Universidade Federal do Rio Grande do Norte

Centro de Ciências Exatas e da Terra

Departamento de Informática e Matemática Aplicada

Programa de Pós-Graduação em Sistemas e Computação

Daniel Sabino Amorim de Araújo

Algoritmos de Agrupamento Aplicados a Dados de Expressão Gênica de Câncer: Um Estudo Comparativo

Natal

2008

Daniel Sabino Amorim de Araújo

Algoritmos de Agrupamento Aplicados a Dados de Expressão Gênica de Câncer: Um Estudo Comparativo

Dissertação de mestrado submetida ao Programa de Pós-Graduação em Sistemas e Computação do Departamento de Informática e Matemática Aplicada da Universidade Federal do Rio Grande do Norte como parte dos requisitos para a obtenção do grau de Mestre em Sistemas e Computação (MSc.).

Orientador:
Marcílio C. P. de Souto (DIMAp/UFRN)

Co-orientador:
Ivan G. Costa Filho (CIn/UFPE)

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS EXATAS E DA TERRA
DEPARTAMENTO DE INFORMÁTICA E MATEMÁTICA APLICADA
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E COMPUTAÇÃO

Natal

2008

Catalogação da Publicação na Fonte. UFRN / SISBI / Biblioteca Setorial Especializada do Centro de Ciências Exatas e da Terra – CCET.

Araújo, Daniel Sabino Amorim de.
Algoritmos de agrupamento aplicados a dados de expressão gênica de câncer : um estudo comparativo / Daniel Sabino Amorim de Araújo. – Natal, 2008.
102 f. : il.

Orientador: Prof. Dr. Marcílio C. P. de Souto.
Co-Orientador: Ivan G. Costa Filho.
Dissertação (Mestrado) – Universidade Federal do Rio Grande do Norte. Centro de Ciências Exatas e da Terra. Departamento de Informática e Matemática Aplicada. Programa de Pós-Graduação em Sistemas e Computação.

1. Inteligência artificial - Dissertação. 2. Bioinformática - Dissertação. 3. Aprendizado de máquina - Dissertação. 4. Análise de agrupamentos - Dissertação. 5. Expressão Gênica. I. Souto, Marcílio C. P. de. II. Costa Filho, Ivan G. III. Título.

RN/UF/BSE-CCET CDU: 004.8

Agradecimentos

Agradeço, primeiramente, à minha família, sobretudo meus pais, por me apoiarem incondicionalmente nesta e em todas as jornadas realizadas em minha vida.

A Jarinete, por estar sempre ao meu lado e pela compreensão desprendida em todos os momentos, desde os mais tensos e nervosos aos mais tranqüilos.

Agradeço também a Marcílio, meu orientador, que forneceu meios e oportunidades fundamentais para a conclusão deste e de outros trabalhos.

A Ivan, que no papel de co-orientador contribuiu em pontos cruciais durante a construção desta dissertação.

Aos meus, já amigos, companheiros de mestrado e universidade pelos momentos de ajuda e descontração necessários para a conclusão da longa caminhada do mestrado.

Agradeço também a Rodrigo pela colaboração nos trabalhos relacionados a esta dissertação.

Por fim, agradeço a CAPES pelo apoio financeiro fornecido.

Resumo

O uso de técnicas de agrupamento na descoberta de subtipos de câncer tem atraído grande atenção da comunidade científica. Enquanto bioinformatas propõem novas técnicas de agrupamento que levam em consideração características dos dados de expressão gênica, a comunidade médica prefere utilizar as técnicas "clássicas" de agrupamento. De fato, não existem trabalhos na literatura que realizam uma avaliação em grande escala de técnicas de agrupamento nesse contexto. Diante disso, este trabalho apresenta o primeiro estudo em grande escala de sete técnicas de agrupamento e quatro medidas de proximidade para a análise de 35 conjuntos de dados de expressão gênica. Mais especificamente, os resultados mostram que a técnica mistura finita de gaussianas, seguida pelo *k-means*, apresentam os melhores resultados em termos de recuperação da estrutura natural dos dados. Esses métodos também apresentam a menor diferença entre o número real de classes e o número de grupos presente na melhor partição. Além disso, os métodos de agrupamento hierárquico, que vêm sendo bastante utilizados pela comunidade médica, apresentaram os piores resultados quando comparados com os outros métodos investigados. Este trabalho também apresenta, como uma referência estável para a avaliação e comparação de diferentes algoritmos de agrupamento para dados de expressão gênica de câncer, um conjunto de bases de dados (*benchmark data sets*) que pode ser compartilhado entre pesquisadores e usado na comparação de novos métodos.

Abstract

The use of clustering methods for the discovery of cancer subtypes has drawn a great deal of attention in the scientific community. While bioinformaticians have proposed new clustering methods that take advantage of characteristics of the gene expression data, the medical community has a preference for using "classic" clustering methods. There have been no studies thus far performing a large-scale evaluation of different clustering methods in this context. This work presents the first large-scale analysis of seven different clustering methods and four proximity measures for the analysis of 35 cancer gene expression data sets. Results reveal that the finite mixture of Gaussians, followed closely by *k-means*, exhibited the best performance in terms of recovering the true structure of the data sets. These methods also exhibited, on average, the smallest difference between the actual number of classes in the data sets and the best number of clusters as indicated by our validation criteria. Furthermore, hierarchical methods, which have been widely used by the medical community, exhibited a poorer recovery performance than that of the other methods evaluated. Moreover, as a stable basis for the assessment and comparison of different clustering methods for cancer gene expression data, this study provides a common group of data sets (*benchmark data sets*) to be shared among researchers and used for comparisons with new methods.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

ALL	*acute lymphoblastic leukemia*
MLL	*mixed-lineage leukemia*
AML	*acute myeloid leukemia*
AC	*adenocarcinoma*
SCC	*squamous cell carcinoma*
SCLC	*small-cell lung cancer*
MPM	*malignant pleural mesothelioma*
CG	*classic glioblastoma*
NG	*non-classic glioblastoma*
CO	*classic oligodendroglioma*
NO	*non-classic oligodendroglioma*
MD	*medulloblastoma*
MGLIO	*malignant glioma*
AT/RT	*atypical teratoid/rhabdoid tumour*
PNET	*primitive neuroectodermal tumour*
DLBCL	*diffuse large B-cell lymphoma*
FL	*follicular lymphoma*
ER	*estrogen receptor*
CLL	*chronic lymphocytic leukaemia*
GBM	*glioblastoma*
OG	*oligodendroglioma*
AT	*astrocytic tumour*
HCC	*hepatocelullar carcinoma*
LCLC	*large cell lung carcinoma*
EWS	*Ewing tumour*

RMS *rhabdomyosarcoma*

NB *neuroblastoma*

BL *Burkitt lymphoma*

PS *papillary serous*

EM *endometrioid*

CC *clear cell*

EPI *benign epithelium*

PIN *prostatic intraepithelial neoplasia*

PCA *prostate cancer*

MET *metastatic prostate cancer*

Sumário

1 *Introdução*

Atualmente, as técnicas de *microarray* tornam possível, entre outras coisas, a medição de assinaturas (perfis) moleculares de células com câncer (D'HAESELEER, 2005). Com esse tipo de dados é possível fazer uma série de análises como, por exemplo, a identificação de genes diferencialmente expressos em condições distintas, o que pode indicar potenciais genes a serem estudados mais detalhadamente em nível molecular. Ainda nesse contexto, é possível a construção de classificadores, por meio do uso de técnicas de aprendizado de máquina, que podem ajudar no processo de diagnóstico de pacientes com câncer (TUSHER; TIBSHIRANI; CHU, 2001; SPANG, 2003).

Um outro tipo de análise que vem sendo comumente realizada é o agrupamento de amostras de células de pacientes com câncer (QUACKENBUSH, 2001; D'HAESELEER, 2005; XU; WUNSCH, 2005). O objetivo desse tipo de investigação é encontrar grupos de objetos a partir de um conjunto de dados. A hipótese é que objetos em um mesmo grupo compartilhem características comuns, o que pode levar a descoberta de novos tipos de câncer ou, até mesmo, a descoberta de subtipos de um dado câncer.

O tipo de estudo descrito no parágrafo anterior foi realizado primeiramente por Golub et al. (1999) e Alizadeh et al. (2000). Desde então, as técnicas de análise de agrupamento vêm atraindo grande atenção da comunidade de Bioinformática (D'HAESELEER, 2005). De fato, várias dessas técnicas vêm sendo propostas, cada uma levando em consideração características específicas dos dados de expressão gênica (XU; WUNSCH, 2005). No entanto, apesar da contribuição dos bioinformatas no aprimoramento e criação de técnicas de agrupamento visando à análise de dados de expressão gênica, a comunidade médica ainda prefere utilizar técnicas de

agrupamento “tradicionais” como, por exemplo, o algoritmo de agrupamento hierárquico, para a investigação de tais dados (D’HAESELEER, 2005).

Uma das razões para isso é o fato de que os agrupamentos gerados pelas técnicas hierárquicas, por terem um número crescente de partições aninhadas, assemelham-se com as árvores filogenéticas, as quais a comunidade médica já possui experiência na análise (QUACKENBUSH, 2001). Além disso, as técnicas de agrupamento hierárquico possuem uma grande quantidade de implementações disponíveis. Elas, em geral, fazem parte de vários pacotes de *softwares* utilizados para análise de dados de *microarray* (HOON et al., 2004; EISEN et al., 1998). Um outro aspecto relevante nessas técnicas é o fato de precisarem de poucos parâmetros para serem ajustados, tornando seu uso viável para não especialistas em computação.

Por exemplo, nesta dissertação são utilizados 35 conjuntos de dados que foram coletados a partir de 24 artigos. Em relação a esses artigos, quando a análise de agrupamentos foi realizada, em 95% dos casos foi empregado o algoritmo hierárquico (ARMSTRONG et al., 2002; BHATTACHARJEE et al., 2001; CHOWDARY et al., 2006; DYRSKJOT et al., 2003; GOLUB et al., 1999; GORDON et al., 2002; LAIHO et al., 2007; NUTT et al., 2003; POMEROY et al., 2002; RAMASWAMY et al., 2001; SHIPP et al., 2002; SU et al., 2001; WEST et al., 2001; YEOH et al., 2002; ALIZADEH et al., 2000; BITTNER et al., 2000; BREDEL et al., 2005; CHEN et al., 2002; GARBER et al., 2001; KHAN et al., 2001; LAPOINTE et al., 2004; LIANG et al., 2005; RISINGER et al., 2003; TOMLINS et al., 2007).

1.1 Motivação e Objetivos

O uso de técnicas de agrupamento na descoberta de subtipos de câncer tem atraído grande atenção da comunidade científica (D’HAESELEER, 2005; QUACKENBUSH, 2001). Como mencionado anteriormente, enquanto os bioinformatas têm proposto novas técnicas de agrupamento, as quais exploram as características dos dados de expressão gênica, a comunidade médica prefere usar as técnicas “clássicas” de agrupamento. De fato, até o momento, não existe nesse contexto nenhum estudo na literatura que faz uma comparação em grande escala de dife-

rentes técnicas de agrupamento.

Motivado por essas limitações, este trabalho realiza o primeiro estudo em grande escala de várias técnicas de agrupamento combinadas com diferentes medidas de proximidade aplicadas a vários conjuntos de dados de expressão gênica de câncer obtidos através de técnicas de *microarray*. Mais especificamente, o objetivo deste trabalho é fornecer uma orientação (para comunidade científica/médica) na escolha de técnicas de agrupamento para serem usadas na análise de dados de expressão gênica.

No trabalho são analisados, com base em 35 conjuntos de dados, sete técnicas de agrupamento e quatro medidas de proximidade. Todos os conjuntos de dados foram obtidos utilizando tecnologias de *microarray* dos tipos *Affymetrix* e cDNA (MONTI et al., 2003; QUACKENBUSH, 2001; SLONIM, 2002). Foram avaliadas tanto técnicas de agrupamento "clássicas", tais como, os métodos hierárquicos aglomerativos e *k-means* (JAIN; DUBES, 1988), como também algumas técnicas mais recentes que são especializadas em dados de alta dimensionalidade como, por exemplo, *Spectral Clustering* (NG; JORDAN; WEISS, 2001) e *Shared Nearest Neighbors* (ERTOZ; STEINBACH; KUMAR, 2002). Além disso, quando possível, foram aplicadas quatro medidas de proximidade: distância euclidiana, correlação de Pearson, coeficiente de correlação de Spearman e cosseno (JAIN; DUBES, 1988). No caso da distância euclidiana, que é sensível a diferenças de escalas nos atributos, os dados foram submetidos a três tipos de transformação: padronização, escalonamento e *ranking* (MILLIGAN; COOPER, 1988; SOUTO et al., 2008a).

Por fim, este trabalho também apresenta um conjunto de bases de dados (*benchmark data sets*) que pode ser compartilhado entre pesquisadores e usado como uma referência estável para a avaliação e comparação de diferentes algoritmos de agrupamento para dados de expressão gênica de câncer ou na comparação de novos métodos.

1.2 Organização do Trabalho

O restante do trabalho está dividido em quatro capítulos:

- Capítulo 2: apresenta uma introdução dos conceitos de Biologia Molecular, expressão gênica e das tecnologias utilizadas para obter os dados de expressão gênica, além de fazer uma revisão sobre os trabalhos relacionados com o tema.

- Capítulo 3: aborda aspectos importantes de análise de agrupamentos, descrevendo as técnicas de agrupamento, medidas de proximidade e metodologia de validação de agrupamentos usadas no trabalho.

- Capítulo 4: fornece informações sobre os conjuntos de dados e os filtros aplicados a eles. Também descreve o projeto e avaliação dos experimentos, apresentando resultados e realizando uma discussão a respeito deles.

- Capítulo 5: por fim, são mostradas as conclusões obtidas com a realização do trabalho e possíveis extensões.

2 *Expressão Gênica*

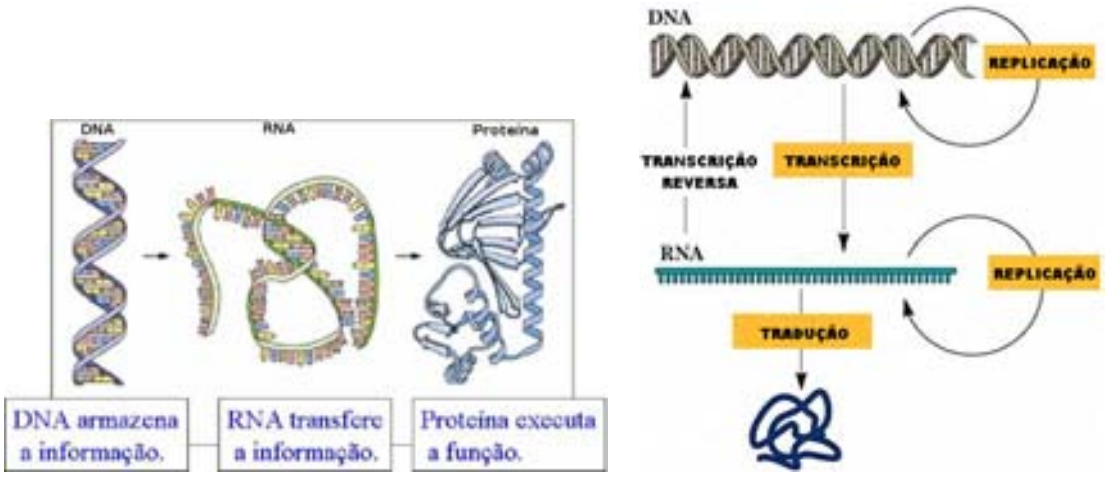
Em um ser vivo, todas as células possuem a informação genética utilizada para codificar todas as proteínas necessárias para a definição e funcionamento de um organismo (ALBERTS et al., 1997). Porém, em cada tipo de célula, dependendo da sua função, somente uma parte dessa informação é utilizada (expressa), ou seja, somente trechos do genoma (genes) são utilizados para sintetizar proteínas.

A criação de projetos de pesquisa científica, como o projeto Genoma, tem levado ao aumento do conhecimento sobre os genes dos seres humanos e outros seres vivos (QUACKENBUSH, 2001). Várias técnicas foram desenvolvidas com o intuito de extrair informações gênicas como, por exemplo, as técnicas de *microarrays* que permitem medir a expressão de milhares de genes simultaneamente. Essas informações podem ser usadas para desenvolver um entendimento mais completo da função, regulação e interação dos genes.

2.1 Expressão Gênica

A expressão gênica diz respeito ao processo pelo qual as sequências de DNA (*Desoxirribonucleic Acid* ou ácido desoxirribonucleico) que representam genes são interpretadas na síntese de proteínas. Mais especificamente, a informação genética está armazenada no DNA em uma sequência de quatro tipos de nucleotídeos: adenina (A), timina (T), guanina (G) e citosina (C). Parte do chamado dogma central da biologia molecular diz respeito ao direcionamento do fluxo de informação da célula, afirmando que a molécula de DNA é utilizada como molde para construção de uma molécula de RNA (*Ribonucleic Acid* ou ácido ribonucleico), a qual será usada

como molde para síntese de proteínas (ALBERTS et al., 1997). A Figura 2.1 (a) mostra o dogma central proposto inicialmente, em que o fluxo de informação é unidirecional no sentido DNA a proteína; e o modelo mais atual, mostrado na Figura 2.1 (b), em que existe a possibilidade de formação de DNA a partir de moléculas de RNA, utilizando a enzima transcriptase reversa, ou até mesmo a replicação do RNA.



(a) (b)

Figura 2.1: Dogma central da biologia molecular[1].

O processo de replicação do DNA é o mecanismo de transmissão da informação genética durante a divisão celular. Na replicação cada uma das fitas do DNA serve de molde para criação de uma nova fita completa. Já as etapas de transcrição e tradução do dogma central regem a expressão gênica da célula. A transcrição é referente a produção de RNA a partir da cópia de regiões específicas (genes) do genoma. O processo de transcrição se assemelha ao processo de duplicação do DNA (replicação) no sentido de que uma das fitas do DNA é utilizada como molde para formação da nova fita simples de RNA. O papel do RNA gerado, entre outras coisas, é guardar toda a informação que foi copiada do DNA durante o processo de transcrição. A Figura 2.2 ilustra o processo de transcrição.

Apesar de o RNA ser feito a partir de uma das fitas do DNA, eles possuem algumas diferenças. Por exemplo, o RNA também é composto por uma sequência de nucleotídeos, entretanto, em vez da desoxirribose do DNA o açúcar presente é a ribose. Além disso, O RNA possui a

[1]Figura retirada de http://faculty.uca.edu/ johnc/rnaprot1440.htm, acessado em 11/11/2008
[2]Figura retirada de (ALBERTS et al., 1997).

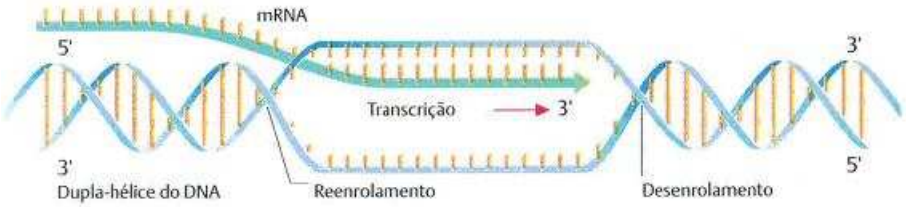


Figura 2.2: Processo de transcrição[2].

base Uracila (U) em lugar da base Timina (T). Transcritos de RNA que direcionam a síntese de proteínas são denominados RNA mensageiros (mRNA). Além desse tipo de RNA, ainda existem, pelo menos, mais dois tipos: os RNA transportadores (tRNA) e os RNA ribossômicos (rRNA).

Para um gene ativo, grandes quantidades de transcritos podem ser produzidos a partir de uma mesma sequência. Como cada molécula de mRNA pode ser traduzida em milhares de cópias de cadeias peptídicas (proteínas), um pequeno trecho do DNA pode ser responsável pela síntese de milhões de cópias de uma proteína específica (ALBERTS et al., 1997).

Nos seres procariotos (aqueles que não apresentam seu material genético delimitado por uma membrana) o processo de transcrição é mais simples que nos eucariotos (possuem membrana delimitadora do material genético). Enquanto que nos procariotos as sequências que formam um mRNA são cópias sem alterações de um trecho do DNA, nos eucariotos as regiões codificadoras (íntrons) do DNA são intercaladas com regiões não-codificadoras (exons). Essas últimas regiões devem ser removidas para que uma sequência apta a ser traduzida seja formada.

O próximo passo da síntese de proteínas é a leitura das sequências de nucleotídeos do mRNA, em grupos de três, e a tradução em aminoácidos. As regras pelos quais as trincas de nucleotídeos são traduzidos em proteínas é conhecido como código genético. O código genético faz um mapeamento entre cada trio de nucleotídeos, o códon, e um aminácido que forma a proteína. Assim, se os mRNA são compostos por quatro nucleotídeos diferentes (uracila, adenina, guanina e citosina) e cada códon contém três nucleotídeos, então, 64 aminoácidos diferentes são possíveis. Entretanto, somente 20 aminoácidos são encontrados nas proteínas. Isso acontece porque a maioria dos aminoácidos são codificados por mais de um códon (código genético

degenerado).

Os códons não reconhecem diretamente os aminoácidos que codificam. Na verdade, o processo da síntese de proteínas é feita pelo ribossomo (complexo de proteínas associadas a rRNA). O ribossomo se move ao longo da cadeia de mRNA traduzindo as sequências de nucleotídeos, um códon de cada vez, com o auxílio moléculas de tRNA para adicionar os aminoácidos correspondentes à extremidade da nova cadeia protéica que está sendo formada. Todos os processos que estão envolvidos na síntese de proteínas de um ser eucarioto, incluindo a tradução, são mostrados na Figura 2.3.

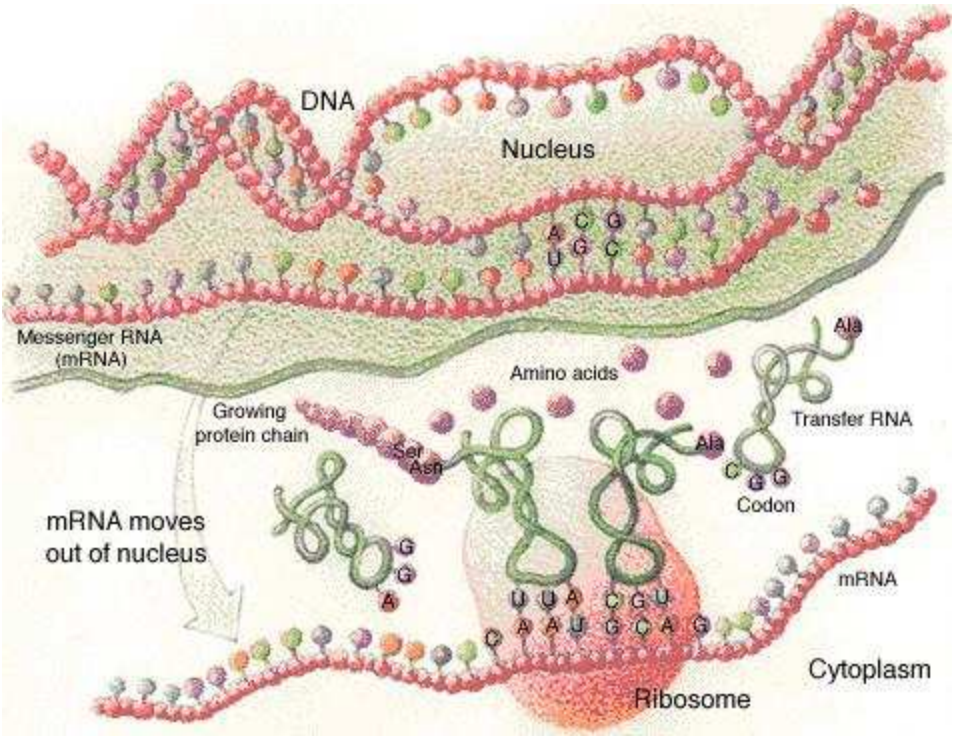


Figura 2.3: Síntese de proteínas em um organismo eucarioto.[3]

Apesar de existirem diversos mecanismos de segurança que evitam erros nos processos de replicação, transcrição e tradução, a maquinaria celular, às vezes, comete erros, como o posicionamento errado de um nucleotídeo durante a replicação. Qualquer mudança devido a um erro genético é denominada mutação. Mesmo com a alteração de apenas um nucleotídeo, as mutações podem ter extensas consequências, levando tanto ao surgimento de anomalias no organismo, como doenças, ou, muito raramente, a criação de novos genes com um função

[3]Figura retirada de http://fai.unne.edu.ar/biologia/macromoleculas/adn.htm, acessado em 11/11/2008

melhorada ou nova.

2.2 Dados de *Microarray*

A habilidade de identificar e extrair informações funcionais de genes em nível de RNA tem concentrado esforços em todo o mundo (DUGGAN et al., 1999; QUACKENBUSH, 2001; D'HAESELEER, 2005). Como mencionado anteriormente, uma rota comum para explorar a função de um gene é determinar seu perfil de expressão (DUGGAN et al., 1999). De modo geral, a análise de expressão gênica fornece informações importantes sobre o funcionamento de uma célula, já que alterações genéticas modificam diretamente a produção de proteínas necessárias para o funcionamento da célula (ALBERTS et al., 1997).

Microarray de DNA é uma ferramenta poderosa para identificação e quantificação de ácidos nucleicos. As tecnologias baseadas em *microarrays* de DNA fornecem um meio relativamente simples para analisar simultaneamente a expressão de milhares de genes (JAIN, 2001; SLONIM, 2002; QUACKENBUSH, 2001). Esse paralelismo aumenta consideravelmente a velocidade dos experimentos e permite comparações significativas entre os produtos gênicos representados no *microarray* (SCHENA et al., 1998).

Entre outros usos, a análise dos dados produzidos por *microarray* se tornou a técnica padrão para monitorar níveis de expressão gênica em laboratórios de biologia molecular (LAUSTED et al., 2004). Tais *microarrays* podem ser feitos através de um posicionamento (*spotting*) de produtos de DNA pré-sintetizados ou pela síntese *de novo*[4] de oligonucleotídeos[5] em um substrato sólido, geralmente uma espécie de lâmina de vidro (SCHENA et al., 1998).

Microarrays podem ser fabricados de diferentes maneiras, utilizando várias tecnologias, dependendo do número de elementos sob análise e custos, por exemplo. Segundo Jain (2001) e Schena et al. (1998), qualquer análise de dados de *microarray* segue, pelo menos, cinco passos: construção do *array*, preparação das amostras alvo, hibridização, escaneamento e detecção, e

[4]Síntese de moléculas complexas a partir de moléculas mais simples.
[5]Fragmento curto de uma cadeia de DNA ou RNA.

normalização e análise dos dados. Dentre esses passos, os *microarays* se diferenciam basicamente com relação a fabricação e detecção/escaneamento. Nas próximas seções serão dados detalhes sobre técnicas e tecnologias para fabricação e escaneamento de *microarrays*.

2.2.1 Fabricação de *microarrays*

Microarrays podem ser fabricados usando uma variedade de tecnologias, incluindo a impressão com pinos em lâminas de vidro, fotolitografia usando máscaras pré-feitas ou eletroquímica em *arrays* de microeletrodos (LAUSTED et al., 2004). Neste trabalho serão explorados as técnicas conhecidas por *spotted microarrays*, que usam a primeira tecnologia para fazer impressão de produtos gênicos pré-sintetizados; e *microarrays* de oligonucleotídeos, que usam fotolitografia para sintetização *de novo* de sondas.

Spotted microarrays

O termo *spotted microarrays* engloba um conjunto de técnicas de deposição que permite a produção automática de *microarrays* através da impressão de pequenas quantidades de substâncias químicas previamente feitas em uma superfície sólida (SCHENA et al., 1998).

Em *spotted microarrays*, as sondas são, em geral, pequenos fragmentos que correspondem a um mRNA (geralmente associados a um gene). Essas sondas são sintetizadas previamente e, então, colocadas (*spotted*) na superfície do *array*. A impressão é contemplada pelo contato direto da superfície de impressão e do mecanismo de entrega que contém um arranjo de pinças, pinos ou capilares que servem para transferir as sondas selecionadas para a superfície (SCHENA et al., 1998). Como resultado desse processo, tem-se um *grid* de sondas representando os perfis de DNA das sondas preparadas e prontas para serem hibridizadas com as sequências complementares de cDNA ou cRNA (alvos), derivadas de experimentos ou amostras clínicas.

Algumas das vantagens dos *spotted microarrays* incluem a facilidade de prototipagem e, consequentemente, a rapidez na implementação, o baixo custo e a versatilidade da técnica (SCHENA et al., 1998). Por exemplo, pesquisadores podem escolher quais sondas e que lo-

cais do *array* vão imprimir, sintetizar essas sondas no próprio laboratório e produzir o *array*. Depois, eles podem gerar suas próprias amostras rotuladas, hibridizá-las com as sondas do *array* e, por fim, escanear o produto usando seus próprios equipamentos. Tais vantagens fazem esse tipo de técnica ser usada por uma grande quantidade de pesquisadores para produzir *microarrys* "domésticos" em seus próprios laboratórios. O fato de se ter um experimento de baixo custo e que pode ser customizado para cada estudo, evita os custos da compra de um *array* comercial de valor muito elevado e que pode representar um grande número de genes que o pesquisador não está interessado.

Dentre as desvantagens desse tipo de técnicas está o fato de que cada sonda utilizada tem que ser sintetizada, purificada e armazenada antes da fabricação do *array* (SCHENA et al., 1998). Além disso, alguns estudos indicam que os *spotted microarrays* podem não fornecer o mesmo nível de sensibilidade quando comparado com os *arrays* de oligonucleotídeos comerciais, possivelmente pelo uso de recursos limitados quando comparado com os produtores industriais de *arrays* (BAMMLER et al., 2005).

Microarrays de oligonucleotídeos

Arrays de oligonucleotídeos para o monitoramento de expressão gênica são projetados e sintetizados baseados somente em informações das sequências, sem a necessidade de intermediários como cDNAs ou produtos de PCR[6]. Nesse método, as sondas são pequenas sequências projetadas para serem compatíveis com partes das sequências conhecidas nas amostras. Mais precisamente, usando uma pequena sequência com 200 ou 300 bases de um gene ou cDNA, oligonucleotídeos contendo 25 nucleotídeos (25-mer) são selecionados (sem sobreposição, quando possível), para servir como detectores únicos de sequências específicas (LIPSHUTZ et al., 1999). Essas sequências são sintetizadas diretamente na superfície do *array*, em lugar de depositar sequências previamente prontas. Como os *arrays* são projetados *in silico*, não é necessário preparar, verificar, quantificar e catalogar um grande número de cDNA ou outros produtos gênicos e, assim, eliminando os riscos de erros de identificação.

[6] *Polymerase Chain Reaction* - processo de replicação de DNA *in vitro*.

Um ponto-chave desse tipo de técnica é o uso de redundância nas sondas. Ou seja, o uso de múltiplas sequências projetadas para hibridizar com diferentes regiões do RNA. Assim, o uso de múltiplos detectores para a mesma molécula aumenta a relação sinal/ruído e a acurácia na quantificação de RNA e reduz a taxa de falsos positivos. A primeira técnica de sucesso para síntese *de novo* de oligonucleotídeos em *chips* foi desenvolvida pela *Affymetrix* usando técnicas fotolitográficas emprestadas da indústria de semicondutores.

2.2.2 Escaneamento e detecção

Dentre os modos existentes para escanemanto e detecção de *microarrays*, somente dois tipos deles serão abordados neste trabalho. *Microarrays Double-channel* (*two-color*) e *arrays* do tipo *single-channel* (*one-color*).

double-channel

Microarrays double-channel ou *double-color* são normalmente hibridizados utilizando cDNA de duas amostras diferentes a serem analisadas e comparadas, por exemplo, amostras coletadas de tecidos com alguma doença e de tecido saudável. Essas amostras são rotuladas com dois corantes fluorescentes diferentes. Os corantes normalmente utilizados nesses experimentos incluem o Cy3, que tem comprimento de onda de 570 nm (correspondendo a parte verde do espectro da luz) e Cy5 com emissão de fluorescência com comprimento de onda de 670 nm (parte vermelha do espectro de luz). Em um único *microarray*, as duas amostras com rótulos fluorescentes diferentes são colocadas para hibridizar com as sondas presentes no *microarray*. As lâminas são, então, escaneadas com *lasers* que podem detectar cada um dos dois corantes fluorescentes usados para rotular as amostras.

Apesar de níveis individuais absolutos de expressão gênica poderem ser medidos em *microarrays double-channel*, a diferença relativa entre os níveis de expressão normalmente é utilizada. Para isso, as imagens obtidas durante a fase de escaneamento de cada uma das fluorescências são alinhadas, de modo que os pontos correspondentes fiquem sobrepostos. A quantidade de pontos e as suas intensidades são, então, medidas subtraindo a intensidade do fundo e ge-

rando uma imagem equivalente a expressão gênica relativa entre as duas amostras. Além disso, medidas de controle externas são usadas para corrigir rotulagens e melhorar a velocidade de detecção dos marcadores luminosos (JAIN, 2001). Assim, esse tipo de técnica fornece meios para medir diferenças relativas de expressão gênica entre amostras de diferentes tecidos ou diferentes amostras de um mesmo tecido. A Figura 2.4 resume todo o processo em um experimento de *microarrays double-channel* e Figura 2.5 mostra a imagem final gerada após o experimento.

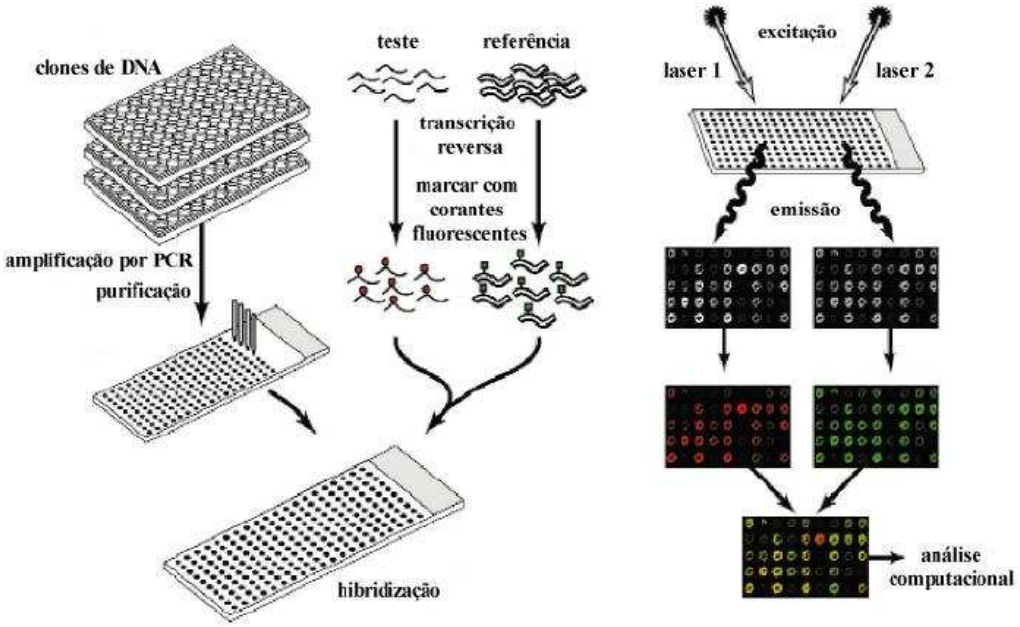


Figura 2.4: Experimentos de *microarray double-channel* [7].

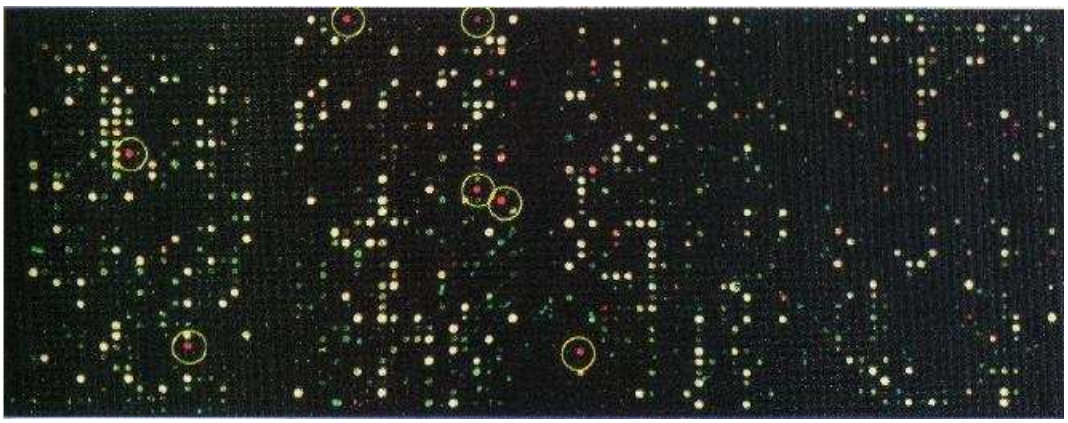


Figura 2.5: Lâmina contendo cadeias de cDNA hibridizados[8].

.

[7]Figura retirada de (DUGGAN et al., 1999).
[8]Figura retirada de (ALBERTS et al., 1997)

single-channel

Os *microarrays single-channel* são, de modo geral, semelhantes aos *double-channel*, mas, nesse caso, somente um corante fluorescente é utilizado por vez. Ou seja, os *arrays* são projetados para dar estimativas sobre os níveis absolutos de expressão gênica. Assim, para a comparação de duas condições são necessárias duas hibridizações separadas.

Uma característica principal desse tipo de técnica é o fato de que amostras com ruído não afetam os dados derivados de outras amostras, pois, em cada *array* somente é utilizado dados de uma única amostra. Isso é uma vantagem em relação a técnica *double-channel*, que pelo fato de trabalhar com mais de uma amostra por vez, se uma única amostra de má qualidade for colocada na hibridização em conjunto com a demais, pode afetar toda a precisão do resultado, mesmo que o restante das amostras sejam todas de alta qualidade.

Como desvantagem dos sistemas *single-channel*, quando comparado como o sistema *double-channel*, tem-se a necessidade de o dobro de *microarrays* para comparar amostras em um experimento. Entre os sistemas atuais disponíveis do tipo *single-channel*, destaca-se o *Affymetrix* "*Gene chip*".

Todos os conjuntos de dados de expressão gênica utilizados neste trabalho foram obtidos usando *spotted microarrays* em conjunto com escaneamento *double-channel*, denominados por cDNA a partir daqui, ou através de *microarrays* de oligonucleotídeos utilizando detecção *single-channel*. Nesse último caso, todos os *arrays* eram produzidos pela *Affymetrix*, sendo, daqui em diante, chamados de *Affymetrix*.

2.3 Análise Computacional

A implementação de uma análise de expressão gênica bem sucedida requer o desenvolvimento de vários protocolos laboratoriais, bem como o desenvolvimento de bases de dados e ferramentas de *software* para uma eficiente coleta e análise dos dados. Apesar de vários desses protocolos terem sido publicados de forma detalhada, as ferramentas computacionais necessá-

rias para tal análise estão evoluindo a cada dia, embora, sem um consenso de qual seria a melhor técnica para encontrar padrões na expressão gênica. De fato, está ficando cada vez mais claro que pode nunca existir uma melhor abordagem e que a aplicação de várias técnicas pode fornecer diferentes perspectivas dos dados explorados. Assim, a escolha de quais técnicas/algoritmos utilizar para fazer a análise dos dados é uma parte crucial no projeto dos experimentos envolvendo expressão gênica (QUACKENBUSH, 2001).

Como será mostrado na Seção 2.4, muitos dos trabalhos desenvolvidos no contexto de expressão gênica focam nos aspectos de identificação de genes diferencialmente expressos, que poderiam indicar possíveis genes alvos a serem estudos mais detalhadamente em nível molecular, ou no uso de técnicas de aprendizado de máquina para melhorar o diagnóstico/resultado de pacientes com câncer. Nesses contextos, os algoritmos de agrupamento, principalmente as chamadas técnicas clássicas, vêm se mostrado ferramentas valiosas na análise exploratória dos dados. Elas organizam os dados em grupos (*clusters*), fornecendo um meio de explorar e verificar estruturas presentes nos dados (D'HAESELEER, 2005).

2.4 Trabalhos Relacionados

Nesta Seção são resumidos os estudos realizados pelos autores dos trabalhos de onde os conjuntos de dados construídos/utilizados nesta dissertação foram retirados. Os estudos aparecem em ordem alfabética ascendente do sobrenome do primeiro autor do trabalho e são identificados pelos seus respectivos títulos.

Alizadeh et al. (2000) - Distinct types of diffuse large B-cell lymphoma identified by gene expresion profile. Em doenças como o linfoma difuso de grandes células B (*diffuse large B-cell lymphoma* - DLBCL), tentativas de definir subgrupos baseados na sua morfologia, em geral, falharam. Apesar de existirem alguns estudos anteriores, não haviam evidências fortes para afirmar que o tipo de linfoma em questão (DLBCL) abriga, pelo menos, duas subclasses diferentes. Os autores focam o estudo na determinação de quando o perfil de expressão gênica pode subdividir essa classe de linfoma em duas subclasses com comportamentos moleculares mais

homogêneos e sugerem que essas subclasses devem ser consideradas duas doenças diferentes.

Para isso, um algoritmo hierárquico (*weighted pair-group method with centroid average*), usando uma variação da correlação de Pearson como medida de proximidade, foi utilizado para agrupar os genes baseado na similaridade da expressão deles ao longo de todas as amostras. O mesmo tipo de algoritmo, mas usando a correlação de Pearson, foi utilizado para agrupar as amostras de tumores e células com base na similaridade da expressão de tais genes. Como resultado, foi definido que os dois tipos de DLBCL são diferenciados um do outro por centenas de diferentes genes, muitos dos quais podem contribuir para o comportamento maligno do tumor.

Armstrong et al. (2002) - MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Um subgrupo de leucemias agudas com uma translocação cromossômica envolvendo o gene de leucemia de linhagem mista (*mixed-lineage leukemia* - MLL), tipicamente tendo morfologia linfoblástica, vem sendo classificado como leucemia linfoblástica aguda (*acute lymphoblastic leukemia* - ALL). Entretanto, diferentemente de outros tipos de leucemia, a presença da translocação MLL resulta em rápidos relapsos após o tratamento com quimioterapia. Os autores sugerem que elas constituem uma doença a parte, denotada por MLL, e mostram que existem diferenças no nível de expressão gênica robustas o suficiente para classificar corretamente as leucemias entre MLL, ALL e AML (*acute myeloid leukemia* - leucemia mielóide aguda). Mais precisamente, técnicas de aprendizado de máquina supervisionado e análise estatística foram utilizadas e revelaram que leucemias linfoblásticas com translocação MLL podem ser claramente separadas das leucemias linfoblásticas agudas e mielóides agudas.

Bhattacharjee et al. (2001) - Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Carcinoma de pulmão corresponde a mais de 150.000 mortes todos os anos nos Estados Unidos, excedendo a taxa de mortalidade dos cânceres de mama, próstata e cólon juntos. A atual classificação desse câncer é baseada em características clinicopatológicas e tem como principal subdivisão os carcinomas de células pequenas (*small-cell lung carcinoma* - SCLC) e carcinoma de células não-pequenas

(*non-small-cell lung carcinomas* - NSCLC). O conhecimento mais profundo da base molecular e classificação de carcinoma de pulmão poderia ajudar a prever o resultado dos pacientes. Sabendo disso, os autores realizaram uma análise de expressão gênica envolvendo amostras de carcinoma de pulmão.

Embora os métodos hierárquicos de agrupamento forneçam uma abordagem poderosa para descobrimento de subclasses, as mesmas não oferecem meios para determinação de um grau de confiança das classe descobertas. Contudo, a combinação de técnicas de agrupamento probabilístico de *bootstrap* com métodos hierárquicos permitiram que os autores medissem a "força" dos agrupamentos formados. Além disso, diferentes métodos de agrupamento identificaram subclasses distintas de adenocarcinoma de pulmão.

Bittner et al. (2000) - Molecular classification of cutaneous malignant melanoma by gene expression profiling. Os tipos mais comuns de cânceres em humanos são os neoplasmas da pele. Com exceção de esforços para identificar preditores de resultados para o melanoma, não existem marcadores moleculares ou patológicos que definam subconjuntos desse neoplasma. Assim, os autores apresentam a descoberta de uma subclasse de melanoma identificada através de análise matemática de expressão gênica em um conjunto de amostras de pacientes com esse tipo de câncer.

Vários métodos foram utilizados para visualizar os padrões de relacionamento de expressão gênica entre as amostras de tumores de melanoma. Por exemplo, um dendrograma foi construído utilizando um algoritmo de agrupamento hierárquico (ligação média) que fazia uso da correlação de Pearson para medir similaridade entre todos os pares de amostras. Outros recursos foram utilizados para visualização, como uma plotagem tridimensional de um *multidimensional scaling* (MDS) e a aplicação de um algoritmo de agrupamento não-hierárquico CAST (*cluster affinity search technique*), para definir grupos experimentais. Essa última técnica usa uma abordagem baseada em grafos e não faz qualquer suposição sobre a função de similaridade usada ou o número de grupos procurado.

Bredel et al. (2005) - Functional network analysis reveals extended gliomagenesis

pathway maps and three novel MYC-Interacting genes in human gliomas. Os autores usaram o método *relevance network analysis* (método exploratório não-supervisionado) como um *framework* conceitual para explorar a patobiologia de gliomas humanos, baseados na suposição de que o comportamento das células de glioma é um atributo contextual de padrões distintos de interações entre múltiplos genes. Para isso, apresentaram uma abordagem integrada que combina a determinação de perfil (*profiling*) de expressão gênica em um conjunto de amostras de glioma humano usando estatística inferencial e descritiva para análise de funções-chaves e rotas (*pathways*) associadas com a gliomagênese.

A técnica SAM (*significance analysis of microarrays*) não-pareada foi usada para identificar genes diferencialmente expressos entre tecidos de cérebro normais e tecido com glioma ou entre tecidos normais e diferentes tipos de glioma. Como resultado, um conjunto de genes foram considerados significantemente ligados ao gliomagênese. Também foi feito, uma análise usando PCA (principal component analysis - análise de componentes principais) com os três primeiros componentes que mostrou uma clara separação entre as classes glioblastoma (GBM), oligodendroglial (OG) e normal (B).

Chen et al. (2002) - Gene Expression Patterns in Human Live Cancers. Carcinoma hepatocelular (*hepatocelullar carcinoma* - HCC) é o tipo de tumor maligno de fígado mais comum e está entre as cinco principais causas de morte por câncer no mundo. A atual classificação patológica clínica padrão de HCC é limitada quando relacionada ao fator previsão de resultado do tratamento. Os autores alegam que há uma necessidade de identificação de marcadores moleculares para ajudar em um diagnóstico mais precoce e preciso e na classificação do HCC. Assim, o objetivo do autores é caracterizar os programas de expressão gênica associados ao HCC como um passo para um melhor entendimento da patofisiologia molecular e para melhores métodos para detecção, diagnóstico e classificação do HCC.

Para isso, os autores aplicaram um algoritmo hierárquico com ligação média, utilizando a correlação de Pearson como medida de proximidade, tanto nas amostras como nos genes. Com isso, quando as amostras de tecidos foram agrupadas hierarquicamente, HCC e amostras

de tecidos não-tumorais foram claramente divididos em dois ramos separados do dendrograma gerado. Os padrões de expressão gênica nas amostras de HCC também foram distintas daquelas associadas a tumores metastáticos para o fígado.

Chowdary et al. (2006) - Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. O uso de RNA como material inicial indica que os tecidos devem ser congelados instantaneamente (*snap-frozen*) ou armazenados em uma solução que pode manter a integridade do RNA (*RNA preservative*). Estudos anteriores demonstrando o valor prognóstico de *microarrays* usaram tecidos congelados instantaneamente, não exploraram o uso de tecidos armazenados em solução conservante de RNA em conjunto com um algoritmo de prognóstico. Com base nisso, os autores compararam pares de amostras de tecidos armazenados das duas formas mencionadas anteriormente de tumores de mama e do cólon.

A análise foi feita avaliando a correlação dos perfis de expressão e a predição de recorrência baseado em dois algoritmos de prognóstico. As amostras de tecido armazenadas em solução conservante de RNA apresentaram altas correlações comparadas com aquelas produzidas por tecidos congeladas instantaneamente. A similaridade entre tecidos dos dois tipos de armazenamento foi medida utilizando um algoritmo hierárquico de agrupamento. Como resultado, as amostras congeladas instantaneamente e armazenadas em solução conservante do mesmo paciente são, na maioria dos casos, mais correlacionadas umas com as outras do que com amostras de outros pacientes.

Dyrskjot et al. (2003) - Identifying distinct classes of bladder carcinoma using microarrays. Câncer de bexiga é uma doença caracterizada por frequentes recorrências. O estágio da doença no diagnóstico é fundamental para determinar o curso do indivíduo afetado. Assim, os autores identificaram subclasses clinicamente relevantes de carcinoma de bexiga usando análise de expressão gênica de tumores desse tipo de câncer.

Na análise foi utilizado um algoritmo hierárquico com ligação média, e uma correlação de Pearson modificada como medida de similaridade, identificando três estágios maiores da doença, denominados Ta, T1 e T2-4, com os tumores Ta sendo posteriormente divididos em

subgrupos. Também foi construído um classificador que foi capaz de discriminar entre tumores benignos e tumores metastáticos.

Garber et al. (2001) - Diversity of gene expression in adenocarcinoma of the lung. Quatro principais subtipos histológicos de câncer de pulmão são geralmente distinguíveis pela morfologia do tumor através do uso de microscópios: adenocarcinoma (AC), carcinoma de células escamosas (*squamous cell carcinoma* - SCC), carcinoma de pulmão de células grandes (*large cell lung carcinoma* - LCLC) e carcinoma de pulmão de células pequenas (*small cell lung carcinoma* - SCLC). Pacientes com tumores de células não-pequenas (SCC, AC e LCLC) são tratados de maneira diferente daqueles com SCLC. A distinção patológica entre os cânceres de pulmão do tipo SCLC e de células não-pequenas é, então, muito importante.

A biologia dos tumores, incluindo a morfologia, é determinada em grande parte por perfis de expressão gênica nas células que englobam o tumor. Os autores apresentam evidências de que a análise dos padrões de expressão gênica pode funcionar como uma base para classificação de câncer de pulmão que reafirma e estende a divisão tradicional existente. Mais precisamente, foram encontrados, utilizando agrupamento hierárquico, padrões de expressão gênica correspondentes a maioria das classes morfológicas de câncer de pulmão. Além disso, definiu-se três subgrupos de AC que se diferenciavam não apenas na expressão gênica, mas também em propriedades clínicas e patológicas, incluindo a sobrevivência do paciente.

Golub et al. (1999) - Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Apesar de a classificação de câncer ter sido melhorada nos últimos anos, não existe uma abordagem genérica para identificar novas classes de câncer (descoberta de classes) ou para atribuir tumores a classes conhecidas (predição de classes). Motivado por isso, os autores apresentam uma abordagem genérica para classificação de câncer baseado no monitoramento da expressão gênica obtida por *microarrays* de cDNA. Como caso de teste, essa abordagem foi descrita e aplicada para um conjunto de amostras de leucemia.

Para a tarefa de predição de classes os autores abordaram três questões: a existência de genes cujo padrão de expressão está fortemente correlacionado com a diferenciação das classes;

como usar uma coleção de amostras conhecidas para criar um preditor de classes capaz de atribuir uma nova amostra a uma das classes; e, por fim, como testar a validade dos preditores de classes. Depois disso, os autores voltaram-se para a questão do descobrimento de classes. Ou seja, eles exploraram quando classes de câncer podem ser descobertas automaticamente. Nessa tarefa, duas questões são discutidas: o desenvolvimento de algoritmos para agrupar tumores usando dados de expressão gênica e a determinação de quando supostas classes encontradas por tais algoritmos tem sentido biológico. Desse modo, os autores optaram por usar como método de agrupamento o *self-organizing maps* (SOM). O método foi aplicado tanto no contexto de distinção de tipos diferentes de leucemia (AML e ALL) quanto no contexto de refinamento de classes (subdivisão da classe ALL em ALL do tipo B (B-ALL) e ALL do tipo T(T-ALL)).

Gordon et al. (2002) - Translation of microarray data into clinically relevant cancer diagnosis tests using gene expression ration in lung cancer and mesothelioma. A distinção patológica entre mesotelioma pleural maligno (*malignant pleural mesothelioma* - MPM) e adenocarcinoma (ADCA) de pulmão pode ser ineficiente usando os métodos já estabelecidos. As estratégias mais recentes de tratamento dependem de um diagnóstico patológico correto. Não raramente, a distinção entre MPM e ADCA é complicada tanto do ponto de vista clínico quanto patológico. Com vista nisso, os autores exploraram uma abordagem alternativa usando expressão gênica para prever parâmetros clínicos em câncer.

Mais precisamente, os autores exploraram a viabilidade de um teste simples e com ampla aplicabilidade que usa os níveis de expressão gênica e racionalmente escolhe limiares para distinguir precisamente tecidos diferentes geneticamente. Esse tipo de abordagem permite a análise de amostras individuais sem referenciar um conjunto de dados adicional de treinamento.

Khan et al. (2001) - Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Tumores são, em geral, diagnosticados através de sua morfologia (histopatologia) e expressão de proteínas. Entretanto, cânceres pouco diferenciados podem ser difíceis de diagnosticar por histopatologia tradicional. Além disso, a aparência histológica do tumor não revela os processos biológicos que contribuíram para o pro-

cesso maligno. Então, os autores desenvolveram um método para classificação de diagnóstico de câncer baseado em perfis de expressão gênica.

Para abordar o problema os autores usaram modelos baseados em redes neurais artificiais (RNA) lineares (perceptrons). O modelo calibrado utilizado classificou corretamente todos os exemplos de treinamento e não mostrou nenhuma evidência de *over-training*, demonstrando a robustez dessa técnica. Os autores também identificaram os genes que mais contribuíam para essa classificação, sendo possível definir um conjunto mínimo deles capaz de classificar corretamente as amostras dentro das suas categorias. Também foi aplicado um algoritmo hierárquico de agrupamento que agrupou corretamente todas as amostras presentes no conjunto de testes.

Laiho et al. (2007) - Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. Carcinomas colorretais do tipo *serrated* (*serrated colorectal carcinomas* - CRC) são morfologicamente diferentes das CRCs normais e seguem um caminho de formação diferente. Entretanto, não está claro onde estão fundamentalmente essas diferenças biológicas. A hipótese que os CRCs do tipo *serrated* são biologicamente diferentes dos CRCs convencionais se baseiam, principalmente, em evidências vindas de dados morfológicos e, considerando a heterogeneidade dos adenomas, seria de fundamental importância esclarecer quando, de fato, os CRCs do tipo *serrated* formam um novo tipo de CRC. Assim, os autores realizaram uma análise imunohistoquímica e de expressão gênica de microarray para investigar a base molecular do CRC do tipo *serrated*.

O estudo mostrou que os CRC do tipo *serrated* se diferenciam dos CRCs convencionais tanto em nível morfológico quanto molecular. Mais precisamente, o uso de um algoritmo hierárquico de agrupamento (usando correlação de spearman) e de uma técnica supervisionada como preditor (*k-nearest neighbor* - *k*-NN) mostrou que os CRCs do tipo *serrated* tem um perfil de expressão gênica distinto do adenocarcinoma convencional.

Lapointe et al. (2004) - Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Câncer de próstata é o terceiro tipo mais comum de câncer e representa cerca de 6% das causas de morte por câncer em homens. Esse tipo de câncer mostra um

comportamento clínico heterogêneo, variando desde relativamente indolente até uma doença metastática agressiva. Assim, tornou-se uma importante questão clínica saber quando e quão agressivo deve ser o tratamento de tais pacientes com câncer de próstata localizado. Os autores reportam um estudo, baseado em *microarray*, de câncer de próstata na perspectiva de expressão gênica. Esse estudo levou a identificação de subtipos, clinicamente e biologicamente relevantes, do tumor em questão. Além disso, eles demostraram que níveis de expressão de proteínas para dois genes, servindo como marcadores para subtipos do tumor, são fortes preditores de recorrência.

Para explorar o relacionamento entre as amostras e características da expressão gênica, os autores aplicaram um algoritmo de agrupamento hierárquico tanto nas amostras quanto nos genes. Os resultados mostraram claramente uma separação entre amostras de tumor e tecido normal na árvore hierárquica gerada pelo algoritmo de agrupamento utilizado. Notavelmente, o algoritmo de agrupamento também dividiu as amostras de tumor em três subgrupos maiores (subtipos I, II e III), que também foi visível em uma plotagem das maiores componentes de uma PCA. Tumores avançados e de alto nível, bem como tumores associados com a recorrência, foram desproporcionalmente representados entre dois dos três subgrupos encontrados (II e III).

Liang et al. (2005) - Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. Glioblastoma multiforme (GBM) é a forma mais comum de glioma, caracterizado por instabilidade genética, variabilidade intratumoral e comportamento clínico imprevisível. Apesar de recentes estudos de gliomas documentarem padrões na expressão gênica associados a específicos graus clínicos de oligodendroglioma (ODG) e astrocitoma (*astrocytoma* - OAC), perfis de expressão gênica associados com o comportamento clínico heterogêneo do GBM não foram explorados e, ainda existe a falta um marcador molecular confiável que consiga prever a sobrevivência de pacientes com esse tipo de tumor. Tendo em vista isso, os autores, utilizando microarrays de cDNA, analisaram padrões de expressão gênica em uma série de amostras de GBM e identificaram prováveis marcadores relacionados a sobrevivência do paciente.

Para explorar os relacionamentos entre as amostras de tumores e os genes que elas expressam, foi realizada uma análise de agrupamento com um algoritmo hierárquico aglomerativo nas amostras e genes. Os padrões de expressão gênica de GBM mostraram alta variabilidade e separaram esse tipo de tumor em dois grupos, onde um dos quais estava mais relacionado com as amostras de OAC/ODG e amostras normais do que amostras de GBM. Na análise de sobrevivência dos paciente, inicialmente, a hipótese era de que as amostras de GBM mais similares aos OAC/ODG e de cérebro normal poderiam variar clinicamente em relação aos tumores restantes. Entretanto, nenhuma das características clínicas abordadas diferenciaram-se significantemente. Como alternativa, os autores agruparam hierarquicamente as amostras que se tinha informações relativas ao tempo de sobrevivência. O resultado revelou uma diferença significativa entre as sobrevivências dos dois grupos: um com média de 25 meses e o outro com 4 meses em média de sobrevivência.

Nutt et al. (2003) - Gene expression-based classification of malignatn gliomas correlates better with survival than histological classification. O sistema de classificação de tumores cerebrais mais usado, o WHO (*World Health Organization*), classifica os gliomas de acordo com características histológicas definidas a partir de uma célula normal. Tumores de histologia clássicas claramente mostram essas características e assemelham-se com imagens vistas em livros textos. Esses casos poderiam ser diagnosticados por qualquer patologista. Entretanto, há situações em que o sistema WHO é problemático, principalmente, porque o diagnóstico patológico permanece subjetivo. Sabendo que existe uma necessidade crítica de um método de classificação de gliomas objetivo e clinicamente relevante, os autores investigaram se perfis de expressão gênica, em conjunto com métodos de predição computacionais, podem ser usados para definir subgrupos de gliomas de altos graus de maneira mais objetiva e consistente que a patologia padrão.

Foi construído um modelo preditor de classes com o uso de um algoritmo de classificação *k*-NN. Esse método foi efetivo não somente para classificar gliomas de alto grau, como também, aparentemente, forneceu um preditor de diagnóstico mais preciso.

Pomeroy et al. (2002) - Prediction of central nervous system embryonal tumor outcome based on gene expression. Tumores embrionários do sistema nervoso central (*central nervous system* - CNS) representam um grupo heterogêneo de tumores sobre o qual pouco é conhecido biologicamente, e o diagnóstico, com base somente na aparência morfológica, é controverso. Os autores abordam esse problema desenvolvendo um sistema de classificação baseado nos dados de expressão gênica derivados de amostras de 99 pacientes.

Os autores procuraram tratar três problemas distintos: diferenciar os tumores embrionários CNS uns dos outros; heterogeneidade dentro dos tumores meduloblastomas; e heterogeneidade em resposta ao tratamento de meduloblastoma. A separação encontrada na plotagem das maiores componentes de uma PCA foi confirmada usando um algoritmo hierárquico de agrupamento. Por último, foram utilizadas técnicas de aprendizado supervisionadas para identificar os genes mais correlacionados com a distinção dos tipos de tumores. No segundo problema, sabe-se que a principal subdivisão da classe meduloblastoma é a meduloblastoma desmoplástica, entretanto, seu diagnóstico é extremamente subjetivo. Os autores abordaram o problema a partir do ponto de vista de expressão gênica e conseguiram comprovar a existência de tal subclasse pela detecção precisa e estatisticamente significante de uma assinatura de expressão gênica de histologia desmoplástica. O último problema, foi primeiro abordado com técnicas de agrupamento, utilizando SOM e, depois, usando o método *k*-NN para diferenciar entre pacientes vivos após tratamento (sobreviventes) e aqueles que morreram (falhas), conseguindo resultados satisfatórios, ou seja, a distinção entre as classes eram estatisticamente significantes.

Ramaswamy et al. (2001) - Multiclass cancer diagnosis using tumor gene expression signatures. A classificação de câncer é baseada na interpretação subjetiva de informações tanto clínica quanto histopatológica buscando atribuir tumores à categorias aceitas atualmente baseado no tecido de origem do tumor. Entretanto, informações clínicas podem ser incompletas ou enganosas. Diagnósticos moleculares oferecem a promessa de uma classificação precisa, objetiva e sistemática, mas não são amplamente aplicadas porque marcadores moleculares para os mais sólidos tumores ainda não foram identificados. Tendo em vista isso, os autores criaram uma base de dados de expressão gênica contendo amostras de quatorze classes de câncer mais

comuns e, utilizando um método analítico inovador, demonstraram que é possível, de fato, uma classificação precisa de várias classes de câncer.

Os autores utlizaram como primeira alternativa um algoritmo hierárquico de agrupamento (ligação média) e SOM para agrupar as amostras de todos os tipos de câncer. Mas, os resultados levaram a concluir que aprendizado não-supervisionado não captura a distinção dos tecidos de origem dos tumores. A segunda alternativa, um método supervisionado inovador baseado no algoritmo *support vector machine* (SVM) para fazer classificação multiclasses, mostrou resultados indicando que muitos cânceres mantêm a identidade do tecido de origem durante o processo de metástase, sugerindo que abordagens baseadas em expressão gênica para o diagnóstico de metástases de origem desconhecida pode ser possível.

Risinger et al. (2003) - Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer. Estudos anteriores de oncogenes e de alterações em genes supressores de tumores sugeriram que existem diferenças na patogênese dos vários tipos histológicos de câncer no endométrio. A patologia molecular do câncer de endométrio não é completamente entendida. Então, para elucidar tal patologia, os autores examinaram padrões de expressão gênica em cânceres endometróides, não-endometróides e endométrios normais.

A análise consistiu inicialmente em procurar similaridades entre os perfis de expressão das amostras cancerígenas e normais utilizando a técnica *multidimensional scaling* não-supervisionado. Esta análise indicou que os padrões de expressão gênica entre as amostas foram de alguma forma variáveis e que essas diferenças foram suficientes para agrupar a maioria dos casos nos seus respectivos grupos. Posteriormente, para entender melhor a biologia por trás desses grupos, foram agrupados os genes que mais se diferenciavam entre os grupos utilizando um algoritmo hierárquico de agrupamento. Além disso, ainda foi realizada uma abordagem supervisionada para classificar diferentes tipos de câncer de endométrio.

Shipp et al. (2002) - Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. Os linfomas difusos de grandes células

B (DLBCL) são os neoplamas linfóides mais comuns, representando 30-40% do linfomas não-Hodgkin adultos. Apesar de um grupo de pacientes de DLBCL serem curados com os métodos de quimioterapia atuais, muitos não resistem a doença. Isso leva a uma necessidade de identificação de abordagens de tratamento mais racionais e molecularmente definidas. Com base nisso, os autores descrevem um algoritmo de classificação supervisionado que define leucemias agudas que surgem de diferentes linhagens (linfóide x mielóide) usando dados de expressão gênica de *microarray*.

O estudo demonstra o potencial do reconhecimento baseado em *microarray* de padrões de expressão gênica para predição de resultados em pacientes com DLBCL. O trabalho também mostra a diferença de abordagens supervisionadas e não-supervisionadas nesse contexto. Um algoritmo hierárquico de agrupamento foi utilizado para investigar a conexão entre os modelos de predição e resultados da doença.

Singh et al. (2002) - Gene expression correlates of clinical prostate cancer behavior. Tumores na próstata estão entre os cânсers mais heterogêneos, tanto histologicamente quando clinicamente. Exite uma necessidade de marcadores de prognósticos robustos capazes de identificar pacientes com riscos de recorrência. Os dados dos autores sugerem que modelos baseados em expressão gênica podem ajudar a identificar pacientes com grande risco de recorrência e facilitar a aplicação racional das terapias atuais.

Os autores buscaram verificar a possibilidade de usar as diferenças na expressão gênica de amostras de tecido de células normais e com câncer para prever a identificação de novas amostras de tecido de próstata. Para isso, foi construído um preditor usando um algoritmo de aprendizado de máquina supervisionado (k-NN). A conclusão foi que diferenças na expressão podem ser usadas para prever a identidade de amostras desconhecidas de próstatas. Uma assinatura de expressão gênica foi encontrada usando um algoritmo hierárquico de agrupamento nos dados analisados com o objetivo de saber se existem padrões na expressão gênica que podem descrever ou prever as diferenças entre o comportamento clínico entre amostra de tumores. Por último, uma tentativa de criar um preditor de recorrência após prostatectomia levou a bons

resultados, porém com suspeitas de superotimização.

Su et al. (2001) - Molecular classification of human carcinomas by use of gene expression signatures. O tratamento eficiente de pacientes de câncer depende fundamentalmente do conhecimento do sítio anatômico primário da origem do tumor. Então, a classificação de cânceres em grupos distintos baseada no seus tecidos de origem e aparência histopatológica é importante para o gerenciamento ótimo dos pacientes. Baseado nisso, os autores construíram um esquema de classificação molecular para carcinoma de múltiplos cânceres, os quais correspondem a cerca de 70% das mortes relacionadas a câncer nos Estados Unidos.

Análises iniciais dos dados por métodos não-supervisionados, como o algoritmo hierárquico de agrupamento, mostraram que é possível agrupar cânceres de alguns sítios anatômicos baseado apenas nos padrões dos genes mais variavelmente expressos. Entretanto, para alguns cânceres a distinção ficou difícil de ser feita baseada em aprendizado não-supervisionado. Então, foram feitos vários preditores multiclasses baseados em métodos de correlação ponderada ou métodos de aprendizado supervisionado. Foi constatado que técnicas, como o SVM, que não faz suposições sobre a distribuição dos dados, tem um desempenho significativamente superior. Por fim, concluiu-se que é possível, para o conjunto de dados utilizado, identificar subconjuntos de genes com expressão estritamente relacionada a uma classe de tumor.

Tomlins et al. (2007) - Integrative molecular concept modeling of prostate cancer progression. - Os autores analisaram a progressão do câncer de próstata usando um recurso alternativo, o *Molecular Concept Map* (MCM), um *framework* analítico para explorar as redes de relacionamento entre um conjunto de "conceitos moleculares" (conjuntos de genes biologicamente relacionados).

Para identificar assinaturas na expressão gênica, os autores carregaram os conjuntos de dados no Oncomine (RHODES et al., 2004), utilizando as ferramentas e recursos disponíveis em tal repositório. Para identificar correlações moleculares na progressão do câncer de próstata, as assinaturas foram analisadas utilizando o MCM. Combinando uma determinação de perfil com uma análise integrada, os autores identificaram conceitos correlacionados com transições

histológicas observadas na progressão do câncer de próstata.

West et al. (2001) - Predicting the clinical status of human breast cancer by using gene expression profiles. Fatores prognósticos e preditivos são ferramentas indispensáveis no tratamento de pacientes com doença neoplástica. Na maioria dos casos, tais fatores residem em uma característica de superfície ou histologia de célula específica. Os métodos tradicionais de caracterização fenotípica são frequentemente limitados e não possuem habilidades suficientes para discernir pequenas diferenças que podem ser importantes para um melhor entendimento do tumor e para o avanço de terapias para o tratamento da doença. Então, os autores desenvolveram um modelo de regressão bayesiano que fornece capacidade preditiva baseado em dados de expressão gênica derivados de *microarray* de DNA de uma série de amostras de tumores primários de câncer de mama.

O estudo demonstrou que fenotipos clinicamente relevantes podem ser determinados para amostras de tumor de mama primário através de análise de expressão gênica. Um ponto chave do trabalho é a capacidade de identificação não somente de genes altamente expressos, mas de genes cuja expressão está correlacionada com o fenótipo sem levar em consideração o nível de expressão.

Yeoh et al. (2002) - Classification, subtype discovery, and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling. Leucemia linfoblástica aguda (ALL) pediátrica é um dos grande sucessos de terapia moderna de câncer, com tratamentos atingindo uma sobrevivência, como um todo, em cerca de 80%. Um ponto crítico no sucesso da abordagem usada para o tratamento é a correta atribuição dos pacientes a grupos de risco específicos. Infelizmente, essa atribuição é um processo árduo e caro, requerendo estudos laboratoriais intensos. Mais precisamente, esses diagnósticos necessitam de um número de profissionais especialistas que, muitas vezes, não estão disponíveis em determinados locais.

Para determinar se perfis de expressão gênica poderiam identificar subgrupos de ALL, foi utilizado um algoritmo hierárquico de agrupamento, tanto nas amostras quanto no genes, sendo identificado seis subtipos maiores de ALL. Essa separação também pode ser vista usando um

procedimento de análise discriminante com variância (*discriminant analysis with variance* - DAV). Como um dos principais objetivos do trabalho era determinar se o perfil de expressão gênica pode identificar com precisão importantes subtipos de leucemia, foram usados algoritmos de aprendizado supervisionado. Entre outros métodos, primeiro, foi usada uma técnica de árvore de decisão e depois um SVM.

2.5 Considerações Finais

Como mencionado anteriormente, o uso de técnicas de agrupamento na descoberta de subtipos de câncer tem atraído grande atenção da comunidade científica (D'HAESELEER, 2005; QUACKENBUSH, 2001). Apesar de existem métodos de agrupamento especializados em tratar dados de expressão gênica, o uso de tais técnicas é, em geral, restrito às técnicas clássicas de agrupamento, que muitas vezes não se adequam bem a tais tipos de dados.

Até o momento, não existe nesse contexto nenhum estudo na literatura que faz uma comparação em grande escala de diferentes técnicas de agrupamento. Visando isso, este trabalho busca fazer um estudo comparativo de técnicas de agrupamento na análise de dados de expressão gênica. Mais precisamente, foram selecionadas sete técnicas de agrupamento, de modo que tais técnicas explorem diferentes aspectos dos dados. Desse modo, são analisados tanto técnicas "clássicas" de agrupamento quanto novos algoritmos que levam em consideração características inerentes a dados de alta dimensionalidade.

Além disso, esta dissertação traz como contribuição a disponibilização de um conjunto de bases de dados que podem ser usadas e compartilhadas por pesquisadores como uma base estável e confiável para avaliação e comparação de diferentes técnicas de aprendizado de máquina.

3 *Análise de Agrupamentos*

A capacidade de obter dados de expressão gênica ultrapassou a habilidade humana de analisá-los manualmente. O uso de técnicas de agrupamento permite dividir os dados em um número menor de categorias, refinando-os até um nível em que sua análise se torna viável (D'HAESELEER, 2005).

O objetivo da análise de agrupamentos é encontrar a estrutura subjacente aos dados, colocando as observações (instâncias ou objetos) mais semelhantes em grupos (HAIR et al., 2005). Para isso, devemos abordar pelo menos quatro questões básicas:

- Como medir a similaridade entre os objetos?
- Como formar os agrupamentos?
- Quantos grupos formar?
- Como validar os agrupamentos formados?

A resposta para o primeiro questionamento pode ser feita com base nas diversas medidas de proximidade existentes na literatura. Basicamente três tipos de medidas de proximidade são utilizadas para medir similaridade no contexto de análise de agrupamentos: as medidas correlacionais, as medidas baseadas em distâncias e as medidas de associação (HAIR et al., 2005). Cada um desses métodos representa uma perspectiva particular de similaridade. Tanto as medidas correlacionais quanto as medidas de distância requerem dados métricos (numéricos), ao passo que as medidas de associação são para dados não-métricos (categóricos). Este trabalho

se restringe ao uso das medidas correlacionais e as baseadas em distância, uma vez que os dados são obtidos a partir de *microarrays*, o que implica que são dados completamente numéricos.

A segunda pergunta diz respeito a que procedimento deve ser utilizado para formar grupos a partir dos dados. Basicamente, exitem duas categorias de técnicas de agrupamento: as hierárquicas e as particionais (não-hierárquicas). As técnicas hierárquicas montam uma estrutura hierárquica dos dados em forma de árvore. Por outro lado, os algoritmos particionais subdividem os dados em um número de subgrupos sem que haja qualquer tipo de relacionamento hierárquico entre os grupos formados (JAIN; DUBES, 1988; D'HAESELEER, 2005; HAIR et al., 2005). Neste trabalho, são utilizados tanto técnicas hierárquicas quanto particionais.

A resposta para a terceira pergunta, ou seja, o número de grupos que deve estar presente na solução final gerada por uma técnica de agrupamento, não é respondida facilmente, pois, em geral, não se tem esse conhecimento *a priori*. Existem algumas heurísticas presentes na literatura que guiam a escolha da quantidade de grupos baseada na estrutura do agrupamento formado (HAIR et al., 2005). Entretanto, como neste trabalho o objetivo é fazer uma comparação do desempenho das técnicas de agrupamento, o número de classes para cada conjunto de dados já é conhecido e, portanto, a descoberta do número exato de grupos não é prioridade na discussão. Em vez disso, um outro tipo de análise é feita: o impacto de agrupar os dados com mais grupos que o número real de classes (cobertura reduzida).

Por fim, para responder a última questão são necessárias medidas de qualidade para avaliar os agrupamento formados. Em geral, são utilizados dois tipos de critérios para isso: os índices internos (baseados nas propriedades intrínsecas dos dados) e os índices externos (utilizam informações adicionais sobre os dados que não são utilizadas no processo de agrupamento) (D'HAESELEER, 2005). Em geral, os índices internos são tendenciosos para algum tipo de algoritmo, ou seja, eventualmente, algum algoritmo maximiza o critério utilizado pelo índice (JAIN; DUBES, 1988; MILLIGAN; COOPER, 1986). Por esse motivo, é utilizado neste trabalho somente o índice externo *correted Rand* (JAIN; DUBES, 1988) que não beneficia nenhum tipo de técnica de agrupamento.

O restante deste capítulo é dedicado a detalhar as questões de similaridade, técnicas de agrupamento e validação de agrupamento abordadas anteriormente. Mas, primeiramente, algumas definições básicas devem ser levadas em consideração para o entendimento das próximas seções. A unidade básica de dados, como mencionado anteriormente, é chamada instância (objeto), que é denotada por um vetor com d dimensões, cujos componentes são escalares, chamados de atributos (características). Assim, a i-ésima instância é denotada por um vetor $\mathbf{x}_i$ e o atributo j da instância i é denotada por x^*_{ij}. O "*" indica que os dados não sofreram qualquer tipo de transformação. Então, se um conjunto de dados possui n instâncias, a matriz de dados X^* está no formato $n \times d$, onde cada linha é uma instância $\mathbf{x}$ com d atributos. De maneira mais formal:

$$X^* = [\mathbf{x}^*_1 \quad \mathbf{x}^*_2 \quad \ldots \quad \mathbf{x}^*_n]^T = \begin{bmatrix} x^*_{11} & x^*_{12} & \ldots & x^*_{1d} \\ x^*_{21} & x^*_{22} & \ldots & x^*_{2d} \\ \ldots & \ldots & \ldots & \ldots \\ x^*_{n1} & x^*_{n2} & \ldots & x^*_{nd} \end{bmatrix}$$

3.1 Medidas de Proximidade

Para que as técnicas de agrupamento possam formar grupos, colocando instâncias próximas em um mesmo grupo, é necessário que essa proximidade seja medida de alguma maneira. Na literatura de expressão gênica é comum utilizar medidas de distância ou correlação para medir proximidade (QUACKENBUSH, 2001; D'HAESELEER, 2005).

As medidas de distância se concentram na magnitude dos valores. Elas medem proximidade baseado na diferença entre os valores que representam as instâncias. Já as medidas correlacionais se baseiam no comportamento dos atributos ao longo da instância. Este trabalho utiliza a distância euclidiana, correlação de Pearson, coeficiente de correlação de Spearman e Cosseno como medidas de similaridade/dissimilaridade.

Para fins de notação, no contexto das próximas seções, a palavra ponto (ou vetor) será

utilizada significando uma instância do conjunto de dados. Cada um desses pontos é descrito por um conjunto de atributos cujos valores representam níveis de expressão gênica. Também serão mostrados alguns exemplos para um melhor entendimento das diferenças e semelhanças entre tais medidas.

3.1.1 Medidas Correlacionais

As medidas correlacionais representam proximidade pela correspondência de padrões ao longo dos atributos. Uma medida correlacional de proximidade não analisa a magnitude, mas sim o padrão de comportamento entre os atributos de duas instâncias (HAIR et al., 2005). A seguir, são descritos os tipos de medida de correlação usados nesta dissertação: cosseno, correlação de Pearson e coeficiente de correlação de Spearman.

Cosseno

O cosseno usa a separação angular entre os vetores de instâncias como medida de similaridade. Tal medida pode ser definida como:

$$Cos(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}^T\mathbf{y}}{\| \mathbf{x} \| \| \mathbf{y} \|} \tag{3.1}$$

em que $\| \mathbf{x} \| = \sqrt{\sum_{i=1}^{d} x_i^2}$ e $\| \mathbf{y} \| = \sqrt{\sum_{i=1}^{d} y_i^2}$ são as normas dos vetores $\mathbf{x}$ e $\mathbf{y}$, respectivamente. Essa medida é invariante a rotações, mas não a transformações lineares no espaço dos vetores.

Correlação de Pearson

A correlação de Pearson é definida como:

$$Pearson(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}_e^T \quad \mathbf{y}_e}{\| \mathbf{x}_e \| \| \mathbf{y}_e \|} \tag{3.2}$$

em que $\mathbf{x}_e = [x_1 - \bar{x}, ..., x_d - \bar{x}]$ e $\mathbf{y}_e = [y_1 - \bar{y}, ..., y_d - \bar{y}]$ são conhecidos como vetores de diferença, e $\bar{x} = \frac{1}{d}\sum_{i=1}^{d} x_i$ e $\bar{y} = \frac{1}{d}\sum_{i=1}^{d} y_i$.

A correlação de Pearson mede a similaridade direcional entre dois pontos, não sendo sensível à magnitude dos valores dos atributos. O valor da correlação varia entre -1 e 1, com 1 significando que os dois pontos possuem exatamente a mesma "forma" (mesmo comportamento); 0, que são completamente não relacionados; e -1 significando que são inversamente relacionados.

Se os dois pontos em questão representam níveis de expressão gênicas, quando dois genes são co-expressos (variam igualmente ao longo do tempo), o valor da correlação de Pearson entre esses pontos deve ficar próximo a 1. Esse é um dos motivos pelo qual a correlação de Pearson vem sendo amplamente utilizada na análise de dados de expressão gênica.

Uma diferença desse tipo de correlação para o cosseno é que a correlação de Pearson não depende diretamente de **x** e **y**, mas dos seu vetores de diferença.

Coeficiente de Correlação de Spearman

O coeficiente de correlação de Spearman é uma medida de correção não-paramétrica, ou seja, não faz nenhum tipo de suposição sobre a distribuição de frequência dos dados. É um caso especial da correlação de Pearson, em que valores dos pontos são convertidos para um *ranking* antes de ser calculada a correlação. No entanto, ao contrário da correlação de Pearson o coeficiente de correlação de Spearman não requer a suposição que a relação entre os pontos é linear. Essas características tornam esse coeficiente mais robusto a *outliers* que outras medidas. A Equação 3.3 define o coeficiente de correlação de Spearman.

$$Spearman(\mathbf{x},\mathbf{y}) = 1 - \frac{6\sum_{i=1}^{d}(r_{x_i} - r_{y_i})^2}{d^3 - d} \tag{3.3}$$

em que, r_{x_i} e r_{y_i} são as posições de x_i e y_i no *ranking* de **x** e **y**.

3.1.2 Medidas de Distância e Transformações dos Dados

Distância Euclidiana

A distância euclidiana funciona como medida de dissimilaridade quando utilizada no contexto de técnicas de agrupamento. Ela mede a distância absoluta entre dois pontos. Assim, quanto mais distantes os pontos, menos similares eles são; e quanto mais próximos, mais similares. A distância euclidiana leva em consideração a magnitude dos valores que compõem o ponto, ou seja, atributos com maiores valores (níveis de expressão mais altos) têm maior influência no cálculo da similaridade. A Equação 3.4 define a distância euclidiana (JAIN; DUBES, 1988).

$$Dist_{Eucl}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \tag{3.4}$$

em que, x_i e y_i são as i-ésimas coordenadas de **x** e **y**.

Exemplo de Aplicação de Medidas

Para ficar clara a diferença entre os tipos de medidas de proximidade, nessa seção será ilustrado um exemplo do uso dessas medidas. Assim, se considerarmos como exemplo, dados obtidos a partir de três funções: **a** = seno(**x**), **b** $= 0.5a$ e **c** $= a - 0.2$; e tais dados forem plotados em duas dimensões, obtem-se o gráfico mostrado na Figura 3.1.

Se as correlações de Pearson, Spearman e cosseno, bem como a distância euclidiana, forem calculadas para esses pontos, tem-se os seguintes valores:

Pearson(**a**, **b**) = 1.00 *Spearman*(**a**, **b**) = 1.00 *Cosseno*(**a**, **b**) = 1.00 $Dist_{Eucl}$(**a**, **b**) = 2.80

Pearson(**a**, **c**) = 1.00 *Spearman*(**a**, **c**) = 1.00 *Cosseno*(**a**, **c**) = 0.96 $Dist_{Eucl}$(**a**, **c**) = 1.58

Pearson(**b**, **c**) = 1.00 *Spearman*(**b**, **c**) = 1.00 *Cosseno*(**b**, **c**) = 0.96 $Dist_{Eucl}$(**b**, **c**) = 3.22

Como dito anteriormente, as medidas correlacionais não analisam a magnitude dos valores, mas sim o padrão dos atributos ao longo das instâncias. Desse modo, quando analisados do

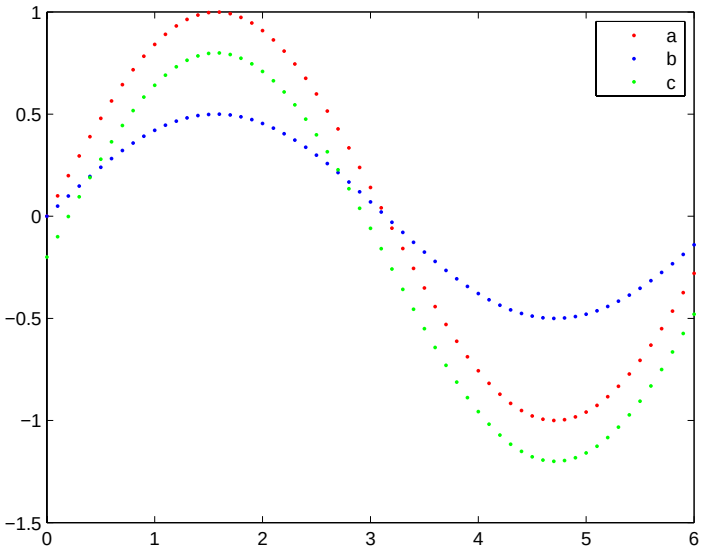


Figura 3.1: Pontos usados para exemplo do cálculo de proximidade.

ponto de vista correlacional, os três pontos (**a**, **b** e **c**) estão altamente correlacionados, ou seja, estão muito próximos uns dos outros. Entretanto, se a distância euclidiana for usada, a perspectiva de proximidade muda. Por exemplo, a correlação entre os pontos **b** e **c** é máxima (1.00), indicando a maior proximidade que dois pontos podem estar. Em contraste, utilizando a distância euclidiana para esses mesmos pontos, a distância calculada é a maior entre todos os pontos existentes, significando a menor proximidade possível entre todos os pontos do exemplo.

Transformações dos Dados

Em muitas situações práticas, um conjunto de dados pode apresentar instâncias cujos atributos estão em diferentes escalas de valores (JAIN; DUBES, 1988; MILLIGAN; COOPER, 1988). Em geral, dados de expressão gênica, principalmente no caso dos obtidos a partir de *chips Affymetrix*, possuem essa característica. Desse modo, para medidas de proximidade, tais como a distância euclidiana, atributos com maiores valores tem maior influência sobre os demais, mesmo que isso não necessariamente implique que são mais importantes para determinação dos agrupamentos.

Em geral, para contornar esse tipo de problema é feita uma transformação nos dados de

modo a deixar os atributos dentro de uma mesma escala. Na literatura existem diversas abordagens para realização desse procedimento, tanto para valores numéricos como para valores categóricos (JAIN; DUBES, 1988; MILLIGAN; COOPER, 1988). Como os dados de expressão gênica estudados aqui são todos obtidos através de *microarray* e, portanto, não possuem valores categóricos, serão considerados somente os casos que tratam valores numéricos. Mais precisamente, são analisados três tipos de transformação de dados: padronização, escalonamento e *ranking* (SOUTO et al., 2008a).

As duas primeiras são amplamente utilizadas em técnicas de agrupamento (JAIN; DUBES, 1988; MILLIGAN; COOPER, 1988). Uma rotaciona e escalona os eixos de modo que os valores dos atributos tenham média igual a 0 e desvio padrão unitário (padronização), e a outra escalona os valores para o intervalo [0,1] ou [-1,1] (escalonamento). O terceiro procedimento apresentado transforma os valores dos atributos em um *ranking* e utiliza-o em lugar dos dados originais. Tal procedimento é mais robusto a *outliers* que as outras duas primeiras (MILLIGAN; COOPER, 1988).

As técnicas de transformação denominadas padronização, escalonamento e *ranking* são apresentadas a seguir.

Padronização

A padronização pode ser definida como:

$$x_{ij} = \frac{x^*_{ij} - \bar{x_j}}{s_j} \tag{3.5}$$

em que, $\bar{x_j} = \frac{1}{n}\sum_{i=i}^{n} x_{ij}$ e $s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2}$ são, respectivamente, a média e o desvio padrão dos valores do atributo j para todas as instâncias.

Esse tipo de transformação, que usa a fórmula *z-score*, faz uma rotação e escalonamento dos eixos de modo que os valores transformados x_{ij} ficam com média igual a zero e desvio padrão unitário. A padronização pode não funcionar adequadamente caso existam diferenças

substanciais entre os desvios-padrões dos grupos presentes no conjunto de dados (MILLIGAN; COOPER, 1988).

Para fins de nomenclatura, essa transformação será referida como Z_1.

Escalonamento

O segundo procedimento utiliza os valores máximos e mínimos dos atributos para realizar a transformação dos dados:

$$x_{ij} = \frac{x^*_{ij} - min(j)}{max(j) - min(j)} \tag{3.6}$$

em que, $min(j)$ e $max(j)$ são, respectivamente, os menores e maiores valores do atributo j. Considerando somente valores positivos os valores de x_{ij} ficam no intervalo [0,1]. Nesse caso, a média e o desvio padrão, diferentemente da padronização, não são constantes para os valores do atributo. No caso de existirem valores negativos no conjunto de dados, é feita uma adaptação da equação anterior. Assim, a nova equação é definida por:

$$x_{ij} = \frac{x^*_{ij} - min(j)}{max(j) - min(j)} \times 2 - 1 \tag{3.7}$$

em que, nesse caso os valores de x_{ij} ficam no intervalo [-1,1].

Ambos os procedimentos de transformação dos dados, padronização e escalonamento, podem ser afetados pela presença de *outliers* no conjunto de dados, o primeiro pelo fato de usar a média do atributos e o segundo por usar os valores máximos e mínimos do atributo.

Ranking

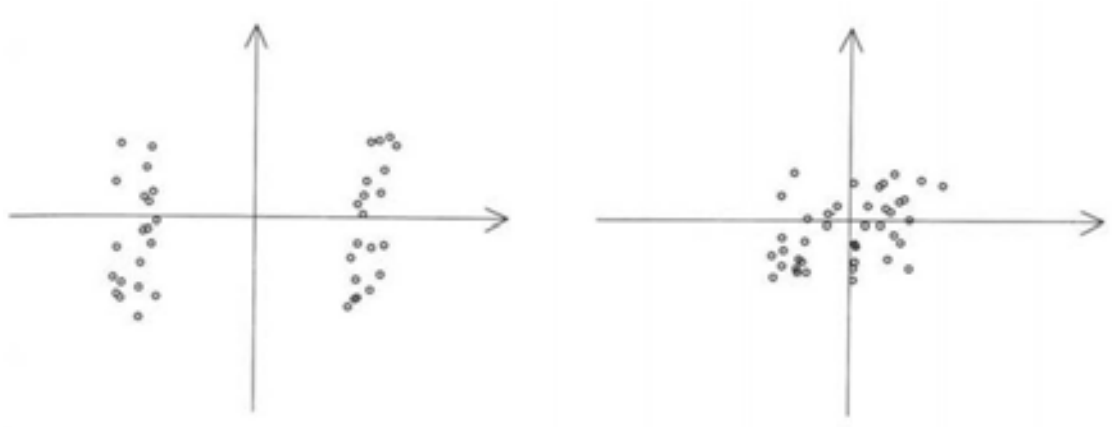
Este procedimento, por usar o conceito de *ranking*, é mais robusto a *outliers* que os dois anteriores. A idéia básica deste procedimento é transformar os valores dos atributos em um *ranking* através da Equação 3.8:

$$x_{ij} = rank(x^*_{ij}) \tag{3.8}$$

Ou seja, é feito um *ranking* com todos os valores do atributo j e, então, é atribuído a x_{ij} a posição de x^*_{ij} nesse *ranking*. Essa equação leva a um atributo transformado com média $(n+1)/2$, variância $(n+1)[((2n+1)/6)-((n+1)/4)]$ para todos os atributos e ficam dentro do intervalo [1,$n-1$].

Impacto Negativo do Uso de Transformações nos Dados

É importante notar que nem sempre o uso de algum tipo de transformação nos dados é vantajosa. Em muitos casos, a discrepância presente entre os valores dos atributos de objetos no conjunto de dados pode indicar a presença de grupos realmente separados. Nesse caso, o uso de, por exemplo, uma padronização pode pode mudar a distância entre os pontos e alterar a separação natural dos grupos (JAIN; DUBES, 1988). Tal exemplo, pode ser visualizado na Figura 3.2.



(a) (b)

Figura 3.2: Impacto negativo do uso de uma transformação nos dados. (a) Dados originais antes da padronização. (b) Dados padronizados[2].

3.2 Técnicas de Agrupamentos

O objetivo das técnicas de agrupamento é dividir o conjunto de dados em grupos de tal modo que objetos similares fiquem em um mesmo grupo, enquanto que os dissimilares fiquem em grupos diferentes (JAIN; DUBES, 1988).

[2]Figura retirada de (JAIN; DUBES, 1988).

Embora na literatura de agrupamentos existam diferentes técnicas de agrupamento que propõem realizar essa tarefa, não existe uma única que seja adequada para todas as situações. Cada técnica impõe suas próprias tendências no agrupamentos que formam. E, apesar de muitas técnicas mostrarem soluções semelhantes para problemas triviais, é comum que se comportem de maneira bem diferente para situações mais complexas - caso dos problemas envolvendo dados de expressão gênica.

Motivado por essas razões, este trabalho dá ênfase a várias técnicas de diferentes classes. Assim, as próximas seções serão dedicadas a explicação dessas técnicas.

3.2.1 Técnicas de Agrupamento Hierárquicas

Um agrupamento hierárquico é uma seqüência de partições, na qual, cada partição é aninhada a partição vizinha na seqüência (JAIN; DUBES, 1988). As técnicas hierárquicas podem ser aglomerativas ou divisivas. Um técnica aglomerativa começa com as instâncias formando grupos unitários disjuntos (*singletons*), ou seja, cada uma das *n* instâncias no conjunto de dados vai ser atribuída a um grupo (*cluster*) diferente; a cada passo, os grupos mais próximos são unidos formando partições aninhadas. Esse processo se repete até que se forme uma única partição, chamada partição conjunta, contendo todas as instâncias da base de dados. Já em uma técnica hierárquica divisiva o processo se dá em ordem inversa à aglomerativa (JAIN; DUBES, 1988). Pelo fato de as técnicas hierárquicas divisivas serem mais computacionalmente custosas e retornarem resultados equivalentes aos métodos aglomerativos (JAIN; DUBES, 1988; HAIR et al., 2005), este trabalho é limitado ao uso de técnicas aglomerativas.

Uma das vantagens de se usar técnicas hierárquicas é que elas não assumem um número predefinidos de grupos, em vez disso, pode-se obter partições com número de grupos variados. Para isto, basta "cortar" a árvore hierárquica (dendrograma) no nível apropriado (JAIN; DUBES, 1988; BARBARA, 2000). A Figura 3.3 mostra um dendrograma formado a partir de um conjunto de dados com cinco objetos. A linha horizontal divide o dendrograma formando uma partição com dois grupos disjuntos.

[3]Figura retirada de (BARBARA, 2000).

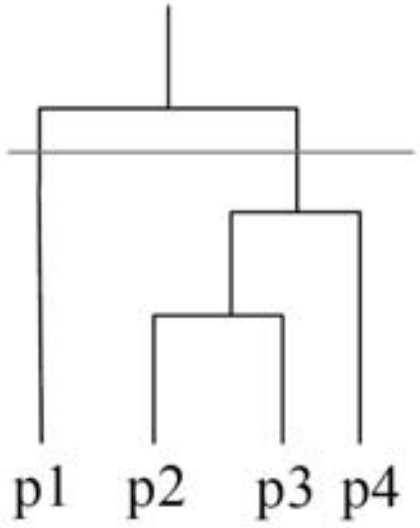


Figura 3.3: Dendrograma formado a partir de um conjunto de dados[3].

O pseudocódigo de um algoritmo hierárquico genérico pode ser descrito como segue (BARBARA, 2000):

1. Computar a matriz de similaridade.
2. Unir o par de grupos mais similares.
3. Atualizar a matriz de similaridade com o novo grupo.
4. Repetir os passos 2 e 3 até que reste somente um grupo.

Os algoritmos hierárquicos podem variar na forma de medir a similaridade entre grupos diferentes. Neste trabalho são investigadas três dessas variações bastante utilizadas (JAIN; DUBES, 1988):

- Algoritmo hierárquico com ligação simples.
- Algoritmo hierárquico com ligação média.
- Algoritmo hierárquico com ligação completa.

No algoritmo hierárquico com ligação simples, a similaridade entre dois grupos é dada pela maior similaridade entre quaisquer objetos (instâncias) dos dois grupos. Esse tipo de algoritmo

hierárquico é indicado para agrupar dados com formato não-esféricos, porém é um algoritmo sensível a ruídos (JAIN; DUBES, 1988).

Já no algoritmo hierárquico com ligação completa, a menor similaridade entre duas instâncias quaisquer de dois grupos determina a similaridade entre esses grupos. Tais algoritmos são menos suscetíveis a ruídos e *outliers*, mas, podem separar grupos naturalmente grandes e enfrenta problemas com grupos que tem formatos convexos (JAIN; DUBES, 1988).

Por fim, no caso do algoritmo hierárquico com ligação média a similaridade entre dois grupos é dada pela similaridade média entre todos os objetos dos dois grupos em questão. Essa é uma abordagem intermediária entre a ligação simples e a completa. As técnicas hierárquicas que utilizam ligação média tendem a gerar grupos com pequena variação interna. Elas também tendem a produzir grupos com mesma variância (HAIR et al., 2005).

Os algoritmos hierárquicos são visivelmente simples, o que levou ao desenvolvimento de diversas implementações disponibilizadas livremente na internet. Além disso, os dendrogramas gerados pelas técnicas hierárquicas assemelham-se com as árvores filogenéticas, as quais a comunidade médica já possui experiência na análise (QUACKENBUSH, 2001). Por isso, dentre outras razões, tornaram-se extremamente populares na análise de dados de expressão gênica, sendo preferido, inclusive, a outras técnicas mais recentes especializadas para tais tipos de dados.

É importante salientar que essa classe de técnicas de agrupamento sofre de uma série de limitações. Algumas já citadas (formato e densidade dos grupos) e outras como, por exemplo, a falta de uma função objetivo global (BARBARA, 2000). Os algoritmos hierárquicos usam vários critérios para decidir localmente quais grupos devem ser unidos a cada passo - tipicamente, os mais similares; em que, similaridade é definida por uma medida específica utilizada no agrupamento.

Barbara (2000) sumariza as limitações dos algoritmos hierárquicos da seguinte forma:

1. Falta de função objetivo global.

2. Decisões de união de grupos são finais: uma vez que dois grupos são aninhados, as instâncias daqueles grupos não podem mais ser atribuídas a outros grupos.

3. Boas decisões locais em aninhar grupos podem não gerar bons resultados globais.

4. Apresentam, pelo menos, um dos seguintes problemas: tratar ruídos e *outliers*, grupos com formatos não-convexos, tendência em separar grupos grandes.

3.2.2 *k-means*

Enquanto que as técnicas hierárquicas organizam os dados em uma sequência aninhada de grupos (árvore hierárquica que pode ser cortada em vários níveis diferentes), as chamadas técnicas particionais (não-hierárquicas), nas quais o *k-means* está incluído, geram uma única partição na tentativa de recuperar a estrutura original dos dados (JAIN; DUBES, 1988). O *k-means* faz parte da classe de técnicas de agrupamento particionais baseada em centro, ou seja, os grupos formados por essas técnicas são representados por um ponto central do grupo (centróide).

Os passos que descrevem o *k-means* podem ser definidos como segue (BARBARA, 2000):

1. Selecione *k* instâncias para serem os centróides inicias dos grupos.

2. Atribua todos as instâncias ao centróide mais próximo.

3. Recalcule o centróide para cada grupo.

   Calcule a média de todas as instâncias do grupo.

4. Repita os passos 2 e 3 até que os centróides não mudem.

Um ponto-chave que determina o desempenho desse algoritmo é a escolha dos centróides iniciais. Escolhas aleatórias, apesar de serem o procedimento mais comum, em geral, levam a mínimos locais.

O *k-means* busca minimizar o erro quadrático dos pontos em relação aos centros de seus grupos. Embora isso seja um critério razoável e leva a um algoritmo simples, também implica em certas limitações e problemas. Por exemplo, o *k-means* apresenta problemas para agrupar dados com grupos de tamanhos diferentes (Figura 3.4) e grupos com formatos convexos (Figura 3.5). Com relação a essas figuras, as linhas de contorno representam os grupos reais e os diferentes símbolos são os grupos formados pelo *k-means*.

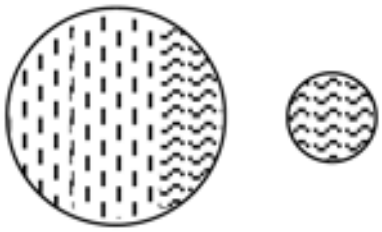


Figura 3.4: *k-means*: dados contendo classes com diferentes tamanhos[4].

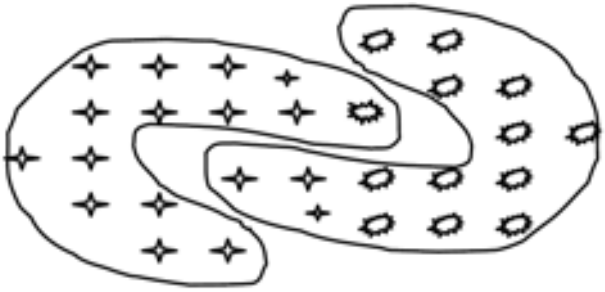


Figura 3.5: *k-means*: dados contendo classes com formato convexo[5].

Tais dificuldades são ocasionadas pela inadequação da sua função objetivo nesses casos. A função objetivo é otimizada (minimizada) para grupos com formato esférico e com mesmo tamanho ou para grupos bem separados.

3.2.3 Mistura Finita de Gaussianas

Segundo Barbara (2000), a idéia das técnicas baseadas em mistura de modelos é descrever os dados como uma mistura de distribuições de probabilidade, cada uma representando um grupo diferente. Em geral, a distribuição mais utilizada é a distribuição normal multivariada,

[4]Figura retirada de (BARBARA, 2000).
[5]Figura retirada de (BARBARA, 2000).

por ser já bem compreendida, relativamente fácil de trabalhar e produzir bons resultados em muitas situações. (JAIN; DUBES, 1988)

O problema de agrupamento em tal contexto é alocar cada instância do conjunto de dados a uma das distribuições, ou seja, atribuir cada elemento a um dos grupos. De maneira mais formal, seja $p(\mathbf{x}|\omega_i, \theta_i)$ a função densidade de probabilidade para um determinado grupo ω_i, em que $\mathbf{x}$ representa uma instância retirada do conjunto de dados e θ_i é o conjunto de parâmetros, ainda desconhecidos, para ω_i. Além disso, seja $P(\omega_i)$ seja a probabilidade *a priori* do grupo ω_i. Então, a distribuição da mistura pode ser definida como (JAIN; DUBES, 1988):

$$p(\mathbf{x}|\theta) = \sum_{i=1}^{k} p(\mathbf{x}|\omega_i, \theta_i)P(\omega_i) \tag{3.9}$$

em que, $\theta = (\theta_1, \theta_2, ..., \theta_k)$. Nesse contexto, as densidades condicionadas a um grupo $p(\mathbf{x}|\omega_i, \theta_i)$ são chamadas de densidade do componente e a probabilidade a priori $P(\omega_i)$ são conhecidas como parâmetros de mistura (JAIN; DUBES, 1988). Note que: $\sum_i^k P(\omega_i) = 1$ e $P(\omega_i) > 0$.

O objetivo do algoritmo é usar as instâncias retiradas do conjunto de dados para estimar o vetor de parâmetros θ. Em geral, os valores de θ são estimados iterativamente usando o algoritmo *Expectation Maximization* (EM). Uma vez que θ é conhecido pode-se decompor a mistura em componentes e atribuir cada uma das instâncias do conjunto de dados a componente com maior probabilidade.

Como já mencionado, é prática comum assumir que as distribuições das componentes sejam Gaussianas multivariadas (abordagem utilizada neste trabalho), com diferentes médias e matrizes de covariância, entretanto, outras distribuições podem ser utilizadas.

3.2.4 Dados com Alta Dimensionalidade

Agrupar conjuntos de dados cujos grupos tem diferentes formatos, tamanhos e densidades é uma tarefa desafiadora. Muitas das técnicas de agrupamento "tradicionais", como mencionado

anteriormente, possuem limitações que dificultam o cumprimento dessa tarefa. Por exemplo, o *k-means* apresenta dificuldade para agrupar dados com formatos arbitrários (ERTOZ; STEINBACH; KUMAR, 2002).

Quando os dados estão em alta dimensionalidade as questões ligadas as características dos dados (tamanho e densidade dos grupos, por exemplo) são intensificadas. Parte desse problema surge por causa de problemas com a noção de distância/similaridade em altas dimensões. Por exemplo, alguns algoritmos de agrupamento utilizam o conceito de distância euclidiana como critério de similidade, embora, a noção de distância euclidiana se torne sem sentido com o aumento da dimensionalidade. Mais especificamente, quando em alta dimensionalidade, pontos tendem a ter baixa similaridade, e então, pode haver casos em que pontos em grupos diferentes podem estar mais próximos que pontos no mesmo grupo. De fato, alguns estudos mostram que 15-20% dos *k* vizinhos mais próximos a um ponto pertencem a outra classe (ERTOZ; STEINBACH; KUMAR, 2002).

Uma outra maneira de tratar esse problema é utilizar medidas baseadas em densidade. Entretanto, a noção de densidade é, às vezes, mais problemática do que a de distância quando se está no contexto de alta dimensionalidade. Em particular, a noção de densidade utilizando a distância euclidiana, que é baseada na quantidade de pontos por unidade de volume, se torna sem sentido, pois, a medida que a dimensão aumenta o volume cresce rapidamente e, a não ser que o número de pontos cresça exponencialmente com o número de dimensões, a densidade vai tender a zero. Alguns algoritmos na literatura trabalham com esse tipo de medida e só são indicados para dados de baixa a média dimensionalidade.

Algumas técnicas de agrupamento mais recentes foram desenvolvidas com o objetivo de implementar medidas de similaridade que tratam as questões referentes a alta dimensionalidade. Nas próximas seções serão mostradas técnicas desse tipo que são utilizadas neste trabalho.

SNN

Shared Nearest Neighbor (SNN) é uma técnica de agrupamento que utiliza uma abordagem baseada em densidade para encontrar pontos representativos no conjunto de dados e usá-los para formar os agrupamentos em torno deles (ERTOZ; STEINBACH; KUMAR, 2002).

Como já foi mencionado, as noções de densidade "tradicionais", como as baseadas em distância euclidiana, por exemplo, não são indicadas para dados em alta dimensão. A noção de densidade que o SNN utiliza é conhecida como densidade probabilística e não possui os problemas que as medidas tradicionais de densidade possuem. Na densidade probabilística, se um ponto tem um grande número de vizinhos próximos a ele, então, esse ponto deve estar em uma região que possui uma densidade probabilística relativamente alta. Assim, pontos que possuem um grande número de vizinhos próximos estão em uma região mais "densa" que os pontos cujos os vizinhos estão distantes.

Na prática, a densidade de um ponto é dada pela soma das similaridades dos vizinhos mais próximos a ele. Quanto maior a densidade, maior é a probabilidade de um ponto ser um ponto representativo; e quanto menor a densidade, maior é a probabilidade daquele ponto ser um ruído ou um *outlier*.

O conceito de similaridade que o SNN utiliza é dado em termos do compartilhamento de vizinhos mais próximos introduzido por Jarvis e Patrick (1973). Segundo Jarvis e Patrick (1973), se um ponto **p** é próximo a outro ponto **q** e se ambos são próximos a um conjunto de pontos C, então, pode-se afirmar que **p** e **q** são próximos com maior confiança já que a similaridade entre eles é "confirmada" pelos pontos em C.

Dados esses conceitos, para construir um agrupamento o SNN cria, primeiro, uma matriz de similaridade, depois, a partir dessa matriz cria uma lista, contendo os k vizinhos mais próximos, e um grafo. No grafo, uma aresta é criada entre dois pontos **p** e **q** se, e somente se, **p** e **q** possuem um ao outro na suas listas de vizinhos mais próximos. Esse processo é denominado *sparsification*. Os pesos das arestas podem ser atribuídos de diferentes maneiras, uma delas é simplesmente o número de vizinhos que os dois pontos compartilham. Outra maneira mais

robusta leva em consideração o grau de similaridade dos vizinhos mais próximos.

O algoritmo do SNN pode ser resumido pelos seguintes passos (ERTOZ; STEINBACH; KUMAR, 2002):

1. Construa a matriz de similaridade.
2. Inicie o processo de *sparsification* da matriz de similaridade.
3. Construa o grafo dos vizinhos mais próximos a partir da matriz gerada pelo processo de *sparsification*.
4. Para cada vértice (ponto) do grafo, calcule o peso total das arestas que saem do vértice.
5. Identificar os pontos representativos escolhendo os pontos que possuem altos pesos totais.
6. Identificar os pontos ruidosos escolhendo os pontos que têm baixos pesos totais.
7. Remover todas as arestas que possuem peso menor que um limiar definido previamente.
8. Utilizar os componentes de pontos conectados para formar grupos, de forma que cada ponto em um grupo ou é um ponto representativo ou está conectado a um ponto representativo.

Diferentemente da maioria das técnicas de agrupamento, o número de grupos que vai estar na partição gerada pelo algoritmo não é um parâmetro informado pelo usuário. Dependendo da natureza dos dados, o algoritmo é capaz de encontrar a estrutura real dos dados.

Embora o SNN não necessite do número de grupos como parâmetro de entrada (comum às técnicas de agrupamento), requer uma série de outros parâmetros para o seu funcionamento, tais como o número de vizinhos mais próximos (*NN*), parâmetros referentes a definição das conexões do grafo dos vizinhos mais próximos compartilhados, como o *strong* que diz se uma conexão é considerada forte (alto peso total) e *merge* usado para juntar dois grupos distintos; ainda existem parâmetros referentes a limiares utilizados para definição de pontos representativos e ruídos: o *topic* para definir se um ponto é representativo e *noise* para definição de pontos

considerados ruídos. Essa grande quantidade de parâmetros pode ser vista como um empecilho para o seu uso, já que diversas variáveis estão envolvidas no processo experimental.

Spectral Clustering

Spectral Clustering é uma classe genérica de algoritmos de agrupamento. Eles são caracterizados por adotar o espectro da matriz de similaridade para reduzir a dimensionalidade do conjunto de dados e, então, aplicar uma técnica básica de agrupamento (*k-means*, por exemplo) a esses dados transformados (NG; JORDAN; WEISS, 2001).

No contexto desta dissertação, a fim de se agrupar um conjunto de dados $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ no $\Re^l$ em k grupos, por meio do *spectral clustering*, os seguintes passos são adotados (NG; JORDAN; WEISS, 2001):

1. Construa a matriz de afinidade $A \in \Re^{n \times n}$ definida por $A_{ij} = exp(-d^2/2\sigma^2)$ se $i \neq j$, e $A_{ii} = 0$; d é a proximidade entre dois objetos $\mathbf{x}_i$ e $\mathbf{x}_j$

2. Defina D como sendo a matriz diagonal cujo elemento (i,i) é a soma da i-ésima linha de A, e construa a matriz $L = D^{-1/2}AD^{-1/2}$.

3. Encontre os k maiores autovetores de L $(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k)$ e forme a matriz $V = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k] \in \Re^{n \times k}$ colocando os autovetores em colunas.

4. Forme a matriz Y a partir de V normalizando as linhas de V usando $Y_{ij} = V_{ij}/(\sum_j V_{ij}^2)^{1/2}$, de modo que fiquem com norma unitária.

5. Trate cada ponto de Y como um ponto no $\Re^k$ e agrupe-os em k grupos utilizando um algoritmo de agrupamento (o *k-means*, por exemplo).

6. Finalmente, atribua o ponto original $\mathbf{x}_i$ ao grupo j se, e somente se, a linha i da matriz Y foi atribuída ao grupos j.

O parâmetro σ controla a velocidade com que os valores da matriz de similaridade S decrescem.

Uma vantagem desse tipo de técnica é que ela não faz nenhum tipo de suposição sobre o formato dos dados. Em contraste com outros algoritmos de agrupamento, como o *k-means* que tendem a geram grupos esféricos, o *Spectral clustering* pode resolver problemas, em que os grupos presentes tem formatos espirais, por exemplo. É importante ressaltar que essa classe de algoritmo é muito sensível ao tipo de medida utilizada para construir a matriz de similaridade (NG; JORDAN; WEISS, 2001) e a transformação aplicada a essa matriz.

3.3 Validação de Agrupamentos

Para que seja possível a avaliação das partições geradas pelas técnicas de agrupamento é necessário o uso de índices de validação. Tais índices visam medir o sucesso de uma técnica na recuperação da estrutura real dos dados presentes em um conjunto de dados. Nesse contexto, dois tipos de índices são mais frequentemente utilizados na literatura: os índices internos e os índices internos (JAIN; DUBES, 1988).

Os índices internos medem a qualidade de um agrupamento com base apenas em informações intrínsecas dos dados, como, por exemplo, a homogeneidade intra-grupo e a separação entre grupos. Entretanto, a maior parte dos índices internos presentes na literatura favorecem partições que contém grupos com certos formatos e com determinado número de instâncias e, tendendo a favorecer, dessa maneira, algumas técnicas.

A idéia dos índices externos é comparar uma partição resultante da aplicação de um algoritmo de agrupamento, U, com uma partição independente dos dados, construída com base no conhecimento *a priori* sobre a estrutura real dos dados.

De fato, por não apresentar tendências em favorecer qualquer técnica de agrupamento, o índice externo *corrected Rand* (cR) (JAIN; DUBES, 1988; KUNCHEVA, 2004) foi escolhido para ser utilizado neste trabalho.

Formalmente, seja $U = \{u_1, ..., u_R\}$ a partição gerada por um algoritmo de agrupamento, e $V = \{v_1, ..., v_C\}$ seja a partição formada com conhecimento *a priori* dos dados, independente da

partição U. O *corrected Rand* pode ser definido dessa maneira:

$$cR = \frac{\sum_i^R \sum_j^C \binom{n_{ij}}{2} - \binom{n}{2}^{-1} \left[\sum_i^R \binom{n_{i.}}{2} \sum_j^C \binom{n_{.j}}{2}\right]}{\frac{1}{2}\left[\sum_i^R \binom{n_{i.}}{2} + \sum_j^C \binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1} \left[\sum_i^R \binom{n_{i.}}{2} \sum_j^C \binom{n_{.j}}{2}\right]} \tag{3.10}$$

em que, n_{ij} representa o número de objetos comuns aos grupos u_i e v_j; $n_{i.}$ indica o número de objetos no grupo u_i; $n_{.j}$ indica o número de objetos no grupo v_j; n é o número total de objetos; e $\binom{a}{b}$ é o coeficiente binomial $\frac{a!}{b!(a-b)!}$.

O cR pode assumir valores entre -1 e 1, com 1 indicando uma concordância perfeita entre as partições e valores próximos a 0 ou negativos correspondendo a concordâncias encontradas ao acaso.

4 Material e Experimentos

4.1 Conjuntos de Dados

Trinta e cinco conjuntos de dados são usados neste trabalho. Cada conjunto de dados possui suas próprias características, diferenciando-se entre si em vários aspectos, tais como, tipo de *chip* de *microarray*, tipo de tecido, número de instâncias, número de classes, distribuição das instâncias dentro das classes, dimensionalidade e dimensionalidade após o processo de filtragem. As Tabelas 4.1 e 4.2 sumarizam as principais características dos conjuntos de dados *Affymetrix* e cDNA, respectivamente. Nesse contexto, n representa o número de instâncias presentes na base de dados, #C o número de classes, m a quantidade de atributos antes do processo de filtragem e d o número de atributos após o filtro (será feita uma descrição do processo de filtragem na Seção 4.2.3).

Como mencionado anteriormente, dentre as diversas tecnologias na literatura para medição de expressão gênica, duas delas são mais frequentemente utilizadas: *single-channel* e *double-channel* (QUACKENBUSH, 2001). Como representantes dessas categorias se destacam o *Affymetrix* e cDNA, respectivamente. Todos os conjuntos de dados utilizados aqui são de um desses dois tipos.

Apesar de as técnicas *Affymetrix* e cDNA serem utilizadas para medir a expressão de genes, elas usam metodologias diferentes para fazer essa medição. Isso acaba tornando os conjunto gerados por elas com características diferentes. Por exemplo, os *chips Affymetrix* medem a expressão de um gene através de uma estimativa da quantidade de cópias de RNA encontradas na amostra celular, enquanto que os valores obtidos a partir de um *microarray* cDNA são relações

	Conj. de Dados	***Chip***	**Tecido**	***n***	**#C**	**Distribuição das classes**	***m***	***d***
1	`Armstrong-v1` (ARMSTRONG et al., 2002)	Affy	Sangue	72	2	24,48	12582	1081
2	`Armstrong-v2` (ARMSTRONG et al., 2002)	Affy	Sangue	72	3	24,20,28	12582	2194
3	`Bhattacharjee` (BHATTACHARJEE et al., 2001)	Affy	Pulmão	203	5	139, 17, 6, 21, 20	12600	1543
4	`Chowdary` (CHOWDARY et al., 2006)	Affy	Mama, Colon	104	2	62,42	22283	182
5	`Dyrskjot`(DYRSKJOT et al., 2003)	Affy	Bexiga	40	3	9,20,11	7129	1203
6	`Golub-v1` (GOLUB et al., 1999)	Affy	Medula	72	2	47,25	7129	1877
7	`Golub-v2` (GOLUB et al., 1999)	Affy	Medula	72	3	38,9,25	7129	1877
8	`Gordon` (GORDON et al., 2002)	Affy	Pulmão	181	2	31,150	12533	1626
9	`Laiho` (LAIHO et al., 2007)	Affy	Colon	37	2	8,29	22883	2202
10	`Nutt-v1` (NUTT et al., 2003)	Affy	Cérebro	50	4	14,7,14,15	12625	1377
11	`Nutt-v2` (NUTT et al., 2003)	Affy	Cérebro	28	2	14,14	12625	1070
12	`Nutt-v3` (NUTT et al., 2003)	Affy	Cérebro	50	2	7,15	12625	1152
13	`Pomeroy-v1` (POMEROY et al., 2002)	Affy	Cérebro	34	2	25,9	7129	857
14	`Pomeroy-v2` (POMEROY et al., 2002)	Affy	Cérebro	42	5	10,10,10,4,8	7129	1379
15	`Ramaswamy` (RAMASWAMY et al., 2001)	Affy	Multi-tecido	190	14	11,10,11,11,22,10,11,10,30,11,11,11,11,20	16063	1363
16	`Shipp` (SHIPP et al., 2002)	Affy	Sangue	77	2	58,19	7129	798
17	`Singh` (SINGH et al., 2002)	Affy	Próstata	102	2	58,19	12600	339
18	`Su` (SU et al., 2001)	Affy	Multitecido	174	10	26,8,26,23,12,11,7,27,6,28	12533	1571
19	`West` (WEST et al., 2001)	Affy	Mama	49	2	25,24	7129	1198
20	`Yeoh-v1` (YEOH et al., 2002)	Affy	Medula	248	2	43,205	12625	2526
21	`Yeoh-v1` (YEOH et al., 2002)	Affy	Medula	248	6	15,27,64,20,79,43	12625	2526

Tabela 4.1: Descrição dos conjuntos de dados *Affymetrix*.

	Conj. de Dados	***Chip***	**Tecido**	***n***	**#C**	**Distribuição das classes**	***m***	***d***
1	`Alizadeh-v1` (ALIZADEH et al., 2000)	cDNA	Sangue	42	2	21,21	4022	1095
2	`Alizadeh-v2` (ALIZADEH et al., 2000)	cDNA	Sangue	62	3	42,9,11	4022	2093
3	`Alizadeh-v3` (ALIZADEH et al., 2000)	cDNA	Sangue	62	4	21,21,9,11	4022	2093
4	`Bittner` (BITTNER et al., 2000)	cDNA	Pele	38	2	19,19	8067	2201
5	`Bredel` (BREDEL et al., 2005)	cDNA	Cérebro	50	3	31,14,5	41472	1739
6	`Chen` (CHEN et al., 2002)	cDNA	Fígado	179	2	104,75	22699	85
7	`Garber` (GARBER et al., 2001)	cDNA	Pulmão	66	4	17,40,4,5	24192	4553
8	`Khan` (KHAN et al., 2001)	cDNA	Multitecido	83	4	29,11,18,25	6567	1068
9	`Lapointe-v1` (LAPOINTE et al., 2004)	cDNA	Próstata	69	3	11,39,19	42640	1625
10	`Lapointe-v2` (LAPOINTE et al., 2004)	cDNA	Próstata	110	4	11,39,19,41	42640	2496
11	`Liang` (LIANG et al., 2005)	cDNA	Cérebro	37	3	28,6,3	24192	1411
12	`Risinger` (RISINGER et al., 2003)	cDNA	Endométrio	42	4	13,3,19,7	8872	1771
13	`Tomlins-v1` (TOMLINS et al., 2007)	cDNA	Próstata	104	5	27,20,32,13,12	20000	2315
14	`Tomlins-v2` (TOMLINS et al., 2007)	cDNA	Próstata	92	4	27,20,32,13	20000	1288

Tabela 4.2: Descrição dos conjuntos de dados cDNA.

entre o número de cópias de RNA na célula alvo e na amostra de uma célula utilizada para controle. Desse modo, os valores obtidos com *Affymetrix* estão em escalas diferentes daqueles obtidos com cDNA.

De acordo com alguns trabalhos, nos conjuntos de dados *Affymetrix*, quando a medição das intensidades dos níveis de expressão gênica está sendo feita, pode haver algum tipo de saturação no processamento da imagem onde os valores que representam esses níveis são muito baixos ou muito altos (MONTI et al., 2003; STEGMAIER et al., 2004). Para não correr riscos desse tipo de informação (*outlier*) estar presente nos dados utilizados neste trabalho, limiares são estabelecidos para valores máximos e mínimos de expressão gênica. Assim, seguindo os procedimentos realizados em Monti et al. (2003) e Stegmaier et al. (2004), valores abaixo de 10 são ajustados para o limiar mínimo de 10 e valores acima de 16.000 assumem o valor do limite superior 16.000.

4.2 Descrição dos Conjuntos de Dados

Nessa seção, são descritos brevemente os conjuntos de dados utilizados neste trabalho. Para cada base são esclarecidas, quando aplicáveis, as modificações com relação a base de dados original e de que maneira elas foram arranjadas para formar o total de 35 conjuntos de dados que compõem a análise realizada neste trabalho. Em geral, as bases construídas contém somente amostras relativas a expressão gênica de câncer, porém, algumas bases de dados também possuem amostras de tecido normal, pois a remoção desse tipo de amostra implicaria em uma base de dados com um número muito pequeno de amostras ou com somente uma classe. A fim de fornecer uma melhor organização ao trabalho, as descrições foram divididas para as bases *Affymetrix* e cDNA, sendo mostradas em ordem alfabética do nome do primeiro autor de cada trabalho. As siglas utilizadas nessa seção estão definidas na Lista de Siglas contida na parte inicial desta dissertação.

4.2.1 *Affymetrix*

`Armstrong-v1` e `Armstrong-v2` (ARMSTRONG et al., 2002) - Foram feitos dois conjuntos de dados utilizando amostras de diferentes tipos de leucemia. Cada amostra é descrita por 12.582 genes. O primeiro conjunto, `Armstrong-v1`, contém 72 amostras, divididas em duas classes: ALL (24 amostras) e MLL (48 amostras); o outro conjunto, `Armstrong-v2`, formado a partir da divisão da classe MLL de `Armstrong-v1`, possui três classes distintas: ALL (24 amostras), MLL (20 amostras) e AML (28 amostras).

`Bhattacharjee` (BHATTACHARJEE et al., 2001) - Com um total de 203 amostras, entre tumores de pulmão e células normais, esse conjunto de dados é formado por 139 amostras da classe AC, 21 do tipo SCC, 20 COID, 6 SCLC e 17 de tecido normal. Cada amostra na base de dados contém 12.600 genes.

`Chowdary` (CHOWDARY et al., 2006) - Esse conjunto de dados é formado 42 amostras de tumor de cólon e 62 de tumores de mama, totalizando 104 amostras de câncer. As amostras são descritas por 22.283 genes.

`Dyrskjot` (DYRSKJOT et al., 2003) - A base de dados contém 40 amostras relacionadas a estágios diferentes de câncer de bexiga. As amostras são distribuídas em: 9 amostras da classe T2+, 20 amostras de TA e 11 amostras T1. Cada amostra contém 7.129 genes.

`Golub-v1` e `Golub-v2` (GOLUB et al., 1999) - Duas bases de dados distintas foram formadas a partir do mesmo conjunto de amostras. Mais especificamente, a base `Golub-v1` foi construída considerando apenas dois tipos de leucemia, sendo formada por 47 amostras da classe ALL e 25 da classe AML; a segunda base de dados (`Golub-v2`) é feita a partir de uma subdivisão da classe ALL (B-ALL e T-ALL) presente na primeira base, sendo distribuída, então, em: 38 amostras da classe B-ALL, 9 da classe T-ALL e 25 amostras de AML. Todas as amostras contém 7.129 genes.

`Gordon` (GORDON et al., 2002) - 181 amostras de tumores de pulmão, distribuídas em duas classes, compõem a base de dados. Cada amostra contém 12.600 genes e encontram-se

dispostas da seguinte forma dentro do conjunto de dados: 31 amostras são da classe MPM e 150 são da classe AD.

Laiho (LAIHO et al., 2007) - Amostras de dois tipos de carcinoma colorretais formam essa base de dados. De um total de 37 amostras, cada uma contendo 22.883 genes, 29 são do tipo convencional e 8 são do tipo *serrated*.

Nutt-v1, Nutt-v2 e Nutt-v3 (NUTT et al., 2003) - Para construção desses três conjuntos de dados foram utilizadas 50 amostras de células de glioma de alto grau. A primeira, Nutt-v1, considera todas as 50 amostras, sendo divididas em quatro classes: 14 amostras do tipo CG, 14 do tipo NG, 7 amostras da classe CO e 15 da classe NO; a segunda, Nutt-v2, contém somente as amostras de glioblastoma: 14 CG e 14 NG; e, por último, a Nutt-v3 é formada pelas amostras de oligodendroglioma: 7 CO e 15 NO. Cada amostra é formada por 12.625 genes.

Pomeroy-v1 e Pomeroy-v2 (POMEROY et al., 2002) - As amostras utilizadas para formação dessas duas bases de dados, apesar de coletadas do mesmo trabalho, foram utilizadas originalmente para análises diferentes e, portanto, não há sobreposição de amostras entre os dois conjuntos de dados. Assim, a Pomeroy-v1 possui 42 amostras entre diferentes tipos de tumores de cérebro e tecido de células normais, sendo distribuídas em: 10 MD, 10 MGLIO, 10 AT/RT, 8 PNET e 4 tecidos normais de cerebelo. A Pomeroy-v2 contém apenas amostras de meduloblastomas: 9 amostras de AM e 25 do tipo CM, totalizando 34 amostras. Ambas as bases contém 7.129 genes.

Ramaswamy (RAMASWAMY et al., 2001) - Esse conjunto de dados une as 144 amostras de câncer do conjunto de treinamento usado no trabalho original e as 46 amostras de tumor primário de um dos conjuntos de teste, não sendo consideradas, com o intuito de formar uma base de dados contendo apenas amostras de tumores primários, as 20 amostras fracamente diferenciadas e as 8 amostras de tumores mestastáticos. Assim, com 16.063 genes, as amostras de câncer, de 14 classes diferentes, são divididas da seguinte maneira: 11 amostras de câncer de mama, 10 de câncer de próstata, 11 de câncer de pulmão, 11 de adenocarcinoma colorretal, 22 linfomas, 10 melanomas, 11 carcinomas de bexiga, 10 amostras de câncer no útero, 30 de

leucemia, 11 de tumores renais, 11 amostras de câncer no pâncreas, 11 adenocarcinomas de ovário, 11 mesoteliomas pleurais e 20 amostras de câncer no sistema nervoso central.

`Shipp` (SHIPP et al., 2002) - O conjunto de dados contém um total de 77 amostras de linfoma, sendo 58 amostras da classe DLBCL e 19 amostras da classe FL . Cada amostra é composta por 7.129 genes.

`Singh` (SINGH et al., 2002) - Essa base de dados é composta por 102 amostras, cada uma com 12.600 genes. 52 amostras são de tecido com câncer de próstata e 50 são de tecido normal.

`Su` (SU et al., 2001) - Composta pela junção dos conjuntos de treinamento e teste originais, essa base de dados contém amostras de dez tipos de câncer diferentes, cada uma contendo 12.533 genes: 26 de próstata, 8 de bexiga, 26 de mama, 23 colorretais, 12 de gastroesôfago, 11 de rins, 7 de fígado, 27 de ovário, 6 de pâncreas e 28 de pulmão.

`West` (WEST et al., 2001) - As amostras de tumores de mama são divididas em positivas para ambos os receptores de estrogênio e progesterona ou negativas para ambos os receptores. Assim, com cada amostra contendo 7.129 genes, o conjunto de dados possui 25 amostras da classe ER+ e 24 amostras da classe ER-.

`Yeoh-v1` e `Yeoh-v2` (YEOH et al., 2002) - Neste trabalho foram usadas somente as amostras de subgrupos conhecidos de B-ALL e as amostras de T-ALL para formar dois conjuntos de dados distintos. O primeiro conjunto (`Yeoh-v1`) visa a distinção dos tipos de leucemia B-ALL e T-ALL e tem as amostras distribuídas da seguinte maneira: 43 do tipo T-ALL e 205 do tipo B-ALL. O segundo conjunto (`Yeoh-v2`) desmembra a classe B-ALL, sendo, então, distribuído em seis subclasses: 15 BCR-ABL, 27 E2A-PBX1, 64 *hyperdiploid* >50, 20 MLL, 79 TEL-AML1 e 43 T-ALL. Com relação a base original, foram dispensadas as amostras do subgrupo identificadas nos experimentos e rotuladas como *OTHER*. Para as duas bases, todas as amostras contém 12.625 genes.

4.2.2 cDNA

`Alizadeh-v1`, `Alizadeh-v2` e `Alizadeh-v3` (ALIZADEH et al., 2000) - Três versões dessa base foram construídas a partir de amostras dos principais tipos de linfoma, todas contendo 4.022 genes. A primeira, `Alizadeh-v1`, possui apenas amostras referentes ao tipo DLBCL formando duas subclasses distintas: DLBCL1 (21 amostras) e DLBCL2 (21 amostras); o segundo conjunto, `Alizadeh-v2`, foi feito adicionando as amostras de outros dois tipos de linfoma, totalizando 62 amostras: 21 amostras de DLBCL1, 21 amostras de DLBCL2, 9 amostras de FL e 11 amostras de CLL. Por último, `Alizadeh-v3` é constituída por 42 amostras do tipo DLBCL (DLBCL1 + DLBCL2), 9 amostras do tipo FL e 11 amostras da classe CLL, em um total de 62 amostras.

`Bittner` (BITTNER et al., 2000) - São utilizadas 31 amostras de pacientes com melanoma. A base formada contém 19 amostras da classe de melanoma denominada por ML1 e 12 amostras da classe ML2 com 8.067 genes. Sete amostras de controle que o conjunto de dados original continha não foram consideradas no estudo feito por esta dissertação.

`Bredel` (BREDEL et al., 2005) - Esse conjunto de dados é constituído por 50 amostras de glioma, cada uma com 41.472 genes, distribuídas da seguinte maneira: 31 amostras do tipo GBM, 14 do tipo OG, e 5 amostras da classe AT. Foram retiradas as amostras de tecido normal incluídas no conjunto de dados para a análise feita em Bredel et al. (2005).

`Chen` (CHEN et al., 2002) - Os autores fizeram vários tipos de análises utilizando conjuntos diferentes de amostras de tumores de fígado, tecidos normais e outros tumores. A base de dados utilizada neste trabalho foi construída com base em uma versão do conjunto de dados usado na análise de amostras de HCC e amostras de tecido normal de fígado. Assim, a base `Chen` contém 179 amostras no total, sendo 104 de HCC e 75 de tecidos de fígado normal, cada uma com 22.699 genes.

`Garber` (GARBER et al., 2001) - Essa base de dados foi construída a partir de 66 amostras de câncer de pulmão: 40 AC, 17 SCC, 4 LCLC e 5 SCLC, todas contendo 24.192 genes. O conjunto de dados `Garber` usado nos experimentos dessa dissertação se diferencia do descrito

no trabalho original, que continha 73 amostras, em três aspectos: por retirar as amostras de tecido normal (5 amostras) e de feto (1 amostra); considerar uma das amostras da classe SCLC como sendo da classe SCC; e, por não conter uma amostra da classe AC, que não estava presente no conjunto de dados disponibilizado pelos autores.

Khan (KHAN et al., 2001) - Uma base de dados com quatro classes foi formada pela união das amostras de mesmo tipo presentes em conjuntos de treinamento e teste usados na análise original, totalizando 83 amostras de cânceres diferentes (29 EWS, 25 RMS, 18 NB e 11 BL). As cinco amostras de controle presentes no conjunto de dados original foram retiradas para formação desta base de dados. Todas as amostras presentes no conjunto de dados contém 6.567 genes.

Lapointe-v1 e Lapointe-v2 (LAPOINTE et al., 2004) - As versões dessa base de dados são formadas, primeiro, somente por amostras de subtipos de câncer de próstata, 11 do tipo I, 39 do tipo II e 19 do tipo III, denominando esse conjunto por Lapointe-v1; segundo, pela adição de 41 amostras de tecido de próstata normal a esse conjunto, que agora é chamado de Lapointe-v2, totalizando 110 amostras. As amostras dos dois conjuntos contém 42.640 genes.

Liang (LIANG et al., 2005) - O conjunto de dados usado nesta dissertação consistiu em 37 amostras, entre células de tumores e células normais de cérebro. Dessas 37 amostras, 28 são da classes GBM, 6 são rotuladas como OG e 3 são de células normais. O número total de genes em cada amostra é 24.192. O estudo original continha 38 amostras no total, porém, a versão disponibilizada pelos autores não continha uma amostra da classe GBM.

Risinger (RISINGER et al., 2003) - Um total de 42 amostras, tanto de câncer endometrial quanto de tecido normal de endométrio, são usadas nessa base. As amostras são distribuídas da seguinte maneira: 13 PS, 19 E, 3 CC e 7 normais (N), cada uma com 8.872 genes.

Tomlins-v1 e Tomlins-v2 (TOMLINS et al., 2007) - 104 amostras representando cada uma das fases de evolução do câncer de próstata dão origem a duas bases diferentes. A primeira, Tomlins-v1, contém 27 amostras de EPI, 13 do tipo PIN, 32 de PCA, 20 MET e 12 STROMA; a outra, Tomlins-v2, é formada pela remoção das amostras das amostras rotuladas

como STROMA (92 amostras) da primeira base.

4.2.3 Filtros

Com o objetivo de remover atributos (genes) não informativos, ou seja, remover atributos irrelevantes ao processo de agrupamento, filtros são aplicados aos conjuntos de dados. Procedimentos semelhantes são usados para os dados de chips *Affymetrix* e cDNA. Para *Affymetrix*, o seguinte procedimento é aplicado: para cada atributo (gene) j, é computada a média m_j ao longo de todas as instâncias - no cálculo da média não são considerados os 10% maiores e menores valores. Baseado nessa média, cada valor x^*_{ij}, ou seja, o valor do atributo j da instância i, sofre uma transformação:

$$y_{ij} = \log_2(x^*_{ij}/m_j) \tag{4.1}$$

São selecionados para análise posterior os atributos (genes) cujos níveis de expressão diferenciam-se em, pelo menos, l-vezes do nível de expressão médio daquele gene ao longo de, pelo menos, c instâncias.

Neste trabalho, com somente algumas exceções, os valores de l e c foram escolhidos, empiricamente, de modo a produzir conjuntos de dados com cerca de 10% do número original de genes presentes nos conjuntos de dados originais. Vale a pena ressaltar que os dados transformados, usando a equação anterior, só são utilizados durante o processo de filtragem, sendo usado nos experimentos com as técnicas de agrupamento os dados sem transformação.

Para os dados cDNA, um procedimento semelhante foi realizado. Como os dados de expressão gênica contidos nas bases cDNA são valores relativos de expressão entre duas amostras, já estão em forma de logaritmo, dispensando o passo da transformação dos dados aplicado nas bases *Affymetrix*. O restante do processo de seleção de genes com uma maior variabilidade ao longo das amostras é realizado da mesma forma que as bases *Affymetrix*.

Uma característica dos conjuntos dados obtidos a partir de *microarray* de cDNA é a possível ausência de valores de expressão gênica. Para garantir a consistência dos dados foram descartados os genes que continham mais de 10% dos seus valores faltando. Para os genes que

permaneceram no conjunto de dados, mas ainda assim continham valores faltando (aqueles que tinham menos de 10% de valores faltosos), tiveram tais valores substituídos pelo valor médio da expressão daquele gene.

4.3 Projeto e Avaliação dos Experimentos

Nos experimentos realizados foram utilizadas sete técnicas de agrupamento: Hierárquico com ligação simples (SL), hierárquico com ligação média (AL), hierárquico com ligação completa (CL), *k-means* (KM), Mistura Finita de Gaussianas (MFG), *Spectral clustering* (SPC) e *Shared Nearest Neighbor* (SNN); e, quando aplicável, usando quatro diferentes medidas de proximidade: correlação de Pearson (P), Cosseno(C), coeficiente de correlação de Spearman (SP) e distância Euclidiana (E) - em algumas das implementações não está disponível o uso de todas as medidas de proximidade utilizadas; e na técnica MFG não se utiliza o conceito proximidade no processo de agrupamento. Essas configurações foram aplicadas aos 35 conjuntos de dados de dados já citados. Além disso, para a distância Euclidiana os dados foram usados em quatro diferentes versões: dados originais (Z_0), padronizados (Z_1), normalizados (Z_2) e ordenados (*ranked*) (Z_3).

Para cada técnica de agrupamento, foram desenvolvidos experimentos variando o número de grupos no intervalo $[k, \sqrt{\lceil n \rceil}]$, em que k representa o número real de classes para um determinado conjunto de dados com n instâncias. A fim de construir partições a partir das árvores geradas pelos algoritmos hierárquicos, foram consideradas as i ($i = k, ..., \sqrt{\lceil n \rceil}$) primeiras subárvores. A recuperação da estrutura dos grupos é medida usando o índice *corrected Rand* (cR, Seção 3.3), através da comparação das classes conhecidas *a priori* com a estrutura de grupos gerada pelas técnicas de agrupamento.

Em cada combinação de técnica de agrupamento, medida de proximidade e procedimento de transformação de dados, foi calculada a média do cR para todos os conjuntos de dados em dois diferentes contextos: levando-se em consideração, para cada conjunto de dados, somente a partição com o número de grupos igual ao número real de classes do respectivo conjunto de

dados (**contexto A**); e, considerando-se, para cada conjunto de dados, a partição que apresenta o melhor cR, independentemente do número de grupos (**contexto B**).

Outro aspecto investigado é a influência do uso de cobertura reduzida pelas técnicas de agrupamento. A idéia é identificar as técnicas que, geralmente, produzem suas melhores partições, em termos de cR, com um número de grupos maior que o número real de classes da base de dados. Assim, para uma dada técnica aplicada a um conjunto de dados, é medida a diferença entre o número de grupos presentes na partição com maior cR e o número real de classes.

Para verificar se as diferenças entre os desempenhos dos algoritmos em termos de cR são estatisticamente significantes foi realizado um teste estatístico, o teste *t*. Dado que o principal interesse é comparar os métodos de agrupamento, foram selecionados para cada método somente os resultados com a medida de proximidade que obteve a maior média de cR. É importante observar que os valores da média do cR para uma técnica de agrupamento e medida de proximidade apresenta grande variância. Isso acontece, principalmente, por causa das diferentes características de cada base de dados, o que torna algumas bases mais difíceis de agrupar que outras.

4.4 Resultados e Discussão

A discussão dos resultados é feita sob duas perspectivas: primeiro, as técnicas de agrupamento, a partir das partições geradas, são avaliadas em relação ao seu desempenho na recuperação da estrutura natural dos dados. Isso é feito utilizando o índice *corrected Rand* (cR), descrito na Seção 3.3, de modo que quanto maior o valor do cR para uma partição, mais essa partição está de acordo com as classes reais dos dados (estrutura natural). A segunda perspectiva explora o uso de cobertura reduzida pelas técnicas de agrupamento. Ou seja, é analisado o impacto de uma técnica agrupar os dados com um número de grupos maior que o número real de classes contidas no conjunto de dados.

Dentro da primeira perspectiva, o desempenho das técnicas, como já foi ressaltado, é medido em termos de cR. Para que seja possível a análise desse desempenho com relação a todos

os conjuntos de dados ao mesmo tempo, é feita uma média dos valores do cR das partições geradas por cada técnica de agrupamento para todos os conjuntos de dados sob os pontos de vista do **contexto A** e **contexto B**.

A análise de resultados é feita de forma idêntica para todos os conjuntos de dados, mas para um melhor entendimento, os resultados são mostrados separadamente para os conjuntos de dados dos tipos *Affymetrix* e cDNA. Tais resultados são apresentados em tabelas e, com o objetivo de melhorar a visualização dos resultados, para cada tabela existe um gráfico, em que as médias e desvios-padrões são representados por barras. Assim, as Tabelas 4.3, 4.4, 4.5 em conjunto com as Figuras 4.1, 4.2 e 4.3 são referentes aos conjuntos de dados *Affymetrix*, e as Tabelas 4.6, 4.7, 4.8 e as Figuras 4.4, 4.5, 4.6 são referentes aos conjuntos de dados cDNA. Nessas tabelas, nos casos onde há "n/a" significa que a medida de proximidade não é aplicável ou não está disponível para a técnica de agrupamento. Além disso, o maior valor de cR para cada algoritmo está destacado em negrito.

4.4.1 Recuperação de Tipos de Câncer

Baseado nas informações contidas nas Tabelas 4.3 (*Affymetrix*), e indicado pelo teste de hipóteses, é possível afirmar que a técnica MFG obteve um melhor desempenho, em termos de cR, do que o SL, AL, CL e SNN. Isso é válido tanto para o **contexto A** quanto para o **contexto B**.

Para o caso das bases cDNA, nos dois contextos investigados, o teste estatístico também indicou que o MFG e KM, independentemente da medida de proximidade usada por esse último, obtiveram um cR maior que SL, AL, CL e SNN.

Os resultados também mostraram que KM e MFG tiveram, em média, a menor diferença entre o número de grupos da partição com maior cR e o número real de classes. É importante observar que o SNN mostrou um comportamento consistente, em termos de cR, ao longo das várias medidas de proximidade, embora, com valores de cR menores que os apresentados por KM e MFG. Na verdade, o SNN apresentou um desempenho similar ao SPC.

Para evitar uma interpretação errada dos resultados, a diferença entre o número de grupos presente nas melhores partições obtidas pelos algoritmos e o número real de classes (Tabelas 4.5 e 4.8) deve ser analisada em conjunto com os valores médios de cR dessas partições (Tabelas 4.4 e 4.7). Por exemplo, de acordo com a Tabela 4.5 o SPC utilizando E_{Z_3} e KM usando C apresentaram, respectivamente, uma média da diferença entre o número real de classes e o número de grupos na melhor partição de 0,52 e 0,95 grupos. Entretanto, esse último obteve um cR médio de 0,51 enquanto que o primeiro um cR médio de 0,09.

Os resultados também mostraram que as técnicas hierárquicas, em geral, apresentaram piores desempenhos que os outros métodos avaliados. Esse resultado já havia sido observado no contexto de agrupamento de genes (D'HAESELEER, 2005). Mais especificamente, dentro dessa classe de técnicas o SL obteve os piores resultados. O teste estatístico indicou que ele teve menor cR médio que cada um de todos os outros métodos. No que diz respeito ao uso de medidas de proximidade pelos algoritmos hierárquicos, a correlação de Pearson (P) e o cosseno (C) levaram a melhores resultados. Isso também está de acordo com trabalhos realizados pela comunidade clínica/biológica (EISEN et al., 1998).

Em geral, para obter maiores valores de cR, em comparação ao KM e MFG, as técnicas hierárquicas necessitam de um maior uso de cobertura reduzida, ou seja, um número maior de grupos que o presente na estrutura real dos dados. Por exemplo, de acordo com a Tabela 4.3 o AL usando P, para o número de grupos igual ao número real de classes, apresentou um cR médio de 0,24. Em contraste, se forem consideradas partições com um número maior de grupos, o cR médio aumenta para 0,39. Tal aumento foi atingido utilizando, em média, 1,52 grupos a mais que o presente na estrutura real dos dados (Tabela 4.4).

Um resultado não esperado foi o bom desempenho atingido com a utilização de Z_3 (dados ordenados), principalmente no caso das técnicas hierárquicas. Nesse contexto, o uso de tal transformação em conjunto com a distância euclidiana levou a resultados próximos aos obtidos usando P, C e SP, especialmente no caso das bases *Affymetrix*. Uma possível razão para esse comportamento é a presença de observações atípicas nos dados (*outliers*) e, como já menci-

onado anteriormente, a transformação Z_3 diminui o impacto dessas observações no processo de agrupamento (MILLIGAN; COOPER, 1988). Isso é uma outra evidência que esse tipo de técnica é mais sensível a *outliers* que as outras técnicas investigadas.

A técnica SPC se mostrou bastante sensível ao tipo de medida de proximidade usada. Em geral, as partições geradas por tal método atingem altos valores de cR (> 0.40) para P, C e SP, porém baixos valores (< 0.15) quando a distância euclidiana (E) é usada em qualquer caso. De fato, como mencionado na Seção 3.2.4, existem evidências na literatura que apontam para a sensibilidade do SPC relacionada a seleção da matriz de similaridade usada (LUXBURG, 2007). E, no entanto, até o momento, não existe qualquer tipo de regra para auxiliar tal seleção.

Em outro tipo de análise é investigado o impacto do uso de cobertura reduzida no desempenho dos algoritmos. Como mencionado anteriormente, esse impacto foi mais significante para o caso das técnicas hierárquicas de agrupamento. Embora, como existem muitas bases de dados extremamente desbalanceadas (número de amostras diferentes para cada grupo), esse comportamento também atingiu o *k-means*, mas com menos intensidade. De fato, todos os métodos, mesmo aqueles que supostamente lidam bem com grupos desbalanceados, beneficiaram-se do uso de cobertura reduzida.

A seguir, para melhor ilustrar algumas questões importantes discutidas nessa seção, serão apresentados os resultados obtidos com o método hierárquico e *k-means* para as bases de dados `Alizadeh-v2` e `Nutt-v3`.

Alg.	P	C	SP	E_{Z_0}	E_{Z_1}	E_{Z_2}	E_{Z_3}
SL	0.00 ± 0.03	0.05 ± 0.20	$\mathbf{0.07 \pm 0.21}$	0.02 ± 0.04	0.01 ± 0.03	0.01 ± 0.04	0.00 ± 0.03
AL	0.24 ± 0.27	$\mathbf{0.29 \pm 0.32}$	0.24 ± 0.29	0.05 ± 0.09	0.03 ± 0.08	0.05 ± 0.10	0.22 ± 0.28
CL	$\mathbf{0.30 \pm 0.29}$	0.29 ± 0.29	0.25 ± 0.29	0.13 ± 0.20	0.05 ± 0.09	0.14 ± 0.25	0.22 ± 0.23
KM	$\mathbf{0.46 \pm 0.33}$	0.46 ± 0.34	n/a	0.36 ± 0.29	0.36 ± 0.32	0.41 ± 0.34	0.43 ± 0.28
MFG	n/a	n/a	n/a	0.41 ± 0.31	0.45 ± 0.32	0.47 ± 0.31	$\mathbf{0.51 \pm 0.30}$
SPC	$\mathbf{0.41 \pm 0.34}$	0.40 ± 0.32	0.39 ± 0.32	0.07 ± 0.06	0.08 ± 0.07	0.13 ± 0.13	0.08 ± 0.06
SNN	$\mathbf{0.32 \pm 0.28}$	$\mathbf{0.32 \pm 0.28}$	n/a	0.29 ± 0.22	0.27 ± 0.24	0.26 ± 0.23	0.28 ± 0.19

Tabela 4.3: *Affymetrix*: média do cR para o **contexto A**.

Alg.	P	C	SP	E_{Z_0}	E_{Z_1}	E_{Z_2}	E_{Z_3}
SL	0.16 ± 0.24	$\mathbf{0.19 \pm 0.25}$	0.18 ± 0.28	0.13 ± 0.19	0.10 ± 0.18	0.11 ± 0.19	0.11 ± 0.21
AL	0.39 ± 0.28	$\mathbf{0.41 \pm 0.31}$	0.35 ± 0.30	0.19 ± 0.22	0.15 ± 0.21	0.16 ± 0.21	0.35 ± 0.30
CL	0.38 ± 0.29	$\mathbf{0.39 \pm 0.29}$	0.32 ± 0.29	0.32 ± 0.23	0.28 ± 0.26	0.26 ± 0.24	0.33 ± 0.23
KM	0.50 ± 0.30	0.51 ± 0.30	n/a	$\mathbf{0.51 \pm 0.27}$	0.49 ± 0.27	0.49 ± 0.28	0.50 ± 0.24
MFG	n/a	n/a	n/a	0.54 ± 0.25	$\mathbf{0.61 \pm 0.26}$	0.58 ± 0.26	0.57 ± 0.25
SPC	0.46 ± 0.30	0.45 ± 0.29	$\mathbf{0.48 \pm 0.27}$	0.09 ± 0.08	0.10 ± 0.07	0.15 ± 0.13	0.09 ± 0.07
SNN	0.43 ± 0.23	$\mathbf{0.44 \pm 0.22}$	n/a	0.36 ± 0.21	0.31 ± 0.24	0.31 ± 0.22	0.35 ± 0.18

Tabela 4.4: *Affymetrix*: média do cR para o **contexto B** .

Alg.	P	C	SP	E_{Z_0}	E_{Z_1}	E_{Z_2}	E_{Z_3}
SL	3.57 ± 3.41	3.00 ± 3.07	$\mathbf{2.43 \pm 2.87}$	4.38 ± 3.92	4.19 ± 3.82	4.10 ± 3.56	3.62 ± 3.60
AL	$\mathbf{1.52 \pm 1.66}$	1.67 ± 2.03	3.33 ± 2.13	3.38 ± 3.38	2.90 ± 2.51	2.95 ± 3.25	2.62 ± 2.92
CL	$\mathbf{1.71 \pm 1.98}$	2.19 ± 2.38	1.90 ± 2.17	2.86 ± 2.71	3.38 ± 3.02	2.71 ± 2.76	2.90 ± 2.96
KM	1.00 ± 1.48	$\mathbf{0.95 \pm 1.12}$	n/a	2.24 ± 2.07	2.52 ± 3.44	1.95 ± 2.16	1.67 ± 2.01
MFG	n/a	n/a	n/a	1.67 ± 1.77	1.67 ± 2.11	2.14 ± 2.26	$\mathbf{1.14 \pm 1.90}$
SPC	1.33 ± 1.62	1.43 ± 1.78	1.71 ± 2.47	1.29 ± 1.49	0.71 ± 0.96	1.24 ± 1.61	$\mathbf{0.52 \pm 0.81}$
SNN	1.95 ± 2.36	$\mathbf{1.52 \pm 1.75}$	n/a	2.05 ± 2.13	2.57 ± 2.38	2.33 ± 2.52	1.81 ± 1.72

Tabela 4.5: *Affymetrix*: média da diferença entre o número de grupos encontrado pela partição com melhor cR e o número real de classes.

Alg.	P	C	SP	E_{Z_0}	E_{Z_1}	E_{Z_2}	E_{Z_3}
SL	0.06 ± 0.12	0.05 ± 0.12	$\mathbf{0.10 \pm 0.23}$	-0.01 ± 0.04	0.01 ± 0.04	0.00 ± 0.05	-0.01 ± 0.05
AL	0.22 ± 0.22	$\mathbf{0.26 \pm 0.22}$	0.22 ± 0.23	0.18 ± 0.10	0.05 ± 0.13	0.02 ± 0.13	0.13 ± 0.25
CL	$\mathbf{0.25 \pm 0.20}$	0.21 ± 0.15	0.19 ± 0.13	0.17 ± 0.22	0.10 ± 0.12	0.13 ± 0.14	0.16 ± 0.13
KM	$\mathbf{0.51 \pm 0.21}$	0.42 ± 0.27	n/a	0.43 ± 0.23	0.35 ± 0.27	0.37 ± 0.25	0.33 ± 0.25
MFG	n/a	n/a	n/a	$\mathbf{0.46 \pm 0.24}$	0.37 ± 0.22	0.38 ± 0.23	0.42 ± 0.21
SPC	$\mathbf{0.32 \pm 0.31}$	0.31 ± 0.31	0.28 ± 0.23	0.08 ± 0.03	0.09 ± 0.06	0.10 ± 0.09	0.08 ± 0.04
SNN	0.30 ± 0.24	$\mathbf{0.31 \pm 0.23}$	n/a	0.24 ± 0.16	0.20 ± 0.12	0.22 ± 0.14	0.29 ± 0.22

Tabela 4.6: cDNA: média do cR para **contexto A**.

Alg.	P	C	SP	E_{Z_0}	E_{Z_1}	E_{Z_2}	E_{Z_3}
SL	$\mathbf{0.17 \pm 0.22}$	0.16 ± 0.23	0.12 ± 0.22	0.03 ± 0.06	0.04 ± 0.07	0.03 ± 0.06	0.05 ± 0.08
AL	$\mathbf{0.35 \pm 0.17}$	0.35 ± 0.22	0.31 ± 0.17	0.16 ± 0.24	0.14 ± 0.23	0.13 ± 0.23	0.25 ± 0.21
CL	$\mathbf{0.32 \pm 0.15}$	0.29 ± 0.13	0.24 ± 0.12	0.27 ± 0.20	0.20 ± 0.15	0.19 ± 0.16	0.20 ± 0.12
KM	$\mathbf{0.53 \pm 0.20}$	0.48 ± 0.23	n/a	0.52 ± 0.20	0.47 ± 0.20	0.45 ± 0.20	0.40 ± 0.21
MFG	n/a	n/a	n/a	$\mathbf{0.54 \pm 0.21}$	0.47 ± 0.18	0.46 ± 0.17	0.46 ± 0.18
SPC	$\mathbf{0.42 \pm 0.26}$	0.41 ± 0.26	0.35 ± 0.20	0.10 ± 0.04	0.10 ± 0.06	0.13 ± 0.10	0.10 ± 0.06
SNN	$\mathbf{0.39 \pm 0.23}$	0.38 ± 0.24	n/a	0.26 ± 0.16	0.23 ± 0.12	0.25 ± 0.14	0.32 ± 0.21

Tabela 4.7: cDNA: média do cR para **contexto B**.

Alg.	P	C	SP	E_{Z_0}	E_{Z_1}	E_{Z_2}	E_{Z_3}
SL	2.64 ± 2.24	2.14 ± 2.51	$\mathbf{1.93 \pm 2.23}$	1.93 ± 1.94	2.21 ± 2.64	2.36 ± 2.56	2.43 ± 2.47
AL	2.14 ± 1.51	2.00 ± 2.08	$\mathbf{1.50 \pm 1.91}$	2.00 ± 1.92	3.29 ± 2.37	2.64 ± 2.24	2.64 ± 2.47
CL	$\mathbf{1.86 \pm 2.21}$	2.07 ± 2.02	1.93 ± 2.02	2.07 ± 2.67	3.43 ± 2.28	3.36 ± 2.10	3.36 ± 2.34
KM	$\mathbf{0.36 \pm 0.84}$	0.64 ± 0.84	n/a	1.79 ± 2.01	2.07 ± 1.86	1.57 ± 1.55	1.21 ± 1.67
MFG	n/a	n/a	n/a	$\mathbf{1.43 \pm 1.50}$	2.36 ± 2.13	1.93 ± 1.77	1.71 ± 2.02
SPC	2.29 ± 2.30	2.64 ± 2.06	1.93 ± 1.90	$\mathbf{0.79 \pm 0.97}$	0.86 ± 1.29	1.86 ± 2.88	1.14 ± 2.14
SNN	1.36 ± 1.78	2.14 ± 2.07	n/a	$\mathbf{0.93 \pm 1.49}$	2.14 ± 2.32	2.00 ± 2.22	1.87 ± 1.77

Tabela 4.8: cDNA: média da diferença entre o número de grupos encontrado pela partição com melhor cR e o número real de classes.

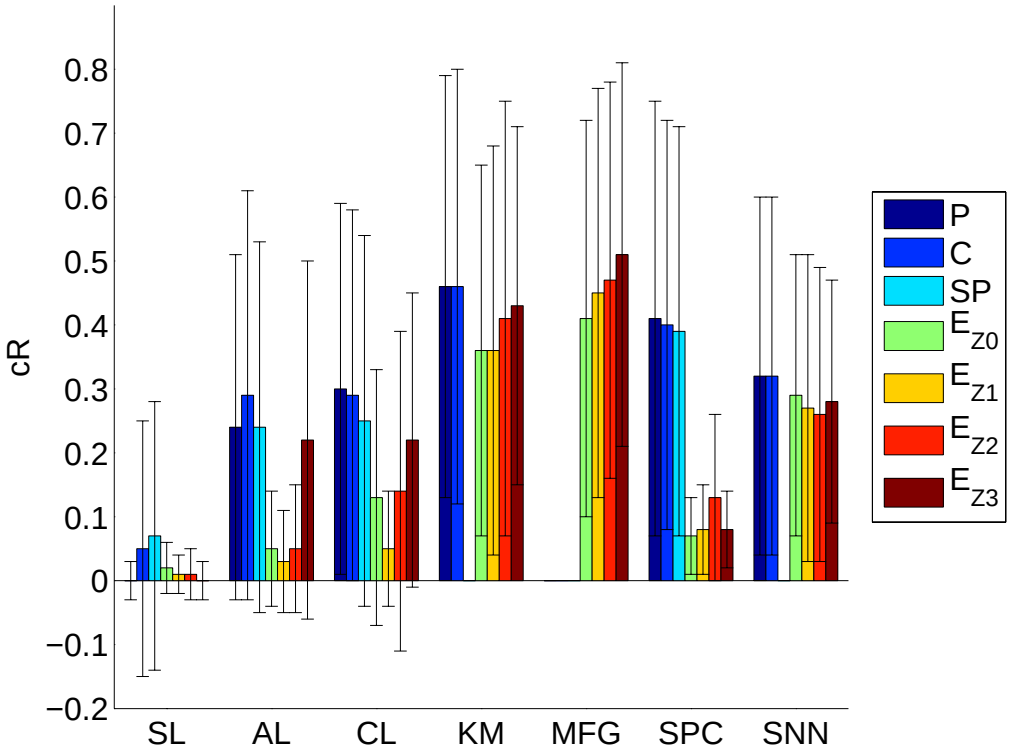


Figura 4.1: *Affymetrix*: Média do cR para o **contexto A**.

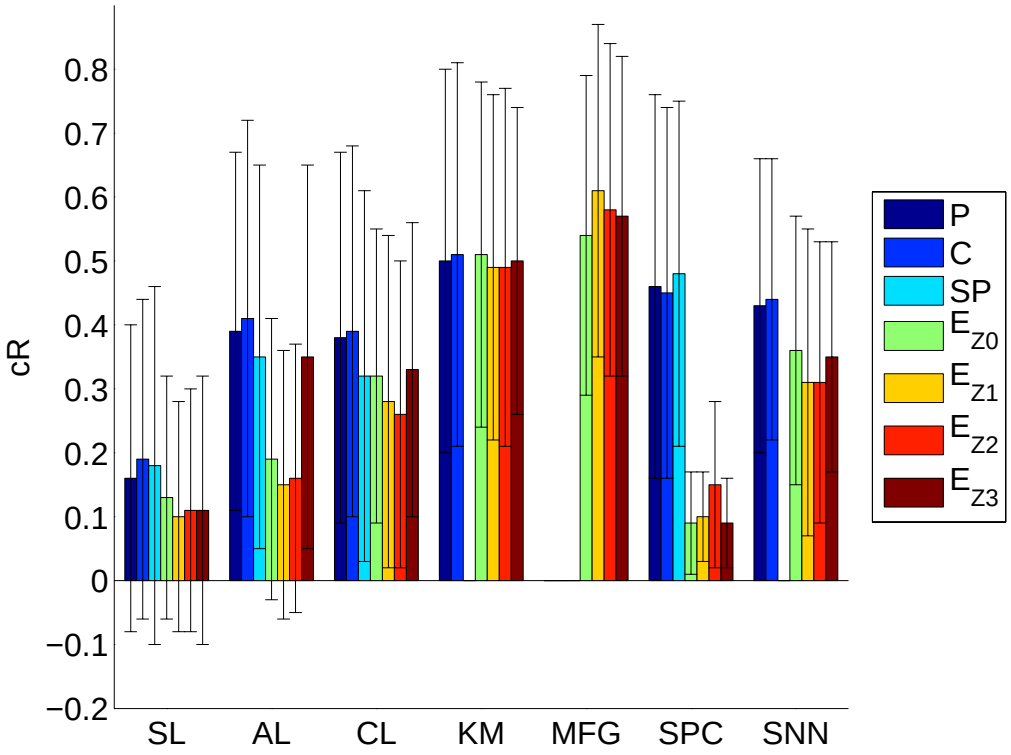


Figura 4.2: *Affymetrix*: Média cR para o **contexto B**.

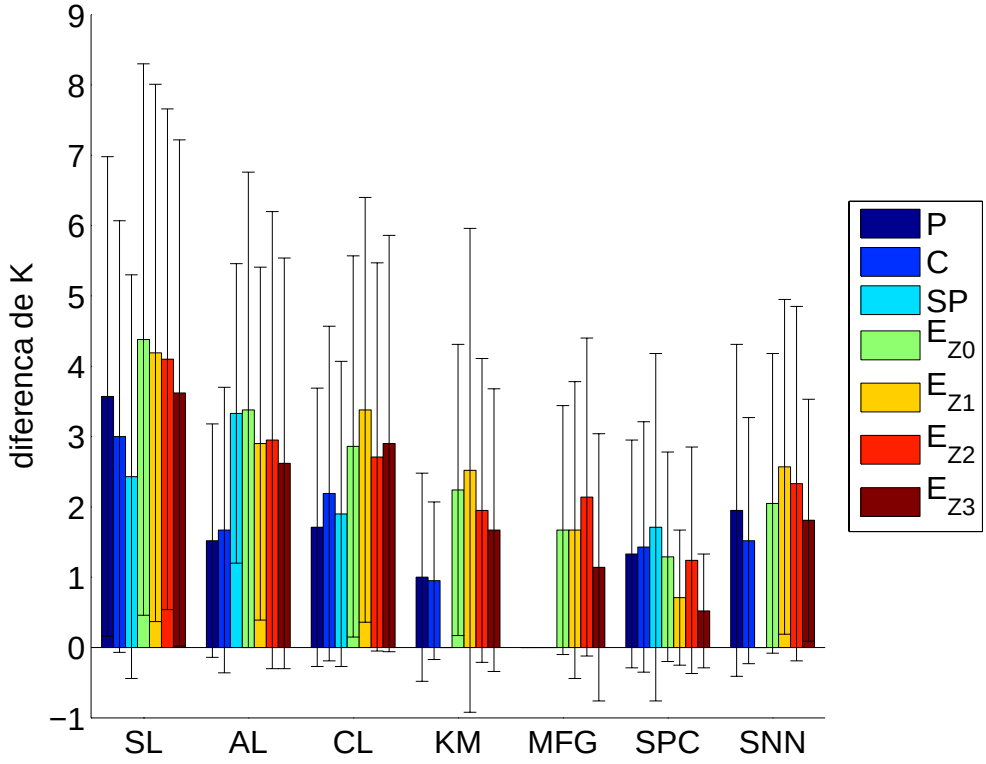


Figura 4.3: *Affymetrix*: Média da diferença entre o número de grupos encontrado pela partição com maior cR e o número real de classes.

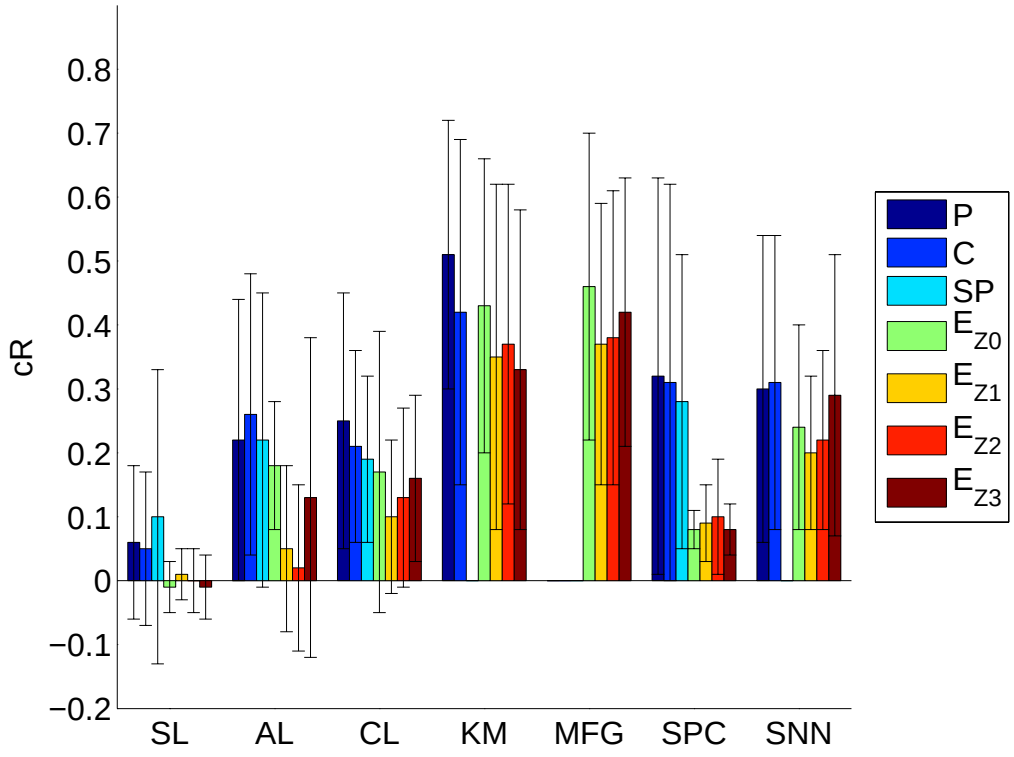


Figura 4.4: cDNA: Média do cR para o **contexto A**.

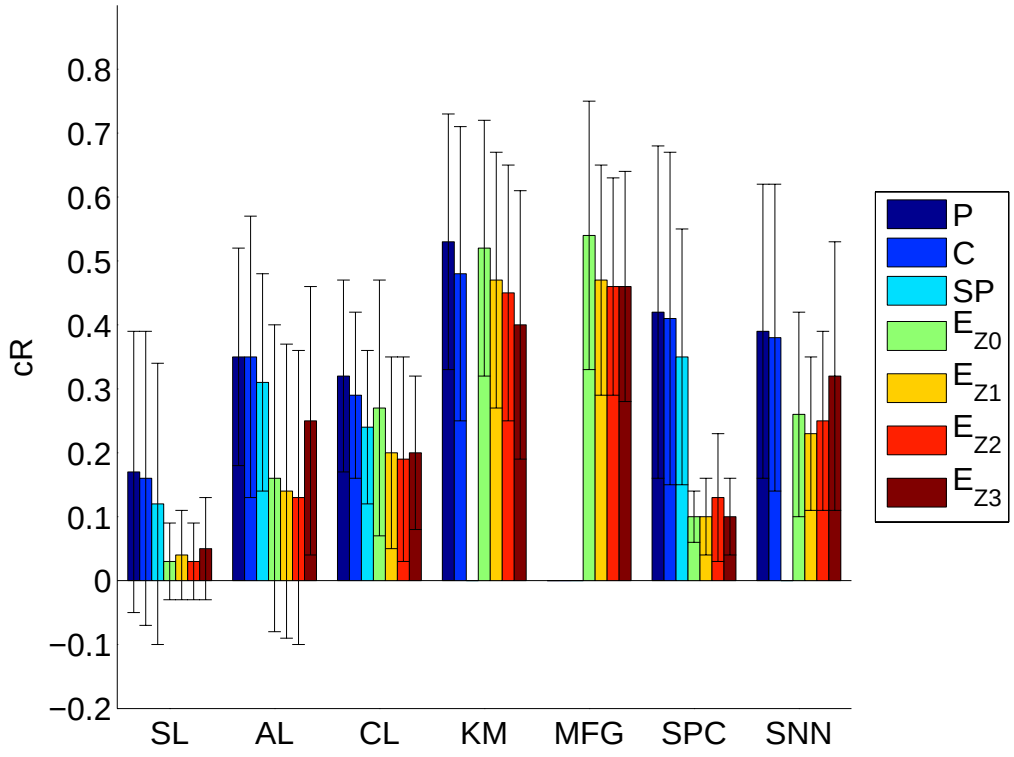


Figura 4.5: cDNA: Média do cR para o **contexto B**.

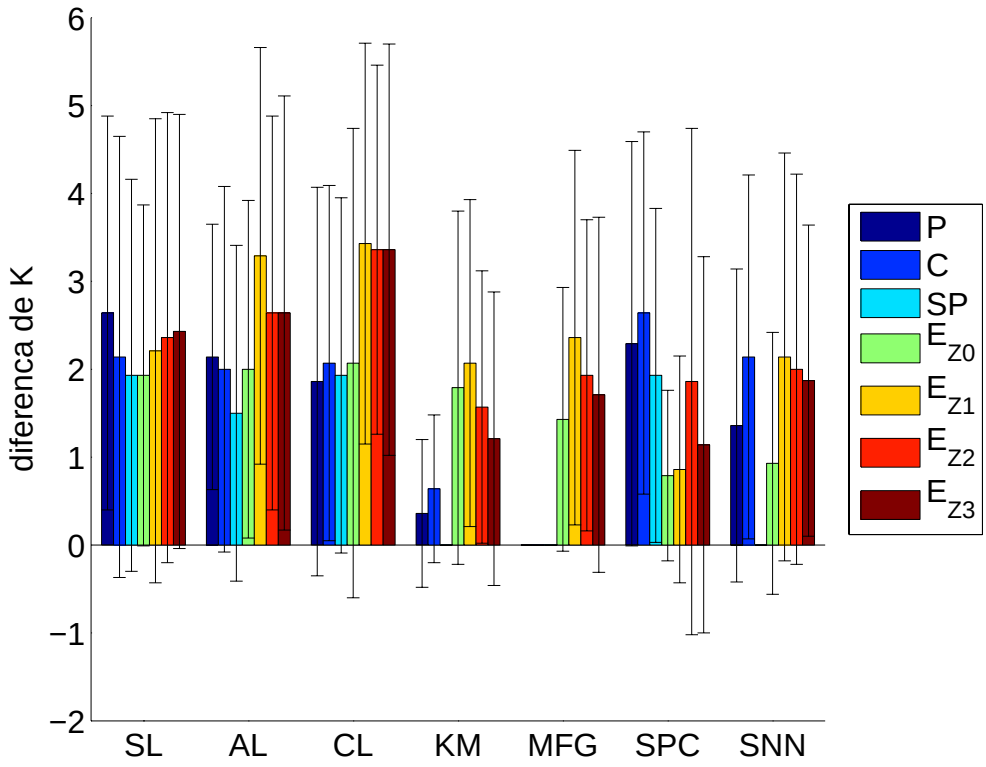


Figura 4.6: cDNA: Média da diferença entre o número de grupos encontrado pela partição com maior cR e o número real de classes.

4.5 Comparação: Agrupamento Hierárquico e *k-means*

4.5.1 `Alizadeh-v2`

A primeira base de dados selecionada para análise foi a `Alizadeh-v2`: a mesma usada por Alizadeh et al. (2000) e também usada neste trabalho. Tal conjunto de dados possui amostras de três tipos de câncer: 42 amostras de *diffuse large B-cell lymphoma* (DLCL), nove amostras de *follicular lymphoma* (FL) e onze amostras de *chronic lymphocytic leukemia* (CLL).

No caso da técnica hierárquica analisada, foi utilizado o mesmo algoritmo usado por Alizadeh et al. (2000), ou seja, o algoritmo hierárquico com ligação média e correlação de Pearson como medida de proximidade. Além disso, para uma melhor visualização da plotagem dos dados de *microarray* referentes a essa base de dados, as folhas da árvore gerada foram rearranjadas de acordo com um procedimento padrão (BAR-JOSEPH; GIFFORD; JAAKKOLA, 2001).

A Figura 4.7(a) mostra a partição obtida pelo algoritmo hierárquico e a Figura 4.7(b) mostra a partição com três grupos gerada pelo *k-means*. A partição gerada pelo *k-means* se aproxima muito da estrutura natural do conjunto de dados. Com exceção de uma única amostra da classe DLCL atribuída erroneamente ao primeiro grupo (todas as outras amostras no grupo são da classe FL), os outros grupos da partição apresentam somente amostras de uma única classe.

No caso do algoritmo hierárquico, cortando o dendrograma no nível três, ou seja, formando três sub-árvores ou três grupos, a subárvore vermelha contém somente amostras da classe DLCL, a subárvore azul contém uma combinação de amostras das classes FL e CLL e a subárvore preta é formada por somente uma amostra de DLCL. A mesma amostra DLCL atribuída erradamente no caso do *k-means* apareceu na subárvore azul, em lugar de aparecer na subárvore vermelha. A terceira subárvore (preta) é formada por somente uma amostra da classe DLCL. Desse modo, para a separação das classes FL e CLL seria necessário cortar a árvore no nível 25, o que levaria a formação de 25 grupos distintos.

Uma das razões para esse tipo de problema é que os algoritmos hierárquicos são baseados em decisões locais, formando, a cada iteração, um único grupo a partir dos dois grupos com-

pactos mais próximos. O critério de compacticidade (quão próximos são os grupos) é definido pelo tipo de ligação e medida de proximidade que o algoritmo utiliza. O *k-means*, por outro lado, usa um critério de agrupamento que maximiza tanto a compacticidade quanto a separação dos grupos.

Esse problema também pode ser ilustrado em uma representação tridimensional dos dados a partir das três maiores componentes principais obtidas depois de um PCA - *Principal Component Analysis* - (Figura 4.8). A partir dessa ilustração é possível notar a presença de três classes distintas na estrutura dos dados. Entretanto, apesar de a classe formada pelas amostras de cor vermelha estar bem separada das outras duas classes, ela se distribui em uma região não-compacta do espaço. Assim, como os algoritmos hierárquicos têm um viés para grupos compactos, tende a unir por último, os pontos em regiões menos compactas. De fato, na Figura 4.7(a), se a árvore hierárquica for seguida seria primeiro sugerido a subdivisão do grupo com as amostras DLCL para depois dividir os grupos com amostras das classes FL e CLL.

Desse modo, em um cenário hipotético onde não se tem informação *a priori* sobre a diferença das classes FL e CLL, cortando-se a árvore no nível três, o uso somente de algoritmos hierárquicos não indicaria a existência dessas duas classes como separadas. Por outro lado, essa divisão pode ser claramente detectada usando técnicas como o *k-means* ou visualizadas usando PCA.

4.5.2 Nutt-v3

Em um outro exemplo ilustrativo, os mesmos algoritmos analisados anteriormente lidam de maneira diferente para uma outra base de dados. Nesse caso uma versão da base de dados usada por Nutt et al. (2003) e utilizada neste trabalho que possui duas classes: *classic anaplastic oligodendroglioma* (CO) e *nonclassic anaplastic oligodendroglioma* (NO). A versão da base usada neste exemplo passou pela transformação Z_3 descrita na Seção 3.1.2.

Utilizando-se, respectivamente, o algoritmo hierárquico com ligação média e distância euclidiana e o algoritmo *k-means* com a mesma medida de proximidade, é possível notar, de

acordo com a Figura 4.9(a), que se árvore gerada pelo algoritmo hierárquico for cortada no nível dois, gerando dois grupos, as amostras das duas classes (CO e NO) são completamente separadas. De fato, a plotagem de expressão gênica mostra uma clara diferença entre os padrões de expressão gênica das duas classes envolvidas. Apesar disso, o algoritmo *k-means* não consegue realizar essa separação de maneira efetiva. A Figura 4.9(b) também mostra a partição gerada pelo *k-means*, onde em um dos grupos gerados, existe somente elementos de uma classe, porém, no outro grupo reside uma mistura das demais amostras de diferentes classes.

O uso da representação tridimensional a partir das três maiores componentes de uma PCA feita sobre a base de dados (Figura 4.10(a)) e o sobre a partição gerada pelo *k-means* (Figura 4.10(b) - os símbolos "*" na figura representam os centróides de cada grupo encontrado) ajudam a entender os desempenhos obtidos pelos algoritmos. Primeiramente, é possível perceber que as classes possuem formatos não-esféricos e que os objetos encontram-se dispersos dentro dos classes. Por exemplo, na classe CO existe um elemento que se encontra bastante afastado dos demais pontos (indicação de um possível *outlier*) e na classe NO nota-se a formação de uma sub-região compacta e separada dos restante das classes, indicando a possibilidade de existência de um subgrupo. Tais indicações levariam a uma possível explicação para o baixo desempenho do *k-means* que, como mencionado na Seção 3.2.2, tem dificuldades para agrupar conjuntos de dados que contém grupos com formato não-esféricos e de tamanhos (densidades) diferentes, devido a inadequação sua função objetivo para tais casos.

De fato, diferentemente do hierárquico que usa um critério local para agrupar os dados, a função objetivo do *k-means* otimiza um critério global. Mais precisamente, a cada iteração do algoritmo os centróides dos grupos são recalculados de acordo com seus elementos atuais buscando maximizar a compacticidade dos grupos (minimizar a variância interna). Como no conjunto de dados existe a presença de sub-região separadas (possíveis subgrupos ou *outliers*), isso implica em um reposicionamento dos centróides para uma posição mais próxima à essas regiões, com isso, levando a uma atribuição errônea de elementos de classes diferentes ao mesmo grupo. Por exemplo, de acordo com a Figura 4.10(b), pode-se ver que objetos da classe NO (azul) que estão mais próximos da classes CO (vermelha) foram atribuídos, de maneira errada,

a essa classe.

A Figura 4.11 mostra o dendrograma feito somente com as amostras da classe NO. Existe um padrão na expressão dos genes que dividem o conjunto de amostras em dois grupos distintos (subárvores azul e vermelha). Essa subdivisão pode ser confirmada analisando a Figura 4.10(a) que mostra que a classe NO é subdividida em dois grupos: uma contendo cinco amostras e o outro com o restante das amostras da classe NO.

No caso dos algoritmos hierárquicos, a existência de subregiões dividiu o conjunto de dados em áreas mais homogêneas, tornando-se o cenário ideal para o critério local de agrupamento utilizado pelos algoritmos hierárquicos. É possível visualizar isso na Figura 4.9(a), onde, em geral, elementos de regiões diferentes são mesclados a cada passo do algoritmo.

Um forte indício de presença de *outliers* nesse conjunto de dados é o aumento significativo de desempenho do algoritmo hierárquico após a aplicação da transformação Z_3. Mais especificamente, o desempenho do algoritmo hierárquico, em termos de cR, subiu de 0.10, quando usado o conjunto de dados sem qualquer tipo de transformação (Z_0), para 1.00 (combinação perfeita com a estrutura natural dos dados) quando usado os dados com a transformação Z_3. Alguns estudos já apontam para a sensibilidade do hierárquico a *outliers* e o impacto da transformação Z_3 nessa classe de algoritmos (SOUTO et al., 2008a).

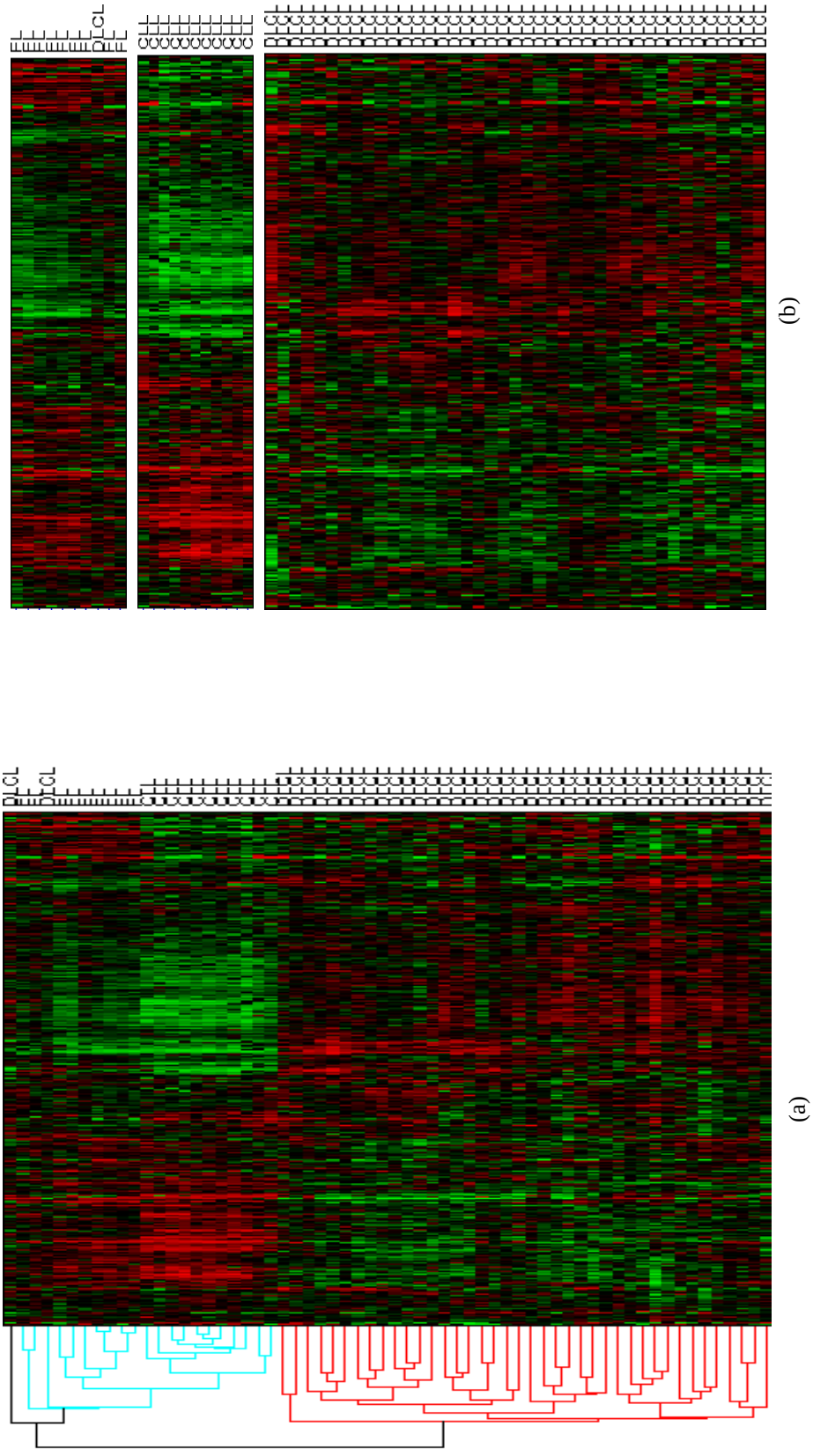


Figura 4.7: `Alizadeh-v2`: (a) partição gerada pelo algoritmo hierárquico; (b) partição gerada pelo *k-means*.

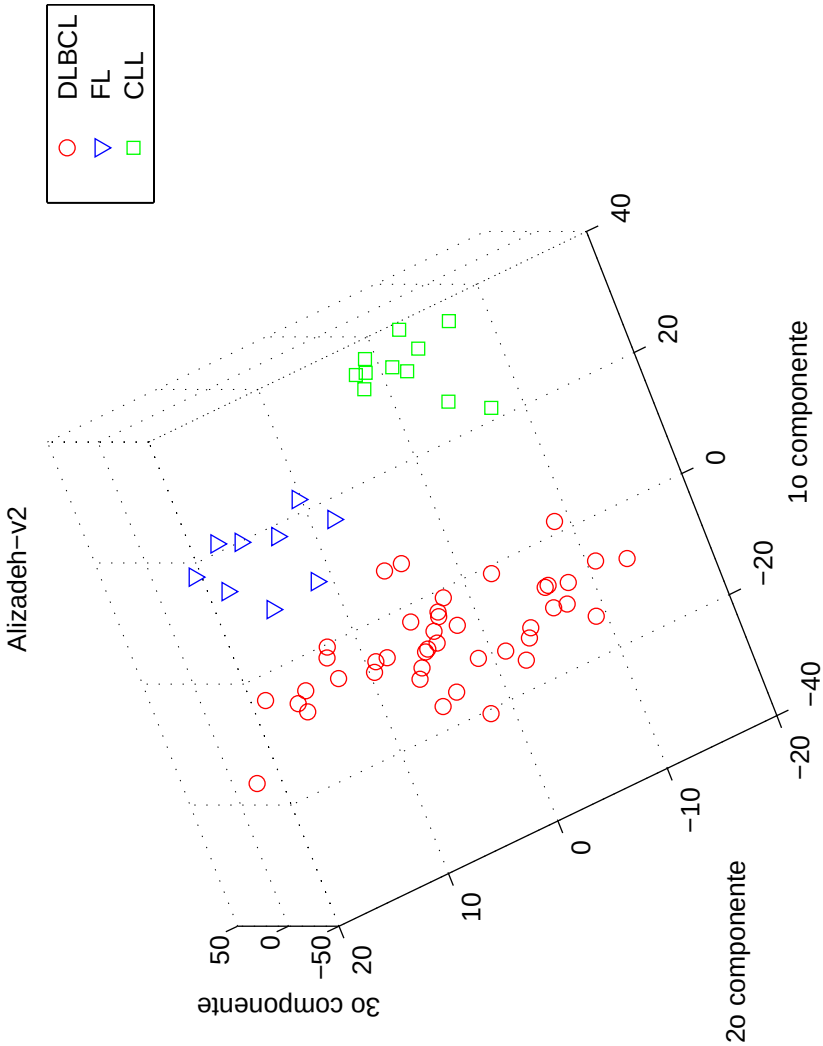


Figura 4.8: `Alizadeh-v2`: visualização 3D do conjunto de dados.

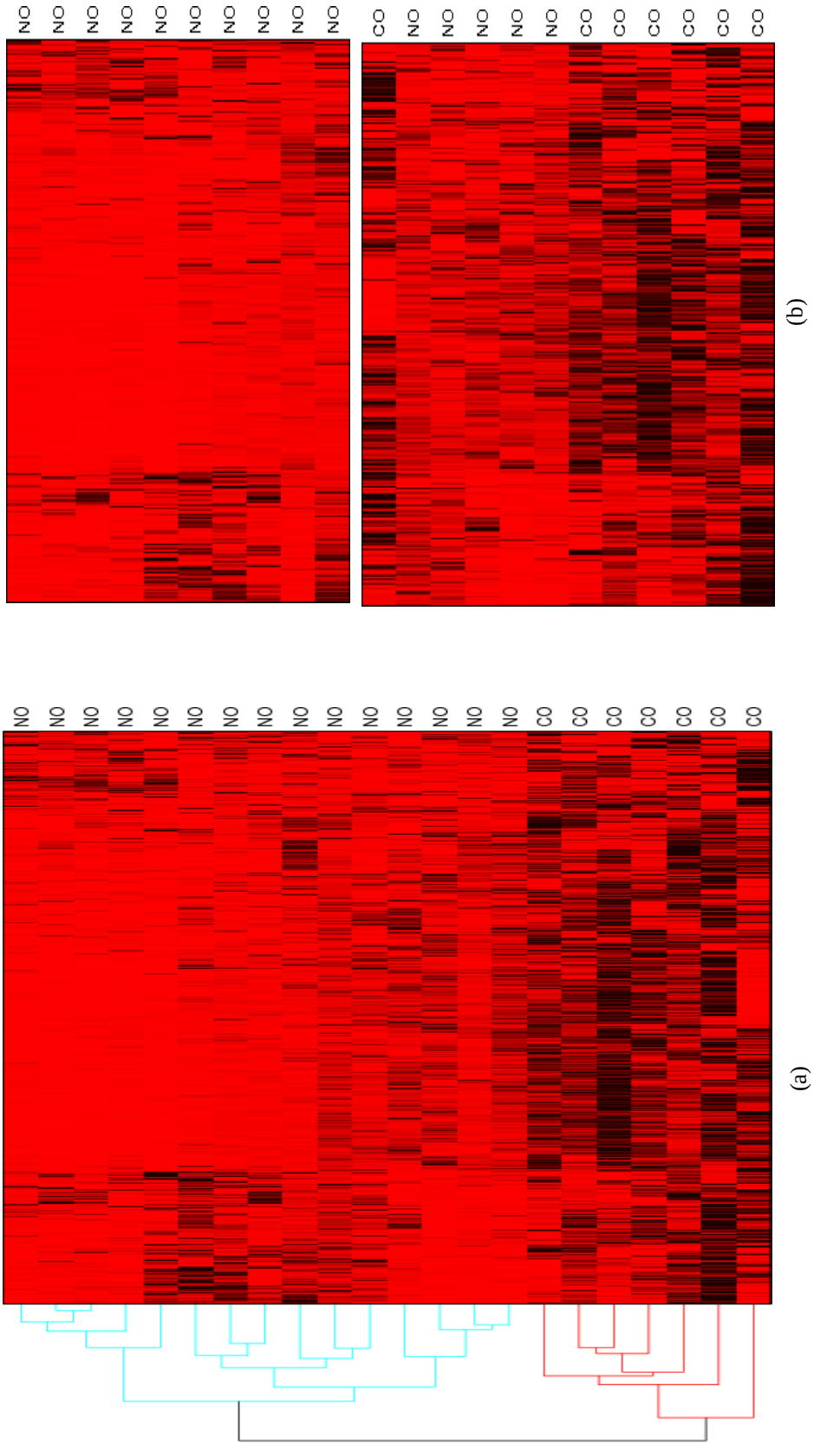


Figura 4.9: `Mutt-v3`: (a) partição gerada pelo algoritmo hierárquico; (b) partição gerada pelo *k-means*.

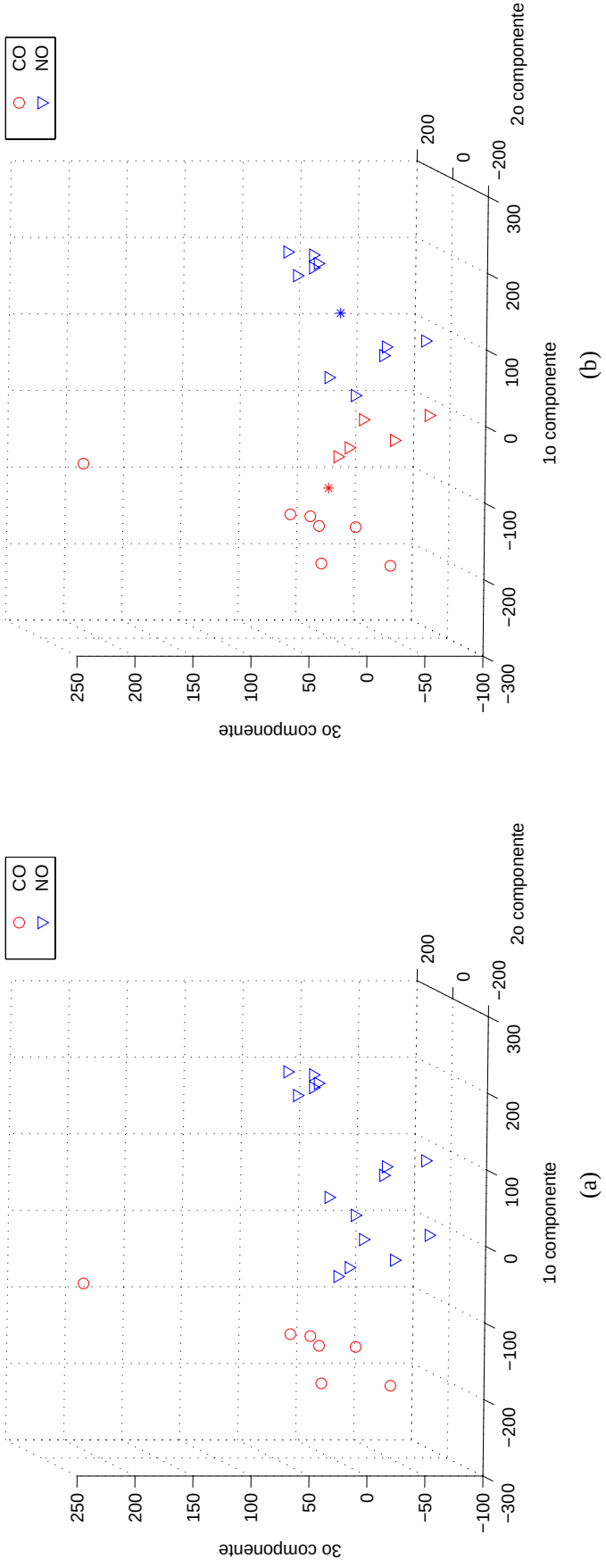


Figura 4.10: `Mutt-v3`: (a) visualização 3D do conjunto de dados; (b) visualização 3D da partição gerada pelo *k-means*.

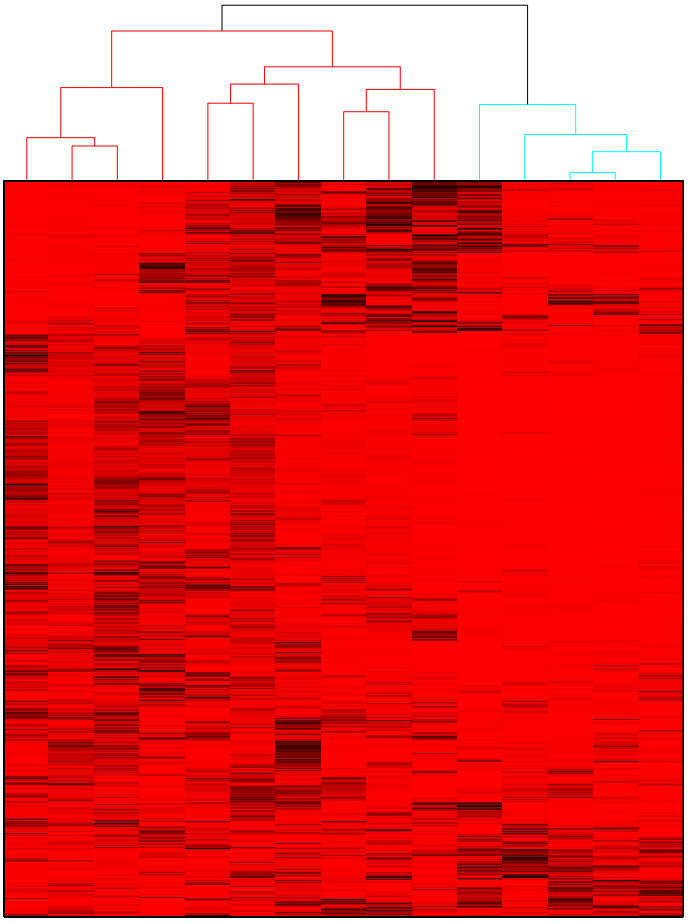


Figura 4.11: `Nutt-v3`: dendrograma somente das amostras da classe NO da base.

5 *Conclusões*

5.1 Conclusões e Trabalhos Futuros

Neste trabalho é apresentado um primeiro estudo comparativo em grande escala de sete algoritmos de agrupamento e quatro medidas de proximidade aplicados a 35 bases de dados de expressão gênica. Os resultados são apresentados em termos de recuperação da estrutura natural dos dados encontrados por diferentes configurações de técnica de agrupamento e medida de proximidade. Também foi investigado o uso de cobertura reduzida pelos algoritmos de agrupamento. Parte dos resultados obtidos nesta dissertação foram publicados em (SOUTO et al., 2008a). Algumas das tendências (diretrizes para agrupamento de dados de expressão gênica de câncer) que surgiram durante a análise deste estudo comparativo são apresentadas a seguir.

1. De uma maneira geral, os resultados mostraram que, para as 35 bases de dados investigadas, a técnica Mixtura Finita de Gaussianas exibiu o melhor desempenho, seguido de perto pelo *k-means*, em termos de recuperação da estrutura natural dos dados, independentemente da medida de proximidade utilizada por esse último algoritmo.

2. Para a maioria dos algoritmos, existe uma clara relação entre o uso de cobertura reduzida e um aumento na habilidade de agrupar corretamente o conjunto de dados (maiores valores de cR).

3. O baixo desempenho das técnicas hierárquicas utilizadas pode ser notado facilmente, como também foi apontado por vários outros estudos relacionados a agrupamento de

genes a partir de dados de expressão gênica (D'HAESELEER, 2005; COSTA; CARVALHO; SOUTO, 2004; DATTA; DATTA, 2003). Uma das razões para esse baixo desempenho é a sensibilidade desse tipo de técnicas a ruídos nos dados (D'HAESELEER, 2005; JAIN; DUBES, 1988; MILLIGAN; COOPER, 1988).

4. Dentro dessa classe de algoritmos, o hierárquico usando ligação simples apresentou os piores resultados.

5. Com relação ao uso de medidas de proximidade pelas técnicas de agrupamento hierárquico, a correlação de Pearson e o cosseno levaram a melhores resultados.

6. Para apresentar valores de cR compatíveis com KM e MFG, a classe de técnicas de agrupamento hierárquico, em geral, necessitam de um uso muito maior de cobertura reduzida.

7. A técnica *Spectral clustering* mostrou ser sensível ao tipo de medida de proximidade usada.

É importante ressaltar que, embora, em média, os algoritmos MFG e KM tenham apresentado melhores resultados, em termos de *corrected Rand*, do que as demais técnicas investigadas, isso não implica que essas técnicas devem ser sempre a melhor escolha quando se quer agrupar dados de expressão gênica. De fato, para certas bases de dados o SNN usando P, por exemplo, apresenta um maior cR que todos os outros algoritmos.

Uma maneira para tratar esse problema de prever qual seria a melhor técnica para aplicar a uma determinada base de dados com determinadas características (por exemplo, número de amostras, dimensionalidade das amostras, tipo de *microarray* utilizado etc.) é o uso de abordagens de meta-aprendizado (CARRIER; VILALTA; BRAZDIL, 2004). Por exemplo, Souto et al. (2008b) mostraram resultados preliminares de uma abordagem de meta-aprendizado levando em consideração somente algumas estatísticas descritivas do conjunto de dados como entrada, e gerando um *ranking* dos melhores métodos de agrupamento para serem usados naquela base de dados específica.

Uma outra contribuição desta dissertação é a disponibilização de um conjunto de bases de dados que podem ser usadas como uma base estável e confiável para avaliação e comparação de diferentes técnicas de aprendizado de máquina. Recentemente, estudos propondo *frameworks* de referência (*benchmark frameworks*) foram introduzidos no contexto de aprendizado de máquina e na literatura de bioinformática (BLOCKEEL; VANSCHOREN, 2007; STATNIKOV et al., 2005). Por exemplo, Blockeel e Vanschoren (2007) propôs um sistema baseado na *Web* para armazenar resultados obtidos a partir de métodos de classificação para várias bases de dados de referência (*benchmark data sets*) usadas pela comunidade de aprendizado de máquina.

De fato, como trabalhos futuros, pode ser feita a construção de um repositório central para avaliação de métodos de agrupamento no contexto de dados de expressão gênica de câncer. Nesse repositório, qualquer novo método de agrupamento poderia ser avaliado com os conjuntos de dados disponíveis, tendo seus resultados armazenados e disponíveis na base de dados *Web*.

Por fim, nesta dissertação são utilizados tanto algoritmos "tradicionais" quanto algoritmos criados para lidar com dados em alta dimensionalidade, embora, dessa última classe de algoritmos apenas dois métodos foram investigados: SNN e *Spectral clustering*. Como um outro trabalho futuro, podem ser investigados mais métodos específicos para tratar dados de expressão gênica (dados em altas dimensões), como também outras medidas de proximidade.

Referências Bibliográficas

ALBERTS, B. et al. *Biologia Molecular da Célula*. 3. ed. Porto Alegre: Editora Artes Médicas, 1997.

ALIZADEH, A. A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, v. 403, n. 6769, p. 503–511, fev. 2000. Disponível em: <http://dx.doi.org/10.1038/35000501>.

ARMSTRONG, S. A. et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, v. 30, n. 1, p. 41–47, jan. 2002. Disponível em: <http://dx.doi.org/10.1038/ng765>.

BAMMLER, T. et al. Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods*, v. 2, n. 5, p. 351–356, maio 2005. ISSN 1548-7091. Disponível em: <http://dx.doi.org/10.1038/nmeth754>.

BAR-JOSEPH, Z.; GIFFORD, D. K.; JAAKKOLA, T. S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, Laboratory for Computer Science, MIT, 545 Technology Square, Cambridge, MA 02139, USA, v. 17 Suppl 1, 2001. ISSN 1367-4803. Disponível em: <http://view.ncbi.nlm.nih.gov/pubmed/11472989>.

BARBARA, D. *An introduction to cluster analysis for data mining*. [S.l.], 2000. Acessado em 20/07/2008. Disponível em: <http://www.cs.umn.edu/ han/dmclass/cluster_survey_10_02_00.pdf>.

BHATTACHARJEE, A. et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*, v. 98, n. 24, p. 13790–13795, nov. 2001. Disponível em: <http://dx.doi.org/10.1073/pnas.191502998>.

BITTNER, M. et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, v. 406, n. 6795, p. 536–540, ago. 2000. Disponível em: <http://dx.doi.org/10.1038/35020115>.

BLOCKEEL, H.; VANSCHOREN, J. Experiment databases: Towards an improved experimental methodology. In: *Machine Learning. Lecture Notes in Computer Science*. [S.l.: s.n.], 2007. p. 6–17.

BREDEL, M. et al. Functional network analysis reveals extended gliomagenesis pathway maps and three novel myc-interacting genes in human gliomas. *Cancer Res*, v. 65, n. 19, p. 8679–8689, out. 2005. Disponível em: <http://dx.doi.org/10.1158/0008-5472.CAN-05-1204>.

CARRIER, C. G.; VILALTA, R.; BRAZDIL, P. Introduction to the special issue on meta-learning. *Machine Learning*, v. 54, n. 3, p. 187–193, 2004.

CHEN, X. et al. Gene expression patterns in human liver cancers. *Mol. Biol. Cell*, v. 13, n. 6, p. 1929–1939, 2002.

CHOWDARY, D. et al. Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *J Mol Diagn*, v. 8, n. 1, p. 31–39, fev. 2006.

COSTA, I. G.; CARVALHO, F. A. T. de; SOUTO, M. C. P. de. Comparative analysis of clustering methods for gene expression time course data. *Genetics and Molecular Biology*, v. 27, n. 4, p. 623–631, 2004. Disponível em: <http://www.scielo.br/pdf/gmb/v27n4/22434.pdf>.

DATTA, S.; DATTA, S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, v. 19, p. 459–466, 2003.

D'HAESELEER, P. How does gene expression clustering work? *Nat Biotech*, v. 23, n. 12, p. 1499–1501, dez. 2005. ISSN 1087-0156. Disponível em: <http://dx.doi.org/10.1038/nbt1205-1499>.

DUGGAN, D. J. et al. Expression profiling using cDNA microarrays. *Nature Genetics*, v. 21, p. 10–14, 1999.

DYRSKJOT, L. et al. Identifying distinct classes of bladder carcinoma using microarrays. *Nat Genet*, v. 33, n. 1, p. 90–96, jan. 2003. Disponível em: <http://dx.doi.org/10.1038/ng1061>.

EISEN, M. B. et al. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, Department of Genetics, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305, USA., v. 95, n. 25, p. 14863–14868, dez. 1998. ISSN 0027-8424. Disponível em: <http://dx.doi.org/10.1073/pnas.95.25.14863>.

ERTOZ, L.; STEINBACH, M.; KUMAR, V. A new shared nearest neighbor clustering algorithm and its applications. In: *Workshop on Clustering High Dimensional Data and its Applications*. [S.l.: s.n.], 2002. p. 105–115.

GARBER, M. E. et al. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A*, v. 98, n. 24, p. 13784–13789, nov. 2001. Disponível em: <http://dx.doi.org/10.1073/pnas.241500798>.

GOLUB, T. R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, v. 286, n. 5439, p. 531–537, out. 1999.

GORDON, G. J. et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res*, v. 62, n. 17, p. 4963–4967, set. 2002.

HAIR, J. F. et al. *Análise multivariada de dados.* 5. ed. Porto Alegre: Bookman, 2005.

HOON, M. J. L. de et al. Open source clustering software. *Bioinformatics*, v. 20, n. 9, p. 1453–1454, jun. 2004. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bth078>.

JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988. ISBN 013022278X.

JAIN, K. K. Biochips for gene spotting. *Science*, v. 294, n. 5542, p. 621–623, out. 2001. Disponível em: <http://dx.doi.org/10.1126/science.294.5542.621>.

JARVIS, R. A.; PATRICK, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.*, IEEE Computer Society, Washington, DC, USA, v. 22, n. 11, p. 1025–1034, 1973. ISSN 0018-9340.

KHAN, J. et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, v. 7, n. 6, p. 673–679, jun. 2001. Disponível em: <http://dx.doi.org/10.1038/89044>.

KUNCHEVA, L. I. *Combining Pattern Classifiers: Methods and Algorithms*. [S.l.]: Wiley-Interscience, 2004. Hardcover. ISBN 0471210781.

LAIHO, P. et al. Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. *Oncogene*, v. 26, n. 2, p. 312–320, jan. 2007. Disponível em: <http://dx.doi.org/10.1038/sj.onc.1209778>.

LAPOINTE, J. et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A*, v. 101, n. 3, p. 811–816, jan. 2004. Disponível em: <http://dx.doi.org/10.1073/pnas.0304146101>.

LAUSTED, C. et al. POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer. *Genome Biol*, The Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103, USA. clausted@systemsbiology.org, v. 5, n. 8, 2004. ISSN 1465-6914. Disponível em: <http://dx.doi.org/10.1186/gb-2004-5-8-r58>.

LIANG, Y. et al. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proc Natl Acad Sci U S A*, v. 102, n. 16, p. 5814–5819, abr. 2005. Disponível em: <http://dx.doi.org/10.1073/pnas.0402870102>.

LIPSHUTZ, R. J. et al. High density synthetic oligonucleotide arrays. *Nat Genet*, Affymetrix, Inc., Santa Clara, California 95051, USA. rob_lipshutz@affymetrix.com, v. 21, n. 1 Suppl, p. 20–24, jan. 1999. ISSN 1061-4036. Disponível em: <http://dx.doi.org/10.1038/4447>.

LUXBURG, U. von. A tutorial on spectral clustering. *Statistics and Computing*, v. 17, n. 4, p. 395–416, dez. 2007. Disponível em: <http://dx.doi.org/10.1007/s11222-007-9033-z>.

MILLIGAN, G. W.; COOPER, M. C. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, v. 21, n. 4, p. 441–458, 1986.

MILLIGAN, G. W.; COOPER, M. C. A study of standardization of variables in cluster analysis. *Journal of Classification*, v. 5, p. 181–204, 1988.

MONTI, S. et al. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, v. 52, p. 91–118, 2003.

NG, A.; JORDAN, M.; WEISS, Y. *On spectral clustering: Analysis and an algorithm*. 2001. Disponível em: <http://citeseer.ist.psu.edu/ng01spectral.html>.

NUTT, C. L. et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*, v. 63, n. 7, p. 1602–1607, abr. 2003.

POMEROY, S. L. et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, v. 415, n. 6870, p. 436–442, jan. 2002. Disponível em: <http://dx.doi.org/10.1038/415436a>.

QUACKENBUSH, J. Computational analysis of cDNA microarray data. *Nature Reviews*, v. 6, n. 2, p. 418–428, 2001.

RAMASWAMY, S. et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*, v. 98, n. 26, p. 15149–15154, dez. 2001. Disponível em: <http://dx.doi.org/10.1073/pnas.211566398>.

RHODES, D. R. et al. Oncomine: a cancer microarray database and integrated data-mining platform. *Neoplasia*, Department of Pathology, University of Michigan Medical School, Ann Arbor, MI 48109, USA., v. 6, n. 1, p. 1–6, 2004. ISSN 1522-8002. Disponível em: <http://view.ncbi.nlm.nih.gov/pubmed/15068665>.

RISINGER, J. I. et al. Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer. *Cancer Res*, v. 63, n. 1, p. 6–11, jan. 2003.

SCHENA, M. et al. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol*, Department of Biochemistry, Beckman Center, Stanford University Medical Center, CA 94305, USA., v. 16, n. 7, p. 301–306, jul. 1998. ISSN 0167-7799. Disponível em: <http://view.ncbi.nlm.nih.gov/pubmed/9675914>.

SHIPP, M. A. et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med*, v. 8, n. 1, p. 68–74, jan. 2002. Disponível em: <http://dx.doi.org/10.1038/nm0102-68>.

SINGH, D. et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, v. 1, n. 2, p. 203–209, mar. 2002.

SLONIM, D. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, v. 32, p. 502–508, 2002.

SOUTO, M. C. P. de et al. Comparative study on normalization procedures for cluster analysis of gene expression datasets. In: *Proc. of IEEE International Joint Conference on Neural Networks*. [S.l.: s.n.], 2008. p. 2792–2798. ISSN 1098-7576.

SOUTO, M. C. P. de et al. Ranking and selecting clustering algorithms using a meta-learning approach. In: *Proc. of IEEE International Joint Conference on Neural Networks*. Hong Kong, China: [s.n.], 2008. p. 3729–3735. ISSN 1098-7576.

SPANG, R. Diagnostic signatures from microarrays: a bioinformatics concept for personalized medicine. *BIOSILICO*, v. 1, n. 2, p. 64–68, maio 2003.

STATNIKOV, A. et al. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, Oxford University Press, Oxford, UK, v. 21, n. 5, p. 631–643, 2005. ISSN 1367-4803.

STEGMAIER, K. et al. Gene expression-based high-throughput screening(ge-hts) and application to leukemia differentiation. *Nature Genetics*, v. 36, n. 3, p. 257–263, 2004.

SU, A. I. et al. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res*, v. 61, n. 20, p. 7388–7393, out. 2001.

TOMLINS, S. A. et al. Integrative molecular concept modeling of prostate cancer progression. *Nat Genet*, v. 39, n. 1, p. 41–51, jan. 2007.

TUSHER, V. G.; TIBSHIRANI, R.; CHU, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, v. 98, n. 9, p. 5116–5121, 2001. Disponível em: <http://www.pnas.org/cgi/content/abstract/98/9/5116>.

WEST, M. et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A*, v. 98, n. 20, p. 11462–11467, set. 2001. Disponível em: <http://dx.doi.org/10.1073/pnas.201162998>.

XU, R.; WUNSCH, D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, v. 16, n. 3, p. 645–678, 2005.

YEOH, E.-J. et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, v. 1, n. 2, p. 133–143, mar. 2002.