, 28, 35, 52, 53, 62, 68, 84, e 95.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30. Citado nas páginas 51 e 82.
- D’haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology*, 23:1499–1501. Citado nas páginas 28, 74, e 95.
- Ding, C. e Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205. Citado nas páginas 29, 30, e 33.
- Dougherty, E. R. (2001). Small sample issues for microarray-based classification. *Comparative and Functional Genomics*, 2(1):28–34. Citado nas páginas 82 e 83.
- Dougherty, E. R., Sima, C., Hua, J., Hanczar, B., e Braga-Neto, U. M. (2010). Performance of error estimators for classification. *Current Bioinformatics*, 5:53–67. Citado na página 83.
- Dudoit, S., Fridlyand, J., e Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87. Citado nas páginas 2, 29, 31, 32, e 33.
- Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., e Trent, J. M. (1999). Expression profiling using cdna microarrays. *Nature Genetics*, 21:10–14. Citado na página 15.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26. Citado na página 82.
- Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial & Applied Mathematics. Citado na página 43.
- Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331. Citado nas páginas 82 e 83.
- Eisen, M. B., Spellman, P. T., Brown, P. O., e Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868. Citado nas páginas 13, 27, e 33.
- Ernst, J., Nau, G. J., e Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, 21(suppl 1):i159–i168. Citado nas páginas 64, 65, e 66.

- Estelberger, W. e Reibnegger, G. (1995). The rank correlation coefficient: an additional aid in the interpretation of laboratory data. *Clinica Chimica Acta*, 239(2):203–207. Citado na página 39.
- Esteves, G. H. (2007). *Métodos Estatísticos para a Análise de Dados de cDNA microarray em um Ambiente Computacional Integrado*. PhD thesis, Universidade de São Paulo. Citado na página 21.
- Everitt, B., Landau, S., e Leese, M. (2001). *Cluster analysis*. Oxford University Press, 4 edition. Citado nas páginas 25 e 28.
- Faceli, K., Carvalho, A. A. C. d., e Silva Jr, W. A. (2004). Evaluation of gene selection metrics for tumor cell classification. *Genetics and Molecular Biology*, 27:651 – 657. Citado nas páginas 52 e 68.
- Freeman, W. M., Robertson, D. J., e Vrana, K. E. (2000). Fundamentals of DNA hybridization arrays for gene expression analysis. *Biotechniques*, 29(5):1042–1055. Citado na página 16.
- Fujita, A., Sato, J. R., Demasi, M. A. A., Sogayar, M. C., Ferreira, C. E., e Miyano, S. (2009). Comparing pearson, spearman and hoeffding’s d measure for gene expression association analysis. *Journal of Bioinformatics and Computational Biology*, 7(4):663–684. Citado nas páginas 3, 39, e 40.
- Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., e Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11(12):4241–4257. Citado nas páginas 67 e 68.
- Gibbons, F. D. e Roth, F. P. (2002). Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Research*, 12(10):1574–1581. Citado nas páginas 27 e 67.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., e Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537. Citado nas páginas 13, 26, 29, e 30.
- Goodman, L. e Kruskal, W. (1954). Measures of association for cross-classifications. *Journal of the American Statistical Association*, 49:732–764. Citado na página 40.
- Guyon, I. e Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182. Citado nas páginas 28 e 29.

- Guyon, I., Weston, J., Barnhill, S., e Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422. Citado nas páginas 29 e 30.
- Hardin, J., Mitani, A., Hicks, L., e VanKoten, B. (2007). A robust measure of correlation between two genes on a microarray. *BMC Bioinformatics*, 8(1):220. Citado nas páginas 2, 34, e 35.
- Harrington, C. A., Rosenow, C., e Retief, J. (2000). Monitoring gene expression using dna microarrays. *Current Opinion in Microbiology*, 3(3):285 – 291. Citado nas páginas xxi, 13, 15, 16, e 19.
- Heyer, L. J., Kruglyak, S., e Yooseph, S. (1999). Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, 9(11):1106–1115. Citado nas páginas 2, 3, 22, 24, 26, 27, 32, 34, 39, 42, e 43.
- Hruschka, E. R., Campello, R. J., e de Castro, L. N. (2006). Evolving clusters in gene-expression data. *Information Sciences*, 176(13):1898 – 1927. Citado na página 78.
- Hua, Y.-J., Tu, K., Tang, Z.-Y., Li, Y.-X., e Xiao, H.-S. (2008). Comparison of normalization methods with microRNA microarray. *Genomics*, 92(2):122 – 128. Citado na página 23.
- Hubert, L. e Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218. Citado na página 50.
- Human Genome Program (1992). *Primer on Molecular Genetics*. U.S. Department of Energy, Washington, D.C. Citado nas páginas xxi, 6, 7, 9, e 11.
- Jain, A. K. e Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. Citado nas páginas 2, 25, 27, 28, 34, 47, e 50.
- Jaskowiak, P. A., Campello, R. J. G. B., Covões, T. F., e Hruschka, E. R. (2010). A comparative study on the use of correlation coefficients for redundant feature elimination. In *11th Brazilian Symposium on Neural Networks (SBRN 2010)*, São Bernardo do Campo, São Paulo, Brazil, October 23-28, 2010, pages 13–18. Citado na página 79.
- Ji, X., Li-Ling, J., e Sun, Z. (2003). Mining gene expression data using a novel approach based on hidden markov models. *FEBS Letters*, 542(1-3):125 – 131. Citado na página 1.
- Jiang, D., Tang, C., e Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386. Citado nas páginas 3, 22, 23, 26, 27, 28, e 33.
- Kaufman, L. e Rousseeuw, P. J. (1990). *Finding Groups in Data: an introduction to cluster analysis*. Wiley. Citado nas páginas 51, 65, e 78.

- Kendall, M. G. (1970). *Rank correlation methods [by] Maurice G. Kendall*. Griffin, London, 4 edition. Citado nas páginas 39 e 40.
- Kerr, G., Ruskin, H. J., Crane, M., e Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in Biology and Medicine*, 38(3):283–293. Citado nas páginas 26, 27, 33, e 39.
- Kim, K., Zhang, S., Jiang, K., Cai, L., Lee, I.-B., Feldman, L., e Huang, H. (2007). Measuring similarities between gene expression profiles through new data transformations. *BMC Bioinformatics*, 8(1):29. Citado nas páginas 3, 34, e 35.
- Kotlyar, M., Fuhrman, S., Ableson, A., e Somogyi, R. (2002). Spearman correlation identifies statistically significant gene expression clusters in spinal cord development and injury. *Neurochemical Research*, 27(10):1133–1140. Citado nas páginas 3, 34, e 40.
- Krause, R., von Mering, C., Bork, P., e Dandekar, T. (2004). Shared components of protein complexes—versatile building blocks or biochemical artefacts? *BioEssays*, 26(12):1333–1343. Citado na página 32.
- Kuo, W. P., Kim, E.-Y., Trimarchi, J., Jenssen, T.-K., Vinterbo, S. A., e Ohno-Machado, L. (2004). A primer on gene expression and microarrays for machine learning researchers. *Journal of Biomedical Informatics*, 37(4):293 – 303. Citado nas páginas 13, 14, 16, 22, e 24.
- Kuramochi, M. e Karypis, G. (2005). Gene classification using expression profiles: A feasibility study. *International Journal on Artificial Intelligence Tools*, 14(4):641–660. Citado na página 31.
- Langley, P. (1994). Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance*, pages 140–144. AAAI Press. Citado na página 29.
- Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., e Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 94(24):13057–13062. Citado nas páginas xxi e 18.
- Lee, J. W., Lee, J. B., Park, M., e Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869 – 885. Citado nas páginas 24, 31, e 32.
- Lesk, A. M. (2008). *Introduction to Bioinformatics*. Oxford University Press, 3 edition. Citado na página 5.

- Li, T., Zhang, C., e Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437. Citado nas páginas 29 e 30.
- Liu, H. e Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502. Citado nas páginas 28, 29, e 31.
- Liu, Q., Sung, A., Xu, J., Liu, J., e Chen, Z. (2006). Microarray gene expression classification based on supervised learning and similarity measures. In *Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on*, volume 6, pages 5094–5099. Citado na página 29.
- Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., e Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680. Citado nas páginas 17, 18, e 19.
- Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, L., e Darnell, J. (2003). *Molecular Cell Biology*. W. H. Freeman, 5 edition. Citado nas páginas xxiii, 5, 6, 7, 8, 10, 11, e 13.
- Lu, Y. e Han, J. (2003). Cancer classification using gene expression data. *Information Systems*, 28(4):243–268. Citado na página 31.
- MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. e Neyman, J., editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press. Citado na página 26.
- Milligan, G. W. e Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21(4):441–458. Citado na página 50.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math. Citado na página 28.
- Mitra, P., Murthy, C. A., e Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312. Citado na página 30.
- Molinaro, A. M., Simon, R., e Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307. Citado na página 83.

- Möller-Levet, C., Klawonn, F., Cho, K.-H., Yin, H., e Wolkenhauer, O. (2005). Clustering of unevenly sampled gene expression time-series data. *Fuzzy Sets and Systems*, 152(1):49 – 66. Citado nas páginas 34, 35, 45, 46, e 68.
- Mount, D. W. (2004). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, 2 edition. Citado na página 5.
- Orchard, M. (1991). A fast nearest-neighbor search algorithm. In *International Conference on Acoustics, Speech, and Signal Processing, 1991. ICASSP-91.*, volume 4, pages 2297 –2300. Citado na página 79.
- Pakhira, M. K., Bandyopadhyay, S., e Maulik, U. (2005). A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. *Fuzzy Sets and Systems*, 155(2):191 – 214. Citado nas páginas 57, 60, e 69.
- Pal, N. R. e Bezdek, J. C. (1995). On cluster validity for the fuzzy c-means model. *IEEE Trans. on Fuzzy Systems*, 3(3):370–379. Citado nas páginas 57, 60, e 69.
- Pearson, K. (1895). Contributions to the mathematical theory of evolution. iii. regression, heredity, and panmixia. *Proceedings of the Royal Society of London*, 59:69–71. Citado na página 38.
- Piatetsky-Shapiro, G. e Tamayo, P. (2003). Microarray data mining: facing the challenges. *SIGKDD Explorations Newsletter*, 5(2):1–5. Citado nas páginas 14, 24, e 77.
- Qin, Z. S. (2006). Clustering microarray gene expression data using weighted chinese restaurant process. *Bioinformatics*, 22(16):1988–1997. Citado na página 68.
- Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3:1371–1382. Citado na página 81.
- Romualdi, C., Campanaro, S., Campagna, D., Celegato, B., Cannata, N., Toppo, S., Valle, G., e Lanfranchi, G. (2003). Pattern recognition in gene expression profiling using dna array: a comparative study of different statistical methods applied to cancer classification. *Hum. Mol. Genet.*, 12(8):823–836. Citado na página 83.
- Saeys, Y., Inza, I., e Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517. Citado na página 29.
- Schena, M., Shalon, D., Davis, R. W., e Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470. Citado na página 15.
- Setubal, J. e Meidanis, J. (1997). *Introduction to computational molecular biology*. PWS Publishing Company. Citado nas páginas xxi, 5, 6, 7, 8, e 9.

- Sherlock, G. (2001). Analysis of large-scale gene expression data. *Briefings in Bioinformatics*, 2(4):350–362. Citado nas páginas 1, 21, 22, 23, e 24.
- Silverman, B. W. e Jones, M. C. (1989). E. fix and j.l. hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on fix and hodges (1951). *International Statistical Review*, 57(3):pp. 233–238. Citado na página 79.
- Smyth, G. K. e Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, 31(4):265 – 273. Citado nas páginas 21 e 23.
- Son, Y. S. e Baek, J. (2008). A modified correlation coefficient based similarity measure for clustering time-course gene expression data. *Pattern Recognition Letters*, 29(3):232 – 242. Citado nas páginas 3, 34, 35, 43, e 45.
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471. Citado na página 39.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., e Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297. Citado nas páginas 67 e 68.
- Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., e Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643. Citado na página 83.
- Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59:1–34. Citado nas páginas 26, 47, e 48.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., e Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6):2907–2912. Citado na página 68.
- Tan, P.-N., Steinbach, M., e Kumar, V. (2005). *Introduction to Data Mining*. Addison Wesley, 1 edition. Citado nas páginas 23, 30, 31, 79, e 81.
- Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R., e Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Computational Biology*, 3(6):e116. Citado nas páginas 17 e 22.
- Tarca, A. L., Romero, R., e Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology*, 195(2):373 – 388. Citado nas páginas 14, 15, e 16.

- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., e Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature genetics*, 22(3):281–285. Citado na página 34.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., e Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525. Citado nas páginas 23 e 24.
- Tuikkala, J., Elo, L., Nevalainen, O., e Aittokallio, T. (2008). Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC Bioinformatics*, 9(1):202. Citado nas páginas 23 e 24.
- VanGuilder, H. D., Vrana, K. E., e Freeman, W. M. (2008). Twenty-five years of quantitative pcr for gene expression analysis. *BioTechniques*, 44(5):619–626. Citado nas páginas 13 e 14.
- Velculescu, V. E., Zhang, L., Vogelstein, B., e Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270(5235):484–487. Citado na página 13.
- Vendramin, L., Campello, R. J. G. B., e Hruschka, E. R. (2009). On the comparison of relative clustering validity criteria. In *SIAM International Conference on Data Mining*, pages 733–744, Sparks/USA. Citado nas páginas 51 e 65.
- Vendramin, L., Campello, R. J. G. B., e Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3:209–235. Citado nas páginas 51, 57, e 65.
- Wilson, K. E., Ryan, M. M., Prime, J. E., Pashby, D. P., Orange, P. R., OâBeirne, G., Whateley, J. G., Bahn, S., e Morris, C. M. (2004). Functional genomics and proteomics: application in neurosciences. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(4):529–538. Citado nas páginas xxi e 20.
- Witten, I. H. e Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2 edition. Citado na página 81.
- Wu, X., Kumar, V., Ross, Ghosh, J., Yang, Q., Motoda, H., Mclachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z.-H., Steinbach, M., Hand, D., e Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37. Citado na página 48.
- Yang, A., Hsu, H.-H., e Lu, M.-D. (2009). Outlier filtering for identification of gene regulations in microarray time-series data. In *Complex, Intelligent and Software Intensive Systems, 2009. CISIS '09. International Conference on*, pages 854–859. Citado na página 22.

- Yang, C., Wan, B., e Gao, X. (2006). Effectivity of internal validation techniques for gene clustering. In Maglaveras, N., Chouvarda, I., Koutkias, V., e Brause, R., editors, *Biological and Medical Data Analysis*, volume 4345, pages 49–59. Springer Berlin / Heidelberg. Citado nas páginas 51 e 65.
- Ye, L., Wang, X., Keogh, E. J., e Mafra-Neto, A. (2009). Autocannibalistic and anyspace indexing algorithms with application to sensor data mining. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*, pages 85–96. Citado na página 79.
- Yeung, K. Y., Haynor, D. R., e Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318. Citado nas páginas 26, 27, e 28.
- Yu, L. e Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of The Twentieth International Conference on Machine Learning (ICML-03)*. Citado na página 30.
- Zhang, A. (2006). *Advanced analysis of gene expression microarray data*. World Scientific Publishing Company, 1 edition. Citado nas páginas xxi, 1, 2, 5, 7, 8, 10, 11, 15, 18, 21, 22, 26, 28, 33, 38, 39, e 43.

Tabelas dos Experimentos de Agrupamento de Amostras

No presente apêndice são apresentados os resultados completos obtidos a partir da realização de experimentos em agrupamento amostras. Na Seção A.1 são apresentadas as Tabelas A.1 à A.4, as quais contém os valores de Rand Ajustado para cada algoritmo, coeficiente de correlação e base de dados utilizadas para o cenário no qual o número de grupos é fixo. Na Seção A.2 são apresentadas nas Tabelas A.5 à A.8 os valores de Rand Ajustado para cada algoritmo, medida de correlação e base de dados utilizadas para o cenário no qual o número de grupos é variável. Adicionalmente, são apresentados nas Tabelas A.9 à A.12 os números de grupos encontrados para este cenário. Finalmente, na Seção A.3 são apresentados os valores de Rand Ajustado (Tabelas A.13 à A.16) e de número de grupos (Tabelas A.17 à A.20), para cada algoritmo, coeficiente de correlação e base de dados, considerando o cenário no qual o número de grupos foi estimado com a Silhueta. Nas tabelas em que são apresentados os números de grupos obtidos, a coluna k^* denota o número de grupos da partição de referência.

A.1 Número de Grupos Fixo

Tabela A.1: Valores de Rand Ajustado para o algoritmo *Single-Linkage* com k fixo.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	alizadeh-2000-v2	-0.04	-0.04	0.79	0.79	0.79	0.79	0.79
	alizadeh-2000-v3	0.41	0.41	0.41	0.41	0.41	0.41	0.41
	bittner-2000	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	bredel-2005	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
	chen-2002	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	garber-2001	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
	khan-2001	0.20	0.20	0.01	-0.01	0.01	0.00	0.02
	lapointe-2004-v1	0.02	0.01	0.04	0.04	0.04	0.01	0.04
	lapointe-2004-v2	0.03	0.00	0.01	0.01	0.01	0.03	0.07
	liang-2005	0.16	0.16	0.16	0.16	0.16	0.16	0.16
	risinger-2003	0.08	0.12	0.03	0.03	0.03	0.03	0.03
	tomlins-2006	0.00	0.00	0.01	0.01	0.01	0.01	0.01
	tomlins-2006-v2	0.00	0.00	0.02	0.00	0.00	0.00	0.00
Affymetrix	armstrong-2002-v1	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.03
	armstrong-2002-v2	-0.01	-0.01	0.00	-0.00	-0.00	-0.01	0.01
	bhattacharjee-2001	-0.01	0.02	0.32	0.35	0.35	0.33	0.35
	chowdary-2006	0.02	0.02	0.92	0.92	0.92	0.92	0.71
	dyrskjot-2003	-0.06	-0.06	0.05	0.05	0.05	0.05	0.04
	golub-1999-v1	0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	golub-1999-v2	-0.00	-0.02	-0.02	-0.00	-0.00	-0.02	-0.02
	gordon-2002	-0.01	-0.01	-0.01	-0.01	-0.01	0.04	-0.01
	laiho-2007	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
	nuttt-2003-v1	-0.00	-0.00	-0.00	-0.01	-0.01	-0.00	0.01
	nuttt-2003-v2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	nuttt-2003-v3	-0.05	-0.05	0.10	0.10	0.10	0.10	-0.05
	pomeroy-2002-v1	0.10	0.10	-0.01	-0.01	-0.01	0.10	-0.01
	pomeroy-2002-v2	-0.00	0.00	0.03	0.00	0.00	0.01	0.00
	ramaswamy-2001	0.02	0.05	0.05	0.01	0.05	0.05	0.00
	shipp-2002-v1	0.05	-0.02	-0.02	0.05	-0.02	0.05	-0.02
	singh-2002	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00
	su-2001	0.01	0.01	0.14	0.13	0.13	0.14	0.16
	west-2001	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00
	yeoh-2002-v1	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	yeoh-2002-v2	-0.00	-0.00	0.01	0.00	0.01	-0.00	0.01
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.2: Valores de Rand Ajustado para o algoritmo *Average-Linkage* com k fixo.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.00	0.37	0.43	0.43	0.43	0.57	0.09
	alizadeh-2000-v2	0.79	0.79	0.73	0.73	0.79	0.79	0.79
	alizadeh-2000-v3	0.40	0.40	0.29	0.29	0.32	0.31	0.32
	bittner-2000	0.02	0.02	-0.03	-0.02	-0.02	-0.03	0.02
	bredel-2005	0.22	0.27	0.36	0.36	0.36	0.36	0.36
	chen-2002	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	garber-2001	0.30	0.30	0.26	0.26	0.26	0.29	0.26
	khan-2001	0.46	0.46	0.00	0.00	0.00	0.21	0.03
	lapointe-2004-v1	0.01	0.01	0.12	0.12	0.12	0.12	0.12
	lapointe-2004-v2	0.01	0.08	0.37	0.37	0.37	0.29	0.04
	liang-2005	0.16	0.16	0.16	0.16	0.16	0.16	0.16
	risinger-2003	0.10	0.20	0.05	0.05	0.06	0.08	0.11
	tomlins-2006	0.29	0.29	0.26	0.26	0.29	0.30	0.25
	tomlins-2006-v2	0.22	0.22	0.23	0.23	0.23	0.22	0.23
Affymetrix	armstrong-2002-v1	0.21	0.24	0.27	0.27	0.27	0.24	0.89
	armstrong-2002-v2	0.54	0.56	0.59	0.56	0.61	0.56	0.92
	bhattacharjee-2001	0.33	0.32	0.41	0.48	0.47	0.35	0.41
	chowdary-2006	0.78	0.92	0.71	0.92	0.71	0.92	0.71
	dyrskjot-2003	0.11	-0.03	0.27	0.22	0.27	0.27	0.54
	golub-1999-v1	-0.05	-0.05	-0.02	-0.02	-0.02	-0.06	0.08
	golub-1999-v2	0.83	0.88	0.14	0.14	0.14	0.90	0.14
	gordon-2002	-0.04	-0.03	0.90	0.92	0.92	-0.02	0.92
	laiho-2007	0.08	-0.04	-0.11	-0.04	-0.04	-0.04	-0.09
	nutt-2003-v1	-0.00	-0.00	0.10	0.08	0.08	0.09	-0.01
	nutt-2003-v2	0.00	0.00	-0.03	-0.03	-0.03	0.00	0.00
	nutt-2003-v3	-0.05	-0.05	-0.02	-0.06	-0.05	-0.09	-0.06
	pomeroy-2002-v1	0.24	-0.01	-0.01	-0.01	0.11	-0.01	-0.01
	pomeroy-2002-v2	0.42	0.42	0.41	0.43	0.43	0.47	0.24
	ramaswamy-2001	0.38	0.34	0.43	0.31	0.40	0.37	0.03
	shipp-2002-v1	0.18	0.05	0.10	0.18	0.13	0.17	-0.10
	singh-2002	0.03	0.03	0.04	0.04	0.04	0.02	0.05
	su-2001	0.40	0.40	0.65	0.66	0.67	0.46	0.67
	west-2001	0.56	0.56	0.00	0.00	0.00	-0.00	0.00
	yeoh-2002-v1	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01
	yeoh-2002-v2	0.19	0.31	-0.01	0.01	-0.01	-0.01	0.01
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.3: Valores de Rand Ajustado para o algoritmo *Complete-Linkage* com k fixo.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.01	0.16	-0.02	-0.02	-0.02	0.13	0.13
	alizadeh-2000-v2	0.45	0.45	0.41	0.41	0.43	0.38	0.41
	alizadeh-2000-v3	0.27	0.27	0.24	0.24	0.25	0.33	0.24
	bittner-2000	0.53	0.53	-0.02	0.11	-0.00	-0.02	-0.01
	bredel-2005	0.38	0.38	0.29	0.29	0.29	0.26	0.23
	chen-2002	0.03	0.29	0.03	0.03	0.55	0.33	-0.00
	garber-2001	0.09	0.09	0.20	0.20	0.20	0.13	0.21
	khan-2001	0.69	0.57	0.00	0.00	0.00	0.19	0.04
	lapointe-2004-v1	0.12	0.00	0.12	0.12	0.12	0.16	0.15
	lapointe-2004-v2	-0.01	0.09	0.15	0.15	0.11	0.04	0.05
	liang-2005	0.16	0.16	0.16	0.16	0.16	0.16	0.16
	risinger-2003	0.10	0.10	0.13	0.13	0.11	0.13	0.08
	tomlins-2006	0.23	0.24	0.25	0.25	0.25	0.19	0.27
	tomlins-2006-v2	0.21	0.21	0.17	0.20	0.20	0.18	0.20
Affymetrix	armstrong-2002-v1	0.18	0.18	0.18	0.24	0.24	0.24	0.84
	armstrong-2002-v2	0.49	0.57	0.88	0.81	0.64	0.80	0.56
	bhattacharjee-2001	0.33	0.32	0.43	0.20	0.20	0.52	0.26
	chowdary-2006	0.92	0.92	0.71	0.92	0.71	0.75	0.68
	dyrskjot-2003	0.11	0.11	0.29	0.33	0.33	0.30	0.36
	golub-1999-v1	0.40	0.69	0.64	0.40	0.02	0.48	-0.02
	golub-1999-v2	0.50	0.77	0.39	0.23	0.25	0.55	0.11
	gordon-2002	0.92	0.90	1.00	1.00	1.00	0.97	0.48
	laiho-2007	-0.02	0.02	-0.11	-0.11	-0.11	-0.11	-0.10
	nuttt-2003-v1	0.05	0.05	0.02	0.07	0.09	-0.01	0.06
	nuttt-2003-v2	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
	nuttt-2003-v3	0.02	0.03	-0.02	0.03	0.03	0.03	0.17
	pomeroy-2002-v1	0.29	-0.01	0.11	0.11	0.11	-0.01	-0.01
	pomeroy-2002-v2	0.56	0.61	0.43	0.37	0.40	0.41	0.41
	ramaswamy-2001	0.42	0.43	0.50	0.61	0.61	0.43	0.25
	shipp-2002-v1	0.01	-0.10	0.06	0.07	0.16	0.15	0.07
	singh-2002	0.04	0.02	0.02	0.02	0.02	0.02	0.02
	su-2001	0.42	0.47	0.59	0.60	0.46	0.54	0.70
	west-2001	0.56	0.63	0.02	0.01	0.01	0.63	-0.01
	yeoh-2002-v1	-0.08	0.08	-0.11	0.03	0.05	0.17	-0.01
	yeoh-2002-v2	0.27	0.34	0.07	0.02	0.01	0.22	-0.01
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.4: Valores de Rand Ajustado para o algoritmo *k-medoids* com *k* fixo.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.50	0.43	0.43	0.37	0.37	0.43	0.37
	alizadeh-2000-v2	0.79	0.79	0.80	0.80	0.80	0.84	0.80
	alizadeh-2000-v3	0.56	0.62	0.62	0.62	0.51	0.62	0.53
	bittner-2000	0.53	0.45	0.38	0.38	0.38	0.45	0.08
	bredel-2005	0.45	0.45	0.46	0.42	0.42	0.42	0.44
	chen-2002	0.60	0.60	0.62	0.62	0.62	0.64	0.69
	garber-2001	0.33	0.38	0.33	0.33	0.36	0.33	0.35
	khan-2001	0.77	0.89	0.40	0.33	0.29	0.65	0.40
	lapointe-2004-v1	0.24	0.16	0.22	0.21	0.22	0.22	0.19
	lapointe-2004-v2	0.29	0.31	0.30	0.30	0.30	0.32	0.34
	liang-2005	0.20	0.16	0.16	0.16	0.16	0.16	0.16
	risinger-2003	0.25	0.25	0.20	0.20	0.47	0.35	0.25
	tomlins-2006	0.37	0.44	0.38	0.34	0.37	0.35	0.34
	tomlins-2006-v2	0.37	0.43	0.32	0.27	0.30	0.33	0.30
Affymetrix	armstrong-2002-v1	0.64	0.30	0.52	0.33	0.52	0.56	0.60
	armstrong-2002-v2	0.81	0.81	0.89	0.89	0.88	0.88	0.70
	bhattacharjee-2001	0.56	0.54	0.40	0.58	0.39	0.57	0.37
	chowdary-2006	0.85	0.81	0.78	0.78	0.78	0.78	0.78
	dyrskjot-2003	0.53	0.62	0.51	0.59	0.52	0.40	0.69
	golub-1999-v1	0.89	0.69	0.30	0.48	0.30	0.69	0.48
	golub-1999-v2	0.89	0.81	0.61	0.55	0.63	0.87	0.48
	gordon-2002	0.82	0.82	0.83	0.85	0.85	0.92	0.83
	laiho-2007	0.38	0.32	0.36	0.51	0.42	0.24	0.24
	nuttt-2003-v1	0.20	0.19	0.22	0.19	0.17	0.17	0.14
	nuttt-2003-v2	-0.01	0.00	0.05	0.05	0.05	0.05	-0.02
	nuttt-2003-v3	0.06	0.03	0.09	0.27	0.09	0.17	0.27
	pomeroy-2002-v1	0.05	0.32	0.33	0.33	0.33	0.39	0.40
	pomeroy-2002-v2	0.62	0.58	0.58	0.56	0.56	0.52	0.59
	ramaswamy-2001	0.56	0.68	0.61	0.66	0.63	0.66	0.01
	shipp-2002-v1	0.39	0.58	0.24	0.21	0.21	0.39	0.02
	singh-2002	0.07	0.06	0.06	0.55	0.06	0.16	0.55
	su-2001	0.60	0.57	0.64	0.60	0.63	0.72	0.66
	west-2001	0.50	0.44	0.44	0.29	0.39	0.56	0.50
	yeoh-2002-v1	0.53	0.59	0.39	0.19	0.22	0.44	0.14
	yeoh-2002-v2	0.46	0.45	0.11	0.11	0.12	0.36	0.03
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

A.2 Número de Grupos Variável

Tabela A.5: Valores de Rand Ajustado para o algoritmo *Single-Linkage* com k variável.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.00	0.00	0.01	0.01	0.01	0.01	0.01
	alizadeh-2000-v2	0.75	0.75	0.79	0.79	0.79	0.79	0.79
	alizadeh-2000-v3	0.41	0.41	0.42	0.42	0.42	0.42	0.42
	bittner-2000	0.10	0.10	0.15	0.00	0.00	0.08	0.14
	bredel-2005	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
	chen-2002	-0.00	-0.01	-0.01	-0.01	-0.01	-0.00	-0.01
	garber-2001	0.26	0.28	0.28	0.28	0.28	0.28	0.30
	khan-2001	0.34	0.28	0.03	0.02	0.03	0.20	0.02
	lapointe-2004-v1	0.02	0.02	0.24	0.24	0.24	0.22	0.06
	lapointe-2004-v2	0.04	0.02	0.03	0.03	0.03	0.04	0.08
	liang-2005	0.23	0.23	0.17	0.17	0.17	0.23	0.23
	risinger-2003	0.10	0.12	0.10	0.10	0.14	0.14	0.14
	tomlins-2006	0.01	0.01	0.01	0.01	0.01	0.02	0.01
	tomlins-2006-v2	0.01	0.01	0.02	0.02	0.02	0.02	0.02
Affymetrix	armstrong-2002-v1	0.13	-0.01	0.26	0.21	0.39	-0.00	0.05
	armstrong-2002-v2	0.49	0.49	0.58	0.51	0.58	0.49	0.04
	bhattacharjee-2001	0.30	0.30	0.35	0.35	0.35	0.33	0.35
	chowdary-2006	0.90	0.66	0.92	0.92	0.92	0.92	0.84
	dyrskjot-2003	-0.01	0.03	0.14	0.14	0.14	0.09	0.45
	golub-1999-v1	0.41	0.56	0.01	0.01	0.01	-0.01	-0.01
	golub-1999-v2	0.50	0.41	-0.01	-0.00	-0.00	-0.01	-0.01
	gordon-2002	0.17	0.16	0.84	0.83	0.83	0.14	0.86
	laiho-2007	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
	nutt-2003-v1	-0.00	-0.00	-0.00	0.00	0.00	-0.00	0.01
	nutt-2003-v2	0.00	0.00	0.01	0.01	0.01	0.03	0.00
	nutt-2003-v3	-0.04	-0.05	0.10	0.10	0.10	0.10	0.04
	pomeroy-2002-v1	0.25	0.23	0.07	0.10	0.07	0.10	0.16
	pomeroy-2002-v2	0.15	0.18	0.03	0.01	0.02	0.03	0.01
	ramaswamy-2001	0.02	0.05	0.05	0.01	0.05	0.05	0.00
	shipp-2002-v1	0.05	0.07	-0.02	0.05	-0.02	0.08	-0.02
	singh-2002	0.05	0.06	0.03	0.01	0.01	0.04	0.04
	su-2001	0.03	0.03	0.18	0.14	0.14	0.14	0.20
	west-2001	0.02	0.54	0.01	0.00	0.01	0.02	-0.00
	yeoh-2002-v1	-0.01	-0.01	0.02	-0.01	0.02	-0.01	0.09
	yeoh-2002-v2	0.00	0.00	0.01	0.01	0.01	0.00	0.01
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.6: Valores de Rand Ajustado para o algoritmo *Average-Linkage* com k variável.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.20	0.37	0.43	0.43	0.43	0.57	0.20
	alizadeh-2000-v2	0.79	0.79	0.73	0.73	0.79	0.79	0.79
	alizadeh-2000-v3	0.42	0.42	0.40	0.40	0.42	0.42	0.42
	bittner-2000	0.18	0.17	0.07	0.06	0.09	0.18	0.06
	bredel-2005	0.37	0.43	0.36	0.36	0.36	0.37	0.36
	chen-2002	0.53	0.54	0.69	0.69	0.66	0.54	0.68
	garber-2001	0.37	0.37	0.30	0.30	0.30	0.33	0.30
	khan-2001	0.55	0.55	0.06	0.06	0.06	0.30	0.07
	lapointe-2004-v1	0.18	0.11	0.21	0.21	0.21	0.19	0.12
	lapointe-2004-v2	0.18	0.09	0.37	0.37	0.37	0.31	0.17
	liang-2005	0.26	0.26	0.26	0.26	0.26	0.26	0.26
	risinger-2003	0.18	0.20	0.09	0.09	0.14	0.18	0.15
	tomlins-2006	0.39	0.39	0.29	0.29	0.31	0.37	0.34
	tomlins-2006-v2	0.25	0.25	0.25	0.25	0.26	0.27	0.26
Affymetrix	armstrong-2002-v1	0.21	0.24	0.60	0.55	0.58	0.45	0.89
	armstrong-2002-v2	0.56	0.56	0.86	0.61	0.75	0.83	0.92
	bhattacharjee-2001	0.35	0.35	0.47	0.48	0.47	0.55	0.42
	chowdary-2006	0.78	0.92	0.85	0.92	0.85	0.92	0.85
	dyrskjot-2003	0.33	0.31	0.32	0.30	0.33	0.33	0.54
	golub-1999-v1	0.61	0.65	0.27	0.22	0.22	0.64	0.30
	golub-1999-v2	0.83	0.88	0.35	0.35	0.35	0.90	0.36
	gordon-2002	0.81	0.82	0.90	0.92	0.92	0.94	0.92
	laiho-2007	0.08	0.04	0.10	0.10	0.10	0.01	0.14
	nuttt-2003-v1	0.12	0.11	0.11	0.11	0.11	0.14	0.11
	nuttt-2003-v2	0.00	0.00	-0.02	-0.02	-0.02	0.00	0.00
	nuttt-2003-v3	-0.02	-0.05	-0.02	-0.02	-0.03	-0.02	0.03
	pomeroy-2002-v1	0.24	0.17	0.18	0.15	0.23	0.17	0.01
	pomeroy-2002-v2	0.49	0.44	0.52	0.52	0.52	0.49	0.26
	ramaswamy-2001	0.38	0.34	0.43	0.31	0.40	0.37	0.03
	shipp-2002-v1	0.18	0.37	0.14	0.22	0.18	0.17	0.03
	singh-2002	0.05	0.07	0.04	0.04	0.04	0.05	0.05
	su-2001	0.50	0.60	0.79	0.86	0.87	0.73	0.67
	west-2001	0.56	0.56	0.12	0.09	0.09	0.53	0.19
	yeoh-2002-v1	0.80	0.91	0.12	0.12	0.16	0.26	0.09
	yeoh-2002-v2	0.38	0.36	0.02	0.01	0.03	0.26	0.01
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.7: Valores de Rand Ajustado para o algoritmo *Complete-Linkage* com k variável.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.18	0.30	0.12	0.12	0.12	0.31	0.13
	alizadeh-2000-v2	0.77	0.77	0.77	0.77	0.77	0.77	0.77
	alizadeh-2000-v3	0.41	0.41	0.43	0.43	0.46	0.41	0.43
	bittner-2000	0.53	0.53	0.13	0.11	0.07	0.04	0.03
	bredel-2005	0.38	0.38	0.29	0.29	0.29	0.26	0.43
	chen-2002	0.32	0.32	0.23	0.23	0.55	0.36	0.35
	garber-2001	0.13	0.13	0.24	0.24	0.24	0.20	0.24
	khan-2001	0.69	0.58	0.11	0.10	0.11	0.20	0.08
	lapointe-2004-v1	0.15	0.08	0.12	0.12	0.12	0.16	0.17
	lapointe-2004-v2	0.04	0.12	0.18	0.18	0.14	0.15	0.15
	liang-2005	0.26	0.26	0.26	0.26	0.26	0.26	0.26
	risinger-2003	0.20	0.39	0.21	0.21	0.21	0.33	0.13
	tomlins-2006	0.30	0.29	0.33	0.33	0.33	0.27	0.34
	tomlins-2006-v2	0.27	0.27	0.25	0.21	0.21	0.23	0.23
Affymetrix	armstrong-2002-v1	0.18	0.48	0.35	0.53	0.53	0.34	0.84
	armstrong-2002-v2	0.63	0.71	0.88	0.81	0.64	0.86	0.63
	bhattacharjee-2001	0.36	0.35	0.45	0.43	0.43	0.63	0.34
	chowdary-2006	0.92	0.92	0.85	0.92	0.85	0.86	0.84
	dyrskjot-2003	0.33	0.22	0.32	0.37	0.37	0.39	0.36
	golub-1999-v1	0.54	0.69	0.64	0.40	0.22	0.48	0.20
	golub-1999-v2	0.73	0.77	0.57	0.34	0.25	0.59	0.26
	gordon-2002	0.92	0.90	1.00	1.00	1.00	0.97	0.59
	laiho-2007	-0.00	0.04	0.01	0.01	0.21	0.03	0.07
	nuttt-2003-v1	0.05	0.05	0.09	0.11	0.11	0.06	0.09
	nuttt-2003-v2	-0.03	-0.02	0.11	0.03	0.03	0.01	0.06
	nuttt-2003-v3	0.02	0.03	-0.02	0.03	0.03	0.03	0.17
	pomeroy-2002-v1	0.29	0.05	0.18	0.28	0.18	0.15	0.01
	pomeroy-2002-v2	0.60	0.66	0.48	0.45	0.41	0.48	0.41
	ramaswamy-2001	0.43	0.43	0.50	0.61	0.62	0.43	0.27
	shipp-2002-v1	0.15	0.23	0.06	0.10	0.16	0.15	0.07
	singh-2002	0.07	0.05	0.03	0.03	0.05	0.07	0.07
	su-2001	0.51	0.49	0.70	0.67	0.54	0.58	0.73
	west-2001	0.56	0.63	0.08	0.06	0.06	0.63	0.04
	yeoh-2002-v1	0.24	0.25	0.04	0.07	0.05	0.17	0.11
	yeoh-2002-v2	0.34	0.40	0.09	0.03	0.04	0.25	0.01
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.8: Valores de Rand Ajustado para o algoritmo *k-medoids* com *k* variável.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.50	0.43	0.43	0.37	0.37	0.43	0.37
	alizadeh-2000-v2	0.79	0.79	0.80	0.80	0.80	0.84	0.80
	alizadeh-2000-v3	0.58	0.62	0.62	0.62	0.62	0.62	0.61
	bittner-2000	0.53	0.45	0.38	0.38	0.38	0.45	0.21
	bredel-2005	0.45	0.45	0.46	0.43	0.43	0.43	0.44
	chen-2002	0.60	0.60	0.62	0.62	0.62	0.64	0.69
	garber-2001	0.37	0.38	0.33	0.33	0.36	0.37	0.36
	khan-2001	0.77	0.89	0.45	0.39	0.43	0.77	0.44
	lapointe-2004-v1	0.24	0.28	0.24	0.27	0.25	0.26	0.25
	lapointe-2004-v2	0.31	0.31	0.33	0.38	0.32	0.33	0.34
	liang-2005	0.36	0.36	0.36	0.36	0.36	0.36	0.36
	risinger-2003	0.29	0.35	0.37	0.39	0.47	0.43	0.31
	tomlins-2006	0.41	0.44	0.38	0.38	0.37	0.43	0.37
	tomlins-2006-v2	0.37	0.43	0.32	0.33	0.32	0.37	0.32
Affymetrix	armstrong-2002-v1	0.64	0.57	0.55	0.56	0.58	0.56	0.60
	armstrong-2002-v2	0.82	0.81	0.89	0.89	0.88	0.88	0.71
	bhattacharjee-2001	0.56	0.54	0.42	0.60	0.56	0.62	0.41
	chowdary-2006	0.87	0.82	0.84	0.84	0.83	0.81	0.79
	dyrskjot-2003	0.58	0.64	0.58	0.61	0.61	0.40	0.69
	golub-1999-v1	0.89	0.69	0.48	0.52	0.57	0.69	0.48
	golub-1999-v2	0.89	0.81	0.61	0.60	0.69	0.87	0.48
	gordon-2002	0.84	0.82	0.85	0.87	0.87	0.92	0.83
	laiho-2007	0.38	0.32	0.36	0.51	0.44	0.24	0.33
	nutt-2003-v1	0.22	0.22	0.22	0.19	0.18	0.20	0.17
	nutt-2003-v2	0.02	0.07	0.08	0.05	0.07	0.05	0.16
	nutt-2003-v3	0.15	0.11	0.18	0.27	0.17	0.24	0.27
	pomeroy-2002-v1	0.22	0.32	0.34	0.33	0.34	0.43	0.40
	pomeroy-2002-v2	0.67	0.61	0.58	0.57	0.62	0.58	0.63
	ramaswamy-2001	0.60	0.68	0.61	0.66	0.64	0.68	0.01
	shipp-2002-v1	0.39	0.58	0.24	0.21	0.27	0.39	0.12
	singh-2002	0.28	0.29	0.36	0.55	0.37	0.35	0.55
	su-2001	0.60	0.60	0.70	0.67	0.70	0.78	0.67
	west-2001	0.50	0.47	0.44	0.36	0.39	0.56	0.50
	yeoh-2002-v1	0.54	0.59	0.39	0.20	0.22	0.44	0.19
	yeoh-2002-v2	0.54	0.49	0.15	0.17	0.15	0.39	0.06
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.9: Número de grupos obtido com o algoritmo *Single-Linkage*.

		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	2	3	3	4	4	4	5	4
	alizadeh-2000-v2	2	4	4	3	3	3	3	3
	alizadeh-2000-v3	2	4	4	3	3	3	3	3
	bittner-2000	2	7	7	5	2	2	7	3
	bredel-2005	2	2	2	2	2	2	2	2
	chen-2002	2	3	4	2	2	2	3	2
	garber-2001	2	7	6	7	7	7	6	8
	khan-2001	2	9	10	7	10	7	5	7
	lapointe-2004-v1	2	3	2	7	7	7	8	6
	lapointe-2004-v2	2	9	7	10	10	10	11	5
	liang-2005	2	5	5	7	7	7	5	5
	risinger-2003	2	7	4	7	7	7	6	7
	tomlins-2006	2	3	3	5	5	5	4	5
	tomlins-2006-v2	2	7	7	6	6	6	6	6
Affymetrix	armstrong-2002-v1	2	9	2	8	5	9	4	6
	armstrong-2002-v2	2	7	7	6	9	6	8	9
	bhattacharjee-2001	2	12	13	11	5	5	5	5
	chowdary-2006	2	3	4	2	2	2	2	4
	dyrskjot-2003	2	6	6	7	7	7	7	5
	golub-1999-v1	2	9	9	9	3	3	2	2
	golub-1999-v2	2	9	9	2	3	3	2	2
	gordon-2002	2	14	11	10	13	13	14	9
	laiho-2007	2	2	2	2	2	2	2	2
	nutt-2003-v1	2	2	2	8	6	6	2	3
	nutt-2003-v2	2	2	2	5	5	5	6	2
	nutt-2003-v3	2	4	2	2	2	2	2	3
	pomeroy-2002-v1	2	4	6	3	5	6	2	6
	pomeroy-2002-v2	2	7	7	5	3	7	6	7
	ramaswamy-2001	2	14	14	14	14	14	14	14
	shipp-2002-v1	2	2	8	2	2	2	9	2
	singh-2002	2	11	9	8	11	10	9	4
	su-2001	2	14	14	11	12	12	9	14
	west-2001	2	7	7	6	7	6	7	7
	yeoh-2002-v1	2	2	2	4	2	3	2	13
	yeoh-2002-v2	2	3	2	9	16	4	2	10
		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.10: Número de grupos obtido com o algoritmo *Average-Linkage*.

		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	2	3	2	2	2	2	2	4
	alizadeh-2000-v2	2	3	3	3	3	3	3	3
	alizadeh-2000-v3	2	3	3	3	3	3	3	3
	bittner-2000	2	6	6	5	7	4	6	5
	bredel-2005	2	2	2	3	3	3	4	3
	chen-2002	2	4	4	3	3	4	4	3
	garber-2001	2	5	5	5	5	5	5	5
	khan-2001	2	8	8	10	8	10	8	10
	lapointe-2004-v1	2	5	4	5	5	5	6	3
	lapointe-2004-v2	2	11	3	4	4	4	7	10
	liang-2005	2	4	4	4	4	4	4	4
	risinger-2003	2	6	4	5	5	7	7	5
	tomlins-2006	2	10	10	9	9	8	11	6
	tomlins-2006-v2	2	7	7	8	8	9	8	10
Affymetrix	armstrong-2002-v1	2	2	2	3	4	4	6	2
	armstrong-2002-v2	2	2	3	6	2	8	7	3
	bhattacharjee-2001	2	2	3	6	5	5	13	11
	chowdary-2006	2	4	2	3	2	3	2	3
	dyrskjot-2003	2	5	6	6	6	7	5	3
	golub-1999-v1	2	3	3	6	6	6	3	5
	golub-1999-v2	2	3	3	7	7	7	3	5
	gordon-2002	2	7	5	2	2	2	3	2
	laiho-2007	2	2	3	6	6	6	3	6
	nutt-2003-v1	2	8	7	5	5	5	7	6
	nutt-2003-v2	2	2	2	5	6	6	2	2
	nutt-2003-v3	2	5	2	2	5	5	5	4
	pomeroy-2002-v1	2	2	3	5	6	4	5	6
	pomeroy-2002-v2	2	7	3	7	7	7	6	7
	ramaswamy-2001	2	14	14	14	14	14	14	14
	shipp-2002-v1	2	2	7	6	3	5	3	8
	singh-2002	2	6	7	2	2	2	4	2
	su-2001	2	14	14	14	14	14	14	10
	west-2001	2	2	2	7	7	7	3	6
	yeoh-2002-v1	2	5	3	12	9	10	9	13
	yeoh-2002-v2	2	10	16	12	9	10	16	16
		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.11: Número de grupos obtido com o algoritmo *Complete-Linkage*.

		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	2	4	4	4	4	4	3	2
	alizadeh-2000-v2	2	2	2	2	2	2	2	2
	alizadeh-2000-v3	2	2	2	8	8	8	2	8
	bittner-2000	2	2	2	5	2	7	5	7
	bredel-2005	2	3	3	3	3	3	3	2
	chen-2002	2	3	4	4	4	2	3	4
	garber-2001	2	2	2	2	2	2	5	2
	khan-2001	2	4	9	10	10	10	9	9
	lapointe-2004-v1	2	4	9	3	3	3	3	5
	lapointe-2004-v2	2	11	7	5	5	9	9	5
	liang-2005	2	4	4	4	4	4	4	4
	risinger-2003	2	7	6	7	7	7	7	6
	tomlins-2006	2	7	7	7	7	7	11	4
	tomlins-2006-v2	2	5	5	10	10	10	9	10
Affymetrix	armstrong-2002-v1	2	2	3	6	3	3	4	2
	armstrong-2002-v2	2	5	5	3	3	3	4	6
	bhattacharjee-2001	2	8	10	4	3	3	4	2
	chowdary-2006	2	2	2	3	2	3	3	3
	dyrskjot-2003	2	4	6	5	5	5	6	3
	golub-1999-v1	2	4	2	2	2	3	2	8
	golub-1999-v2	2	4	3	2	2	8	4	5
	gordon-2002	2	2	2	2	2	2	2	3
	laiho-2007	2	3	3	6	6	4	7	5
	nutt-2003-v1	2	6	4	6	5	5	6	7
	nutt-2003-v2	2	3	5	6	6	6	6	5
	nutt-2003-v3	2	2	2	2	2	2	2	2
	pomeroy-2002-v1	2	2	5	4	4	4	4	5
	pomeroy-2002-v2	2	6	6	4	6	7	4	5
	ramaswamy-2001	2	13	14	14	14	13	14	9
	shipp-2002-v1	2	5	4	9	3	2	2	2
	singh-2002	2	3	4	5	6	4	9	11
	su-2001	2	14	11	13	14	14	14	13
	west-2001	2	2	2	5	5	5	2	6
	yeoh-2002-v1	2	3	3	11	5	2	2	9
	yeoh-2002-v2	2	11	10	12	4	7	5	8
		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.12: Número de grupos obtido com o algoritmo *k-medoids*.

		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	2	2	2	2	2	2	2	2
	alizadeh-2000-v2	2	4	3	3	3	3	3	4
	alizadeh-2000-v3	2	6	4	4	4	5	4	5
	bittner-2000	2	2	2	2	2	2	2	6
	bredel-2005	2	3	3	3	2	2	2	3
	chen-2002	2	3	2	2	2	2	2	2
	garber-2001	2	6	4	4	4	4	6	5
	khan-2001	2	5	4	7	5	10	5	5
	lapointe-2004-v1	2	3	4	4	7	5	6	8
	lapointe-2004-v2	2	5	4	11	6	6	9	4
	liang-2005	2	5	5	5	5	5	5	5
	risinger-2003	2	5	3	2	3	4	3	6
	tomlins-2006	2	8	5	5	6	5	7	9
	tomlins-2006-v2	2	6	4	9	6	8	5	7
Affymetrix	armstrong-2002-v1	2	2	3	3	3	3	2	2
	armstrong-2002-v2	2	4	3	3	3	3	3	5
	bhattacharjee-2001	2	4	7	4	4	4	4	4
	chowdary-2006	2	3	3	3	3	4	3	3
	dyrskjot-2003	2	4	2	4	4	5	3	3
	golub-1999-v1	2	2	2	3	3	3	2	2
	golub-1999-v2	2	3	3	3	4	4	4	3
	gordon-2002	2	3	2	3	3	4	2	2
	laiho-2007	2	2	2	2	2	3	2	3
	nutt-2003-v1	2	6	6	4	4	3	8	7
	nutt-2003-v2	2	6	6	5	4	4	2	4
	nutt-2003-v3	2	3	3	3	2	4	3	2
	pomeroy-2002-v1	2	3	2	3	2	3	3	2
	pomeroy-2002-v2	2	7	7	5	6	6	7	7
	ramaswamy-2001	2	13	14	14	14	12	13	14
	shipp-2002-v1	2	2	2	2	2	4	3	3
	singh-2002	2	3	4	3	2	3	5	2
	su-2001	2	10	11	13	7	11	14	13
	west-2001	2	2	3	2	3	2	2	2
	yeoh-2002-v1	2	3	2	2	3	2	2	8
	yeoh-2002-v2	2	5	4	7	4	7	4	12
		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

A.3 Número de Grupos Estimado

Tabela A.13: Resultados de Rand Ajustado para *Single-Linkage*. O número de grupos foi estimado com o uso da Silhueta.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	alizadeh-2000-v2	0.75	0.75	0.79	0.79	0.79	0.79	0.79
	alizadeh-2000-v3	0.41	0.41	0.42	0.42	0.42	0.42	0.42
	bittner-2000	0.07	0.05	0.14	0.00	0.00	0.08	0.14
	bredel-2005	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
	chen-2002	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	garber-2001	0.26	0.28	0.28	0.28	0.28	0.28	0.30
	khan-2001	0.20	0.20	-0.01	-0.01	-0.01	0.20	-0.01
	lapointe-2004-v1	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	lapointe-2004-v2	0.03	0.01	0.01	0.01	0.01	0.01	0.01
	liang-2005	0.16	0.16	0.16	0.16	0.16	0.16	0.16
	risinger-2003	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	tomlins-2006	0.00	0.00	0.01	0.01	0.01	0.00	0.01
	tomlins-2006-v2	0.00	-0.01	0.01	0.01	0.01	0.01	0.01
Affymetrix	armstrong-2002-v1	-0.01	-0.01	0.22	0.20	0.22	-0.01	0.03
	armstrong-2002-v2	0.49	0.49	0.58	-0.00	0.58	-0.00	0.00
	bhattacharjee-2001	0.30	0.29	0.33	0.35	0.35	0.32	0.35
	chowdary-2006	0.90	0.66	0.92	0.92	0.92	0.92	0.71
	dyrskjot-2003	-0.05	-0.06	0.02	0.02	0.02	0.05	0.45
	golub-1999-v1	0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	golub-1999-v2	0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	gordon-2002	-0.01	-0.01	-0.01	-0.01	-0.01	0.04	-0.01
	laiho-2007	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
	nuttt-2003-v1	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
	nuttt-2003-v2	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00
	nuttt-2003-v3	-0.11	-0.05	-0.02	-0.07	-0.02	0.10	-0.03
	pomeroy-2002-v1	0.10	0.10	-0.01	-0.01	-0.01	0.10	-0.01
	pomeroy-2002-v2	0.15	-0.00	0.00	0.00	0.00	0.00	-0.00
	ramaswamy-2001	0.00	0.00	0.04	0.00	0.00	0.00	-0.00
	shipp-2002-v1	0.05	-0.02	-0.02	0.05	-0.02	0.05	-0.02
	singh-2002	0.00	0.00	0.02	0.01	0.01	0.03	0.04
	su-2001	0.00	0.00	0.09	0.00	0.00	0.10	0.09
	west-2001	0.00	0.54	-0.00	-0.00	-0.00	-0.00	-0.00
	yeoh-2002-v1	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	yeoh-2002-v2	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.14: Resultados de Rand Ajustado para *Average-Linkage*. O número de grupos foi estimado com o uso da Silhueta.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.12	0.16	0.19	0.19	0.19	0.26	0.19
	alizadeh-2000-v2	0.77	0.77	0.71	0.71	0.77	0.77	0.77
	alizadeh-2000-v3	0.41	0.41	0.38	0.38	0.41	0.41	0.41
	bittner-2000	0.17	0.13	-0.03	-0.02	-0.02	-0.03	0.02
	bredel-2005	0.37	0.43	0.28	0.28	0.28	0.28	0.28
	chen-2002	0.53	0.54	0.69	0.69	0.64	0.54	0.68
	garber-2001	0.21	0.21	0.22	0.22	0.22	0.13	0.22
	khan-2001	0.54	0.50	-0.01	-0.01	-0.01	0.16	-0.01
	lapointe-2004-v1	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	lapointe-2004-v2	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	liang-2005	0.16	0.16	0.16	0.16	0.16	0.16	0.16
	risinger-2003	0.03	0.03	0.04	0.04	0.04	0.00	0.07
	tomlins-2006	0.17	0.17	0.14	0.14	0.11	0.14	0.16
	tomlins-2006-v2	0.21	0.21	0.12	0.12	0.12	0.21	0.14
Affymetrix	armstrong-2002-v1	0.21	0.24	0.27	0.27	0.27	0.24	0.89
	armstrong-2002-v2	0.56	0.56	0.58	0.61	0.61	0.61	0.59
	bhattacharjee-2001	0.35	0.33	0.39	0.41	0.41	0.37	0.34
	chowdary-2006	0.78	0.92	0.71	0.92	0.71	0.92	0.71
	dyrskjot-2003	-0.05	-0.05	0.29	0.06	0.29	0.06	0.24
	golub-1999-v1	0.61	0.65	0.22	0.16	0.16	0.64	0.08
	golub-1999-v2	0.83	0.88	0.28	0.30	0.30	0.90	0.16
	gordon-2002	0.75	0.82	0.90	0.92	0.92	0.94	0.92
	laiho-2007	-0.03	-0.04	-0.11	-0.04	-0.04	-0.04	-0.09
	nutt-2003-v1	-0.00	-0.00	0.07	0.07	0.07	-0.00	0.11
	nutt-2003-v2	0.00	0.00	-0.03	-0.03	-0.03	-0.03	0.00
	nutt-2003-v3	-0.05	-0.05	-0.02	-0.06	-0.05	-0.08	-0.06
	pomeroy-2002-v1	0.05	0.02	-0.01	-0.01	0.11	-0.01	-0.01
	pomeroy-2002-v2	0.42	0.44	0.19	0.19	0.19	0.23	0.21
	ramaswamy-2001	0.38	0.34	0.17	0.17	0.17	0.18	0.02
	shipp-2002-v1	0.18	0.05	0.10	0.18	0.13	0.17	0.01
	singh-2002	0.03	0.03	0.04	0.04	0.04	0.02	0.05
	su-2001	0.50	0.59	0.79	0.81	0.82	0.73	0.65
	west-2001	0.56	0.56	0.00	0.00	0.00	0.50	0.00
	yeoh-2002-v1	0.22	0.22	-0.02	-0.01	-0.01	-0.01	-0.01
	yeoh-2002-v2	0.37	0.34	0.00	0.01	0.00	0.00	0.00
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.15: Resultados de Rand Ajustado para *Complete-Linkage*. O número de grupos foi estimado com o uso da Silhueta.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.18	0.19	0.12	0.12	0.12	0.14	0.12
	alizadeh-2000-v2	0.77	0.77	0.77	0.77	0.77	0.77	0.77
	alizadeh-2000-v3	0.41	0.41	0.41	0.41	0.41	0.41	0.41
	bittner-2000	0.39	0.18	-0.02	0.08	0.06	0.04	-0.01
	bredel-2005	0.28	0.28	0.23	0.23	0.23	0.12	0.43
	chen-2002	0.11	0.09	0.12	0.12	0.29	0.27	0.16
	garber-2001	0.13	0.13	0.24	0.24	0.24	0.17	0.24
	khan-2001	0.48	0.49	-0.01	-0.01	-0.01	0.16	-0.01
	lapointe-2004-v1	0.01	0.02	0.02	0.02	0.02	0.15	0.02
	lapointe-2004-v2	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	liang-2005	0.16	0.16	0.16	0.16	0.16	0.16	0.16
	risinger-2003	0.03	0.03	0.03	0.03	0.03	0.03	0.07
	tomlins-2006	0.21	0.27	0.08	0.08	0.08	0.20	0.24
	tomlins-2006-v2	0.21	0.21	0.18	0.16	0.16	0.19	0.19
Affymetrix	armstrong-2002-v1	0.18	0.18	0.18	0.24	0.24	0.24	0.84
	armstrong-2002-v2	0.54	0.54	0.51	0.51	0.51	0.52	0.59
	bhattacharjee-2001	0.31	0.31	0.43	0.41	0.40	0.31	0.34
	chowdary-2006	0.92	0.92	0.71	0.92	0.71	0.75	0.68
	dyrskjot-2003	-0.05	-0.05	0.26	0.26	0.26	0.21	0.31
	golub-1999-v1	0.54	0.53	0.64	0.19	0.14	0.34	0.14
	golub-1999-v2	0.73	0.77	0.57	0.22	0.25	0.55	0.21
	gordon-2002	0.81	0.90	1.00	1.00	1.00	0.97	0.48
	laiho-2007	-0.03	-0.00	-0.11	-0.11	-0.11	-0.11	-0.10
	nutt-2003-v1	0.05	0.04	0.02	0.07	0.06	-0.01	-0.01
	nutt-2003-v2	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
	nutt-2003-v3	-0.04	-0.04	-0.02	-0.04	-0.04	-0.04	0.17
	pomeroy-2002-v1	0.04	0.04	0.11	0.11	0.11	-0.01	-0.01
	pomeroy-2002-v2	0.59	0.66	0.20	0.17	0.23	0.24	0.03
	ramaswamy-2001	0.42	0.43	0.25	0.27	0.25	0.19	0.04
	shipp-2002-v1	0.07	0.17	0.06	0.07	0.16	0.15	-0.02
	singh-2002	0.04	0.02	0.02	0.02	0.02	0.02	0.02
	su-2001	0.46	0.49	0.66	0.60	0.54	0.58	0.73
	west-2001	0.56	0.63	0.02	0.01	0.01	0.63	-0.01
	yeoh-2002-v1	0.07	0.08	-0.04	0.03	0.02	0.15	-0.01
	yeoh-2002-v2	0.33	0.40	-0.01	-0.01	0.00	0.25	0.00
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.16: Resultados de Rand Ajustado para *k-medoids*. O número de grupos foi estimado com o uso da Silhueta.

		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	0.25	0.12	0.20	0.06	0.23	0.28	0.16
	alizadeh-2000-v2	0.71	0.71	0.71	0.71	0.71	0.77	0.71
	alizadeh-2000-v3	0.38	0.38	0.38	0.38	0.38	0.41	0.38
	bittner-2000	-0.03	0.18	-0.02	-0.00	-0.00	-0.02	0.01
	bredel-2005	0.43	0.43	0.43	0.43	0.43	0.43	0.37
	chen-2002	0.59	0.56	0.57	0.57	0.59	0.61	0.55
	garber-2001	0.24	0.24	0.16	0.16	0.16	0.21	0.16
	khan-2001	0.51	0.52	-0.01	-0.01	-0.01	0.16	-0.01
	lapointe-2004-v1	0.04	0.04	0.04	0.04	0.04	0.04	0.04
	lapointe-2004-v2	0.01	0.02	0.02	0.02	0.02	0.02	0.02
	liang-2005	0.16	0.16	0.16	0.16	0.16	0.16	0.16
	risinger-2003	0.10	0.10	0.04	0.04	0.02	0.07	0.04
	tomlins-2006	0.15	0.15	0.18	0.18	0.18	0.16	0.19
	tomlins-2006-v2	0.18	0.18	0.16	0.16	0.16	0.19	0.16
Affymetrix	armstrong-2002-v1	0.18	0.27	0.40	0.33	0.33	0.27	0.33
	armstrong-2002-v2	0.56	0.51	0.58	0.59	0.59	0.50	0.58
	bhattacharjee-2001	0.31	0.31	0.39	0.37	0.38	0.35	0.35
	chowdary-2006	0.85	0.81	0.78	0.78	0.78	0.78	0.78
	dyrskjot-2003	0.18	0.47	0.20	0.20	0.21	0.03	0.40
	golub-1999-v1	0.47	0.54	0.42	0.43	0.57	0.62	0.24
	golub-1999-v2	0.70	0.76	0.43	0.53	0.50	0.87	0.31
	gordon-2002	0.77	0.82	0.83	0.85	0.85	0.92	0.83
	laiho-2007	-0.04	-0.12	-0.11	-0.07	-0.11	-0.04	-0.03
	nutt-2003-v1	-0.00	-0.00	0.06	0.06	0.06	0.07	0.13
	nutt-2003-v2	-0.03	-0.03	-0.02	-0.02	-0.02	0.00	-0.02
	nutt-2003-v3	0.01	-0.05	-0.06	-0.06	-0.06	-0.08	-0.06
	pomeroy-2002-v1	0.08	0.04	0.11	0.11	0.11	0.05	0.05
	pomeroy-2002-v2	0.53	0.51	0.17	0.14	0.15	0.21	0.01
	ramaswamy-2001	0.56	0.68	0.18	0.24	0.24	0.18	-0.00
	shipp-2002-v1	0.08	0.58	0.19	0.21	0.21	0.39	0.00
	singh-2002	0.05	0.05	0.05	0.05	0.05	0.04	0.04
	su-2001	0.49	0.51	0.67	0.60	0.61	0.68	0.60
	west-2001	0.50	0.44	-0.01	-0.01	-0.01	0.56	-0.01
	yeoh-2002-v1	0.12	0.07	-0.01	-0.00	-0.00	0.13	-0.01
	yeoh-2002-v2	0.40	0.44	0.00	0.04	0.04	0.27	0.00
		ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.17: Número de grupos obtido com o algoritmo *Single-Linkage*.

		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	2	2	2	2	2	2	2	2
	alizadeh-2000-v2	3	4	4	3	3	3	3	3
	alizadeh-2000-v3	4	4	4	3	3	3	3	3
	bittner-2000	2	6	5	6	2	2	7	3
	bredel-2005	3	2	2	2	2	2	2	2
	chen-2002	2	2	2	2	2	2	2	2
	garber-2001	4	7	6	6	6	6	6	8
	khan-2001	4	2	2	2	2	2	5	2
	lapointe-2004-v1	3	3	2	2	2	2	2	2
	lapointe-2004-v2	4	4	2	2	2	2	2	2
	liang-2005	3	3	3	3	3	3	3	3
	risinger-2003	4	2	2	2	2	2	2	2
	tomlins-2006	5	2	2	2	2	2	2	2
	tomlins-2006-v2	4	2	2	2	2	2	2	2
Affymetrix	armstrong-2002-v1	2	2	2	3	4	3	2	2
	armstrong-2002-v2	3	6	6	5	2	6	2	2
	bhattacharjee-2001	5	7	8	4	5	5	4	5
	chowdary-2006	2	3	4	2	2	2	2	2
	dyrskjot-2003	3	2	3	2	2	2	3	5
	golub-1999-v1	2	2	2	2	2	2	2	2
	golub-1999-v2	3	2	2	2	2	2	2	2
	gordon-2002	2	2	2	2	2	2	2	2
	laiho-2007	2	2	2	2	2	2	2	2
	nutt-2003-v1	4	2	2	2	2	2	2	2
	nutt-2003-v2	2	2	2	2	4	4	2	2
	nutt-2003-v3	2	3	2	3	5	3	2	5
	pomeroy-2002-v1	2	2	2	2	2	2	2	2
	pomeroy-2002-v2	5	7	2	2	2	2	2	2
	ramaswamy-2001	14	2	2	3	2	2	2	3
	shipp-2002-v1	2	2	2	2	2	2	2	2
	singh-2002	2	2	2	3	6	4	7	3
	su-2001	10	2	2	3	2	2	6	3
	west-2001	2	2	7	2	2	2	2	3
	yeoh-2002-v1	2	2	2	2	2	2	2	2
	yeoh-2002-v2	6	2	2	2	2	2	2	2
		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.18: Número de grupos obtido com o algoritmo *Average-Linkage*.

		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	2	7	7	5	5	5	5	5
	alizadeh-2000-v2	3	2	2	2	2	2	2	2
	alizadeh-2000-v3	4	2	2	2	2	2	2	2
	bittner-2000	2	4	4	2	2	2	2	2
	bredel-2005	3	2	2	2	2	2	2	2
	chen-2002	2	4	4	3	3	3	4	3
	garber-2001	4	2	2	2	2	2	2	2
	khan-2001	4	9	10	2	2	2	3	2
	lapointe-2004-v1	3	2	2	2	2	2	2	2
	lapointe-2004-v2	4	2	2	2	2	2	2	2
	liang-2005	3	3	3	3	3	3	3	3
	risinger-2003	4	2	2	2	2	2	2	2
	tomlins-2006	5	2	2	2	2	2	2	2
	tomlins-2006-v2	4	2	2	2	2	2	2	2
Affymetrix	armstrong-2002-v1	2	2	2	2	2	2	2	2
	armstrong-2002-v2	3	2	2	2	2	2	2	2
	bhattacharjee-2001	5	2	2	2	2	2	2	2
	chowdary-2006	2	2	2	2	2	2	2	2
	dyrskjot-2003	3	2	2	2	2	2	2	2
	golub-1999-v1	2	3	3	9	8	8	3	2
	golub-1999-v2	3	3	3	9	8	8	3	2
	gordon-2002	2	4	4	2	2	2	3	2
	laiho-2007	2	6	2	2	2	2	2	2
	nutt-2003-v1	4	2	2	3	3	3	2	5
	nutt-2003-v2	2	2	2	2	2	2	4	2
	nutt-2003-v3	2	2	2	2	2	2	3	2
	pomeroy-2002-v1	2	5	5	2	2	2	2	2
	pomeroy-2002-v2	5	5	3	2	2	2	2	2
	ramaswamy-2001	14	13	13	3	3	3	3	2
	shipp-2002-v1	2	2	2	2	2	2	2	3
	singh-2002	2	2	2	2	2	2	2	2
	su-2001	10	14	12	14	13	12	14	11
	west-2001	2	2	2	2	2	2	4	2
	yeoh-2002-v1	2	12	7	2	2	2	2	2
	yeoh-2002-v2	6	12	7	2	2	2	2	2
		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.19: Número de grupos obtido com o algoritmo *Complete-Linkage*.

		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	2	5	5	4	4	4	7	4
	alizadeh-2000-v2	3	2	2	2	2	2	2	2
	alizadeh-2000-v3	4	2	2	2	2	2	2	2
	bittner-2000	2	3	7	2	5	4	6	3
	bredel-2005	3	2	2	2	2	2	2	2
	chen-2002	2	13	14	12	12	6	6	11
	garber-2001	4	2	2	2	2	2	2	2
	khan-2001	4	10	10	2	2	2	3	2
	lapointe-2004-v1	3	2	2	2	2	2	2	2
	lapointe-2004-v2	4	2	2	2	2	2	2	2
	liang-2005	3	3	3	3	3	3	3	3
	risinger-2003	4	2	2	2	2	2	2	2
	tomlins-2006	5	2	11	2	2	2	2	2
	tomlins-2006-v2	4	2	2	2	2	2	2	2
Affymetrix	armstrong-2002-v1	2	2	2	2	2	2	2	2
	armstrong-2002-v2	3	2	2	2	2	2	2	2
	bhattacharjee-2001	5	2	2	2	2	2	2	2
	chowdary-2006	2	2	2	2	2	2	2	2
	dyrskjot-2003	3	2	2	2	2	2	2	4
	golub-1999-v1	2	4	3	2	4	8	3	9
	golub-1999-v2	3	4	3	2	4	8	3	9
	gordon-2002	2	3	2	2	2	2	2	2
	laiho-2007	2	5	4	2	2	2	2	2
	nutt-2003-v1	4	5	5	2	4	2	2	2
	nutt-2003-v2	2	4	4	2	2	2	4	2
	nutt-2003-v3	2	4	3	2	3	3	3	2
	pomeroy-2002-v1	2	6	6	2	2	2	2	2
	pomeroy-2002-v2	5	7	6	2	2	2	2	2
	ramaswamy-2001	14	14	14	4	4	4	3	2
	shipp-2002-v1	2	9	9	2	2	2	2	3
	singh-2002	2	2	2	2	2	2	2	2
	su-2001	10	12	11	11	10	14	14	12
	west-2001	2	2	2	2	2	2	2	2
	yeoh-2002-v1	2	16	10	3	2	3	5	2
	yeoh-2002-v2	6	16	10	3	2	3	5	2
		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabela A.20: Número de grupos obtido com o algoritmo *k-medoids*.

		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	2	4	6	5	2	4	4	5
	alizadeh-2000-v2	3	2	2	2	2	2	2	2
	alizadeh-2000-v3	4	2	2	2	2	2	2	2
	bittner-2000	2	2	4	2	2	2	2	3
	bredel-2005	3	2	2	2	2	2	2	2
	chen-2002	2	3	3	3	3	3	3	3
	garber-2001	4	2	2	2	2	2	2	2
	khan-2001	4	10	10	2	2	2	3	2
	lapointe-2004-v1	3	2	2	2	2	2	2	2
	lapointe-2004-v2	4	2	2	2	2	2	2	2
	liang-2005	3	3	3	3	3	3	3	3
	risinger-2003	4	2	2	2	2	2	2	2
	tomlins-2006	5	2	2	2	2	2	2	2
	tomlins-2006-v2	4	2	2	2	2	2	2	2
Affymetrix	armstrong-2002-v1	2	2	2	2	2	2	2	2
	armstrong-2002-v2	3	2	2	2	2	2	2	2
	bhattacharjee-2001	5	2	2	2	2	2	2	2
	chowdary-2006	2	2	2	2	2	2	2	2
	dyrskjot-2003	3	2	3	2	2	2	2	6
	golub-1999-v1	2	3	3	3	3	3	3	2
	golub-1999-v2	3	3	3	3	3	3	3	2
	gordon-2002	2	3	2	2	2	2	2	2
	laiho-2007	2	4	2	2	2	2	2	2
	nutt-2003-v1	4	2	2	2	2	2	3	4
	nutt-2003-v2	2	2	2	2	2	2	2	2
	nutt-2003-v3	2	4	2	2	2	2	3	2
	pomeroy-2002-v1	2	6	6	2	2	2	2	2
	pomeroy-2002-v2	5	6	5	2	2	2	2	2
	ramaswamy-2001	14	12	14	3	4	4	3	2
	shipp-2002-v1	2	9	2	2	2	2	2	2
	singh-2002	2	2	2	2	2	2	2	2
	su-2001	10	13	10	12	14	10	13	9
	west-2001	2	2	2	2	2	2	2	2
	yeoh-2002-v1	2	8	15	2	2	2	5	2
	yeoh-2002-v2	6	7	7	2	2	2	9	2
		k^*	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Tabelas dos Experimentos de Seleção de Atributos

Neste apêndice são apresentadas as tabelas completas obtidas a partir da realização de experimentos de seleção de atributos com o algoritmo *Simplified Silhouette Filter* (SSF). São apresentadas nas Tabelas B.1 à B.35 os erros médios e os desvios padrão (entre parênteses) de classificação obtidos com os classificadores NB e 1NN após a aplicação do algoritmo SSF com cada um dos coeficientes de correlação avaliados. Nestas 35 tabelas são também apresentados os resultados obtidos sem a prévia aplicação do método de seleção de atributos. Nas colunas indicadas por *NAS* são apresentados os números médios de atributos selecionados, enquanto nas colunas *NB* e *1NN* são apresentados os erros médios de classificação obtidos com estes classificadores. Todas as médias apresentadas são referentes ao procedimento de validação cruzada de 10 pastas estratificado.

Tabela B.1: Erros de classificação obtidos para a base de dados *alizadeh-2000-v1*.

Método	NAS	NB	1NN
SSF- ρ	4.00 (0.00)	44.00 (30.07)	40.00 (21.60)
SSF- ϱ	5.40 (0.97)	33.00 (24.97)	28.00 (20.17)
SSF- γ	4.20 (0.63)	31.50 (25.93)	43.50 (23.46)
SSF- τ	4.20 (0.63)	36.50 (33.59)	48.50 (21.35)
SSF- $\hat{\rho}$	4.20 (0.63)	38.50 (24.27)	43.00 (24.74)
SSF- r	4.20 (0.63)	23.50 (22.61)	37.00 (28.40)
SSF- $\hat{\gamma}$	4.00 (0.00)	31.00 (16.30)	38.50 (21.22)
Sem Seleção	1095	2.50 (7.91)	24.50 (16.74)

Tabela B.2: Erros de classificação obtidos para a base de dados *alizadeh-2000-v2*.

Método	NAS	NB	1NN
SSF- ρ	4.60 (0.97)	7.86 (8.33)	14.29 (11.56)
SSF- ϱ	4.80 (1.03)	8.10 (11.62)	11.19 (12.95)
SSF- γ	4.20 (0.63)	8.10 (8.56)	15.00 (16.57)
SSF- τ	4.40 (0.84)	6.43 (8.33)	11.67 (17.66)
SSF- $\hat{\rho}$	4.40 (0.84)	9.52 (8.25)	14.29 (11.56)
SSF- r	4.60 (0.97)	14.29 (11.56)	19.76 (13.24)
SSF- $\hat{\gamma}$	4.40 (0.84)	9.52 (8.25)	13.10 (15.28)
Sem Seleção	2093	4.76 (7.69)	1.43 (4.52)

Tabela B.3: Erros de classificação obtidos para a base de dados *alizadeh-2000-v3*.

Método	NAS	NB	1NN
SSF- ρ	4.60 (0.97)	38.81 (22.45)	41.90 (28.27)
SSF- ϱ	4.40 (0.84)	30.71 (19.55)	37.14 (13.48)
SSF- γ	4.20 (0.63)	39.76 (26.68)	45.24 (18.44)
SSF- τ	4.20 (0.63)	38.57 (21.99)	43.81 (17.97)
SSF- $\hat{\rho}$	4.20 (0.63)	32.14 (20.30)	35.24 (17.56)
SSF- r	4.40 (0.84)	30.71 (21.46)	40.71 (18.59)
SSF- $\hat{\gamma}$	4.40 (1.26)	43.10 (21.60)	43.33 (14.45)
Sem Seleção	2093	10.95 (12.86)	22.62 (11.46)

Tabela B.4: Erros de classificação obtidos para a base de dados *armstrong-2002-v1*.

Método	NAS	NB	1NN
SSF- ρ	59.40 (2.50)	11.25 (8.99)	21.25 (15.53)
SSF- ϱ	31.40 (28.91)	12.68 (12.50)	25.00 (17.50)
SSF- γ	4.00 (0.00)	27.86 (11.76)	25.18 (14.88)
SSF- τ	4.00 (0.00)	8.39 (7.24)	11.07 (5.88)
SSF- $\hat{\rho}$	4.00 (0.00)	1.43 (4.52)	8.21 (7.10)
SSF- r	4.00 (0.00)	26.96 (24.81)	15.36 (10.62)
SSF- $\hat{\gamma}$	4.20 (0.63)	27.32 (18.79)	23.57 (13.38)
Sem Seleção	1081	1.43 (4.52)	9.82 (9.57)

Tabela B.5: Erros de classificação obtidos para a base de dados *armstrong-2002-v2*.

Método	NAS	NB	1NN
SSF- ρ	26.60 (34.74)	30.89 (19.09)	41.61 (10.98)
SSF- ϱ	4.20 (0.63)	29.29 (8.38)	30.71 (14.98)
SSF- γ	4.00 (0.00)	46.96 (16.06)	55.18 (16.85)
SSF- τ	4.00 (0.00)	27.50 (15.25)	35.89 (19.64)
SSF- $\hat{\rho}$	4.00 (0.00)	34.29 (18.71)	31.61 (13.63)
SSF- r	4.00 (0.00)	33.57 (19.30)	30.54 (8.60)
SSF- $\hat{\gamma}$	4.00 (0.00)	34.29 (9.40)	45.54 (11.46)
Sem Seleção	2194	1.25 (3.95)	8.39 (7.24)

Tabela B.6: Erros de classificação obtidos para a base de dados *bhattacharjee-2001*.

Método	NAS	NB	1NN
SSF- ρ	4.00 (0.00)	30.05 (12.42)	24.69 (9.52)
SSF- ϱ	4.20 (0.63)	20.26 (10.12)	20.19 (7.86)
SSF- γ	4.00 (0.00)	36.79 (16.94)	31.40 (12.40)
SSF- τ	4.00 (0.00)	30.90 (11.83)	32.98 (12.11)
SSF- $\hat{\rho}$	4.00 (0.00)	26.48 (8.88)	31.14 (11.96)
SSF- r	4.00 (0.00)	27.55 (7.90)	25.60 (4.39)
SSF- $\hat{\gamma}$	4.00 (0.00)	31.02 (13.62)	31.02 (8.39)
Sem Seleção	1543	7.86 (5.68)	6.38 (6.90)

Tabela B.7: Erros de classificação obtidos para a base de dados *bittner-2000*.

Método	NAS	NB	1NN
SSF- ρ	4.00 (0.00)	32.50 (28.99)	41.67 (28.60)
SSF- ϱ	4.00 (0.00)	45.83 (25.23)	45.83 (25.23)
SSF- γ	4.00 (0.00)	23.33 (25.09)	28.33 (27.55)
SSF- τ	4.00 (0.00)	30.83 (22.92)	30.83 (30.69)
SSF- $\hat{\rho}$	4.00 (0.00)	39.17 (28.34)	41.67 (20.03)
SSF- r	4.00 (0.00)	40.83 (23.72)	40.83 (23.72)
SSF- $\hat{\gamma}$	4.00 (0.00)	29.17 (24.92)	41.67 (28.60)
Sem Seleção	2201	23.33 (25.09)	25.83 (20.58)

Tabela B.8: Erros de classificação obtidos para a base de dados *bredel-2005*.

Método	NAS	NB	1NN
SSF- ρ	4.00 (0.00)	32.00 (23.48)	36.00 (30.98)
SSF- ϱ	4.00 (0.00)	26.00 (16.47)	40.00 (21.08)
SSF- γ	4.00 (0.00)	20.00 (16.33)	32.00 (13.98)
SSF- τ	4.00 (0.00)	22.00 (14.76)	30.00 (17.00)
SSF- $\hat{\rho}$	4.00 (0.00)	26.00 (18.97)	32.00 (23.48)
SSF- r	4.00 (0.00)	24.00 (18.38)	36.00 (22.71)
SSF- $\hat{\gamma}$	4.00 (0.00)	30.00 (17.00)	30.00 (14.14)
Sem Seleção	1739	22.00 (14.76)	20.00 (16.33)

Tabela B.9: Erros de classificação obtidos para a base de dados *chen-2002*.

Método	NAS	NB	1NN
SSF- ρ	4.00 (0.00)	32.94 (9.88)	30.65 (11.95)
SSF- ϱ	4.00 (0.00)	31.24 (6.85)	29.67 (8.62)
SSF- γ	12.40 (7.23)	21.18 (12.45)	17.29 (11.21)
SSF- τ	13.60 (6.65)	18.95 (12.30)	16.73 (10.44)
SSF- $\hat{\rho}$	4.00 (0.00)	25.62 (11.08)	26.73 (11.56)
SSF- r	8.80 (5.75)	19.48 (11.45)	20.07 (12.83)
SSF- $\hat{\gamma}$	17.90 (0.32)	17.78 (14.05)	16.70 (12.80)
Sem Seleção	85	6.70 (5.73)	10.59 (6.09)

Tabela B.10: Erros de classificação obtidos para a base de dados *chowdary-2006*.

Método	NAS	NB	1NN
SSF- ρ	7.20 (3.16)	6.64 (6.28)	4.82 (5.09)
SSF- ϱ	11.60 (3.10)	3.82 (4.94)	2.91 (4.69)
SSF- γ	4.20 (0.63)	16.36 (9.20)	8.73 (7.35)
SSF- τ	4.20 (0.63)	14.45 (9.44)	8.64 (7.07)
SSF- $\hat{\rho}$	4.00 (0.00)	15.55 (11.84)	9.64 (0.47)
SSF- r	5.00 (2.16)	14.36 (8.00)	8.73 (9.73)
SSF- $\hat{\gamma}$	4.20 (0.63)	29.09 (17.05)	20.82 (13.25)
Sem Seleção	182	1.91 (4.03)	3.82 (4.94)

Tabela B.11: Erros de classificação obtidos para a base de dados *dyrskjot-2003*.

Método	NAS	NB	1NN
SSF- ρ	24.80 (15.00)	45.00 (25.82)	35.00 (26.87)
SSF- ϱ	4.00 (0.00)	27.50 (21.89)	25.00 (11.79)
SSF- γ	4.00 (0.00)	50.00 (20.41)	52.50 (27.51)
SSF- τ	4.00 (0.00)	32.50 (16.87)	27.50 (18.45)
SSF- $\hat{\rho}$	4.00 (0.00)	27.50 (14.19)	40.00 (21.08)
SSF- r	4.00 (0.00)	35.00 (21.08)	50.00 (26.35)
SSF- $\hat{\gamma}$	4.00 (0.00)	37.50 (24.30)	20.00 (15.81)
Sem Seleção	1203	30.00 (15.81)	20.00 (19.72)

Tabela B.12: Erros de classificação obtidos para a base de dados *garber-2001*.

Método	NAS	NB	1NN
SSF- ρ	4.00 (0.00)	41.90 (16.81)	50.71 (16.69)
SSF- ϱ	4.00 (0.00)	40.24 (14.35)	46.90 (13.79)
SSF- γ	4.00 (0.00)	32.86 (15.58)	56.43 (18.88)
SSF- τ	4.00 (0.00)	31.19 (18.18)	54.52 (21.40)
SSF- $\hat{\rho}$	4.00 (0.00)	47.62 (19.44)	56.90 (18.25)
SSF- r	4.00 (0.00)	45.00 (18.49)	49.05 (16.23)
SSF- $\hat{\gamma}$	4.00 (0.00)	46.90 (18.17)	53.10 (20.91)
Sem Seleção	4553	31.19 (11.53)	22.86 (16.81)

Tabela B.13: Erros de classificação obtidos para a base de dados *golub-1999-v1*.

Método	NAS	NB	1NN
SSF- ρ	27.80 (6.63)	22.50 (15.59)	31.79 (14.13)
SSF- ϱ	17.20 (9.58)	18.39 (13.63)	13.75 (17.54)
SSF- γ	4.00 (0.00)	25.18 (12.44)	31.96 (15.76)
SSF- τ	4.00 (0.00)	14.11 (11.68)	21.25 (18.22)
SSF- $\hat{\rho}$	4.00 (0.00)	11.25 (11.23)	14.29 (15.06)
SSF- r	4.20 (0.63)	33.39 (23.08)	26.79 (20.96)
SSF- $\hat{\gamma}$	4.60 (0.97)	25.18 (18.91)	30.36 (17.92)
Sem Seleção	1868	1.25 (3.95)	16.79 (14.49)

Tabela B.14: Erros de classificação obtidos para a base de dados *golub-1999-v2*.

Método	NAS	NB	1NN
SSF- ρ	27.00 (14.06)	21.79 (21.57)	38.57 (14.75)
SSF- ϱ	16.60 (10.92)	27.68 (16.82)	29.29 (12.26)
SSF- γ	4.00 (0.00)	36.25 (15.48)	48.57 (14.98)
SSF- τ	4.20 (0.63)	37.32 (17.10)	30.71 (11.57)
SSF- $\hat{\rho}$	4.00 (0.00)	26.43 (19.30)	36.43 (23.63)
SSF- r	4.20 (0.63)	48.93 (18.73)	44.82 (18.14)
SSF- $\hat{\gamma}$	4.80 (1.03)	30.71 (16.10)	35.89 (16.55)
Sem Seleção	1868	8.39 (7.24)	28.93 (17.37)

Tabela B.15: Erros de classificação obtidos para a base de dados *gordon-2002*.

Método	NAS	NB	1NN
SSF- ρ	62.80 (10.88)	4.42 (5.10)	4.42 (4.38)
SSF- ϱ	72.30 (7.24)	2.22 (2.87)	3.89 (3.75)
SSF- γ	4.00 (0.00)	3.89 (5.89)	12.11 (7.61)
SSF- τ	4.00 (0.00)	3.83 (3.63)	10.44 (6.43)
SSF- $\hat{\rho}$	4.00 (0.00)	6.05 (4.80)	10.99 (6.73)
SSF- r	4.00 (0.00)	3.89 (5.27)	5.50 (6.88)
SSF- $\hat{\gamma}$	4.00 (0.00)	6.05 (4.80)	10.44 (6.94)
Sem Seleção	1626	0.56 (1.76)	0.56 (1.76)

Tabela B.16: Erros de classificação obtidos para a base de dados *khan-2001*.

Método	NAS	NB	1NN
SSF- ρ	14.40 (1.58)	16.94 (11.43)	21.53 (15.55)
SSF- ϱ	12.40 (2.63)	21.67 (18.17)	21.67 (7.72)
SSF- γ	4.00 (0.00)	49.03 (18.76)	42.22 (15.29)
SSF- τ	4.00 (0.00)	43.47 (22.18)	38.61 (11.05)
SSF- $\hat{\rho}$	4.00 (0.00)	36.53 (16.16)	53.75 (16.37)
SSF- r	5.80 (0.63)	31.11 (14.98)	42.22 (17.82)
SSF- $\hat{\gamma}$	4.00 (0.00)	47.92 (22.20)	51.53 (24.18)
Sem Seleção	1069	1.25 (3.95)	12.08 (9.87)

Tabela B.17: Erros de classificação obtidos para a base de dados *laiho-2007*.

Método	NAS	NB	1NN
SSF- ρ	7.80 (8.40)	29.17 (18.53)	28.33 (24.91)
SSF- ϱ	5.80 (0.63)	21.67 (20.49)	30.83 (21.17)
SSF- γ	4.20 (0.63)	34.17 (16.41)	34.17 (25.59)
SSF- τ	4.20 (0.63)	16.67 (14.70)	25.00 (15.21)
SSF- $\hat{\rho}$	4.80 (1.03)	25.83 (20.58)	27.50 (12.45)
SSF- r	4.00 (0.00)	38.33 (22.29)	31.67 (19.56)
SSF- $\hat{\gamma}$	4.00 (0.00)	32.50 (22.03)	29.17 (14.30)
Sem Seleção	2202	15.83 (13.86)	5.83 (12.45)

Tabela B.18: Erros de classificação obtidos para a base de dados *lapointe-2004-v1*.

Método	NAS	NB	1NN
SSF- ρ	4.20 (0.63)	30.24 (13.93)	36.19 (25.32)
SSF- ϱ	5.00 (1.05)	37.62 (17.88)	38.10 (19.44)
SSF- γ	4.20 (0.63)	46.19 (12.51)	46.19 (12.51)
SSF- τ	4.00 (0.00)	46.19 (10.54)	46.19 (12.51)
SSF- $\hat{\rho}$	4.20 (0.63)	49.05 (19.96)	37.62 (13.55)
SSF- r	4.00 (0.00)	42.14 (18.58)	43.10 (16.05)
SSF- $\hat{\gamma}$	4.20 (0.63)	53.10 (22.95)	54.76 (19.34)
Sem Seleção	1625	30.48 (14.23)	30.71 (19.06)

Tabela B.19: Erros de classificação obtidos para a base de dados *lapointe-2004-v2*.

Método	NAS	NB	1NN
SSF- ρ	5.00 (1.05)	39.09 (9.63)	39.09 (15.48)
SSF- ϱ	4.40 (0.84)	39.09 (9.63)	47.27 (13.42)
SSF- γ	4.00 (0.00)	55.45 (15.12)	63.64 (10.50)
SSF- τ	4.00 (0.00)	55.45 (15.12)	63.64 (10.50)
SSF- $\hat{\rho}$	4.00 (0.00)	49.09 (17.25)	46.36 (17.38)
SSF- r	4.00 (0.00)	48.18 (14.88)	46.36 (13.85)
SSF- $\hat{\gamma}$	4.00 (0.00)	43.64 (17.03)	49.09 (17.25)
Sem Seleção	2496	31.82 (9.82)	30.91 (12.27)

Tabela B.20: Erros de classificação obtidos para a base de dados *liang-2005*.

Método	NAS	NB	1NN
SSF- ρ	4.00 (0.00)	22.50 (12.45)	23.33 (18.76)
SSF- ϱ	4.00 (0.00)	27.50 (17.15)	21.67 (16.76)
SSF- γ	4.00 (0.00)	12.50 (17.68)	13.33 (18.51)
SSF- τ	4.00 (0.00)	12.50 (17.68)	13.33 (18.51)
SSF- $\hat{\rho}$	4.00 (0.00)	15.00 (17.48)	8.33 (13.61)
SSF- r	4.00 (0.00)	30.00 (18.51)	24.17 (19.02)
SSF- $\hat{\gamma}$	4.00 (0.00)	19.17 (18.02)	7.50 (12.08)
Sem Seleção	1411	12.50 (17.68)	2.50 (7.91)

Tabela B.21: Erros de classificação obtidos para a base de dados *nutt-2003-v1*.

Método	NAS	NB	1NN
SSF- ρ	5.60 (0.84)	40.00 (24.94)	44.00 (24.59)
SSF- ϱ	5.20 (1.03)	44.00 (26.33)	34.00 (26.75)
SSF- γ	4.00 (0.00)	56.00 (20.66)	64.00 (29.51)
SSF- τ	4.00 (0.00)	58.00 (19.89)	52.00 (13.98)
SSF- $\hat{\rho}$	4.00 (0.00)	34.00 (13.50)	50.00 (17.00)
SSF- r	4.20 (0.63)	36.00 (20.66)	48.00 (21.50)
SSF- $\hat{\gamma}$	4.00 (0.00)	52.00 (13.98)	50.00 (21.60)
Sem Seleção	1377	34.00 (9.66)	36.00 (22.71)

Tabela B.22: Erros de classificação obtidos para a base de dados *nutt-2003-v2*.

Método	NAS	NB	1NN
SSF- ρ	6.00 (0.00)	35.00 (31.87)	30.00 (17.21)
SSF- ϱ	6.00 (0.00)	28.33 (27.27)	18.33 (19.95)
SSF- γ	4.40 (0.84)	23.33 (38.65)	23.33 (31.62)
SSF- τ	4.00 (0.00)	21.67 (19.33)	35.00 (31.87)
SSF- $\hat{\rho}$	4.00 (0.00)	13.33 (17.21)	25.00 (23.90)
SSF- r	6.00 (0.00)	28.33 (27.27)	23.33 (22.50)
SSF- $\hat{\gamma}$	4.20 (0.63)	46.67 (21.94)	25.00 (18.00)
Sem Seleção	1070	30.00 (28.11)	40.00 (19.56)

Tabela B.23: Erros de classificação obtidos para a base de dados *nutt-2003-v3*.

Método	NAS	NB	1NN
SSF- ρ	4.00 (0.00)	31.67 (27.72)	25.00 (26.35)
SSF- ϱ	4.00 (0.00)	36.67 (31.23)	20.00 (25.82)
SSF- γ	4.00 (0.00)	31.67 (43.35)	38.33 (31.48)
SSF- τ	4.00 (0.00)	26.67 (23.83)	33.33 (33.33)
SSF- $\hat{\rho}$	4.00 (0.00)	18.33 (24.15)	28.33 (34.29)
SSF- r	4.00 (0.00)	25.00 (36.22)	26.67 (23.83)
SSF- $\hat{\gamma}$	4.00 (0.00)	31.67 (27.72)	23.33 (25.09)
Sem Seleção	1152	8.33 (18.00)	23.33 (25.09)

Tabela B.24: Erros de classificação obtidos para a base de dados *pomeroy-2002-v1*.

Método	NAS	NB	1NN
SSF- ρ	27.00 (12.66)	31.67 (21.08)	25.83 (15.44)
SSF- ϱ	4.40 (0.84)	40.83 (20.20)	32.50 (17.32)
SSF- γ	4.00 (0.00)	45.83 (28.12)	36.67 (25.52)
SSF- τ	4.00 (0.00)	23.33 (23.50)	40.83 (23.39)
SSF- $\hat{\rho}$	4.00 (0.00)	37.50 (18.11)	28.33 (23.31)
SSF- r	4.00 (0.00)	48.33 (33.05)	35.83 (23.59)
SSF- $\hat{\gamma}$	4.40 (0.84)	36.67 (31.72)	35.83 (32.41)
Sem Seleção	857	20.83 (21.61)	27.50 (20.81)

Tabela B.25: Erros de classificação obtidos para a base de dados *pomeroy-2002-v2*.

Método	NAS	NB	1NN
SSF- ρ	45.40 (11.93)	37.50 (24.75)	43.50 (23.46)
SSF- ϱ	68.40 (6.10)	26.50 (25.06)	41.00 (20.92)
SSF- γ	4.20 (0.63)	59.00 (27.47)	60.00 (27.99)
SSF- τ	4.00 (0.00)	64.50 (23.51)	51.50 (21.35)
SSF- $\hat{\rho}$	4.00 (0.00)	69.50 (19.07)	73.50 (14.54)
SSF- r	4.00 (0.00)	69.00 (16.30)	61.50 (17.65)
SSF- $\hat{\gamma}$	4.00 (0.00)	52.50 (26.06)	50.00 (17.32)
Sem Seleção	1379	27.50 (25.30)	50.00 (20.95)

Tabela B.26: Erros de classificação obtidos para a base de dados *ramaswamy-2001*.

Método	NAS	NB	1NN
SSF- ρ	8.60 (1.35)	57.89 (9.92)	60.53 (10.30)
SSF- ϱ	8.60 (1.65)	54.74 (9.67)	56.32 (10.83)
SSF- γ	4.00 (0.00)	75.26 (7.46)	77.37 (4.99)
SSF- τ	4.00 (0.00)	62.11 (10.17)	71.05 (7.94)
SSF- $\hat{\rho}$	4.00 (0.00)	65.79 (11.44)	65.79 (13.19)
SSF- r	6.00 (0.94)	61.05 (8.30)	65.79 (10.89)
SSF- $\hat{\gamma}$	4.00 (0.00)	62.63 (8.75)	65.26 (9.67)
Sem Seleção	1363	29.47 (9.67)	46.32 (4.15)

Tabela B.27: Erros de classificação obtidos para a base de dados *risinger-2003*.

Método	NAS	NB	1NN
SSF- ρ	8.80 (9.44)	47.00 (22.75)	57.50 (21.11)
SSF- ϱ	4.40 (1.26)	62.00 (22.51)	53.00 (19.47)
SSF- γ	4.00 (0.00)	40.50 (22.29)	46.00 (23.43)
SSF- τ	4.00 (0.00)	35.50 (22.42)	38.50 (25.06)
SSF- $\hat{\rho}$	4.00 (0.00)	19.00 (21.71)	37.50 (24.75)
SSF- r	4.00 (0.00)	43.00 (24.74)	43.00 (27.41)
SSF- $\hat{\gamma}$	4.00 (0.00)	36.50 (18.57)	50.50 (25.44)
Sem Seleção	1771	30.00 (23.45)	24.00 (15.24)

Tabela B.28: Erros de classificação obtidos para a base de dados *shipp-2002-v1*.

Método	NAS	NB	1NN
SSF- ρ	6.60 (1.65)	34.46 (15.88)	40.54 (13.34)
SSF- ϱ	4.00 (0.00)	31.61 (16.75)	31.61 (17.15)
SSF- γ	4.00 (0.00)	46.96 (16.35)	34.82 (22.70)
SSF- τ	4.00 (0.00)	25.71 (16.00)	25.00 (19.01)
SSF- $\hat{\rho}$	4.00 (0.00)	23.39 (20.49)	33.93 (16.81)
SSF- r	4.00 (0.00)	28.57 (4.98)	28.57 (15.61)
SSF- $\hat{\gamma}$	4.00 (0.00)	39.64 (25.25)	29.11 (14.46)
Sem Seleção	798	17.32 (13.23)	19.29 (16.73)

Tabela B.29: Erros de classificação obtidos para a base de dados *singh-2002*.

Método	NAS	NB	1NN
SSF- ρ	4.00 (0.00)	34.27 (9.21)	26.64 (13.46)
SSF- ϱ	4.00 (0.00)	35.27 (8.19)	21.45 (8.72)
SSF- γ	4.80 (1.03)	25.36 (15.14)	34.36 (10.84)
SSF- τ	4.20 (0.63)	28.36 (14.23)	31.27 (9.80)
SSF- $\hat{\rho}$	4.00 (0.00)	30.45 (12.96)	30.36 (9.70)
SSF- r	4.40 (0.84)	36.36 (8.45)	22.73 (13.24)
SSF- $\hat{\gamma}$	5.00 (1.41)	36.27 (9.35)	33.36 (10.72)
Sem Seleção	339	21.55 (11.22)	19.64 (13.21)

Tabela B.30: Erros de classificação obtidos para a base de dados *su-2001*.

Método	NAS	NB	1NN
SSF- ρ	49.00 (15.27)	21.80 (7.49)	29.97 (14.32)
SSF- ϱ	52.70 (14.67)	18.95 (10.01)	26.27 (10.43)
SSF- γ	4.00 (0.00)	58.59 (11.28)	66.05 (7.65)
SSF- τ	4.20 (0.63)	59.84 (10.98)	66.14 (5.17)
SSF- $\hat{\rho}$	4.00 (0.00)	60.95 (11.13)	61.41 (11.17)
SSF- r	4.00 (0.00)	63.86 (11.23)	63.76 (7.74)
SSF- $\hat{\gamma}$	4.00 (0.00)	62.39 (15.14)	67.03 (12.81)
Sem Seleção	1571	11.47 (5.44)	12.03 (4.96)

Tabela B.31: Erros de classificação obtidos para a base de dados *tomlins-2006*.

Método	NAS	NB	1NN
SSF- ρ	4.00 (0.00)	35.73 (8.61)	48.27 (15.18)
SSF- ϱ	4.00 (0.00)	50.36 (17.65)	51.09 (9.23)
SSF- γ	4.00 (0.00)	42.09 (13.41)	49.73 (18.18)
SSF- τ	4.00 (0.00)	42.09 (13.41)	48.73 (18.43)
SSF- $\hat{\rho}$	4.00 (0.00)	46.09 (13.17)	60.73 (14.88)
SSF- r	4.00 (0.00)	51.91 (14.31)	52.82 (20.33)
SSF- $\hat{\gamma}$	4.00 (0.00)	47.82 (16.54)	58.82 (10.88)
Sem Seleção	2315	22.09 (13.46)	23.00 (16.53)

Tabela B.32: Erros de classificação obtidos para a base de dados *tomlins-2006-v2*.

Método	NAS	NB	1NN
SSF- ρ	4.00 (0.00)	47.67 (10.71)	60.44 (17.12)
SSF- ϱ	4.00 (0.00)	51.00 (17.66)	57.22 (18.72)
SSF- γ	4.00 (0.00)	55.44 (20.42)	56.56 (12.47)
SSF- τ	4.00 (0.00)	53.22 (17.36)	55.44 (10.74)
SSF- $\hat{\rho}$	4.00 (0.00)	54.44 (17.53)	61.89 (17.61)
SSF- r	4.00 (0.00)	44.44 (16.19)	59.56 (16.85)
SSF- $\hat{\gamma}$	4.00 (0.00)	48.89 (13.81)	50.89 (15.35)
Sem Seleção	1288	25.22 (15.94)	24.67 (14.47)

Tabela B.33: Erros de classificação obtidos para a base de dados *west-2001*.

Método	NAS	NB	1NN
SSF- ρ	15.20 (3.91)	39.00 (22.34)	27.00 (14.94)
SSF- ϱ	42.40 (25.33)	35.00 (17.16)	35.00 (14.34)
SSF- γ	4.40 (0.84)	47.50 (24.64)	39.50 (29.10)
SSF- τ	4.00 (0.00)	29.00 (20.25)	22.50 (14.77)
SSF- $\hat{\rho}$	4.00 (0.00)	24.00 (15.78)	28.50 (16.67)
SSF- r	4.00 (0.00)	38.50 (21.61)	37.00 (18.89)
SSF- $\hat{\gamma}$	4.00 (0.00)	36.50 (20.28)	39.50 (17.71)
Sem Seleção	1198	12.50 (17.20)	24.50 (18.33)

Tabela B.34: Erros de classificação obtidos para a base de dados *yeoh-2002-v1*.

Método	NAS	NB	1NN
SSF- ρ	15.80 (7.97)	25.53 (22.58)	21.77 (7.81)
SSF- ϱ	40.00 (45.82)	16.22 (7.41)	18.62 (12.27)
SSF- γ	4.00 (0.00)	32.60 (13.92)	24.13 (7.39)
SSF- τ	70.20 (45.70)	26.12 (14.66)	18.58 (9.83)
SSF- $\hat{\rho}$	4.00 (0.00)	22.57 (6.55)	27.45 (9.52)
SSF- r	4.00 (0.00)	20.98 (6.25)	25.87 (6.07)
SSF- $\hat{\gamma}$	3.00 (0.00)	9.22 (16.00)	8.05 (8.63)
Sem Seleção	2526	5.25 (3.82)	8.88 (4.95)

Tabela B.35: Erros de classificação obtidos para a base de dados *yeoh-2002-v2*.

Método	NAS	NB	1NN
SSF- ρ	14.80 (8.50)	78.15 (17.24)	72.62 (7.87)
SSF- ϱ	41.40 (48.30)	63.72 (17.26)	65.77 (12.30)
SSF- γ	4.00 (0.00)	87.88 (4.69)	75.82 (8.22)
SSF- τ	70.00 (45.57)	70.10 (8.40)	59.65 (13.17)
SSF- $\hat{\rho}$	4.00 (0.00)	81.12 (8.14)	75.78 (8.29)
SSF- r	4.00 (0.00)	74.12 (9.76)	71.00 (9.11)
SSF- $\hat{\gamma}$	3.00 (0.00)	65.22 (11.58)	57.22 (8.70)
Sem Seleção	2526	41.98 (7.41)	45.53 (7.88)