UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós-graduação em Bioinformática

Saulo Augusto de Paula Pinto

CLUSTERING DE AMOSTRAS DE DADOS DE EXPRESSÃO GÊNICA UTILIZANDO DUAS MÉTRICAS DE SIMILARIDADE BIOLOGICAMENTE INSPIRADAS

Belo Horizonte
2008

Saulo Augusto de Paula Pinto

CLUSTERING DE AMOSTRAS DE DADOS DE EXPRESSÃO GÊNICA UTILIZANDO DUAS MÉTRICAS DE SIMILARIDADE BIOLOGICAMENTE INSPIRADAS

Versão final

Tese apresentada ao Colegiado do Programa de Doutorado em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do grau de Doutor em Bioinformática.

Orientador: Prof. Dr. José Miguel Ortega

Belo Horizonte
2008

043 Pinto, Saulo Augusto de Paula.
Clustering de amostras de dados de expressão gênica utilizando duas métricas de similaridade biologicamente inspiradas [manuscrito] / Saulo Augusto de Paula Pinto. – 2008.
92 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. José Miguel Ortega.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Doutorado em Bioinformática.

1. Bioinformática. 2. Aprendizado de Máquina não Supervisionado. 3. Expressão Gênica. 4. Métrica. I. Ortega, José Miguel. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

Universidade Federal de Minas Gerais -
Programa de Doutorado em Bioinformática/UFMG
Av. Antônio Carlos, 6627 – Pampulha
31270-901 – Belo Horizonte – MG
e-mail: bioinfo@icb.ufmg.br (31)3409-2554

ATA DA DEFESA DA TESE DE DOUTORADO DE SAULO AUGUSTO DE PAULA PINTO. Aos trinta e um dias do mês de Outubro de 2008 às 09h30min, reuniu-se no Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais a Comissão Examinadora da tese de doutorado, indicada no dia quatro de julho de 2008 durante a 53ª reunião do Colegiado do Programa, para julgar, em exame final, o trabalho intitulado "*Clustering* de amostras de dados de expressão gênica utilizando duas métricas de similaridade biologicamente inspiradas" requisito final para a obtenção do grau de Doutor em Ciências, Área de Concentração: Bioinformática. Abrindo a sessão o Presidente da Comissão, Prof. José Miguel Ortega da Universidade Federal de Minas Gerais, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a argüição pelos examinadores, com a respectiva defesa do candidato. Logo após a Comissão se reuniu sem a presença do candidato e do público para julgamento e expedição do resultado final. Foram atribuídas as seguintes indicações: Dr. Guilherme Correa Oliveira do Centro de Pesquisas René Rachou – CPqRR/FIOCRUZ, Belo Horizonte, MG, aprovado; Drª Helena Paula Brentani da Fundação Antônio Prudente, São Paulo, SP, aprovado; Drª Riva de Paula Oliveira da Universidade Federal de Ouro Preto – UFOP, Ouro Preto, MG, aprovado; Drª Gisele Lobo Pappa da Universidade Federal de Minas Gerais, aprovado; Dr. José Miguel Ortega, orientador, da Universidade Federal de Minas Gerais, aprovado. Pelas indicações o candidato foi considerado APROVADO. O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar o Presidente da Comissão encerrou a reunião e lavrou a presente ata que será assinada por todos os membros participantes da Comissão Examinadora. Belo Horizonte, aos trinta e um dias de Outubro de 2008.

Dr. Guilherme Correa Oliveira – CPqRR/FIOCRUZ

Drª Helena Paula Brentani – Fundação Antônio Prudente

Drª Riva de Paula Oliveira – UFOP

Profª. Glória Regina Franco
Coordenadora do Programa de Doutorado
em Bioinformática - UFMG

Drª Gisele Lobo Pappa – UFMG

Dr. José Miguel Ortega – orientador – UFMG

Dedico este trabalho a todos aqueles que nunca desistem. Não importa o tamanho aparente do obstáculo. Ele sempre ficará para trás.
Don't ever give up!

Agradecimentos

Obrigado, Jesus, por nos guiar e nos mostrar (e ser) o Caminho, a Verdade (Realidade) e a Vida – "aquele que faz o *bem* renascerá para a ressurreição da Vida".

Agradeço ao meu orientador, J. Miguel Ortega, por aceitar orientar-me quase que de modo completamente virtual nestes anos. Pela paciência com minhas questões e coisas "de computeiro" e por ter me mostrado o lado mais descontraído de fazer pesquisa. Pena que o ICB não provê uma infraestrutura adequada para que pessoas com necessidades físicas especiais possam conviver todos os dias, por muitas horas, em um ambiente de pesquisa tão rico... Aqui não é lugar, mas senti falta dessa convivência... De qualquer forma, valeu Miguelito!

Não vou agradecer à minha família e meus amigos, pois eles sabem que sou eternamente grato por eles estarem sempre comigo! Mas não dá pra deixar de falar pros meus pais que eu só cheguei aqui porque me apoiei nos ombros deles: um sapateiro e uma costureira, José e Nitail: simplesmente dois gigantes... Nas pessoas deles agradeço a todos!

Mas, em Cristo, não posso deixar de agradecer a você, Renata, meu Anjo de Amor Eterno, minha Inspiração Nele e à Clarice, minha filha linda, meu amorzinho, minha motivação Nele! Agradecimentos no Amor (*Ágape*) são atemporais, eternos... Amo vocês!

Senhor, eu te agradeço; Jesus, eu te agradeço...

Obrigado, obrigado, obrigado...

Interrogado Jesus pelos fariseus sobre quando viria o Reino de Deus, respondeu-lhes Ele: "O Reino de Deus não vem com visível aparência. Nem dirão: "Eilo aqui !" Ou: "Lá está !" Porque o Reino de Deus está dentro de vós."

Lucas 17, 20-21

"Não vos maravilheis disto, porque vem a hora em que os que se encontram nos túmulos sairão: os que tiverem feito o bem, para a ressurreição da vida; e os que tiverem praticado o mal, para a ressurreição do juízo."

João 5, 28-29

Resumo

Os algoritmos de *clustering* estão entre os mais utilizados na análise de dados de expressão gênica. Por ser uma técnica exploratória, o *clustering* permite aos pesquisadores encontrar padrões de expressão similares entre os diversos tecidos amostrados indicando quais condições amostradas são mais similares. O presente trabalho apresenta duas metodologias para o cálculo da similaridade entre amostras inteiras de dados de expressão gênica utilizando uma fração das seqüências mais expressas (MESs) em cada amostra, que originam duas métricas diferentes. Ambas as métricas são computadas com base na ordenação da expressão das várias seqüências presentes nas amostras, sendo que uma privilegia o compartilhamento entre seqüências mais expressas entre amostras (chamada de *similaridade MESs*) e a outra a manutenção da ordem de expressão das seqüências (chamada de *conservação da ordenação MESs*). O *clustering* hierárquico utilizando as métricas de similaridade propostas foi aplicado em 18 séries de dados de expressão gênica, totalizando 612 amostras, e os resultados foram comparados àqueles produzidos utilizando-se métricas tradicionais como a distância euclidiana e correlações de Pearson e Spearman. No geral, a utilização das duas métricas propostas produziu resultados que superaram as demais: a similaridade MESs apresentou uma acurácia de cerca de 89% e a conservação da ordenação MESs de 80%, enquanto a melhor métrica tradicional para os dados utilizados foi a correlação de Pearson que apresentou acurácia de 76%. Os resultados apresentados indicam que as métricas apresentadas são uma alternativa às métricas tradicionais, além de proverem dados que refletem características biologicamente significativas dos sistemas amostrados.

Palavras-chave: Aprendizagem Não-supervisionada. *Clustering*. Métricas de similaridade. Amostras de dados de expressão gênica.

Abstract

The clustering algorithms are among the most utilized techniques in gene expression data analysis. Being an exploratory technique, clustering allows researchers to find out similar expression patterns among the variety of sampled tissues pointing out which sampled conditions are more similar than others. This work presents two methodologies to compute the similarity among whole samples of gene expression data utilizing only a fraction of the most expressed sequences (MESs) in each sample. Both similarity metrics are computed considering the expression ordering of the various sequences present in the samples. One of them privileges the sharing of the most expressed sequences (named *MESs similarity*). The other privileges the keeping of the expression ordering of the sequences (named *MESs ordering conservation*). Hierarchical clustering utilizing the proposed similarity metrics was applied in 18 gene expression data series summing up 612 samples and the results compared to those produced by some traditional metrics like Euclidian distance, Pearson, and Spearman correlations. Overall, the use of the two proposed metrics outperformed the others: the MESs similarity showed 89% accuracy and the MESs ordering conservation 80% whereas the best traditional metric for the same data was Pearson correlation that yielded 76% accuracy. The results presented here indicate that the proposed metrics are an alternative to the traditional ones. Besides, they produce data that reflect biologically significant features of the sampled systems.

Keywords-chave: Unsupervised Learning. Clustering. Similarity metrics. Data expression samples.

Lista de Figuras

Lista de Tabelas

Lista de Principais Siglas e Abreviaturas Utilizadas

ath	*A. thaliana*
hsa	*H. sapiens*
mmu	*M. musculus*
rno	*R. novergicus*

DNA	*DesoxiriboNucleic Acid*
GEO	*Gene Expression Omnibus*
MESs	*Most Expressed Sequences*
mRNA	*messenger RiboNucleic Acid*
NCBI	*National Center for Biotechnology Information*
SAGE	*Serial Analysis of Gene Expression*

Sumário

1 Introdução

"... nunca esgote o assunto de forma que nada seja deixado para o leitor."
("... never exhaust the subject, so that nothing is left to a reader.")
Montesquieu

O homem é um ser essencialmente explorador. Desde a dimensão "Macro" da expressão do Deus Criador medida em milhões e bilhões de anos-luz até a dimensão infinitesimal onde *quarks* e *gluons* e tantas outras partículas se multiplicam, parecendo expressar o bom humor de Deus para com as peripécias exploratórias de suas criaturas. Para entender o "macro", o homem explora o "micro". Em toda a sua existência o homem explora o ambiente à sua volta, coleta dados, adquire informação, conhecimento, aprende. As atividades exploratórias são essenciais à aprendizagem e ao desenvolvimento do ser humano tanto nas experiências comuns à maioria de nós quanto em ambientes de pesquisa, por exemplo. Como uma criança sabe que um coco é mais parecido com uma bola de futebol do que com uma caixa de papelão? De alguma forma, ela extrai características ou atributos dos três objetos e usa uma *métrica de similaridade* para processar as características e decidir quais são mais ou menos parecidos ou similares. Em muitos ambientes de pesquisa, outros seres humanos procuram entender como uma criança executa tais tarefas... Mais especificamente, quando o conjunto de objetos ou quando o número de características é muito grande, nós, seres humanos temos problemas em conseguir explorar tal conjunto e encontrar padrões de características similares entre os objetos de modo a conseguir agrupá-los e, por conseguinte, extrair informações e conhecimento sobre os mesmos. Este é um dos problemas enfrentados por biólogos e bioinformatas que trabalham, por exemplo, com a análise de dados de expressão gênica produzidos em larga escala. Tais dados possuem milhares de objetos ou milhares de características, dependendo do ponto vista em que são estudados, de modo que é virtualmente impossível ao ser humano extrair informações desse tipo de dados sem o auxílio do computador e técnicas especialmente desenvolvidas para tal. Além disso, nem todas dentre as milhares de características são necessariamente relevantes para determinado processo de análise, como no processo de *clustering*

(agrupamento) onde objetos devem ser agrupados em *clusters* (grupos) de acordo com suas similaridades, o que torna necessária a seleção de características relevantes.

1.1 Escopo do Trabalho

O presente trabalho insere-se no contexto da Bioinformática e tem como objetivo principal a definição e avaliação de duas métricas de similaridade adequadas para serem utilizadas na análise exploratória de dados de expressão gênica produzidos em larga escala. Tais dados são compostos por um conjunto razoavelmente pequeno de amostras (<100, tipicamente), onde cada amostra é composta por milhares de características, que são seqüências (genes, normalmente) presentes ou expressas na mesma. As duas métricas de similaridade apresentadas são adequadas quando um pesquisador necessita decidir quais amostras são “mais parecidas”, por exemplo, no *clustering* (agrupamento) de amostras de tecidos de diferentes tipos de câncer.

1.2 Organização do Texto

O restante do texto está organizado como segue:

Capítulo 2: apresenta uma revisão da literatura sobre conceitos relacionados à expressão gênica e sua quantificação, às métricas de similaridade, bem como ao problema de *clustering*.

Capítulo 3: apresenta as métricas propostas, suas definições e formalização, bem como as idéias e algoritmos para sua computação.

Capítulo 4: apresenta resultados da utilização das métricas apresentadas no *clustering* de dados de expressão gênica, bem como uma discussão sobre os mesmos.

Capítulo 5: apresenta conclusões e indicações de trabalhos futuros ligados às métricas apresentadas.

2 Revisão Bibliográfica

Conheça o que já foi feito;
Assim aprende-se, aperfeiçoa-se
E economiza-se tempo.

2.1 Bioinformática, Expressão Gênica e sua Quantificação

Ainda não há consenso sobre a abrangência ou a definição de Bioinformática. Ela pode tanto ser entendida como sendo o estudo da informática "inerente" aos organismos vivos, desde que as interações internas nos seres vivos podem ser vistas como um grande fluxo de informações que deve ser analisada e compreendida. Pode-se, de outro ponto de vista, definir a Bioinformática como o campo da ciência em que a Matemática, Ciência da Computação e Estatística se unem às várias disciplinas das ciências biológicas, como por exemplo a Biologia Molecular, com o objetivo final de produzir conhecimento que possa ser aplicado, de alguma forma, onde quer que haja necessidade. Ainda que não muito precisa, essa definição provê uma noção da interdisciplinaridade inerente ao campo da Bioinformática. Interdisciplinaridade esta que está presente desde os primeiros trabalhos na área, que foram produzidos quando a Bioinformática ainda não era reconhecida como uma ciência, o que aconteceu somente na última década do século passado (Altman, 1998). Suas raízes se estendem, entretanto, aos anos 60 quando Margaret Dayhoff e colaboradores organizaram seqüências de cadeias de proteínas no primeiro banco de dados de seqüências biológicas, o *PIR – Protein Information Resource* (Barker *et al*., 1998). Dessa forma, desde sua origem, a Bioinformática está intimamente ligada à análise de dados de seqüências protéicas e de ácidos nucléicos (Altman, 1998; Luscombe *et al*., 2001). Mais recentemente, áreas de estudo foram impulsionadas de tal forma pela Bioinformática que trabalhos das diversas áreas passaram a ser tratados como se fossem de uma só área. É o caso, por exemplo, da *Genômica*, da *Transcriptômica* e da *Proteômica*, que, simplificadamente, estudam, respectivamente, a estrutura e o funcionamento do *genoma*: conjunto total dos genes de um organismo, o *transcriptoma*: conjuntos dos transcritos (RNA mensageiro, ribossômico e micro) de um dado tipo celular e o *proteoma*: conjunto das proteínas codificadas pelos genes de um organismo. Dessa forma, um dos objetivos fundamentais da Bioinformática é auxiliar a compreensão da *expressão gênica* em células de diferentes organismos, tecidos e em diferentes condições biológicas e como essa expressão está relacionada. É possível supor que

diferentes organismos e distintos tecidos em organismos multicelulares possuam um padrão de expressão gênica que possa ser usado para classificá-los. Em outras palavras, existiria uma informática no sistema vivo no que diz respeito aos padrões de expressão gênica e contribuições para desvendá-la poderiam assim ser definidas, também, como Bioinformática.

2.1.1 Expressão Gênica

De modo simplificado, um *gene* pode ser definido como uma "unidade" de DNA que contém informação para especificar a síntese de uma cadeia polipeptídica ou de RNA funcional (Lodish *et al.*, 2003). A maioria dos genes possui informação da síntese de moléculas de proteína e as cópias de RNA transcritas a partir desses genes *codificadores de proteínas* são o mRNA (RNA mensageiro) que transporta informação dos genes até os ribossomos, onde a molécula de proteína será sintetizada. Além do mRNA, RNA "funcional" é codificado pelos genes como o RNA transportador funcional (Lodish *et al.*, 2003) e o microRNA (Kim & Nam, 2006). Assim, o DNA prepara o sistema celular, por meio dos produtos de seus genes, para controlar as atividades celulares como o metabolismo, crescimento, desenvolvimento e reprodução, ou seja, determina a função celular bem como a função do tecido formado por um grupo de células cujo conjunto de genes expressos é o mesmo. O DNA interage dinamicamente com proteínas conhecidas como fatores de transcrição, para regular o momento e a quantidade de proteína a ser produzida, embora uma parcela do processo seja controlada dinamicamente pelo próprio sistema protéico, que também reage ao ambiente com fosforilações, degradações programadas, etc. Logo, uma coleção complexa de proteínas no núcleo celular controla a produção de proteínas influenciando quais genes são potencialmente *expressos* ou não em um processo (Hunter, 1995), preparando, assim, o sistema celular característico do tecido ou tipo celular.

Desta forma, pode-se definir a *expressão gênica* como o processo conjunto de *transcrição* de um gene em RNA (RNA mensageiro, tipicamente, ou RNA "funcional"), o processamento deste RNA (*splicing* ou "edição", por exemplo, no caso de mRNA) e sua posterior *tradução* em uma seqüência protéica (Berg, 2002), bem como a atividade pós-traducional que culmina com a execução da função da proteína — para genes codificadores de proteínas. Em outras palavras, é o processo pelo qual a informação armazenada em um gene é transformada em estruturas que operam na célula. Por exemplo, quando uma enzima é sintetizada e exerce sua atividade, o gene responsável por iniciar aquela síntese é dito ser *expresso*, pois o caráter causado por ele é expresso e tende a se manifestar. Os genes expressos incluem aqueles que são transcritos e codificam proteínas e

também aqueles que são transcritos e não codificam proteínas, por exemplo, RNA ribossômico, RNA transportador ou os miRNA, que são micro RNAs que "silenciam" genes (Kim & Nam, 2006). Dada a existência de RNA "funcional", ainda que técnicas de caracterização direta do proteoma tecidual já tenham alcançado o nível técnico satisfatório para o estudo do sistema celular (Menck & van Sluys, 2004) , as principais técnicas de observação e "medição" da expressão gênica o fazem no *nível de transcrição*, onde o controle ou regulação da expressão gênica ocorre de maneira proeminente (Berg, 2002; Lodish *et al*., 2003). A aproximação da análise utilizando dados de expressão gênica coletados no nível de transcrição parece bastante apropriada (Brazma & Vilo, 2000) e tem sido largamente utilizada. Assim, o termo *expressão gênica*, embora não corresponda (principalmente em eucariotos, dada à miríade de mecanismos pós-transcricionais e pós-traducionais existente) à transcrição do gene, em última análise depende grandemente do funcionamento do conjunto de genes daquele tipo celular ou do organismo, ou seja, de uma espécie de assinatura tecidual.

2.1.2 Quantificação da Expressão Gênica

Compreender o mecanismo de regulação da expressão de um determinado gene é importante quando se pode atribuir à expressão do mesmo um papel importante na fisiologia celular, no desenvolvimento do organismo ou em quaisquer outras respostas biológicas associadas à formação de seu produto gênico. Por outro lado, é plausível supor que o sistema celular seja reflexo do conjunto da expressão de todos os genes, já que este reflete, também, a interação com o ambiente extracelular e com mediadores químicos do organismo. Assim, para estudar a expressão gênica que caracterize tipos celulares, tecidos e órgãos de maneira eficiente é necessário lançar mão de meios de se observar a expressão de centenas ou milhares de genes ao mesmo tempo. Existem, atualmente, quatro tecnologias principais que possibilitam experimentos desse tipo e produzem dados coletados no nível de transcrição (incluindo-se aqui a meia-vida das moléculas), ou seja, em todas elas é estimada ou medida a quantidade de transcritos em uma população de células. São elas (i) SAGE, (*Serial Analysis of Gene Expression*, análise serial da expressão gênica) (Velculescu, 1995), (ii) os *chips* de DNA (*GeneChips*® da Affymetrix) (Lockhart *et al.*, 1996), (iii) os *microarrays* (Schena *et al.*, 1995) e (iv) as ESTs (*Expressed Sequence Tags*, Adams *et al*., 1993; Adams *et al*., 1999). SAGE é considerada a técnica mais precisa e confiável para medição da expressão gênica (Chu, 2003). No outro extremo encontram-se os *microarrays*, que apresentam relativo baixo custo e uma conspícua sensibilidade (Lu *et al.*, 2004). Os *GeneChips*® também são *microarrays*, mas são construídos de maneira completamente diferente (Lockhart *et al.*, 1996). Existem trabalhos indicando que a

correlação entre dados quantitativos produzidos por *microarrays* e *GeneChips*® é baixa (Kuo *et al.*, 2002) e a comparação direta entre dados de experimentos realizados utilizando-se tecnologias diferentes não é confiável atualmente. Entretanto, existem estudos contrariando tais resultados e que buscam uma melhor qualidade e padronização dos experimentos com *microarrays* (MAQC Consortium, 2006). Os dados de expressão gênica produzidos utilizando-se os *GeneChips*® mostram alta correlação com os produzidos por SAGE, segundo Ishii *et al*. (2000). Entretanto, tal correlação pode não ser generalizável (Lu *et al*., 2004). Além dessas três tecnologias, pode-se conseguir uma indicação do nível de expressão por meio da contagem de ESTs (Franco *et al*., 1997; Stekel *et al*., 2000) em bibliotecas de cDNA. O banco de dados Unigene no NCBI fornece dados de amostragem de ESTs por agrupamentos Unigene para vários organismos, retratando diferentes tipos celulares e estágios de desenvolvimento (Pontius, Wagner & Schuler, 2003). Ainda que a metodologia apresentada neste trabalho seja aplicável independentemente da tecnologia utilizada para geração dos dados de expressão, os *GeneChips*® foram escolhidos por serem "plataformas fechadas" nas quais os mesmos equipamentos e protocolos são utilizados na execução dos experimentos que geram dados de expressão gênica. Além disso, o maior banco de dados de expressão gênica disponível publicamente, o GEO (*Gene Expression Omnibus*), do NCBI (*National Center for Biotechnology Information*), nos Estados Unidos da América (Barrett *et al*., 2005), armazena tais dados de modo uniforme, por serem de plataformas de um mesmo tipo, o que facilita seu processamento em lote.

A Figura 2-1 apresenta as várias etapas de um experimento típico utilizando *GeneChips*®. Inicialmente, o RNA total (apenas mRNA pode ser usado) é extraído de um *pool* de células a serem estudadas. Esse RNA já foi processado para remoção de *introns*, que são seqüências de nucleotídeos não utilizadas na síntese da proteína em questão, restando apenas os *exons*, ou seja, o RNA foi editado ("*spliced*") e foi adicionada cauda de poli-A. A partir do RNA é produzido cDNA, por transcrição reversa, a qual pode ser iniciada por um oligo-dT contínuo a um promotor viral ou este cDNA pode ser clonado em vetores no sentido da extremidade 3' (*downstream*) do referido sítio promotor. Assim, o cDNA será novamente transcrito em RNA, produzindo o chamado cRNA. Esse RNA é marcado com biotina por incorporação de uridina modificada com um grupamento contendo biotina e é fragmentado randomicamente em pedaços de 30 a 400 bases com uma biotina pelo menos em cada fragmento. Um *GeneChip*® possui milhares de *conjuntos de sondas* (*probe sets*). Cada conjunto é composto por oligonucleotídeos de 25 bases (em geral) e cuidados são tomados na escolha da seqüência das bases no tocante a aspectos que podem influenciar nos resultados finais, causando desvios ("*bias*") nos resultados de todos os experimentos, como a eficiência de hibridização

que é função, entre outras, da quantidade de Citosina-Guanina, por exemplo (Mei *et al.*, 2003). Cada oligonucleotídeo representa uma seqüência *única* (ou seja, que não é repetida) do genoma do organismo de interesse, sendo que essa seqüência representa um gene. Assim sendo, o cRNA deverá, idealmente, *hibridizar* (ou *hibridar*)com as sondas que possuírem seqüências complementares à sua.

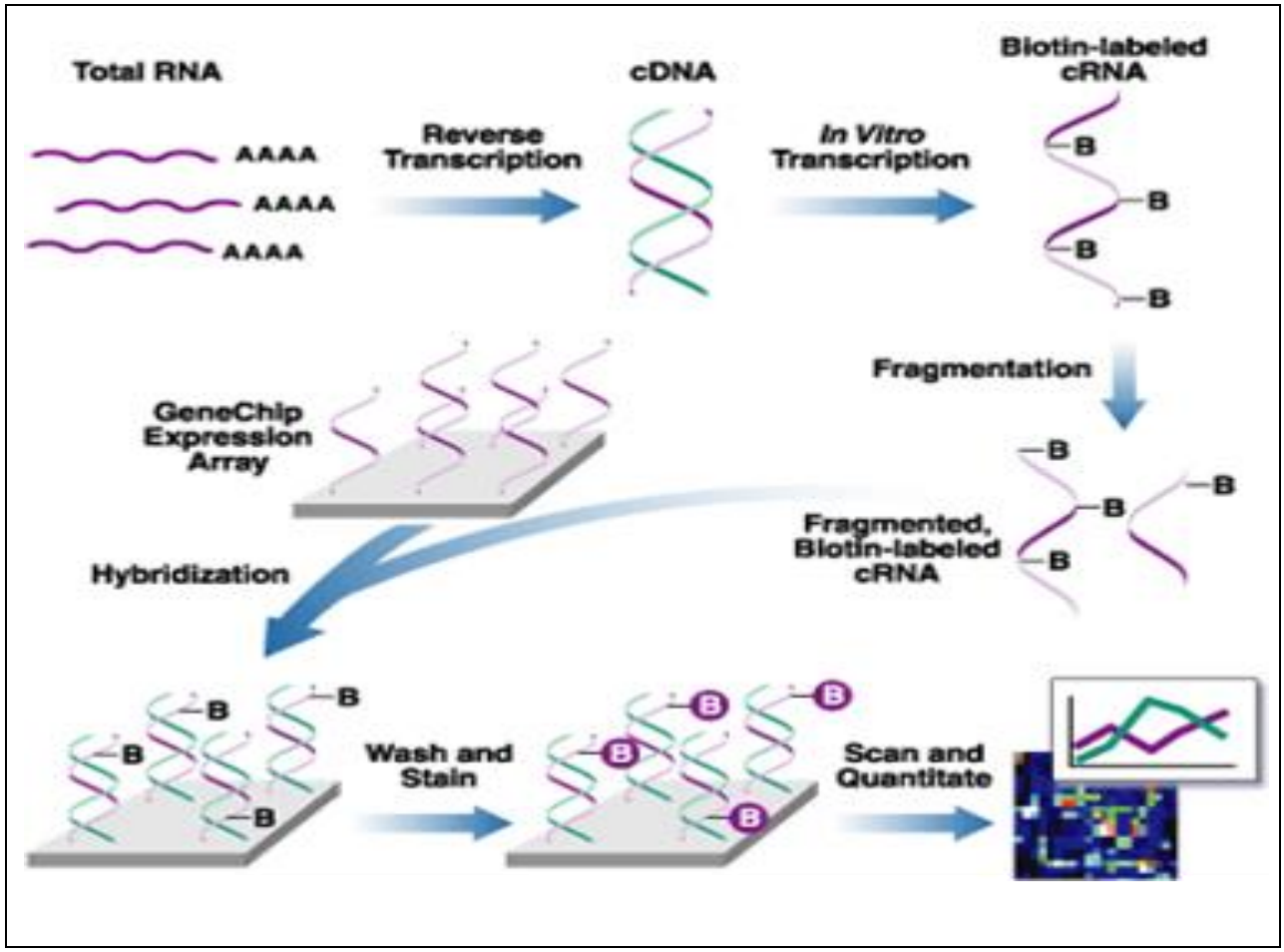


Figura 2-1: **um experimento típico utilizando os GeneChips® da Affymetrix** (disponível em www.affymetrix.com). O cRNA foi marcado com biotina neste exemplo e hibridou com as sondas presentes no *chip*.

Geralmente mais de uma sonda é desenhada para representar regiões consideradas únicas de genes conhecidos, o que é o caso dos GeneChips® (Affymetrix, 2002). Um ponto crucial na tecnologia é que para cada sonda onde deve ocorrer hibridação correta do cRNA (Figura 2-1), existe também uma sonda que difere apenas na base central. Caso haja hibridação correta é dito que ocorreu um *perfect match* (PM), caso haja hibridização do cRNA com a sonda que possui base central errada é dito que ocorreu um *mismatch* (MM). Após a hibridação, o *GeneChip*® é lavado para remover o cRNA que não hibridou e o mesmo é escaneado (escâner a *laser*) para a geração dos níveis de expressão. O nível de expressão de cada seqüência é calculado com base na média da diferença entre

o número de PM e o número de MM (Affymetrix, 2002), nos vários conjuntos de sondas representativos de um gene (já que mais de um deles é manufaturado em cada *GeneChip*® em locais diferentes), o que grandemente reduz a influência do *background* e da hibridação cruzada (Lipshutz *et al.*, 1999). O *software* que gera os níveis de expressão gera também, para cada seqüência representada no *GeneChip*®, um valor de "chamada" (*call*) informando se a seqüência está confiavelmente presente (*call* = 'P'), ausente (*call* = 'A') ou se é marginal (*call* = 'M') — nem presente nem ausente de modo confiável. Assim, uma estimativa quantitativa da expressão é gerada (nível de expressão), bem como um dado qualitativo (chamada) para cada gene representado em um *GeneChip*® utilizado em cada experimento.

Um experimento que envolve análise de expressão gênica utilizando *microarrays* ou *GeneChips*® usa, tipicamente, várias condições experimentais de interesse. Para cada condição de interesse são gerados dados com pelo menos um *chip*, sendo que é comum haver réplicas de uma mesma condição para aumentar a confiabilidade (Su *et al.,* 2002; Bergmann *et al.*, 2004). As condições de interesse são representadas, em muitos casos, por instantes de tempo, formando uma *série temporal* normalmente irregular, ou seja, com instantes de tempo com intervalos irregulares entre os mesmos, onde um sistema (organismo ou tecido) é investigado (Vázquez-Chona *et al*., 2004). Além disso, são comuns experimentos onde *tecidos diferentes* são investigados no tocante à expressão gênica, seja em mesmas condições experimentais para todos os tecidos (Ge *et al*., 2005) ou não (Rock *et al*., 2005). A Tabela 2-1 mostra um pequeno extrato de dados de expressão gênica de oito amostras de tecidos e nove genes, que são um subconjunto bastante pequeno dos dados publicados por Ge *et al*. (2005). Independentemente do tipo de experimento realizado, normalmente o pesquisador utiliza um método de *clustering* (agrupamento), que é o "modo" de aprendizagem não-supervisionada mais comumente utilizado em Inteligência Artificial, com o objetivo de encontrar padrões similares nos dados.

Tabela 2-1: Exemplo de dados de expressão gênica.

Gene Symbol	**Heart**	**Spleen**	**Ovary**	**Adrenal Gland**	**Kidney**	**Liver**	**Fetal Brain**	**Fetal Liver**
UBB	2873.4	2681.6	3017.6	3020.1	2666.1	2614.6	3490.6	2745.4
EEF1A1	3135.5	3776.2	4436.5	4499.8	3682	4520.3	5080.7	5454.5
ACTB	2040.8	1728.2	2633.5	2335.6	1752.3	1464.6	2760.9	1895.8
HSPCA	1321.1	774.3	1623.4	799.2	875.4	737.5	1014.7	1697.4
GAPD	5312.4	3221.9	3202.6	4548.6	4327	3758.5	5273.3	4762.7
RPLP0	2393.5	2799	3925.4	4106.3	3008.9	2288.8	4169.5	5137.8

TUBA1	1681.1	1588	1397.4	1321.9	1464.4	706.6	4251	2694.5
RPL10	592.3	832.3	1128.9	922.8	487.1	615.1	956.8	901.7
TUBB	924	336.8	750.5	453.5	493.1	426.4	4336.8	1066.1

2.2 *Clustering* e Métricas de Similaridade

O termo "Inteligência Artificial" foi proposto por John McCarthy, em 1956, durante um *workshop* no Dartmouth College, ainda que o primeiro trabalho considerado seja o neurônio artificial de McCulloch e Pitts, de 1943 (Russell & Norvig, 2003). Não há consenso em torno de uma definição de Inteligência Artificial, visto que não há consenso em torno de uma definição operacional ou funcional de "inteligência" entre os pesquisadores da área. Entretanto, há consenso sobre as características que um sistema "inteligente" deve ter. Entre elas estão: o processamento de linguagem natural; o raciocínio; o comportamento emergente; a representação de conhecimento e a aprendizagem (ou o aprendizado) (Russell & Norvig, 2003; Luger, 2004), sendo esta última de especial interesse para o presente trabalho.

Em Inteligência Artificial, a *aprendizagem* pode ser definida como sendo as mudanças em um sistema[1] que o capacitam a executar suas tarefas de modo mais eficiente e efetivo da próxima vez, sobre o mesmo conjunto de exemplos disponíveis. É importante observar que um problema de aprendizagem pode ser *irrealizável* caso o conjunto de exemplos utilizado não represente a função ou conceito que o sistema deve aprender (Russell & Norvig, 2003). Considerando que o sistema deve trabalhar de modo *adaptativo*, ou seja, deve melhorar ou aperfeiçoar seu desempenho a cada vez que ele trabalha os dados disponíveis, existem três formas de realimentação do sistema de modo que seu processo de aprendizagem seja orientado e que dão nome ao processo: *aprendizagem por reforço (reinforcement learning)*, *aprendizagem supervisionada (supervised learning)* e *aprendizagem não-supervisionada (unsupervised learning)*.

A *aprendizagem por reforço* é caracterizada pela existência de um reforço positivo, que o sistema aprendiz recebe quando executa uma ação adequada e um reforço negativo, recebido pelo sistema

[1] O termo "sistema" tanto pode se referir a um organismo em sua totalidade quanto a um conjunto de programas computacionais.

quando esse executa uma ação não adequada. A *aprendizagem supervisionada* é a mais bem compreendida pela comunidade de Aprendizagem de Máquina (*Machine Learning*) (Russell & Norvig, 2003). Nela, o sistema é exposto a um conjunto de exemplos, denominado conjunto de treinamento, em que cada exemplo é composto por um par [*entrada*, *saída*]. A *entrada*[2] de um exemplo é, tipicamente, um conjunto de atributos ou características que discriminam o objeto (ou conceito) sendo representado e a *saída* é a próprio objeto (ou conceito). Assim sendo, um sistema em aprendizagem supervisionada pode avaliar seu desempenho visto que ele sabe qual saída é esperada para uma determinada entrada de todos os exemplos disponíveis em seu treinamento. O objetivo é que o sistema possa ser capaz de aprender uma regra ou função que mapeie entradas em saídas corretas de maneira satisfatória para exemplos ainda não-vistos, ou seja, que não fazem parte do conjunto de treinamento. No caso do conceito "caixa", por exemplo, podem ser apresentadas ao sistema várias fotografias (ou representações) de caixas e várias fotografias (ou representações) de objetos parecidos, mas que não são caixas (alguns edifícios ou armários retangulares, por exemplo). O crucial na aprendizagem supervisionada é que, para cada exemplo, é informado ao sistema se o exemplo representa o conceito a ser aprendido ou não. Já a *aprendizagem não-supervisionada* é caracterizada pela não existência ou pela não utilização da *saída* dos exemplos do conjunto de dados, ou seja, não há realimentação. Dessa forma, o sistema deve encontrar "similaridades" ou diferenças entre os vários exemplos do conjunto, utilizando somente a *entrada* dos exemplos, com o objetivo de distinguir ou aprender diferentes conceitos representados nos dados. O tipo mais comum de aprendizagem não-supervisionada é o *clustering* (agrupamento), no qual o sistema deve agrupar os dados em *clusters* (grupos) de modo que cada exemplo em um mesmo *cluster* seja *mais similar* a cada um dos demais exemplos do mesmo *cluster* que a qualquer exemplo de um *cluster* diferente. Assim, *clustering pode ser definido como o processo pelo qual um conjunto de objetos ou conceitos é classificado ou agrupado de acordo com algum critério ou métrica de similaridade entre eles, de tal maneira que objetos (conceitos) em uma mesma classe ou grupo (cluster) sejam mais similares entre si que com qualquer outro de uma classe ou grupo diferente*. Por conseguinte, todos os métodos de *clustering* devem, de alguma maneira, extrair informações dos dados para encontrarem características comuns aos exemplos que compõem o conjunto de dados e agrupá-los de modo que os "mais parecidos", de acordo com o critério de similaridade escolhido, estejam em um mesmo grupo. Além disso, é comum que atributos ou características dos exemplos sejam irrelevantes para o processo de agrupamento, o que torna necessário um passo de seleção de características

[2] A *entrada* pode ser, simplesmente, os parâmetros de entrada de uma função a ser aprendida e que está representada nos dados.

(*feature selection*) executado antes do processo (Jain, Murty & Flynn, 1999), principalmente quando o número de tais atributos é da ordem de milhares, como é o caso no *clustering* de amostras de dados de expressão gênica.

2.2.1 Clustering de Amostras de Dados de Expressão Gênica

O *clustering* de dados tem sido largamente utilizado na análise de dados de expressão gênica produzidos em larga escala por tecnologias como *microarrays*, incluindo *genechips* (Eisen *et al.*, 1998; Jiang *et al*., 2004; Sharan *et al.*, 2002) e SAGE (Chu, 2003). Entre os métodos mais utilizados destacam-se o *K-means* (McQueen, 1967), redes neurais SOM (*Self-Organizing Map*) (Kohonen, 1984), o *clustering* hierárquico (Eisen *et al.*, 1998) e os métodos baseados em grafos como o CLICK (Shamir & Sharan, 2000) e o CAST (Ben-Dor, Shamir & Yakhini, 1999). Uma característica marcante desse tipo de dados é que é significativo aplicar o *clustering* tanto para agrupar genes quanto para agrupar amostras inteiras (Alon *et al*., 1999; Jiang *et al*., 2004), sendo os objetos a serem agrupados genes ou amostras, respectivamente.

A premissa por trás do *clustering* de dados de expressão gênica em que os objetos a serem agrupados são genes é que se dois ou mais deles possuem *padrões de expressão* similares, então eles podem ser co-regulados ou podem participar de um mesmo processo ou função biológica. Um padrão de expressão de um gene é um conjunto de valores de medições da expressão do gene colhidas em várias amostras diferentes. De forma similar, *o padrão de expressão de uma amostra é o conjunto formado por cada valor de medição de cada gene nela presente*. Assim, sendo, o *clustering* de amostras de dados de expressão gênica possui diferenças importantes em relação ao *clustering* de genes. Entre elas, destacam-se (Jiang *et al*., 2004):

1. O número de amostras é, em geral, pequeno (N < 100);
2. O número de características (genes ou seqüências expressas) de cada amostra é alto (da ordem de milhares).
3. A maioria dos genes presentes nas amostras pode não necessariamente ser de interesse para o processo de *clustering* (Golub *et al.* 1999), ou seja, não são *informativos*.

Aliados à esparsidade[3] dos dados, tais características trazem as seguintes dificuldades (Xing & Karp, 2001):

1. Existem muitas maneiras bem fundadas, estatisticamente significas de se agrupar amostras. Nem todos os algoritmos de *clustering* capturarão as partições em grupos que correspondem aos fenótipos de interesse (com câncer ou sem câncer, por exemplo) porque o mesmo conjunto de dados pode apresentar variabilidade quanto à idade, sexo ou outro tipo de doença ou síndrome, por exemplo, que também podem servir de critério para o *clustering*.
2. Normalmente os *microarrays* não são específicos para um fenótipo, o que faz com que muitos genes não interessantes ao estudo estejam interferindo como características irrelevantes à determinação ou distinção dos fenótipos;
3. O objetivo do *clustering* não é apenas agrupar amostras similares, mas também servir como uma referência para que novas amostras sejam corretamente classificadas no futuro. Este processo de classificação, chamado de *generalização* pode ser negativamente influenciado de o número de características utilizado é "grande".

O terceiro ponto, em ambas as listas acima, traz à tona a questão da seleção de características a serem utilizadas no processo de *clustering*: quais genes são *informativos*, ou seja, permitem diferenciar uma amostra outra?

2.2.2 Seleção de Características

Jain, Murty e Flynn (1999) definem a *seleção de características* ou de atributos como o processo de seleção do subconjunto das características originais mais efetivas para serem utilizadas no *clustering*, em contraste com a *extração de características* que é a geração de novas características a partir das originais por meio de uma ou mais *transformações*, sendo que estas novas características serão utilizadas no *clustering*. No contexto do agrupamento de amostras de dados de expressão gênica, tal seleção torna-se importante dado o grande número de características (genes) presentes nas mesmas.

Golub *et al*. (1999) relatam que apenas uma pequena fração (< 5%) dos genes presentes em uma amostra é responsável por sua caracterização, mas não no processo de *clustering* e sim em *classificação* (aprendizagem supervisionada) de amostras, onde a classe de uma amostra deve ser

[3] Em Português, o termo que mais se aproxima de uma tradução de *sparsity* seria "espargimento". Entretanto, é comum na comunidade de Computação utilizar-se o neologismo "esparsidade" que foi mantido aqui.

escolhida dentre algumas classes já conhecidas *a priori*. Xu e Zhang (2005) propõem a utilização de "genes virtuais" (*virtual genes*) que são combinações lineares de subconjuntos de dois ou três "genes reais" presentes em dados de *microarrays*, apresentando resultados da utilização de 20 e 50 genes virtuais. Entretanto, a maioria dos métodos utilizados para seleção de genes é baseada em análise estatística dos dados (Jiang *et al*., 2004), sendo que suposições sobre a distribuição dos dados devem ser feitas, o que limita a aplicabilidade de tais métodos. Por exemplo, no trabalho de Chris Ding (2002), em que é feita a suposição de que genes informativos são aqueles que exibem maior variância, o que, em geral, não é válido para dados de *microarrays* e *genechips* de acordo com Yeung & Ruzzo (2000).

2.2.3 Métricas de Similaridade

Como exposto anteriormente, a similaridade é fundamental na definição de *clustering* e, por conseguinte, de um *cluster*. Assim sendo, cada par de objetos ou conceitos de um *cluster* deve ter, entre si, uma medida de sua similaridade que foi responsável por seu agrupamento. A métrica de similaridade[4] mais comumente utilizada é a distância euclidiana, que é um caso especial da métrica de Minkowski (Jain, Murty & Flynn, 1999), assim como o é a distância Manhattan. Tais métricas, juntamente com o produto escalar de dois vetores de características normalizados, a correlação de Pearson e a correlação de Spearman são as de uso mais difundido entre a comunidade de Bioinformática (Eisen *et al*., 1998), sendo que os algoritmos de *clustering* mais utilizados computam similaridade por meio da distância euclidiana ou pela correlação de Pearson (Jiang *et al*., 2004). Entretanto, cada métrica é mais ou menos adequada para ser utilizada em um processo de agrupamento, dependendo de características próprias ou do conjunto de dados sendo analisado.

As métricas de Minkowski, por exemplo, devem ser utilizadas em dados normalizados de modo a evitar que atributos de alta magnitude dominem os demais (Jain, Murty & Flynn, 1999). Já a correlação de Pearson parece não ser robusta em relação a *outliers*, podendo informar alto grau de similaridade a um par de objetos dissimilares. Além disso, ela assume uma distribuição aproximadamente gaussiana dos pontos e pode não ser robusta em distribuições não-gaussianas (Jiang *et al*., 2004). A correlação de Spearman contorna os problemas da correlação de Pearson substituindo os valores dos atributos, por seu *ranking* (é uma medida não-paramétrica). Ela não requer a suposição

[4] Apesar de algumas métricas indicarem distâncias ou dissimilaridade entre objetos, a conversão dessa em similaridade é direta.

da distribuição gaussiana e é mais robusta com relação a *outliers*. Entretanto, alguns resultados experimentais não demonstraram superioridade de resultados da utilização da correlação de Spearman sobre a correlação de Pearson (Jiang *et al*., 2004).

2.2.4 Alguns Métodos de *Clustering* Comumente Utilizados

Entre os métodos de *clustering* comumente utilizados na análise de dados de expressão gênica estão o *K-Means* (McQueen, 1967; D'haeseleer, 2005) e o *clustering hierárquico*, sendo que esse se destaca por permitir uma visualização bastante informativa dos *clusters* encontrados nos dados (Figura 2-2), sendo o mais freqüentemente utilizado por biólogos em geral.

O *K-Means* funciona da seguinte maneira (Jain, Murty & Flynn, 1999):

1. Escolha *k* centros que coincidem com *k* pontos de dados, escolhidos randômicamente, do conjunto sendo agrupado;
2. Atribua cada ponto do conjunto de dados ao *cluster* correspondente ao centro mais próximo dele;
3. Recompute os centros de cada *cluster* utilizando os pontos de dados incluídos em cada *cluster*;
4. Se o critério de convergência não foi alcançado então retorne ao Passo 2. Os critérios de convergência mais comuns são: nenhum ponto de dados foi movido de um *cluster* para outro ou o erro quadrático médio não diminuiu de modo significativo.

O erro quadrático médio, *E*, é computado da seguinte forma:

$$E = \sum_{i=1}^{k} \sum_{O \in C_i} \left| O - \mu_i \right|^2 \qquad (1)$$

onde,

O é ponto de dados pertencente ao *cluster* C_i

μ_i é o centro (centróide) do *cluster* C_i

O *clustering hierárquico* permite a expressão dos *clusters* formados como um "dendrograma", sendo muito adequados à análise manual com cunho biológico por serem bastante parecidos com as árvores filogenéticas (Eisen *et al*., 1998) (Figura 2-2). Duas abordagens são comuns: os *clusters*

são formados por particionamento do conjunto de dados repetidamente ou por aglomeração, em que elementos são unidos para formar os *clusters*. Na segunda abordagem, conhecida como *clustering* hierárquico "aglomerativo" (Sharan *et al*., 2002), inicialmente cada elemento forma um *singleton* (*cluster* de um único elemento) e são intercalados ou fundidos (*merged*) sucessivamente até que haja um único *cluster*. Sendo *N* elementos a serem agrupados, são formados $N \times N$ pares. A cada iteração, o par *(i, j)*, de maior similaridade, é fundido em um único *cluster* e tratado como um único elemento daí em diante. O número de *clusters* é diminuído para $N - 1$ e a similaridade entre o novo *cluster (i, j)* e os outros elementos, *k*, é atualizada como segue., originando, usualmente, três esquemas diferentes, de acordo com a definição da nova similaridade:

1. *Single-linkage: s[k, (i, j)] = max(s(k, i), s(k, j)).*
2. *Complete-linkage: s[k, (i, j)] = minax(s(k, i), s(k, j))*
3. *Average-linkage:* $s[k, (i, j)] = (n_i s(k, i) + n_j s(k, j)) / (n_i + n_j)$

onde n_i e n_j são o número de elementos nos *clusters i* e *j,* respectivamente.

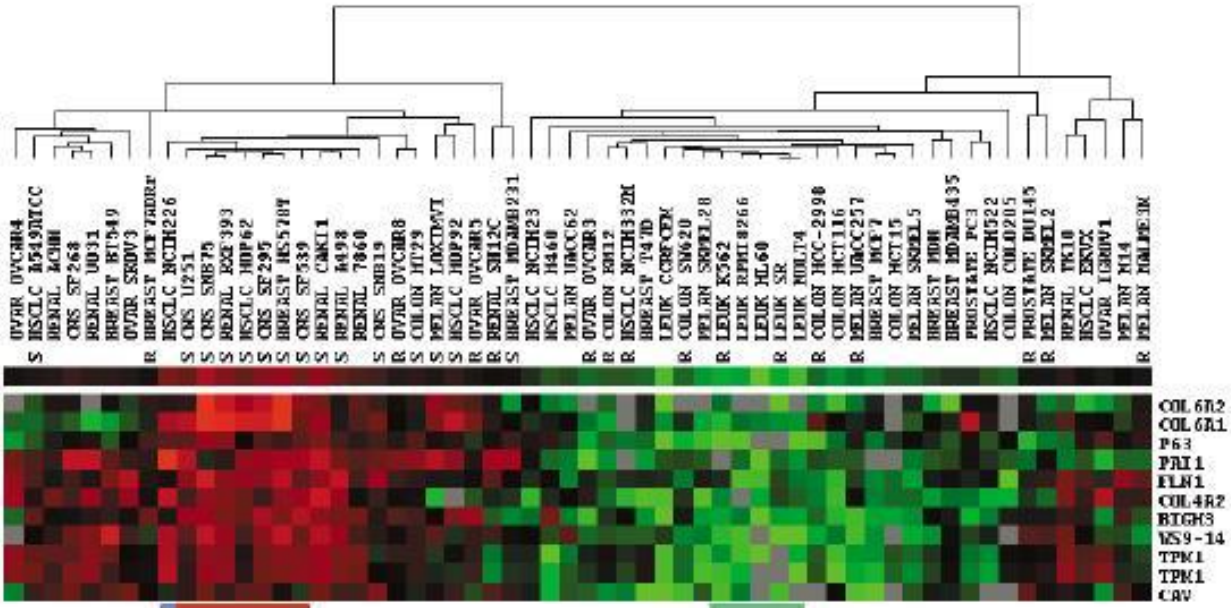


Figura 2-2: **Exemplo de uma figura produzida pelo programa *TreeView***. O exemplo mostra o resultado de um *clustering* hierárquico, onde os objetos a serem agrupados são amostras representadas pelos níveis de expressão de 11 genes gerados por experimentos com *microarrays* (Staunton *et al.* 2001) executado pelo programa *Cluster* (http://rana.lbl.gov/EisenSoftware.htm). Os dois *clusters* principais podem ser visualizados no dendrograma e na figura (um *cluster* em vermelho e outro em verde). Os identificadores das amostras aparecem acima dos níveis de expressão, formando a "base" do dendrograma e os identificadores dos genes aparecem à direita. Figura apresentada por Slonim (2002).

3 Metodologia

E, depois disso, designou o Senhor ainda outros setenta e mandou-os adiante da sua face, de dois em dois, a todas as cidades e lugares aonde ele havia de ir. E dizia-lhes: Grande é, em verdade, a seara, mas os obreiros são poucos; rogai, pois, ao Senhor da seara que envie obreiros para a sua seara.
Lucas 10.1-2

O presente trabalho propõe o desenvolvimento de uma metodologia de *clustering* baseada na premissa de que seria mais informativo agrupar tipos celulares com base nos genes mais expressos. Embora os genes menos expressos possam estar intimamente relacionados com a regulação da expressão gênica e possivelmente sejam muito diferentes dentre tipos celulares distintos, as características mais compartilhadas poderiam representar uma melhor métrica para classificação. Adicionalmente, similaridade de expressão gênica também poderia ser refletida na manutenção da ordem de intensidade de expressão dentre amostras parecidas. Os detalhes da metodologia desenvolvida estão apresentados a seguir.

3.1 Por que MESs?

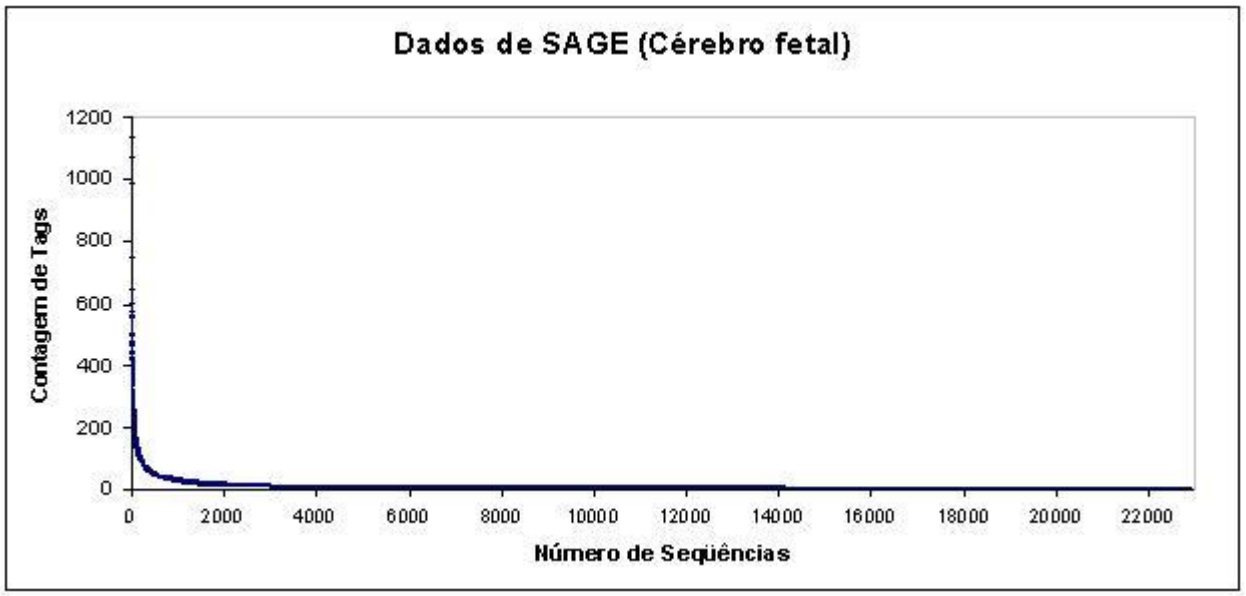
A utilização de métricas de similaridade já tradicionalmente aplicadas no processo de *clustering* possui algumas limitações como exposto no capítulo anterior. Dadas algumas características dos dados de expressão gênica gerados em larga escala, é oportuna a definição de novas métricas de similaridade que considerem a natureza específica de tais tipos de dados.

Ainda que não haja consenso entre pesquisadores, é comumente aceito que quanto mais cópias de RNA mensageiro estiverem presentes em um tipo de célula, mais importante para a sua funcionalidade é o gene correspondente. Assim sendo, o conjunto das seqüências mais expressas (MESs, do Inglês "*Most Expressed Sequences*") em uma célula ou em um conjunto de células pode ser utilizado para caracterizá-la, no sentido de que em diferentes células diferentes seqüências são mais expressas ou, pelo menos, o mesmo conjunto de seqüências é "mais expresso", ainda que o número de cópias dessas varie dentro deste conjunto de um tipo celular para outro. Evidentemente, tal

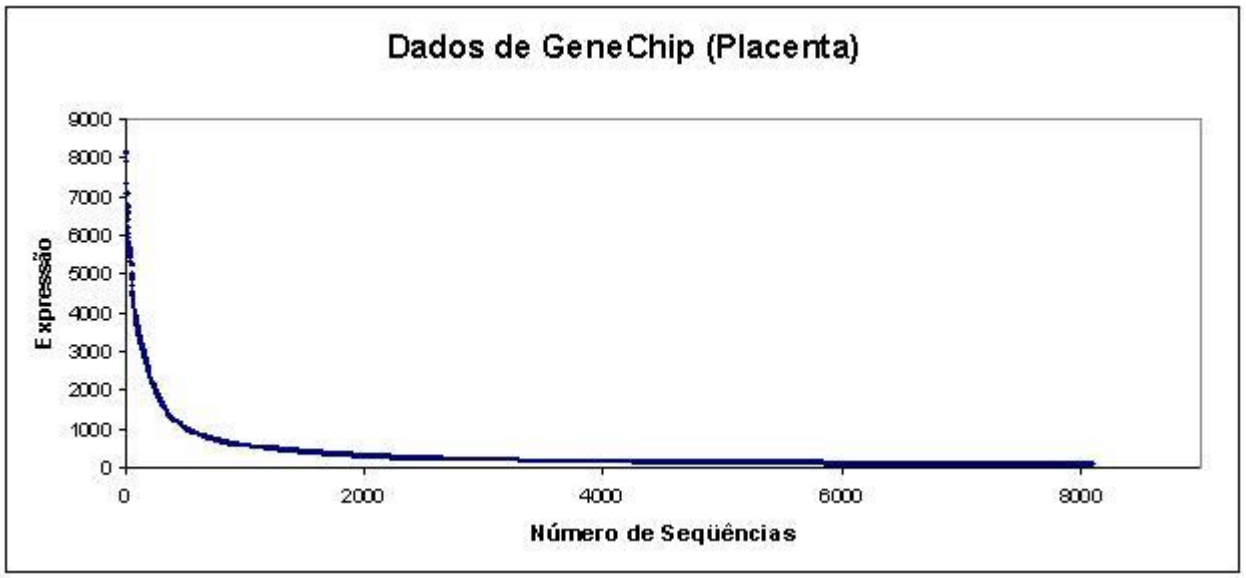
afirmativa não exclui as seqüências menos expressas (LESs, do Inglês "*Least Expressed Sequences*") como sendo importantes ou relevantes para caracterizar um determinado tipo de célula, visto que um conjunto de genes que só é expresso em um determinado tipo celular, mesmo em "pequenas quantidades", certamente o diferenciam. Entretanto, as principais tecnologias de geração de dados de expressão gênica em larga escala (*microarrays*, *genechips*, SAGE e ESTs) produzem dados nos quais a confiabilidade das medições obtidas para as MESs é alta enquanto que a confiabilidade das LESs é baixa (nula, no caso de SAGE e de ESTs). Dessa forma, a utilização das MESs como base para a seleção de características relevantes ao processo de *clustering*, bem como na definição de uma métrica de similaridade adequada aos dados produzidos pelas tecnologias, privilegia aspectos biológicos comumente aceitos. Todavia, é necessário definir quando uma seqüência é considerada "mais expressa" ou não. Isto é, ordenando-se as seqüências presentes (expressas) em uma determinada amostra de dados de expressão gênica, em ordem decrescente de suas expressões, deve-se determinar o número de seqüências que são MESs.

3.2 Definição do Limiar das MESs

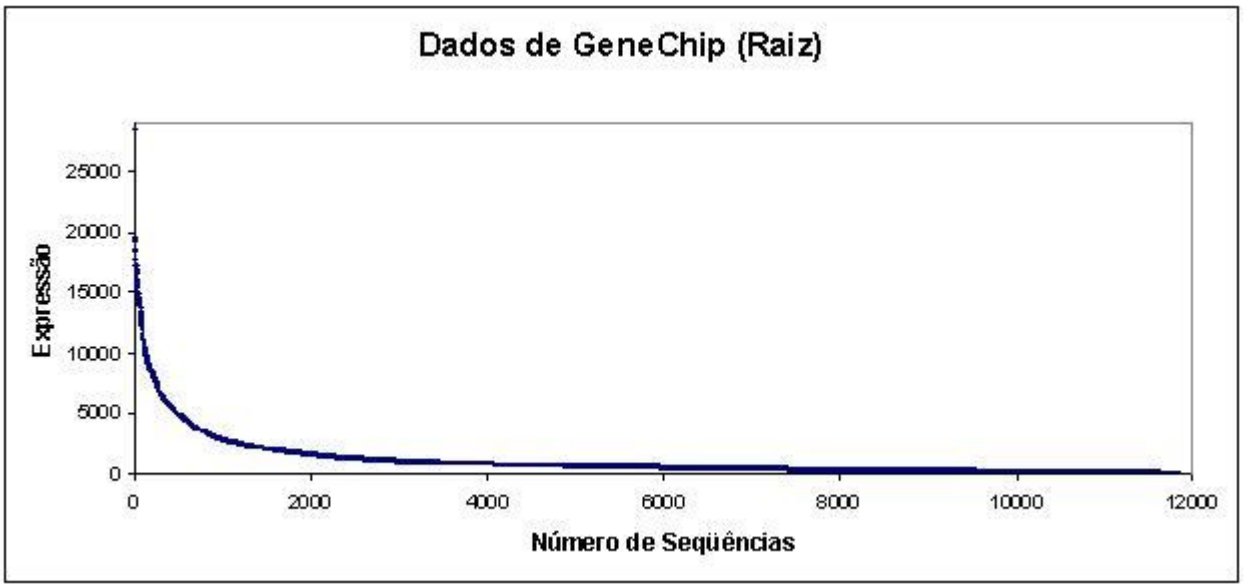
A partir da ordenação de uma amostra de dados de expressão gênica em ordem decrescente dos valores de expressão de suas seqüências, pode-se observar que dados de diferentes organismos e diferentes tecnologias apresentam uma *curva de expressão* bastante similar, seguindo certo padrão: a relação MESs/LESs, mesmo que (aparentemente) não claramente determinável, parece ser baixa, ou seja, o número de genes altamente expressos parece ser bastante reduzido em relação ao número de genes pouco expressos, como ilustrado pelas curvas de expressão exibidas na Figura 3-1.


Dados de SAGE (Cérebro fetal)
Contagem de Tags
1200
1000
800
600
400
200
0
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
Número de Seqüências

(a)


Dados de GeneChip (Placenta)
Expressão
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
2000
4000
6000
8000
Número de Seqüências

(b)



(c)

Figura 3-1: **Curvas de expressão produzidas por dados de duas tecnologias diferentes para três tipos diferentes de tecidos em dois organismos distintos**. (a) Dados de cérebro de feto humano produzidos por SAGE. (b) Dados de placenta humana produzidos por GeneChips. (c) Dados de raiz de *A. thaliana* produzidos por GeneChips.

Dada tal característica das curvas de expressão, a questão a ser respondida é: qual limiar define quais seqüências são MESs e quais não são? *Arbitrariamente, define-se o número de MESs em uma amostra de dados de expressão gênica como o número de seqüências correspondente à ordenada do único ponto da curva de expressão da amostra que é tangente à reta de inclinação igual a -1.* Tal ponto corresponde ao ponto na curva onde a taxa de variação (decaimento) da expressão tende a diminuir e a aproximar-se mais lentamente de zero à medida que o número de seqüências aumenta. Como a curva de expressão não é contínua, a determinação de tal ponto de tangência torna-se inviável. Logo, uma aproximação faz-se necessária. Desta forma, define-se o número de MESs em uma amostra como *o número de seqüências correspondente à média das ordenadas dos dois pontos que definem a reta de inclinação igual (ou mais próxima de) -1 que seja mais próxima do ponto ideal de tangência segundo uma distância arbitrariamente pequena* ε. A Figura 3-2 ilustra o processo de determinação do número de MESs descrita a seguir.

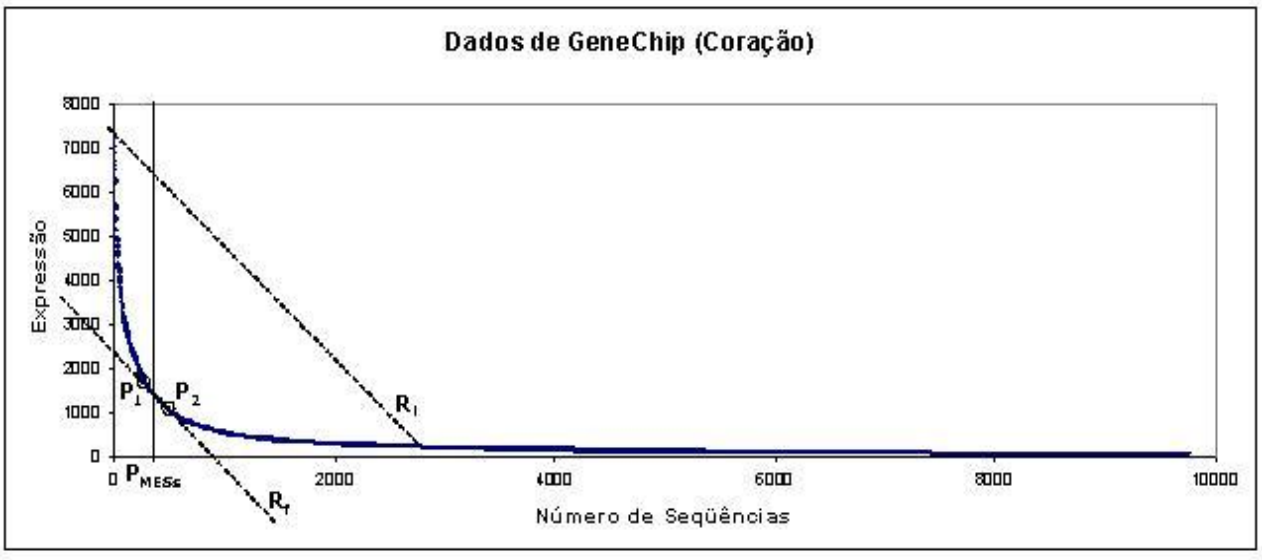


Figura 3-2: **Ilustração da metodologia implementada para a determinação do limiar que define a classe MESs**.

Inicialmente, é determinada a reta R_i (**Erro! Fonte de referência não encontrada.**) que é a reta de i nclinação mais próxima de -1 que tem como *um de seus pontos determinantes* [0, Max_{Exp}] ou [Max_{Seqs}, 0], onde Max_{Exp} é o maior valor de expressão e Max_{Seqs} o número total de seqüências presente na amostra. Em seguida, os valores das coordenadas dos pontos determinantes da reta são diminuídos de modo que a inclinação da reta determinada por eles seja igual a $-1 \pm \tau$, onde τ é uma constante arbitrária igual a 0.1. Este processo é repetido até que a distância entre os pontos determinantes da reta (P_1 e P_2, na Figura 3-2) seja menor que ε, onde ε é uma constante arbitrária que se faz necessária desde que a curva de expressão não é contínua. O limiar das MESs, P_{MESs}, é determinado como sendo a média aritmética das ordenadas dos pontos P_1 e P_2, o que informa o número de seqüências consideradas como sendo as mais expressas na amostra que gerou a curva de expressão.

3.3 Métrica de Similaridade Utilizando as MESs

As MESs foram utilizadas na definição de uma métrica de similaridade que permita quantificar o quão parecidas são duas amostras de modo que estas possam ser agrupadas. O cálculo da métrica pode ser descrito da seguinte forma:

1. Ordene as amostras em ordem decrescente de expressão de suas seqüências;
2. Conte o número de seqüências MESs comuns às duas amostras considerando as *M* MESs e compute a porcentagem de MESs compartilhadas gerando uma parcela da similaridade MESs entre as duas. Esta contagem é feita considerando incrementos ou “janelas” de

tamanho *I* a partir da seqüência MESs na primeira posição da ordenação, ou seja, para cada porção de *I* MESs conte quantas são compartilhadas às duas amostras e divida esse número por *I*, obtendo a porcentagem de MESs comuns (*shared* MESs, ou *sMESs*). Divida tal porcentagem pela distância do incremento considerada a partir da primeira posição da ordenação, ou seja, o primeiro incremento tem distância correspondente a 1, o segundo a 2, o terceiro a 3,... o *M*/*I* a *M*/*I*.

3. Repita o passo 2 para os *M*/*I* incrementos somando as parcelas de similaridade para constituir a similaridade MESs entre as duas amostras.

A Figura 3-3 ilustra os passos 2 e 3 apresentados acima. Por exemplo, considerando o par de amostras do baço (*spleen*) e do ovário (*ovary*), a partir do topo da lista de MESs conte quantas entre as *I* seqüências são comuns ao baço e ovário e divida tal contagem por *I*, obtendo a porcentagem de compartilhamento. Apenas TUBA1 é comum. Logo, a porcentagem é de 0,33. Divida tal porcentagem pela distância do incremento ao topo da lista (que é 1), obtendo a parcela do valor de *sMESs*: 0,33. Avance para o segundo incremento. Apenas HSPCA é comum. Assim sendo, a porcentagem é a mesma que a anterior, mas a parcela, não, já que a distância do incremento ao topo é igual a 2. Logo, a parcela correspondente ao segundo incremento é igual a 0,165. Avance para o próximo incremento. Ao observar apenas as *I* MESs deste incremento, vê-se que nenhuma das seqüências é comum às duas amostras. Entretanto, se, para cada seqüência deste incremento, caminhar-se em direção ao topo, verifica-se que ela é compartilhada entre as duas amostras. Dessa forma, conte tais seqüências. Assim, seis MESs são comuns, sendo a porcentagem 2 e a parcela igual a 0,66. Os *M* / *I* incrementos já foram computados então o valor de *sMESs* para as duas amostras é igual a 0,33 + 0,165 + 0,66 = 1.155.

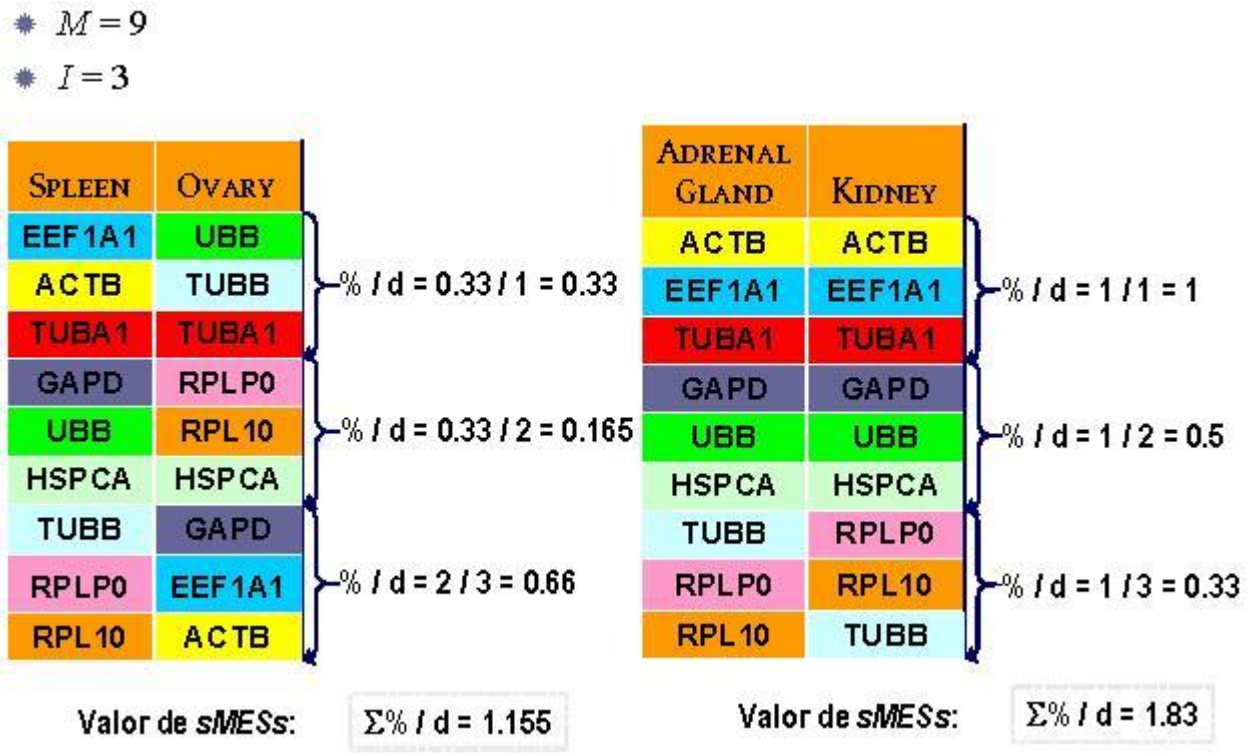


Figura 3-3: **Ilustração do cálculo da similaridade MESs e *sMESs***. Exemplo para dois pares de amostras considerando 9 MESs e incrementos (janelas) de tamanho 3. As seqüências nas amostras estão ordenadas em ordem decrescente (seqüência mais ao alto é a de maior expressão em cada lista).

O procedimento de cálculo da similaridade MESs, *sMESs*, descrito acima, pode ser definido analiticamente da seguinte forma para um par de amostras s_1 e s_2:

$$sMESs = \sum_{d=1}^{M/I} \left[\frac{\left(\sum_{i=0}^{dI} \sum_{j=0}^{dI} C_{i,j}(s_1, s_2) \right)}{I \times d} \right] \qquad (2)$$

Onde,

d é a distância do incremento I.

$C_{i,j}(s_1, s_2)$ é igual a 1 se e somente se a seqüência i na amostra s_1 é igual à seqüência j na amostra s_2 e elas não foram contadas ainda *ou* é igual a 0, caso contrário.

Como descrito na seção anterior, o limiar das MESs é determinado para cada amostra de dados. Entretanto, a métrica de similaridade definida com base nas MESs deve ser utilizada para agrupar várias amostras que constituem uma série de dados. Como cada amostra possui um número de

MESs próprio e normalmente diferente de outra amostra, a média aritmética do limiar de todas as amostras componentes da série sendo agrupada é utilizada. Desta forma, a similaridade MESs pode ser utilizada para agrupar uma série de amostras por meio de um algoritmo de *clustering* hierárquico, por exemplo, como o da Figura 3-4.

Entradas:
1. Um conjunto de N amostras;
2. Um valor de incremento, I;
3. O número de MESs, M.
Saída:
Um clustering hierárquico das N amostras.

0) Ordene os dados em cada amostra em ordem decrescente de expressão;
1) distância ← 1;
2) Para cada valor distância × I menor ou igual a M faça
3) Para cada par de amostras faça
4) Compute a percentagem de MESs comuns ao par considerando as I MESs para trás a partir de distância x I até 0;
5) porcentagem ← porcentagem / distância;
6) Adicione a porcentagem ao valor de ranking do par (*sMESs*);
Fim da repetição para (3)
7) distância ← distância + 1;
Fim da repetição para (2)
8) Ordene os pares de acordo com o valor de *ranking* computado e acumulado nos passos 2 a 7;
9) Com o par no topo do *ranking* faça
9.1) Se suas amostras componentes não foram agrupadas ainda, então intercale-as e as considere como uma só amostra deste ponto em diante;
9.2) Caso contrário, se uma das amostras já foi agrupada, então adicione a outra amostra naquele cluster;
10) Atualize o *ranking*;
11) Repita os passos 8 a 10 até que exista um só *cluster*.

Figura 3-4. **Algoritmo de *clustering* hierárquico projetado para agrupar amostras utilizando a similaridade MESs**.

Os passos 0 a 7 (Figura 3-4) foram descritos acima (Figura 3-3), pois se referem ao cálculo da similaridade entre pares de amostras. Os passos 8 a 10 são os passos de um agrupamento hierárquico propriamente dito. Quando um par de amostras é intercalado (unido), a lista de pares do *ranking* deve ser atualizada para remover os pares envolvendo as amostras componentes deste par. Seja *sMESs(a, b)* a similaridade entre as amostras *a* e *b*. Seja *k* uma amostra que tem dois pareamentos: um com *a* e um com *b*. Assim, a similaridade entre *k* e o par recém-intercalado *(a, b)* é dada por

$$sMESs(k,(a,b)) = \max(sMESs(k,a), sMESs(k,b)) \quad (3)$$

o que é similar à opção *complete linkage* do *clustering* hierárquico (Item 2.2.4) se a similaridade *sMESs* não for interpretada como distância entre os pares (assim, as duas amostras estão à menor distância). Esta é a melhor opção parar a presente abordagem, pois os pares mais bem posicionados no *ranking* serão agrupados primeiro produzindo um agrupamento no qual o número de MESs comuns aos pares (refletido pela similaridade MESs) é o maior possível.

3.4 Métrica de Similaridade de Conservação da Ordenação das MESs

Dadas duas amostras de dados de expressão gênica em que a expressão de suas seqüências componentes está ordenada, qual a taxa de conservação da ordem relativa de expressão dos vários pares de seqüências componentes de ambas? A resposta a tal pergunta dá origem a uma métrica de similaridade, sob a premissa de que amostras similares mantêm tal ordenação em taxas mais altas que amostras diferentes.

Sejam duas amostras, A_1 e A_2 caracterizadas, cada uma, por N seqüências expressas. Seja a ordenação de A_1 e A_2, respectivamente

$$r_1 > r_2 > r_3 > r_4 > ... > r_{N-1} > r_N$$

e

$$s_1 > s_2 > s_3 > s_4 > ... > s_{N-1} > s_N$$

Define-se a taxa de conservação da ordem de expressão entre A_1 e A_2 como *o somatório de todos os pares de seqüências* (seq_1, seq_2) *tais que* $seq_1 > seq_2$ *tanto em* A_1 *quanto em* A_2 *dividido pelo número pares possíveis*. Por exemplo, considerando-se as amostras apresentadas na Figura 3-3, têm-se as seguintes ordenações:

Spleen: EEF1A1 > ACTB > TUBA1 > GAPD > UBB > HSPCA > TUBB > RPLP0 > RPL10

Ovary: UBB > TUBB > TUBA1 > RPLP0 > RPL10 > HSPCA > GAPD > EEF1A1 > ACTB

Os pares conservados são 12, a saber:

EEF1A1 > ACTB, TUBA1 > GAPD, TUBA1 > HSPCA, TUBA1 > RPLP0, TUBA1 > RPL10,

UBB > *HSPCA, UBB* > *TUBB, UBB* > *RPLP0, UBB* > *RPL10, TUBB* > *RPLP0, TUBB* > *RPL10* e *RPLP0* > *RPL10*.

Desde que existem *N(N – 1)* pares possíveis e que *N* = 9, no exemplo acima, tem-se uma taxa de conservação de 12/36 = 30%. Já para o par de amostras *Adrenal Gland* e *Kidney* da Figura 3, existem 34 pares que conservam a ordem de expressão, o que provê uma taxa de conservação de 34/36 = 94,44%.

Considerando-se juntamente a premissa de que as MESs caracterizam uma amostra de dados de expressão gênica mais informativa com a premissa da conservação da ordem de expressão em amostras relacionadas, pode-se avaliar apenas as MESs de cada amostra para definir a *taxa de conservação da similaridade MESs entre duas amostras* A_1 e A_2 como sendo *o somatório dos pares de seqüências MESs cuja ordem de expressão é mantida tanto na amostra* A_1 *quanto na amostra* A_2 *considerando apenas o intervalo médio das MESs da série de dados da qual* A_1 *e* A_2 *fazem parte*.

4 Resultados e Discussão

E falou-lhe de muitas coisas por parábolas, dizendo: Eis que o semeador saiu a semear...
... mas o que foi semeado em boa terra é o que ouve e compreende a palavra;
e dá fruto, e um produz cem, outro, sessenta, e outro, trinta.
Mateus 13.3;23

Este capítulo apresenta os resultados da aplicação da metodologia apresentada no capítulo anterior em dados de experimentos reais disponíveis publicamente. Inicialmente são apresentadas considerações sobre os dados e algoritmos utilizados (seções 4.1 e 4.2, respectivamente). Em seguida, são apresentados os resultados da definição do limiar das MESs (Seção 4.3), da utilização das métricas de similaridade MESs e de conservação da ordenação das MESs no *clustering* de amostras de dados de expressão gênica e sua comparação com métricas tradicionalmente utilizadas (seções 4.4 e 4.5, respectivamente) e entre si (Seção 4.6), além de resultado de um experimento exploratório envolvendo séries de tecidos cancerosos, normais e em estágios diferentes de desenvolvimento e diferenciação (Seção 4.7). O capítulo é fechado com uma discussão sobre os resultados apresentados (Seção 4.8).

4.1 Considerações sobre os Dados e Parâmetros Utilizados

Para demonstrar a utilização das duas medidas de similaridade apresentadas no capítulo anterior, dezoito séries de dados disponíveis publicamente no banco de dados GEO (Barrett *et al*., 2005) do NCBI, totalizando 612 amostras de dados de expressão gênica foram utilizadas processadas por um programa escrito em Java. Apenas séries geradas utilizando-se as plataformas que representam *GeneChips*® da Affymetrix (Lockhart *et al.,* 1996) foram utilizadas e estão listadas na Tabela 4-1. Essa escolha deve-se, principalmente, ao fato de os dados de séries que utilizam tais plataformas serem mais uniformes já que os equipamentos, protocolos a serem seguidos e fabricante dos *chips* utilizados nos experimentos são os mesmos. As séries foram escolhidas sem seguir algum critério pré-estabelecido, mas não foram limitadas a apenas um organismo. Além disso, algumas séries

contendo dados de estudos envolvendo alguns tipos de câncer foram escolhidas com o objetivo de analisar o comportamento da metodologia quando aplicada a séries em condições biológicas distintas. A série gse2361 foi escolhida por ser composta apenas por amostras de tecidos "normais" de humanos (Ge *et al.,* 2005). Todas as séries foram agrupadas separadamente. A série gse1982 (Boni *et al.,* 2005) não foi agrupada individualmente, foi somente utilizada no *clustering* de amostras de tecidos cancerosos (Seção 4.6). A razão para tal exclusão é que os dados disponíveis desta série não estão completos: algumas amostras não estão presentes, bem como algumas informações que permitam uma análise confiável dos resultados produzidos pelo *clustering* de suas amostras.

Tabela 4-1: As séries de dados utilizadas nos testes executados.

Identificador no GEO	**Organismo**	**Número de amostras**	**Referência**	**Principal característica biológica e descrição resumida**
gse3416	*Arabidopsis thaliana* (ath)	18	Bläsing *et al.*, 2005	Como o nível de expressão de genes expressos na folha de plantas muda durante o ciclo circadiano.
gse607	*Arabidopsis thaliana* (ath)	11	Bergmann *et al.,* 2004	Análise da expressão gênica em três estruturas principais da planta: Folha (colhidas 15 dias após a germinação), caule e flor (colhidas no vigésimo nono dia após a germinação).
gse9311	*Arabidopsis thaliana* (ath)	8	Van Hoewyk *et al*., 2008	Estudo da influência (excesso e falta) do selenato no metabolismo da planta, bem como de sua tolerância.
gse1036	*Homo sapiens* (hsa)	12	Addya *et al.,* 2004	*Leucemia*. Linhagem celular K562 da eritroleucemia humana tratada/não tratada com hemin (um indutor do comprometimento eritróide).
gse1432	*Homo sapiens* (hsa)	24	Rock *et al.*, 2005	*Células do sistema nervoso central (SNC).* Resposta de células microgliais humanas ao interferon-γ nos instantes 1, 6 e 24h após o início do tratamento.
gse1493	*Homo sapiens* (hsa)	6	Manfredini *et al.,* 2005	*Células tronco humanas.* Estudos sobre os perfis de expressão de três categorias de células tronco hematopoiéticas (HSC).
gse1541	*Homo sapiens* (hsa)	20	dos Santos *et al.,* 2004	Células epiteliais pulmonárias imortalizadas (A549) submetidas a cinco condições de estudo diferentes em dois instantes de tempo.
gse1614	*Homo sapiens* (hsa)	12	Fleet *et al.,* 2003	*Diferenciação intestinal*. Perfis de expressão de células caco-2 BBe em três diferentes estágios.
gse1982	*Homo sapiens* (hsa)	103	Boni *et al.*, 2005	*Câncer*. Células periféricas mononucleadas do sangue (PBMCs) tiveram seus perfis de expressão coletados em três instantes de tempo no tratamento CCI-779 de 46 sujeitos com câncer real avançado.

gse2361	*Homo sapiens* (hsa)	36	Ge *et al.*, 2005	Perfil de expressão de 36 tecidos humanos normais.
gse2719	*Homo sapiens* (hsa)	54	Yoon *et al.*, 2006	Estudo da expressão de 39 amostras de sarcomas. Quinze amostras de controles normais foram utilizadas.
gse8692	*Homo sapiens* (hsa)	12	Liu *et al.*, 2007	*Câncer*. Análise da expressão de RNA extraído de 12 biópsias de tumores cerebrais primários.
gse1912	*Mus musculus* (mus)	25	Lin *et al.*, 2004	Análise temporal da expressão gênica do ciclo do pelo de camundongos (oito instantes de tempo).
gse2195	*Mus musculus* (mus)	42	Moggs *et al.*, 2004	Análise das mudanças na expressão gênica durante o crescimento do útero induzido por estrogênio (sete instantes de tempo).
gse775	*Mus musculus* (mus)	59	gse775, 2008	Análise temporal da expressão gênica em animais enfartados e não enfartados (oito instantes de tempo).
gse1001	*Rattus norvegicus* (rno)	18	Vázquez-Chona *et al.*, 2004	Análise e definição das mudanças temporais na expressão gênica da retina após lesão.
gse1156	*Rattus norvegicus* (rno)	30	gse1156, 2008	Análise das mudanças específicas do envelhecimento e dependentes do tempo na expressão gênica no hipocampo após a indução de convulsões por meio de kainato.
gse952	*Rattus norvegicus* (rno)	122	Walker *et al.*, 2004	Análise em larga escala de tecidos de ratos *Wistar* e *Sprague Dawley*.

A fim de proporcionar maior confiabilidade na análise dos resultados produzidos pelo *clustering*, as séries de dados foram manualmente inspecionadas (“curadas”) para agrupar amostras de modo a definir os *clusters* esperados e determinar o número deles para cada uma das séries utilizadas nos testes, o que está sumarizado nas tabelas Tabela 4-2 e Tabela 4-3, respectivamente. Isto foi feito de acordo com as informações disponíveis na referência bibliográfica relativa a cada série (Tabela 4-1), bem como no arquivo de dados da série disponível no GEO. Este passo mostrou-se necessário visto não ser conhecido um conjunto de dados de expressão gênica produzidos em larga escala que possa ser utilizado para avaliação do desempenho de algoritmos de *clustering* e que seja de uso generalizado pela comunidade de Bioinformática, como existe em Aprendizagem de Máquina. Por exemplo, aquele disponível em http://archive.ics.uci.edu/ml/. É importante ressaltar que algumas séries de dados possuem mais de um critério que pode ser utilizado na formação de *clusters*, bem como na avaliação dos resultados do processo de agrupamento executado por um dado algoritmo. Por exemplo, a série gse9311 (Van Hoewyk *et al*., 2008) tanto pode ser agrupada segundo o critério “condição de estudo” (condições “controle” e “sob efeito de selenato”), o que provê quatro *clusters* bem definidos, como pode ser agrupada segundo o critério “tecido” (raiz ou broto), o que provê dois clusters bem definidos (tabelas Tabela 4-2 e Tabela 4-3); já a série gse1001 (Vázquez-Chona *et al.*, 2004) pode ser agrupada de acordo com os diferentes tecidos (dois *clusters*) ou de acordo com

os vários instantes de tempo estudados (seis *clusters*). As tabelas no apêndice listam os *clusters* esperados para cada série, segundo os critérios alternativos. Quando mais de um critério podia ser aplicado, os critérios marcados com † na Tabela 4-2 foram os considerados "mais difíceis" ou "mais severos" para o *clustering* de acordo com os resultados apresentados neste trabalho. Evidentemente, o cruzamento dos vários diferentes critérios criaria uma grande quantidade de resultados ao se processar todas as séries utilizadas, assim sendo, a não ser que explicitamente destacado, os resultados apresentados nas seções 4.4 a 4.6 são apenas aqueles referentes às condições mais difíceis e às mais fáceis dentre as várias possibilidades. Desta forma, pode-se ter uma idéia dos limiares inferior e superior para a acurácia dos métodos tendo em vista todo o conjunto de dados utilizado.

Tabela 4-2: Critérios para definição de um pareamento correto entre duas amostras nas séries agrupadas separadamente.

Series	**O que é um pareamento correto?**
gse3416	Amostras que foram coletas em um mesmo instante de tempo. (t) (*)
gse607	As amostras pertencem à mesma estrutura da planta (folha, caule, flor). (d) (*)
gse9311	Amostras que pertencem a um mesmo tecido (raiz ou broto). (d) (*) Amostras de uma mesma condição biológica (controle / selenato). (c) (*)†
gse1036	Duas amostras do mesmo instante de tempo. (t) (*)
gse1432	Duas amostras de controles ou condição de estudo. (c) (*) Duas amostras de um mesmo doador. (d) (*)†
gse1493	Duas réplicas do mesmo tipo celular.(d) (*)
gse1541	Duas amostras da mesma condição de estudo. (c) (*)† Duas amostras do mesmo instante de tempo. (t) (*)
gse1614	Duas réplicas do mesmo instante de tempo. (t) (*)
gse2361	*Indefinido*(**).
gse2719	Amostras pertencentes a um mesmo tipo de tumor(***).(d) (*)
gse8692	Amostras de um mesmo tipo de tumor. (d) (*)
gse1912	Amostras de um mesmo instante de tempo. (t) (*)
gse2195	Duas amostras da mesma condição de estudo. (c) (*) Duas amostras do mesmo instante de tempo. (t) (*)†
gse775	Amostras pertencentes a uma mesma condição (região do infarto / região próxima ao infarto / não infarto). (c) (*) Amostras de um mesmo instante de tempo. (t) (*)†
gse1001	Amostras de uma mesma condição (normal / lesionada). (c) (*) Amostras de um mesmo instante de tempo. (t) (*)†
gse1156	Amostras de uma mesma condição (controle / convulsão).(c) (*) Amostras de um mesmo instante de tempo. (t) (*)†
gse952	Amostras de um mesmo tecido. (d) (*)

† Critérios que apresentam maior dificuldade (mais severos) de agrupamento segundo os resultados apresentados neste trabalho.

(*) (d) refere-se aos *clusters* esperados considerando diferentes tipos de doadores, tecidos, réplicas de experimentos, etc, que podem ser utilizados para agrupar amostras. (t) refere-se ao número de instantes de tempo analisados pelos pesquisadores em estudos de séries temporais e que *podem* ser utilizados para agrupar amostras. (c) refere-se a alguma condição especial de estudo.

(**) A série gse2361 é composta por 36 amostras de 36 tecidos diferentes (ou em estágios de desenvolvimento diferentes), o que impossibilita uma definição razoável de pareamentos corretos. Evidentemente, é esperado, por exemplo, que amostras de tecidos do sistema nervoso central pareiem entre si primeiramente.
(***) A série gse2719 possui 15 amostras de tecidos diferentes, que são utilizadas como controles, o que impossibilita uma definição razoável de pareamentos corretos *para estas 15 amostras*, como no caso da série gse2361.

Tabela 4-3: Número esperado de clusters para as séries da Tabela 4-1.

Série	**Organismo**	**Amostras**	***Clusters* esperados**
gse3416	ath	18	6
gse607	ath	11	3
gse9311	ath	8	2(d)/2(c)(*)
gse1036	hsa	12	6(t) (*)
gse1432	hsa	24	2(c)/4(d) (*)
gse1493	hsa	6	3
gse1541	hsa	20	5(d)/2(t) (*)
gse1614	has	12	3
gse2361	hsa	36	indefinido
gse2719	hsa	54 (39) (**)	8(**)
gse8692	hsa	12	3
gse1912	mus	25	8(t) (*)
gse2195	mus	42	2(d)/7(t) (*)
gse775	mus	59	3(d)/6(t) (*)
gse1001	rno	18 (15) (***)	2(d)/5(t) (*)(**)
gse1156	rno	30	2(c)/5(t)
gse952	rno	122	27

(*) Reporte-se aos comentários da Tabela 4-2
(**) Apenas *clusters* de amostras de tecidos cancerosos foram considerados. Quinze amostras foram excluídas da computação da acurácia. Veja nota da Tabela 4-2.
(***) Um *cluster* de três amostras é formado por amostras de controles e são atemporais, logo foram excluídos na computação da acurácia.

A análise da presença de dados de expressão gênica nas várias amostras componentes das séries nos permitiu classificar as seqüências de cada série em uma das três classes seguintes classes que podem resultar em valores de métricas diferentes, bem como em agrupamentos diferentes:

1. *Seqüências de manutenção*: comumente conhecidas como *housekeeping*, são definidas como aquelas "expressas"[5] em todas as amostras da série. Para uma dada amostra, somente são consideradas as seqüências expressas em todas as amostras da série.
2. *Seqüências de não-manutenção*: são aquelas dadas como *não* expressas em pelo menos uma das amostras componentes da série em questão. Para uma dada amostra,

[5] Uma seqüência é expressa se a tecnologia utilizada para produzir os dados informa que ela está presente em todas as amostras componentes da série em questão.

somente são consideradas as seqüências *não* expressas em pelo menos uma amostra da série.

3. *Todas*: união das duas classes anteriores, ou seja, todas as seqüências expressas em uma amostra, independentemente de serem ou não expressas nas demais.

4.2 Considerações sobre a Comparação entre Diferentes Métodos

A fim de comparar os resultados do *clustering* produzidos pela aplicação da metodologia apresentada no Capítulo 3, foram escolhidas quatro métricas de similaridade comumente utilizadas no *clustering* hierárquico (Tabela 4-4). A correlação de Spearman é de especial interesse, pois é *não-paramétrica*, o que também se aplica à similaridade utilizando as MESs. O mesmo algoritmo de hierarquização implementado para o *clustering* das MESs foi utilizado com as referidas métricas de similaridade. Além disso, três algoritmos de *clustering* bem conhecidos e muito utilizados em *Machine Learning* foram utilizados em algumas das séries trabalhadas: *K-means* (McQueen, 1967), EM (*Expectation Maximization*) (Russel & Norvig, 2003) e *Farthest First* (Hochbaum & Shmoys, 1985). As implementações disponíveis no pacote WEKA (Witten & Frank, 2000) foram utilizadas rodando com parâmetros *default*, exceto o número de *clusters* que foi ajustado de acordo com cada série segundo a Tabela 4-3. A Tabela 4-5 sumariza os principais parâmetros utilizados nos testes para os três algoritmos. É importante ressaltar que o objetivo principal dos testes relativos ao *clustering* não é comparar os algoritmos de *clustering per se*, mas o desempenho dos mesmos utilizando as métricas de similaridade especificadas.

Tabela 4-4: Medidas de similaridade utilizadas nos testes comparativos do *clustering* hierárquico, supondo duas amostras $A = \{a_1, a_2, \ldots a_N\}$ e $B = \{b_1, b_2, \ldots b_N\}$.

Métrica	**Expressão**
Distância Manhattan (*city-block*)	$\sum_{i=1}^{N}(a_i - b_i)$
Distância euclidiana	$\sqrt{\sum_{i=1}^{N}(a_i - b_i)^2}$
Correlação de Pearson	$\frac{\sum_{i=1}^{N}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{N}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{N}(b_i - \bar{b})}}$ $\bar{a}$ e $\bar{b}$ são as médias dos valores a_i e b_i, respectivamente.
Correlação de Spearman	$\frac{\sum_{i=1}^{N}(Pa_i - \overline{Pa})(Pb_i - \overline{Pb})}{\sqrt{\sum_{i=1}^{N}(Pa_i - \overline{Pa})^2}\sqrt{\sum_{i=1}^{N}(Pb_i - \overline{Pb})}}$ Pa_i e Pb_i são a ordem posicional ("*rank*") de a_i e b_i na ordenação das amostras *A* e *B*, respectivamente.

Tabela 4-5: Principais parâmetros utilizados na execução dos algoritmos / métodos utilizados para comparação.

Método	**Parâmetros**	**Medida de similaridade**
K-means	*Seed*: 10	Quadrado da distância euclidiana
EM	Número de iterações: 100 Desvio-padrão mínimo: 1×10^{-6}	Distribuição de probabilidade normal
Farthest First	*Seed*: 1	Distância euclidiana

4.3 Definição do Limiar das MESs

Aplicando-se a metodologia apresentada na Seção 3.1 às séries descritas na Tabela 4-1, obtêm-se os limiares para serem utilizados no processo de *clustering*. A Tabela 4-6 apresenta os limiares

encontrados (P_{MESs}) para as séries utilizadas nos testes considerando as três classes de seqüências utilizadas em alguns testes: de não-manutenção, de manutenção e todas as seqüências, utilizando-se o valor de ε igual a 10. Os limiares mostrados são as médias aritméticas dos limiares de todas as amostras componentes das respectivas séries. As porcentagens são calculadas em relação à média do total de seqüências presentes nas amostras de cada série.

Tabela 4-6: Limiares utilizados para a definição das MESs das séries utilizadas nos testes tanto para seqüências de manutenção quanto de não-manutenção e todas.

Classe de Seqüências	**não-manutenção**		**manutenção**		**todas**			
Série	**Limiar P_{MESs}**	**Porcentagem do total**	**Limiar P_{MESs}**	**Porcentagem do total**	**Limiar P_{MESs}**	**Porcentagem do total**	**Total**	**Número de amostras**
gse3416	91	0.71	982	7.67	1280	10	12801	18
gse607	383	3.15	1728	14.2	1745	14.34	12172	11
gse9311	413	2.75	1625	10.84	1710	11.41	14991	8
gse1036	607	5.49	667	6.04	1108	10.03	11051	12
gse1432	317	3.12	1863	18.32	1845	18.15	10167	24
gse1493	144	2.49	600	10.38	645	11.16	5780	6
gse1541	68	1.6	617	14.55	622	14.66	4242	20
gse1614	267	5.19	2714	52.79	2222	43.22	5141	12
gse1982	556	12.71	140	3.2	620	14.18	4373	103
gse2361	510	6.32	592	7.34	827	10.25	8065	36
gse2719	798	7.59	1761	16.74	1714	16.3	10517	54
gse8692	213	1.47	975	6.73	1449	10	14493	12
gse1912	393	6.94	1194	21.1	1214	21.45	5660	25
gse2195	280	4.31	1263	19.43	1230	18.92	6501	42
gse775	288	4.84	1327	22.28	1314	22.07	5955	59
gse1001	182	5.2	777	22.18	766	21.87	3503	18
gse1156	96	2.58	627	16.87	627	16.87	3717	30
gse952	400	13.75	53	1.82	733	25.2	2909	122
Média	**320**	**4.56**	**1139**	**15.84**	**1238**	**17.4**	**7891**	**34**

O fato de a porcentagem ser mais alta indica que uma proporção mais alta de seqüências é considerada como mais expressa nas amostras componentes de uma dada série. Assim, considerando as todas as seqüências, as amostras possuem, na média, uma maior proporção de MESs do que considerando as seqüências de manutenção e de não-manutenção. Na maioria das séries, a

proporção de MESs determinada é menor que 30%: todas as 18 séries para seqüências de não-manutenção; 12 séries para seqüências de manutenção; e 17 séries para todas as seqüências.

4.4 *Clustering* Utilizando MESs

Como exposto na Seção 3.3, as seqüências mais expressas em amostras de dados de expressão gênica podem ser utilizadas como métrica de similaridade para agrupar amostras inteiras. Assim sendo, a metodologia descrita na referida seção foi aplicada às séries de dados apresentadas na Tabela 4-1. As três classes de seqüências (não-manutenção, manutenção e todas) foram utilizadas nos testes cujos resultados são apresentados a seguir. São apresentados, também, dados comparativos entre o *clustering* hierárquico utilizando MESs e outras métricas de similaridade amplamente utilizadas: distâncias *Manhattan* e euclidiana e as correlações de *Pearson* e *Spearman*.

4.4.1 *Clustering* utilizando Três Classes de Seqüências

Sendo o *clustering* uma abordagem exploratória, foram testadas, inicialmente, as três classes de seqüências descritas no final da Seção 4.1: de não-manutenção, de manutenção e todas as seqüências. O objetivo da utilização das três classes de seqüências é determinar se uma determinada classe é mais adequada que outra para ser utilizada no cálculo da medida de similaridade utilizada para agrupar amostras de dados de expressão gênica. As tabelas Tabela 4-7 e Tabela 4-8 sumarizam as taxas de acerto (acurácia[6]) para as dezessete séries e classes de seqüências. As taxas de acerto foram computadas considerando as definições de "pareamentos corretos" apresentadas na Tabela 4-2.

Excluindo-se *H. sapiens*, a utilização das seqüências de não-manutenção forneceu uma maior acurácia que as outras classes de seqüências considerando os critérios de agrupamento mais difíceis (Tabela 4-7), apresentando uma taxa aproximadamente 5% maior que as demais classes de seqüências. Para os critérios menos severos as taxas gerais de acerto aumentam, como esperado, sendo que a acurácia é maior com a utilização das seqüências de não manutenção para todos os organismos, como apresentado na Tabela 4-8. Isso, evidentemente, não exclui a utilização das seqüências de manutenção e todas elas em experimentos exploratórios, mas indica que tais classes de seqüências são menos informativas para o processo de agrupamento de amostras, no geral, apesar de não o ser para todas as séries de dados analisadas, como a série gse1036, por exemplo. Assim sendo, *alguns* dos resultados apresentados doravante, para o *clustering* utilizando as MESs, são aqueles

[6] Os termos "taxa de acerto(s)" e "acurácia" são utilizados de forma "intercambiada".

produzidos utilizando-se as seqüências de não-manutenção. Dadas as taxas de acerto reportadas, o desempenho do *clustering* utilizando as MESs foi comparado ao daquele utilizando outras métricas de similaridade como exposto a seguir.

Tabela 4-7: Taxas de acerto considerando os critérios *mais severos* de agrupamento de cada série, onde aplicável, de acordo com a Tabela 4-2.

Organismo	Série	Não-manutenção	Manutenção	Todas	Critério
	gse3416	38,89	22,22	16,67	tempo
	gse607	100	100	100	tecido
ath	*gse9311*	*50*	*50*	*50*	*condição*
	Média:	**62,96**	**57,41**	**55,56**	
	gse1036	83,33	83,33	91,67	tempo
	gse1432	*58,33*	*58,33*	*58,33*	*doador*
	gse1493	100	100	100	tecido
	gse1541(*)	*0*	*0*	*0*	*condição*
	gse1614	100	100	100	tempo
	gse2719	69,23	74,36	71,79	condição
hsa	gse8692	83,33	91,67	83,33	condição
	Média:	**82,37**	**84,61**	**84,19**	
	gse1912	88	88	88	tempo
	gse2195	*40,48*	*16,67*	*16,67*	*tempo*
mmu	*gse775*	*76,27*	*76,27*	*71,19*	*tempo*
	Média:	**68,25**	**60,31**	**58,62**	
	gse1001	*46,67*	*53,33*	*46,67*	*tempo*
	gse1156	*83,33*	*56,67*	*53,33*	*tempo*
rno	gse952	95,08	86,07	95,08	tecido
	Média:	**75,03**	**65,36**	**65,03**	
Média geral:		**72.15**	**66,92**	**65,85**	

(*) A série gse1541, segundo o critério "condição", parece ser *unlearnable* ou o problema de agrupá-la *irrealizável* para todas as métricas utilizadas neste trabalho. Ela não foi considerada no cálculo das médias aritméticas para *H. sapiens* (hsa) e geral. Veja os resultados apresentados no Item 4.4.2 e a Seção 4.8 para uma discussão mais detalhada.

Tabela 4-8: Taxas de acerto considerando o critério de agrupamento *menos severo* de acordo com a Tabela 4-2.

Organismo	Série	Não-manutenção	Manutenção	Todas	Critério
	gse3416	38,89	22,22	16,67	tempo
	gse607	100	100	100	tecido
ath	*gse9311*	*100*	*100*	*100*	*tecido*
	Média:	**79,63**	**74,07**	**72,22**	
	gse1036	83,33	83,33	91,67	tempo
	gse1432	*83,33*	*75*	*70,83*	*condição*
	gse1493	100	100	100	tecido
	gse1541	*100*	*90*	*90*	*tempo*
	gse1614	100	100	100	tempo
	gse2719	69,23	74,36	71,79	condição
hsa	gse8692	83,33	91,67	83,33	condição
	Média:	**88,46**	**87,77**	**86,80**	
	gse1912	88	88	88	tempo
	gse2195	*92,86*	*85,71*	*85,71*	*condição*
mmu	*gse775*	*94,92*	*88,14*	*89,83*	*tecido*
	Média:	**91,93**	**87,28**	**87,85**	
	gse1001	*88,89*	*88,89*	*88,89*	*condição*
	gse1156	*93,33*	*66,67*	*70*	*condição*
rno	gse952	95,08	86,07	95,08	tecido
	Média:	**92,43**	**80,54**	**84,66**	
Média geral:		**88,11**	**82,42**	**81,83**	

4.4.2 Comparação do Desempenho da Similaridade MESs com Outras Métricas de Similaridade

Métricas de similaridade comumente utilizadas para agrupar objetos são utilizadas, também, em dados de expressão gênica (Eisen *et al*. 1998). Assim sendo, é necessária a comparação do desempenho destas em relação ao *clustering* utilizando as MESs. As tabelas 4-9 e 4-10 apresentam as taxas de acerto no *clustering* das séries utilizando-se as diferentes métricas apresentadas na Tabela 4-4, bem como as três classes de MESs.

Tabela 4-9: Taxas de acerto utilizando-se medidas de similaridade comumente utilizadas para agrupar dados em comparação àquelas utilizando as três diferentes classes de MESs, para os critérios *mais* severos de agrupamento.

						MESs			
Organismo	**Série**	**Manhattan**	**Euclidiana**	**Pearson**	**Spearman**	**Não-manutenção**	**Manutenção**	**Todas**	**Critério**
ath 37 amostras	gse3416	11,11	11,11	11,11	11,11	38,89	22,22	16,67	tempo
	gse607	100	100	100	100	100	100	100	tecido
	gse9311	*50*	*50*	*50*	*75*	*50*	*50*	*50*	*condição*
	Média:	**53,70**	**53,7**	**53,7**	**62,04**	**62,96**	**57,41**	**55,56**	
hsa 125 amostras	gse1036	69,44	69,44	77,78	52,78	83,33	83,33	91,67	tempo
	gse1432	*61,11*	*56,94*	*65,28*	*61,11*	*58,33*	*58,33*	*58,33*	*doador*
	gse1493	66,67	33,33	100	100	100	100	100	tecido
	gse1541	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*condição*
	gse1614	100	91,67	100	91,67	100	100	100	tempo
	gse2719	63,15	53,89	65	68,71	69,23	74,36	71.79	condição
	gse8692	41,67	75	75	50	83,33	91,67	83.33	condição
	Média:	**67,01**	**63,38**	**80,51**	**70,71**	**82,37**	**84,62**	**84.19**	
mmu 126 amostras	gse1912	93,33	89,33	85,33	85,33	88	88	88	tempo
	gse2195	*11,9*	*11,9*	*14,29*	*23,81*	*40,48*	*16,67*	*16.67*	*tempo*
	gse775	*79,66*	*81,36*	*79,66*	*76,27*	*76,27*	*76,27*	*71.19*	*tempo*
	Média:	**61,63**	**60,86**	**59,76**	**61,8**	**68,25**	**60,31**	**58.62**	
rno 170 amostras	*gse1001*	*50,75*	*34,08*	*61,86*	*56,3*	*46,67*	*53,33*	*46.67*	*tempo*
	gse1156	*60*	*53,33*	*63,33*	*56,67*	*83,33*	*56,67*	*53.33*	*tempo*
	gse952	94,26	89,34	91,8	92,62	95,08	86,07	95.08	tecido
	Média:	**68,34**	**58,92**	**72,33**	**68,53**	**75,03**	**65,36**	**65.03**	
Média geral:		**60,32**	**57,68**	**66,41**	**63,97**	**72,15**	**66,92**	**65,85**	

Tabela 4-10: Taxas de acerto utilizando medidas de similaridade comumente utilizadas para agrupar dados em comparação àquelas utilizando as três diferentes classes de MESs, para os critérios *menos* severos de agrupamento.

						MESs			
Organismo	**Série**	**Manhattan**	**Euclidiana**	**Pearson**	**Spearman**	**Não-manutenção**	**Manutenção**	**Todas**	**Critério**
ath 37 amostras	gse3416	11,11	11,11	11,11	11,11	38,89	22,22	16.67	tempo
	gse607	100	100	100	100	100	100	100	tecido
	gse9311	*50*	*50*	*50*	*75*	*100*	*100*	*100*	*tecido*
	Média:	**53,7**	**53,7**	**53,7**	**62,04**	**79,63**	**74,07**	**72.22**	
hsa 125 amostras	gse1036	69,44	69,44	77,78	52,78	83,33	83,33	91.67	tempo
	gse1432	*79,16*	*74,99*	*83,33*	*79,16*	*83,33*	*75*	*70.83*	*condição*
	gse1493	66,67	33,33	100	100	100	100	100	tecido
	gse1541	*93,33*	*93,33*	*93,33*	*93,33*	*100*	*90*	*90*	*tempo*
	gse1614	100	91,67	100	91,67	100	100	100	tempo
	gse2719	63,15	53,89	65	68,71	69,23	74,36	71.79	condição
	gse8692	41,67	75	75	50	83,33	91,67	83.33	condição
	Média:	**70,02**	**66,39**	**83,52**	**73,72**	**88,46**	**87,77**	**86.8**	
mmu 126 amostras	gse1912	93,33	89,33	85,33	85,33	88	88	88	tempo
	gse2195	*75,39*	*75,39*	*77,78*	*87,3*	*92,86*	*85,71*	*85.71*	*condição*
	gse775	*96,05*	*97,75*	*96,05*	*92,66*	*94,92*	*88,14*	*89.83*	*tecido*
	Média:	**88,26**	**87,49**	**86,39**	**88,43**	**91,93**	**87,28**	**87.85**	
rno 170 amostras	*gse1001*	*53,71*	*37,04*	*64,82*	*59,26*	*88,89*	*88,89*	*88.89*	*condição*
	gse1156	*72,22*	*65,55*	*75,55*	*68,89*	*93,33*	*66,67*	*70*	*condição*
	gse952	94,26	89,34	91,8	92,62	95,08	86,07	95.08	tecido
	Média:	**73,40**	**63,98**	**77,39**	**73,59**	**92,43**	**80,54**	**84.66**	
Média geral:		**71**	**68,05**	**76,78**	**74,34**	**88,11**	**82,42**	**81,83**	

Novamente, a utilização das MESs proporcionou, na média, taxas de acerto mais altas que as obtidas com a utilização das métricas tradicionais. Considerando apenas as MESs de não-manutenção, as taxas de acerto foram consistentemente maiores que as demais no geral e em cada organismo individualmente. Mesmo considerando as outras duas classes de MESs (manutenção e todas), o desempenho foi melhor no geral, ainda que em algumas séries, dependendo do critério de agrupamento, a correlação de Pearson ou a de Spearman tenham proporcionado taxas de acerto ligeiramente superiores aos obtidos com estas duas classes de MESs. Entretanto, é importante ressaltar que a acurácia foi bastante baixa em algumas séries e mesmo em organismos, tendo como referência os critérios mais severos de agrupamento (Tabela 4-9). Por exemplo, em *A. thaliana*, taxas entre 55% e 63% foram obtidas, na média, considerando critérios mais severos, o que é ruim, ainda que superior ao gerado pelas métricas tradicionais. Tal resultado foi forçado pela baixa acurácia resultante nos dados da série gse3416, cujas amostras representam 48.6% daquelas de *A. thaliana* utilizadas nos testes. Independentemente de tais casos, a acurácia obtida pela utilização das

seqüências de não-manutenção foi a mais alta e foi utilizada para comparar, ilustrativamente, o desempenho da abordagem aqui apresentada em relação a três diferentes algoritmos de *clustering* relacionados na Seção 4.2, bem como para gerar a hierarquização das amostras de algumas das séries utilizadas, como segue.

Três séries de dados para as quais é possível definir *clusters* de maneira "fácil" (gse607, gse1493 e gse1614) foram submetidas aos quatro algoritmos de *clustering*. Tais séries foram escolhidas com o objetivo de mostrar que mesmo sobre dados onde os grupos de amostras são facilmente determinados por informações outras que não aquelas contidas nos dados de expressão gênica, alguns algoritmos comumente utilizados podem ter desempenho pobre. Entre os motivos principais de tal desempenho encontram-se o elevado número de atributos (seqüências) das amostras e a complexidade intrínseca dos dados. Além disso, uma série onde os *clusters* esperados são definidos de forma não tão fácil foi incluída: a série gse8692 contém uma amostra que apresenta características fenotípicas que permite agrupá-la em dois *clusters*.

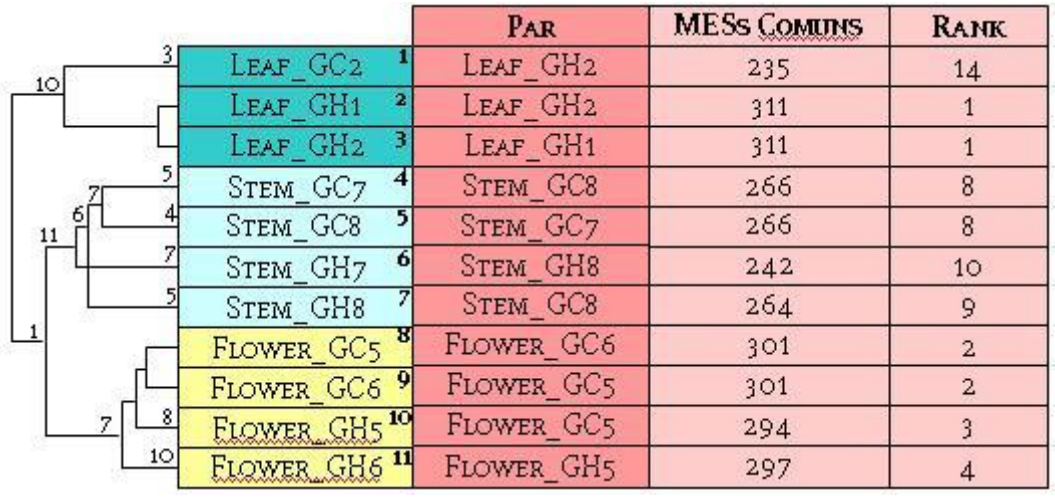
Os resultados são mostrados na Tabela 4-11, que apresenta, também, o número de amostras presentes nos *clusters* de cada uma das quatro séries de dados. No geral, a abordagem das MESs supera os demais, com uma taxa de acerto de cerca de 95%, sendo o método de desempenho mais próximo, o *Farthest First*, com 90% de acertos, ficando os demais com desempenho bastante inferior.

As três séries com *clusters* bem definidos foram agrupadas com cem por cento de acerto usando as MESs, enquanto os métodos tradicionais (EM e *Farthest First*) alcançaram cem por cento apenas em duas delas, sendo que o métodos mais comumente utilizado, *k-Means*, teve o pior desempenho. Além do desempenho ruim, outra deficiência de tais métodos é que o número de *clusters* esperado em cada conjunto de dados submetido a eles deve ser informado *a priori*. Tal deficiência é superada com a utilização do *clustering* hierárquico. A Figura 4-1 apresenta a hierarquia formada pela aplicação do *clustering* utilizando as MESs de não manutenção nas quatro séries de dados da Tabela 4-11.

Tabela 4-11: Taxas de acerto para quatro séries de dados agrupadas por quatro diferentes algoritmos.

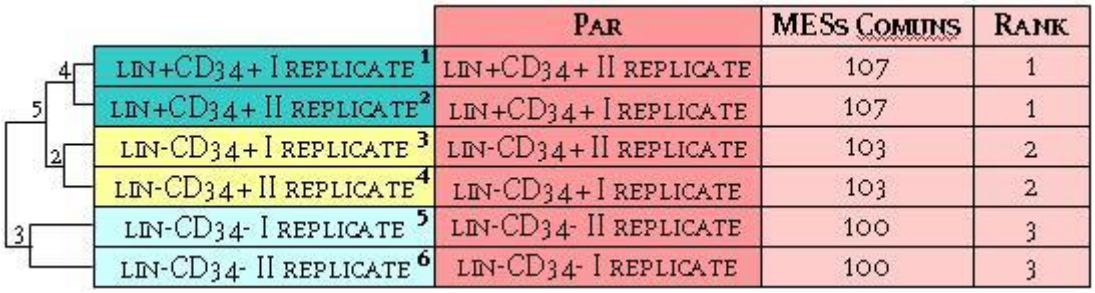
Séries	**Amostras nos grupos**	**MESs (não-manutenção)**	**K-Means**	**EM**	**Farthest First**
gse607	{3, 4, 4}	11 (100%)	8 (72,7%)	9 (81,8%)	11 (100%)
gse1493	{2, 2, 2}	6 (100%)	3 (50%)	6 (100%)	6 (100%)
gse1614	{4, 4, 4}	12 (100%)	9 (75%)	12 (100%)	11 (91,7%)
gse8692	{3, 3, 6}	10 (83,3%)	9 (75%)	6 (50%)	9 (75%)
Média:	—	**95,1%**	**70,7%**	**80,5%**	**90,2%**

A hierarquização facilita o processo de visualização da similaridade entre as várias amostras. É possível enxergar os *clusters* apresentados na Tabela 4-11, bem como é possível extrair informações sobre a similaridade entre os próprios *clusters* e/ou entre amostras destes *clusters*. Por exemplo, na Figura 4-1-(a), é fácil visualizar que as três amostras de folha (*leaf*), as quatro de caule (*stem*) e as quatro de flor (*flower*) formam *clusters* entre si; e que, em seguida, flor e caule se relacionam primeiramente por meio de suas amostras *STEM_GH8* (número 7) e *FLOWER_GH6* (número 11) e que folha se relaciona ao caule por meio do par de amostras formado por *LEAF_GC2* e *FLOWER_GH5* (números 1 e 10, respectivamente). Desta forma, vê-se que os números nas linhas de um dendrograma indicam o par do *cluster* formado pela sub-árvore ou ramo no qual ele está.



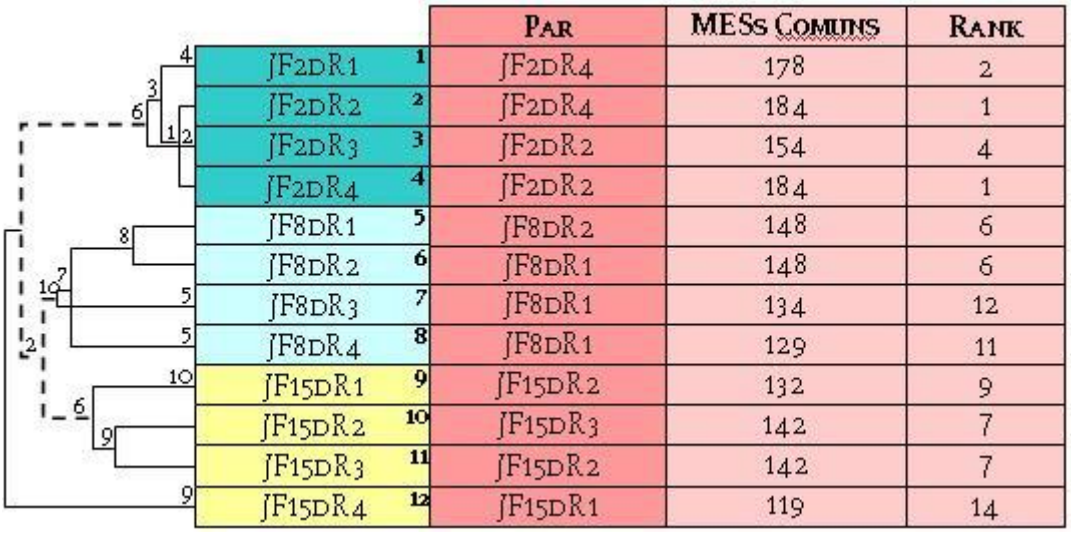
	Par	MESs Comuns	Rank
1 Leaf_GC2	Leaf_GH2	235	14
2 Leaf_GH1	Leaf_GH2	311	1
3 Leaf_GH2	Leaf_GH1	311	1
4 Stem_GC7	Stem_GC8	266	8
5 Stem_GC8	Stem_GC7	266	8
6 Stem_GH7	Stem_GH8	242	10
7 Stem_GH8	Stem_GC8	264	9
8 Flower_GC5	Flower_GC6	301	2
9 Flower_GC6	Flower_GC5	301	2
10 Flower_GH5	Flower_GC5	294	3
11 Flower_GH6	Flower_GH5	297	4

(a) gse607



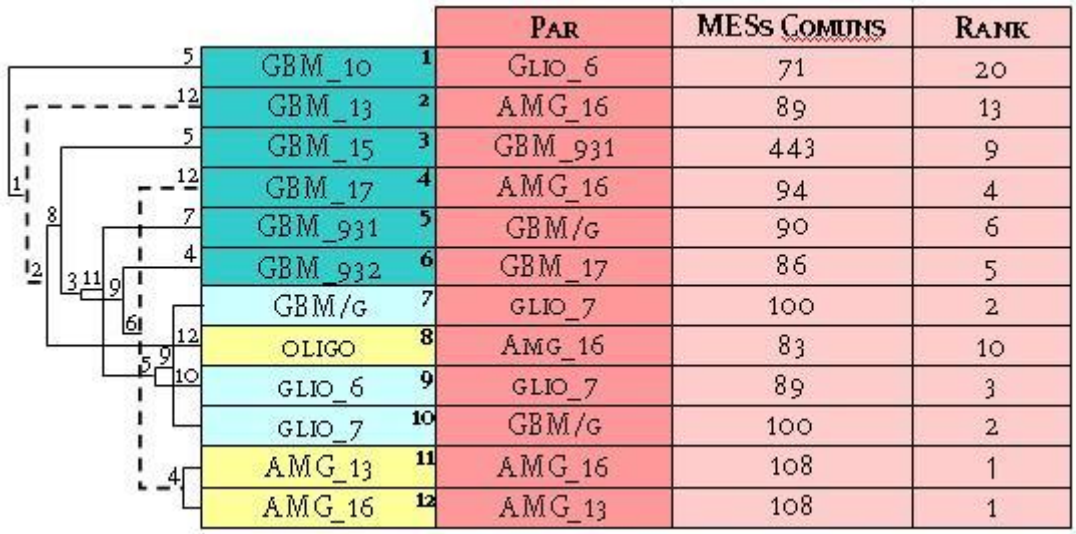
	Par	MESs Comuns	Rank
LIN+CD34+ I REPLICATE 1	LIN+CD34+ II REPLICATE	107	1
LIN+CD34+ II REPLICATE 2	LIN+CD34+ I REPLICATE	107	1
LIN-CD34+ I REPLICATE 3	LIN-CD34+ II REPLICATE	103	2
LIN-CD34+ II REPLICATE 4	LIN-CD34+ I REPLICATE	103	2
LIN-CD34- I REPLICATE 5	LIN-CD34- II REPLICATE	100	3
LIN-CD34- II REPLICATE 6	LIN-CD34- I REPLICATE	100	3

(b) gse1493



	Par	MESs Comuns	Rank
JF2DR1 1	JF2DR4	178	2
JF2DR2 2	JF2DR4	184	1
JF2DR3 3	JF2DR2	154	4
JF2DR4 4	JF2DR2	184	1
JF8DR1 5	JF8DR2	148	6
JF8DR2 6	JF8DR1	148	6
JF8DR3 7	JF8DR1	134	12
JF8DR4 8	JF8DR1	129	11
JF15DR1 9	JF15DR2	132	9
JF15DR2 10	JF15DR3	142	7
JF15DR3 11	JF15DR2	142	7
JF15DR4 12	JF15DR1	119	14

(c) gse1614

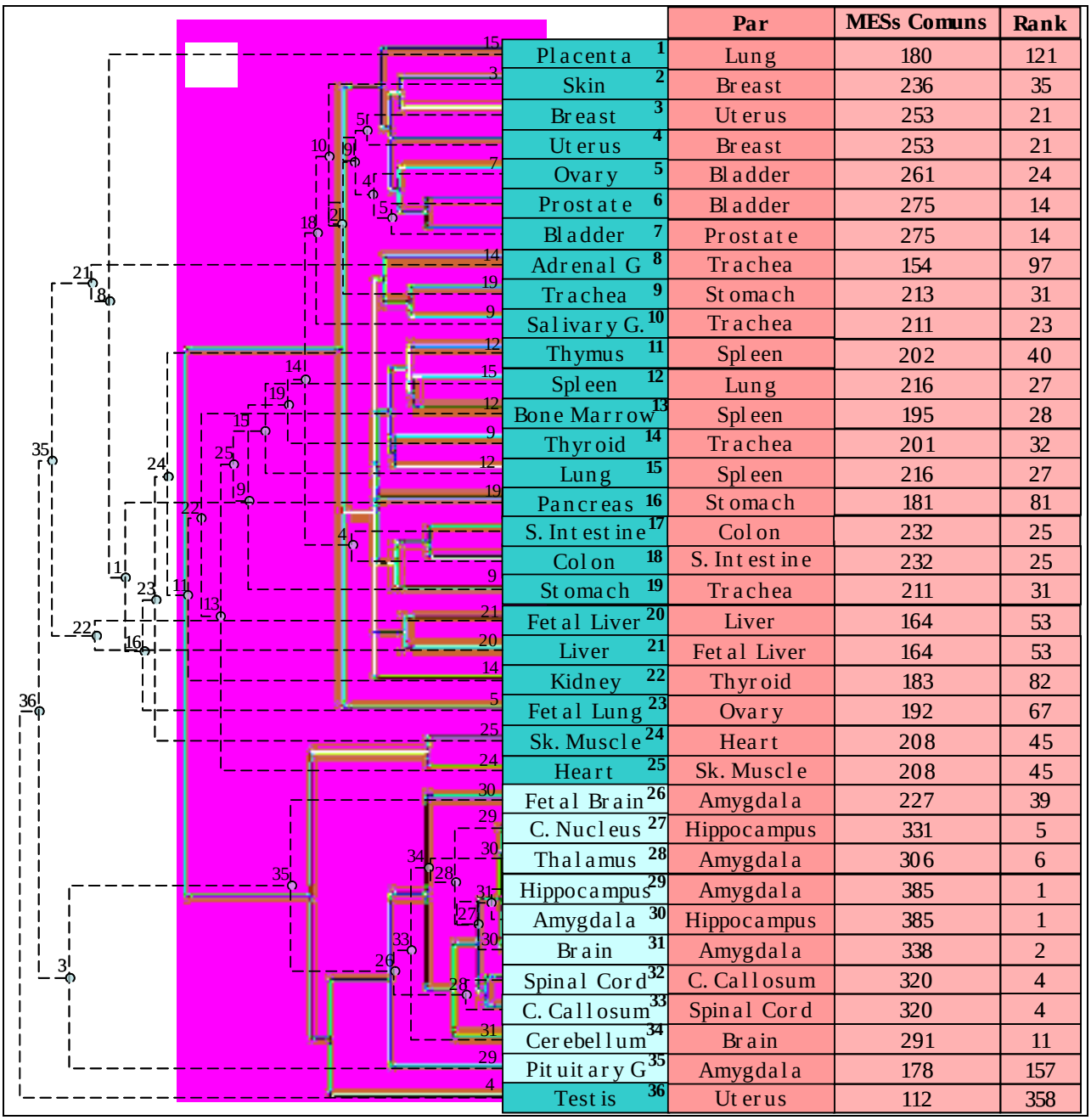


	Par	MESs Comuns	Rank
GBM_10 1	GLIO_6	71	20
GBM_13 2	AMG_16	89	13
GBM_15 3	GBM_931	443	9
GBM_17 4	AMG_16	94	4
GBM_931 5	GBM/G	90	6
GBM_932 6	GBM_17	86	5
GBM/G 7	GLIO_7	100	2
OLIGO 8	AMG_16	83	10
GLIO_6 9	GLIO_7	89	3
GLIO_7 10	GBM/G	100	2
AMG_13 11	AMG_16	108	1
AMG_16 12	AMG_13	108	1

(d) gse8692

Figura 4-1: ***Clustering* utilizando as MESs de não-manutenção das quatro series da Tabela 4-11**. Amostras pertencentes ao mesmo *cluster* (esperado) possuem a mesma cor. As amostras estão numeradas para facilitar a localização do pareamento correto. As linhas tracejadas informam que o pareamento não era esperado naquela ordem de hierarquização. Os dois números em uma mesma linha indicam o par de amostras responsável por unir os *clusters*. (a) gse607: *clusters* esperados: *leaf* (3 amostras), *stem* (4 amostras) e *flower* (4 amostras). (b) gse1493: *clusters* esperados: lin+CD34+, lin-CD34+, lin-CD34- (2 amostras em cada *cluster*). (c) gse1614: *clusters* esperados: 2D, 8D, 15D (representando três instantes de tempo, 4 amostras em cada *cluster*). (d) gse8692: *clusters* esperados: glioblastoma (6 amostras, em verde), glioma (3 amostras, em amarelo) e gliosarcoma (3 amostras, em azul).

A Figura 4-1 apresenta o *clustering* hierárquico usando a abordagem MESs para séries onde os *clusters* são relativamente fáceis de serem enxergados antes do processo de agrupamento das amostras (Figura 4-1 a – c). Entretanto, em muitos estudos não é fácil definir-se *clusters* de amostras e nem é este o objetivo de utilizarem-se tais técnicas, já que o *clustering* é intrinsecamente uma técnica exploratória. A Figura 4-2 apresenta uma superposição das árvores de *clustering* hierárquico geradas para a série gse2361 (Ge *et al.*, 2005), composta por 36 amostras de 36 tecidos humanos "normais". Olhando do ponto de vista fisiológico, como *clusters* de tais amostras poderiam ser definidos? Obviamente, a resposta mais simples é agrupar tecidos que possuem fisiologia conhecida e similar. Por exemplo, tecidos do sistema nervoso central (SNC) poderiam agrupar-se, da mesma forma que tecidos do trato reprodutor. As linhas tracejadas na Figura 4-2 representam a árvore MESs e as cheias o *clustering* hierárquico utilizando a correlação de Pearson como medida de similaridade, como apresentado por Ge *at al*. (2005). Os resultados são similares, mas com diferenças importantes: amostras de tecidos do SNC, do trato reprodutor [útero (*uterus*), próstata (*prostate*), placenta (*placent*), ovário (*ovary*), etc], de tecidos hemapoiéticos [baço (*spleen*), timo (*thymus*), medula óssea (*bone marrow*)] são agrupadas como esperado. Entretanto, testículo (*testis*) não é agrupado com tecidos do SNC e nem o par [coração (*heart*), músculo esquelético (*sk. muscle*)] no *clustering* MESs, o que ocorre no *clustering* utilizando-se a correlação de Pearson.



	Par	MESs Comuns	Rank
Placenta 1	Lung	180	121
Skin 2	Breast	236	35
Breast 3	Uterus	253	21
Uterus 4	Breast	253	21
Ovary 5	Bladder	261	24
Prostate 6	Bladder	275	14
Bladder 7	Prostate	275	14
Adrenal G 8	Trachea	154	97
Trachea 9	Stomach	213	31
Salivary G. 10	Trachea	211	23
Thymus 11	Spleen	202	40
Spleen 12	Lung	216	27
Bone Marrow 13	Spleen	195	28
Thyroid 14	Trachea	201	32
Lung 15	Spleen	216	27
Pancreas 16	Stomach	181	81
S. Intestine 17	Colon	232	25
Colon 18	S. Intestine	232	25
Stomach 19	Trachea	211	31
Fetal Liver 20	Liver	164	53
Liver 21	Fetal Liver	164	53
Kidney 22	Thyroid	183	82
Fetal Lung 23	Ovary	192	67
Sk. Muscle 24	Heart	208	45
Heart 25	Sk. Muscle	208	45
Fetal Brain 26	Amygdala	227	39
C. Nucleus 27	Hippocampus	331	5
Thalamus 28	Amygdala	306	6
Hippocampus 29	Amygdala	385	1
Amygdala 30	Hippocampus	385	1
Brain 31	Amygdala	338	2
Spinal Cord 32	C. Callosum	320	4
C. Callosum 33	Spinal Cord	320	4
Cerebellum 34	Brain	291	11
Pituitary G 35	Amygdala	178	157
Testis 36	Uterus	112	358

Figura 4-2: **Uma superposição das árvores de *clustering* resultantes da aplicação da abordagem MESs (linhas tracejadas) e do *clustering* hierárquico apresentado por Ge *et al.*, (2005) para a série gse2361**. Os tecidos do sistema nervoso central estão destacados e as amostras estão numeradas para facilitar a localização dos pareamentos. Os dois números em uma mesma linha indicam o par de amostras responsável por unir os *clusters*.

4.5 *Clustering* utilizando a Similaridade de Conservação (da Ordem de Expressão)

A similaridade de conservação é uma métrica que indica o número de pares de seqüências que mantêm a ordem relativa de expressão em um par de amostras, considerando apenas as MESs. A abordagem detalhada na Seção 3.4 foi aplicada às 18 séries de dados apresentadas na Tabela 4-1. Esta seção apresenta tais resultados. Inicialmente, as maiores e menores porcentagens de conservação são mostradas. Em seguida, são mostrados resultados comparativos da utilização da conservação em relação a outras métricas tradicionalmente utilizadas. Finalmente, são comparados os resultados produzidos pela aplicação das métricas MESs e conservação.

4.5.1 Taxas de Conservação

A Tabela 4-12 apresenta as taxas de conservação da ordem de expressão gênica máxima e mínima para as séries da Tabela 1. As colunas “Máximo” e “Mínimo” informam, respectivamente, a porcentagem máxima e mínima de conservação da ordem de expressão observada, sendo:

- Máximo: porcentagem de pares de MESs que conservaram a ordem de expressão no par de amostras de *maior* conservação da série;
- Mínimo: porcentagem de pares de MESs que conservaram a ordem de expressão no par de amostras de *menor* conservação da série.

Os valores sombreados em cada coluna representam as maiores e menores porcentagens entre todas as séries processadas de acordo com a classe de seqüência examinada na geração da porcentagem de conservação. Os dois valores em negrito e itálico indicam a maior e a menor taxa de conservação geral, isto é, considerando todos os dados examinados. Assim, o parâmetro conservação da ordem de expressão gênica varia dentro de um intervalo em cada série que permitiria agrupamentos hierárquicos de amostras.

É notável como a conservação é *maior* considerando as seqüências de manutenção em comparação às de não-manutenção e todas. Isso é um padrão observado em todas as séries de dados, independentemente do organismo, conforme o gráfico abaixo (Figura 4-3). Tal comportamento é esperado, em certo sentido, já que as seqüências de manutenção são aquelas expressas em todas as amostras da série e cujas funções celulares correspondentes devem ser comuns ou similares nos vários

tecidos e mesmo em diferentes organismos. Desde que as mesmas seqüências são participantes de um mesmo processo em tecidos ou condições diferentes, pode-se esperar que a ordem de expressão delas deva ser mantida nos vários tecidos ou condições diferentes. Dessa forma, é esperado que as seqüências de manutenção tenham a ordem de expressão mais conservada que as de não-manutenção, o que é condizente com os resultados produzidos.

Tabela 4-12: Taxas de conservação da ordem de expressão gênica de pares de seqüências máximas e mínimas. Maiores e menores taxas são destacadas.

		Não-manutenção		Manutenção		Todas	
Organismo	**Série**	**Máximo**	**Mínimo**	**Máximo**	**Mínimo**	**Máximo**	**Mínimo**
ath	gse3416	89.13	5.49	90.20	75.70	90.75	76.09
	gse607	82.03	4.43	92.89	70.80	92.67	66.25
	gse9311	91.95	***1.00***	92.31	65.28	92.25	57.36
	Média:	**87.70**	**3.64**	**91.80**	**70.60**	**91.89**	**66.57**
hsa	gse1036	92.37	64.52	92.13	62.31	90.36	81.24
	gse1432	79.30	16.53	88.18	73.09	87.36	70.36
	gse1493	76.98	1.96	90.18	75.67	89.67	74.90
	gse1541	85.60	7.77	94.18	77.45	***94.21***	77.40
	gse1614	65.67	16.39	91.80	79.40	88.59	71.10
	gse2361	83.24	6.16	87.75	67.27	85.51	38.54
	gse2719	69.39	8.35	85.73	67.70	80.45	48.31
	gse8692	59.57	15.04	81.91	73.07	81.18	71.47
	Média:	**76.52**	**17.09**	**88.98**	**72.00**	**87.17**	**66.67**
mmu	gse1912	81.11	9.91	91.27	73.52	90.01	64.51
	gse2195	78.15	21.18	90.91	73.49	90.55	70.99
	gse775	82.61	22.24	92.60	63.08	92.25	58.22
	Média:	**80.62**	**17.78**	**91.59**	**70.03**	**90.94**	**64.57**
rno	gse1001	72.14	9.21	85.46	70.58	84.92	67.64
	gse1156	80.75	19.50	87.79	76.53	87.76	76.17
	gse952	90.08	12.55	93.47	64.27	89.31	39.24
	Média:	**80.99**	**13.75**	**88.91**	**70.46**	**87.33**	**61.02**
Média geral:		**80.00**	**14.25**	**89.93**	**71.13**	**88.69**	**65.28**

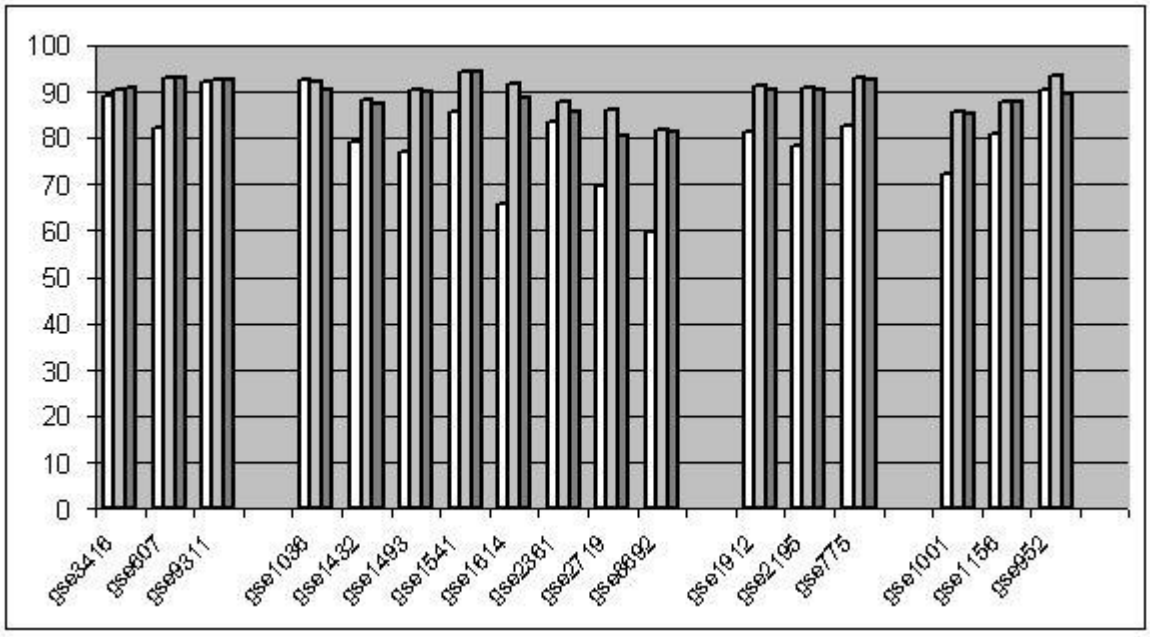


Figura 4-3: **Porcentagens de conservação da ordem de expressão nas séries de dados analisadas considerando-se as três classes de seqüências**.

Considerando-se que *A. thaliana* é um organismo bastante distinto dos outros três, as taxas de conservação mostraram-se praticamente inalteradas considerando as seqüências de manutenção, tanto se examinado os pares de amostras com maior conservação (“Máximo”) quanto os de menor conservação (“Mínimo”). Em relação às seqüências de não-manutenção, vê-se que *A. thaliana* apresenta maior conservação que os demais, o que pode ser explicado, fundamentalmente, pela existência de um número menor de estruturas “teciduais” diferentes neste organismo em relação aos demais, que possuem vasta gama de tecidos distintos, com funções completamente diversas dadas por diferentes padrões de expressão gênica.

Ainda que as seqüências de manutenção apresentem resultados aparentemente superiores, tais resultados foram produzidos pela análise apenas dos extremos, máximo e mínimo, em relação à conservação da expressão. Conforme mostrado a seguir para o processo de agrupamento, a conservação da ordem de expressão das MESs de não-manutenção produz resultados similares aos produzidos com as MESs de manutenção e todas.

4.5.2 Comparação do Desempenho da Similaridade de Conservação com outras Medidas de Similaridade

As tabelas Tabela 4-13 e Tabela 4-14, abaixo, são as correspondentes às tabelas Tabela 4-9 e Tabela 4-10 (Item 4.4.2) e mostram os resultados, em termos de taxas de acerto, entre métricas comumente utilizadas e a métrica de conservação da ordem de expressão das MESs.

Tabela 4-13: Taxas de acerto utilizando-se métricas de similaridade comumente utilizadas para agrupar dados em comparação àquelas utilizando a conservação da ordem de expressão nas três diferentes classes de MESs, para os critérios *mais severos* de agrupamento.

Organismo	Série	Manhattan	Euclidiana	Pearson	Spearman	Não-manutenção	Manutenção	Todas	Critério
ath 37 amostras	gse3416	11.11	11.11	11.11	11.11	33.33	22.22	16.67	tempo
	gse607	100	100	100	100	100	100	100	tecido
	gse9311	50	50	50	75	50	50	50	*condição*
	Média:	**53.7**	**53.7**	**53.7**	**62.04**	**61.11**	**57.41**	**55.56**	
hsa 125 amostras	gse1036	69.44	69.44	77.78	52.78	83.33	83.33	91.67	tempo
	gse1432	61.11	56.94	65.28	61.11	54.16	70.83	83.33	*doador*
	gse1493	66.67	33.33	100	100	100	100	100	tecido
	gse1541	0	0	0	0	0	0	0	*condição*
	gse1614	100	91.67	100	91.67	66.67	100	100	tempo
	gse2719	63.15	53.89	65	68.71	71.79	71.79	74.36	condição
	gse8692	41.67	75	75	50	75	75	83.33	condição
	Média:	**67.01**	**63.38**	**80.51**	**70.71**	**75.16**	**83.49**	**88.78**	
mmu 126 amostras	gse1912	93.33	89.33	85.33	85.33	88	88	84	tempo
	gse2195	11.9	11.9	14.29	23.81	45.24	21.43	11.9	*tempo*
	gse775	79.66	81.36	79.66	76.27	55.93	71.19	74.58	*tempo*
	Média:	**61.63**	**60.86**	**59.76**	**61.8**	**63.06**	**60.21**	**56.83**	
rno 170 amostras	*gse1001*	50.75	34.08	61.86	56.3	35.56	58.88	57.78	*tempo*
	gse1156	60	53.33	63.33	56.67	53.33	66.67	56.67	*tempo*

	gse952	94.26	89.34	91.8	92.62	94.26	88.52	95.08	tecido
	Média:	**68.34**	**58.92**	**72.33**	**68.53**	**61.05**	**71.36**	**69.84**	
Média geral:		**60.32**	**57.68**	**66.41**	**63.97**	**65.09**	**68.12**	**67.75**	

O primeiro dado que chama a atenção na Tabela 4-13 é que a acurácia (média) na utilização da métrica de conservação da ordem de expressão não é significativamente melhor, no geral, que aquela alcançada com a utilização das métricas tradicionais, sendo até mesmo pior para a classe de seqüências de não-manutenção já que ela é superada pela correlação de Pearson. Entretanto, as outras duas classes de seqüências superam as demais, mesmo que por uma margem baixa, sendo que a classe Todas apresentou os melhores resultados no geral e destacou-se nas séries de *H. sapiens*, com 89% de acertos em média. Entretanto, os resultados apresentados na Tabela 4-14 mostram uma diferenciação mais pronunciada.

Pode-se observar na Tabela 4-14 que a utilização da métrica de conservação supera, na média geral, as métricas tradicionais para as três classes de seqüências, sendo que a classe de não-manutenção apresentou resultado absoluto maior, porém, sem significância estatística sobre as outras duas classes. Tal melhoria no desempenho quando o critério de agrupamento muda reforça o fato de que as técnicas de *clustering* são fundamentalmente exploratórias, podendo resultar em *clusters* mais adequados ou não, dependendo do conceito ou aspecto sendo aprendido.

Tabela 4-14: Taxas de acerto utilizando-se métricas de similaridade comumente utilizadas para agrupar dados em comparação àquelas utilizando a conservação da expressão nas três diferentes classes de MESs, para os critérios *menos severos* de agrupamento.

Organismo	Série	Manhattan	Euclidiana	Pearson	Spearman	Não-manutenção	Manutenção	Todas	Critério
ath 37 amostras	gse3416	11.11	11.11	11.11	11.11	33.33	22.22	16.67	tempo
	gse607	100	100	100	100	100	100	100	tecido
	gse9311	50	50	50	75	100	100	100	*tecido*
	Média:	**53.7**	**53.7**	**53.7**	**62.04**	**77.78**	**74.07**	**72.22**	
hsa 125 amostras	gse1036	69.44	69.44	77.78	52.78	83.33	83.33	91.67	tempo
	gse1432	79.16	74.99	83.33	79.16	66.67	83.34	66.66	*condição*
	gse1493	66.67	33.33	100	100	100	100	100	tecido
	gse1541	93.33	93.33	93.33	93.33	95	90	90	*tempo*
	gse1614	100	91.67	100	91.67	66.67	100	100	tempo
	gse2719	63.15	53.89	65	68.71	71.79	71.79	74.36	condição
	gse8692	41.67	75	75	50	75	75	83.33	condição
	Média:	**70.02**	**66.39**	**83.52**	**73.72**	**79.78**	**86.21**	**86.57**	
mmu	gse1912	93.33	89.33	85.33	85.33	88	88	84	tempo

126 amostras	gse2195	75.39	75.39	77.78	87.3	38.1	16.67	11.9	*condição*
	gse775	96.05	97.75	96.05	92.66	84.76	86.45	89.83	*tecido*
	Média:	**88.26**	**87.49**	**86.39**	**88.43**	**70.28**	**63.71**	**61.91**	
rno 170 amostras	*gse1001*	53.71	37.04	64.82	59.26	94.44	88.89	88.89	*condição*
	gse1156	72.22	65.55	75.55	68.89	83.33	93.33	90	*condição*
	gse952	94.26	89.34	91.8	92.62	94.26	88.52	95.08	tecido
	Média:	**73.4**	**63.98**	**77.39**	**73.59**	**90.68**	**90.25**	**91.32**	
Média geral:		**71**	**68.05**	**76.78**	**74.34**	**79.63**	**78.56**	**78.01**	

A Tabela 4-15 mostra as taxas de acerto da utilização da métrica de conservação e dos algoritmos comumente utilizados, de modo similar ao exibido na Tabela 4-11. Na média, a métrica de conservação da ordem de expressão com as classes de seqüências de manutenção e todas supera os demais, enquanto que a utilização da classe de não-manutenção apresenta taxa de acerto superada pela do algoritmo *Farthest First*, ainda que tenha superado *K-Means* e EM.

Tabela 4-15: Comparação das acurácias obtidas pela utilização da taxa de conservação com *clustering* hierárquico e três algoritmos comumente utilizados

Séries	**Amostras nos grupos**	**Não-manutenção**	**Manutenção**	**Todas**	**K-Means**	**EM**	**Farthest First**
gse607	{3, 4, 4}	11 (100%)	11 (100%)	11 (100%)	8 (72,7%)	9 (81,8%)	11 (100%)
gse1493	{2, 2, 2}	6 (100%)	6 (100%)	6 (100%)	3 (50%)	6 (100%)	6 (100%)
gse1614	{4, 4, 4}	12 (66,67%)	12 (100%)	12 (100%)	9 (75%)	12 (100%)	11 (91,7%)
gse8692	{3, 3, 6}	9 (75 %)	9 (75 %)	10 (83,33 %)	9 (75%)	6 (50%)	9 (75%)
Média:	—	**85,41%**	**93,75%**	**95,83%**	**70.7%**	**80,5%**	**90,2%**

4.6 *Clustering* MESs *versus* Conservação da Ordenação

As tabelas Tabela 4-16 e Tabela 4-17 mostram os resultados da utilização das métricas MESs e de conservação da ordem de expressão no *clustering*. A abordagem MESs, com seqüências de não-manutenção, supera a abordagem de conservação da ordem de expressão em todos os organismos, com exceção das séries de *H. sapiens* (Tabela 4-16), em que a métrica de conservação com as seqüências "Todas" apresentou acurácia quase idêntica. Considerando-se os critérios de

agrupamento menos severos (Tabela 4-17), MESs com seqüências de não-manutenção supera a taxa de conservação da ordem de expressão em todos os organismos, sem exceção, apresentando uma acurácia maior em mais de 8 pontos percentuais em relação ao melhor desempenho apresentado pela concorrente (também seqüências de não-manutenção).

Tabela 4-16: Comparação da acurácia obtida pela utilização das MESs e da taxa de conservação das MESs, segundo os critérios *mais severos* de agrupamento.

		MESs			**Conservação MESs**			
Organismo	**Série**	**Não-manutenção**	**Manutenção**	**Todas**	**Não-manutenção**	**Manutenção**	**Todas**	**Critério**
ath 37 amostras	gse3416	38.89	22.22	16.67	33.33	22.22	16.67	tempo
	gse607	100	100	100	100	100	100	tecido
	gse9311	*50*	*50*	*50*	*50*	*50*	*50*	*condição*
	Média:	**62.96**	**57.41**	**55.56**	**61.11**	**57.41**	**55.56**	
hsa 125 amostras	gse1036	83.33	83.33	91.67	83.33	83.33	91.67	tempo
	gse1432	*58.33*	*58.33*	*58.33*	*54.16*	*70.83*	*83.33*	*doador*
	gse1493	100	100	100	100	100	100	tecido
	gse1541	*0*	*0*	*0*	*0*	*0*	*0*	*condição*
	gse1614	100	100	100	66.67	100	100	tempo
	gse2719	69.23	74.36	71.79	71.79	71.79	74.36	condição
	gse8692	83.33	91.67	83.33	75	75	83.33	condição
	Média:	**82.37**	**84.615**	**84.19**	**75.16**	**83.49**	**88.78**	
mmu 126 amostras	gse1912	88	88	88	88	88	84	tempo
	gse2195	*40.48*	*16.67*	*16.67*	*45.24*	*21.43*	*11.9*	*tempo*
	gse775	*76.27*	*76.27*	*71.19*	*55.93*	*71.19*	*74.58*	*tempo*
	Média:	**68.25**	**60.31**	**58.62**	**63.06**	**60.21**	**56.83**	
rno 170 amostras	*gse1001*	*46.67*	*53.33*	*46.67*	*35.56*	*58.88*	*57.78*	*tempo*
	gse1156	*83.33*	*56.67*	*53.33*	*53.33*	*66.67*	*56.67*	*tempo*
	gse952	95.08	86.07	95.08	94.26	88.52	95.08	tecido
	Média:	**75.03**	**65.36**	**65.03**	**61.05**	**71.36**	**69.84**	
Média geral:		**72.15**	**66.92**	**65.85**	**65.09**	**68.12**	**67.75**	

Tabela 4-17: Comparação da acurácia obtida pela utilização das MESs e da taxa de conservação das MESs, segundo os critérios *menos severos* de agrupamento.

		MESs			Conservação MESs			
Organismo	**Série**	**Não-manu-tenção**	**Manutenção**	**Todas**	**Não-manu-tenção**	**Manutenção**	**Todas**	**Critério**
ath 37 amostras	gse3416	38.89	22.22	16.67	33.33	22.22	16.67	tempo
	gse607	100	100	100	100	100	100	tecido
	gse9311	*100*	*100*	*100*	100	100	100	*tecido*
	Média:	**79.63**	**74.07**	**72.22**	**77.78**	**74.07**	**72.22**	
hsa 125 amostras	gse1036	83.33	83.33	91.67	83.33	83.33	91.67	tempo
	gse1432	*83.33*	*75*	*70.83*	66.67	83.34	66.66	*condição*
	gse1493	100	100	100	100	100	100	tecido
	gse1541	*100*	*90*	*90*	95	90	90	*tempo*
	gse1614	100	100	100	66.67	100	100	tempo
	gse2719	69.23	74.36	71.79	71.79	71.79	74.36	condição
	gse8692	83.33	91.67	83.33	75	75	83.33	condição
	Média:	**88.46**	**87.77**	**86.8**	**79.78**	**86.21**	**86.57**	
mmu 126 amostras	gse1912	88	88	88	88	88	84	tempo
	gse2195	*92.86*	*85.71*	*85.71*	38.1	16.67	11.9	*condição*
	gse775	*94.92*	*88.14*	*89.83*	84.76	86.45	89.83	*tecido*
	Média:	**91.93**	**87.28**	**87.85**	**70.28**	**63.71**	**61.91**	
rno 170 amostras	gse1001	*88.89*	*88.89*	*88.89*	94.44	88.89	88.89	*condição*
	gse1156	*93.33*	*66.67*	*70*	83.33	93.33	90	*condição*
	gse952	95.08	86.07	95.08	94.26	88.52	95.08	tecido
	Média:	**92.43**	**80.54**	**84.66**	**90.68**	**90.25**	**91.32**	
Média geral:		**88.11**	**82.42**	**81.83**	**79.63**	**78.56**	**78.01**	

4.7 *Clustering* de Amostras de Tecidos Cancerosos com Normais

Algumas séries utilizadas na produção dos resultados aqui apresentados são de estudos envolvendo alguns tipos de câncer: gse1036 (leucemia), gse1982 (câncer renal), gse8692 (tumores cerebrais). A fim de analisar a similaridade entre tipos de câncer diferentes, bem como entre eles e células ou tecidos em condições diferenciadas e "normais", estas três séries e as séries gse1493 (células-tronco hematopoiéticas), gse1614 (diferenciação intestinal) e gse2361 (tecidos normais), todas de *H. sapiens*, foram agrupadas utilizando-se a abordagem MESs e de conservação de ordem de expressão. Para permitir o agrupamento de amostras com identificadores diferentes, as seqüências presentes nas amostras das várias séries de dados foram agrupadas em seus grupos Unigene e, depois, submetidas ao *clustering*, da seguinte forma: existem *probe sets* nos *genechips* que representam variantes de *splicing* de um mesmo "gene", por exemplo, mas todos os *probe sets* possuem

um Unigene associado. Desta forma, as seqüências (*probe sets*) com um mesmo Unigene foram agrupadas e consideradas como uma seqüência única. O sinal correspondente à expressão foi atribuído como sendo a média aritmética dos sinais das seqüências componentes. Após este processamento, as seqüências de cada amostra de cada série podem ser ordenadas, as MESs determinadas e utilizadas no *clustering* (as MESs neste experimento são unigenes).

4.7.1 *Clustering* Utilizando MESs

Em todas as execuções realizadas (para cada classe de seqüências), cada amostra foi sempre agrupada a uma amostra de sua própria série de dados, ou seja, de um mesmo tipo de tecido em mesmas condições experimentais, reforçando a idéia de utilização da tecnologia de *microarrays* como uma ferramenta no auxílio à determinação da classe de um tipo de câncer (Golub *et al*., 1999). Além disso, o agrupamento de amostras de séries diferentes produziu informações interessantes, como exposto a seguir.

Considerando apenas seqüências de não-manutenção, o agrupamento foi feito com baixo compartilhamento de seqüências [o par com maior taxa de seqüências comuns compartilha apenas 24% de MESs: uma amostra de células leucêmicas (gse1036) com fígado fetal (gse2361)]. Entretanto, um fato marcante é o agrupamento de onze das doze amostras da série gse1036 (células leucêmicas) à amostra de fígado fetal normal, sendo que a amostra restante foi agrupada ao útero normal. Em relação às seqüências de manutenção é notável que todas as 36 amostras de tecidos normais da série gse2361 compartilhem entre 51% (músculo esquelético) e 62% (pulmão) de suas MESs com alguma amostra de tumor cerebral da série gse8692 e que a taxa aumenta de 62% para 70% (medula espinhal) quando todas as MESs são consideradas. É também digno de nota que as amostras da linhagem celular CD34, células tronco hematopoiéticas (gse1493) agrupam-se preferencialmente às amostras de células em diferenciação intestinais (série gse1614).

4.7.2 *Clustering* Utilizando Conservação da Ordenação MESs

De modo similar aos resultados da abordagem MESs, as amostras das séries agruparam-se primeiramente às amostras da própria série. Entretanto, as taxas de conservação da ordem de expressão foram significativamente altas em alguns casos.

Considerando as seqüências de manutenção, as 36 amostras de tecidos normais (série gse2361) foram agrupadas a amostras de tecidos de tumores cerebrais (série gse8692) com taxas de conservação variando entre 68.48% (coração-gliobastoma) e 81.45% (medula espinhal-glioma). Em seguida, amostras de células leucêmicas (gse1036) foram agrupadas a amostras de tumores cerebrais (gse8692) com taxas de conservação entre 61.29% (células leucêmicas-oligodendroglioma) e 68% (células leucêmicas-glioblastoma) e as amostras da linhagem de células-tronco CD34 hematopoiéticas agruparam-se com amostras de células em diferenciação intestinais (gse1614) com taxas de conservação entre 58.2% (Linhagem +CD34+-células em proliferação) e 60.14% (Linhagem -CD34+-células em proliferação). As amostras da série gse1982 (câncer renal) foram as que apresentaram menores taxas de conservação da ordem de expressão para as seqüências de manutenção, porque essa série possui apenas 151 unigenes presentes em todas as amostras da mesma, número extremamente baixo se comparado aos das demais séries. Entretanto, é importante observar que as taxas mais baixas foram todas com amostras de tecidos leucêmicos (gse1036), entre 0.62% e 1.46%, seguidas das amostras de tecidos normais (gse2361), com taxas entre 1.61% (testículo) e 2.15% (tireóide). Já a taxa de conservação mais alta desta série foi com uma amostra de tecido leucêmico da série gse1493 com taxa de 5.45% (Linhagem –CD34+).

Em termos das seqüências de não-manutenção, as taxas de conservação são bastante baixas quando comparadas às taxas relativas às seqüências de manutenção no tocante ao agrupamento de amostras de séries diferentes. As amostras da série gse1036 (leucemia) foram agrupadas a amostras de câncer renal (série gse1982) com taxas de conservação entre 35.39% e 42.69%, sendo estas as de maior conservação da ordem de expressão. Em termos das menores taxas de conservação não há um padrão claro marcante nos dados, sendo que amostras de todas as séries aparecem nos pareamentos formados. Considerando os agrupamentos de amostras de uma mesma série, é notável que amostras da série gse1982 (câncer renal) possuam taxas de conservação de ordem de expressão extremamente altas, sendo a mais alta delas de 96.16%. Entre os pareamentos com mais de 90% de conservação (270 pares de amostras), por exemplo, 95% (256 pares) são de amostras dessa série, sendo os outros 14 da série gse1036 (leucemia).

4.8 Discussão

Duas novas métricas de similaridade foram apresentadas, bem como resultados de suas aplicações no *clustering* de séries de dados de experimentos disponíveis publicamente. As abordagens baseadas

nas MESs (até onde é reportado na literatura) são as primeiras métricas de similaridade que exploram a noção biológica de que genes mais expressos teriam papel proeminente na fisiologia celular e, por conseqüência, na fisiologia do tecido, órgão ou mesmo do organismo como um todo. Tal proeminência sugere que tais genes podem ser utilizados para caracterizar amostras inteiras de dados de expressão gênica, o que pode ser confirmado pelos resultados aqui apresentados. É plausível admitir que genes regulatórios, expressos em quantidades menores, apesar de demonstrarem alta especificidade a cada amostra tecidual, por exemplo, não seriam tão úteis para a quantificação de similaridades utilizadas no agrupamento. Isso se deve, principalmente, por causa da menor confiabilidade dos dados de expressão gerados em larga escala pelas tecnologias disponíveis atualmente.

A utilização de subconjuntos da amostra não é uma novidade. Golub *et al*. (1999) afirmam que uma fração menor que 5% dos genes expressos em uma amostra de dados de expressão gênica é suficiente para caracterizá-la. Xu e Zhang (2005) utilizam 50 genes "virtuais" na caracterização. Tibshirani *et al*. (2002) reduzem o número de genes informativos para 43, enquanto que Alon *et al*. (1999) utilizam 2000 genes. Em todos os trabalhos, com exceção do último, é reportado que a seleção dos genes ou seqüências utilizadas melhorou a acurácia do método utilizado. Considerando os resultados apresentados na Tabela 4-6 para a classe de seqüências de não-manutenção, há uma concordância com a afirmativa de Golub *et al*., na média, sendo que algumas séries apresentam porcentagens claramente superiores ao limiar de 5%, mas a maioria está abaixo. Nota-se uma certa relação entre o número de amostras, o número total de seqüências e a porcentagem para as seqüências de não-manutenção: duas séries com número alto de amostras (> 100) e com número total de seqüências abaixo da média, gse1982 e gse952, apresentam porcentagens de 12,71% e 13,75%, respectivamente. Para as outras duas classes de seqüências a porcentagem média foi cerca de três vezes maior. Considerando-se que estas últimas são as classes de "seqüências de manutenção" e "todas", tais números confirmam que as seqüências de manutenção são mais expressas que as demais, o que está de acordo com resultados anteriores (Mudado & Ortega, 2006; Pinto & Ortega, 2007).

A utilização da métrica de similaridade de MESs com as seqüências de não-manutenção proporcionou uma acurácia de mais de 88% no geral. Tal acurácia representa um ganho de mais de 11% em relação à métrica tradicional de melhor desempenho (correlação de Pearson) e de mais de 20% em relação à distância euclidiana (Tabela 4-10), considerando critérios *menos severos* de *clustering*. Considerando os critérios *mais severos* (Tabela 4-9), o ganho foi menor (cerca de 6%), sendo que a acurácia foi bastante reduzida: apenas 72%. De forma similar, considerando-se um limiar de 92% de acurácia, a similaridade de MESs também foi superior às demais, alcançando tal limiar para quatro séries (critérios *mais severos*) e para nove séries (critérios *menos severos*). Além disso, considerando-

se acurácia total, o que é importante em situações onde não deve haver erros, como em diagnósticos, MESs empata com a correlação de Pearson para critérios *mais severos* (três séries) e a supera nos critérios *mais fáceis* (cinco séries contra três). Além disso, o *clustering* utilizando MESs superou outros algoritmos de *clustering* tradicionais em taxas que variam de 5% (para o *Farthest First*) a 25% (para o *K-Means*), como mostrado na Tabela 4-11. Tais resultados claramente indicam superioridade da métrica de similaridade de MESs sobre as tradicionais. Além disso, o *clustering* hierárquico utilizando MESs produziu resultados mais precisos, do ponto de vista fisiológico, do que o produzido utilizando-se a correlação de Pearson (Figura 4-2). Todos os tecidos do sistema nervoso central, por exemplo, foram agrupados entre si primeiramente, o que não aconteceu com a utilização da correlação de Pearson, que produziu resultados no mínimo inesperados, como o agrupamento de testículo com tecidos do sistema nervoso central.

Eisen *et al*. (1998) afirmam que o coeficiente de correlação padrão (o produto escalar de dois vetores normalizados) se conforma bem com a noção biológica do significado de dois genes serem ditos "co-regulados". Entretanto, é difícil expandir tal noção para determinar a similaridade entre duas amostras de dados de expressão gênica. Além disso, como as amostras devem ser caracterizadas pela expressão das suas seqüências constituintes e como é sabido que diferentes tipos celulares expressam diferentes genes, a utilização de métricas de similaridade que trabalham com um mesmo conjunto de características (seqüências ou genes) para todas as amostras não se mostra a mais adequada. (Ben-Dor *et al*., 2000) alcançaram 88.7% de acurácia na classificação baseada em *clustering* de amostras em duas classes (um problema relacionado ao deste trabalho) utilizando a correlação de Pearson como métrica de similaridade. Os autores utilizaram o CAST (Ben-Dor, Shamir & Yakhini, 1999) para agrupar as amostras. O CAST utiliza um *limiar* arbitrário para decidir a qual *cluster* uma amostra pertence. Um aprimoramento do desempenho foi feito com a utilização de rotulação das amostras em grupos antes do processo de modo supervisionado. Entretanto, o CAST aplicado nos dados de Alon *et al*. (1999) produziu resultados similares (Ben-Dor, Shamir & Yakhini, 1999), ou seja, uma acurácia próxima e abaixo de 90% para a distinção entre duas classes ou *clusters*. Alon *et al*. (1999) aplicaram o *clustering* em 62 amostras de tecidos sendo 40 com câncer de colon e 22 de tecido normal. Os autores utilizaram os 2000 genes com os maiores valores *mínimos* de expressão *em todas as amostras* para o *clustering*. Foi aplicado um algoritmo de *clustering* hierárquico por particionamento do conjunto de dados de forma binária (grosseiramente, o contrário do *clustering* hierárquico utilizado no presente trabalho), chamado de *deterministic annealing algorithm* (Alon *et al*., 1999). A acurácia obtida foi de 87.1%. Golub *et al*. (1999) utilizaram *Self-Oganizing Maps* e obtiveram 89,47% de acertos. Os resultados da utilização da similaridade de MESs aqui apresentados foram equivalentes, em valores absolutos, aos reportados por tais trabalhos. Entretanto, tais resultados

foram obtidos pela aplicação das diversas metodologias em apenas um conjunto de amostras. Os resultados aqui apresentados são possivelmente mais relevantes, desde que os dados utilizados perpassam quatro organismos distintos, 18 séries de dados e um número de amostras quase dez vezes maior que o reportado em tais trabalhos.

Tibshirani *et al*. (2002) apresentam a utilização de "centróides compactados" (*shrunken centroids*), o que reduz o número de genes informativos para apenas 43 para uma determinada série de dados composta por 88 amostras. Os autores apenas informam que o método pode ser aplicado ao *clustering* sem prover quaisquer resultados. O objetivo principal do trabalho foi reduzir ao máximo o número de genes necessários à classificação. Entretanto, mesmo que a redução alcance 100% de acurácia, há o risco de perda de genes biologicamente relevantes.

Qualquer abordagem ao problema de *clustering* de amostras de dados de expressão gênica que faça a redução ou a seleção de características deve ser tratada de modo cuidadoso em aplicações onde os genes participantes de determinado processo biológico investigado devem ser determinados (por exemplo, no *clustering* de diferentes tipos ou subtipos de câncer). Genes relevantes para a investigação e futuro desenvolvimento de fármacos podem ser deixados de fora, por exemplo. Tal fato pode ocorrer, também, com as abordagens aqui apresentadas que consideram apenas a fração dos genes que são mais expressos nas amostras analisadas. Entretanto, dadas as limitações das tecnologias atualmente disponíveis para a geração de dados de expressão gênica, a confiabilidade dos dados é tanto menor quanto menor for a "expressão" dos genes ou seqüências. Assim sendo, a seleção das seqüências ou genes mais expressos privilegia o fato de que a "expressão" é mais confiavelmente quantificada para estes, dadas as atuais tecnologias de geração de dados em larga escala.

A utilização da métrica de similaridade de conservação da ordenação MESs produziu resultados inferiores ao da similaridade de MESs, tanto nos critérios *mais severos* quanto nos *menos severos* (Tabela 4-16 e 4-17), no geral. Apenas nos critérios *mais severos* com as "seqüências de manutenção" e "todas" é que a conservação da ordenação foi ligeiramente superior (cerca de 1%) — Tabela 4-16. Nas demais situações ela foi superada, sendo maior a diferença para as seqüências de não-manutenção: 9% (Tabela 4-17). Em comparação às métricas tradicionais, a conservação da ordenação apresentou resultados superiores em geral, porém com margens menores de melhoria [máximo de 79% de acurácia, o que representa 11% a mais que a pior métrica tradicional (distância euclidiana) e apenas 3% a mais que a melhor tradicional (correlação de Pearson)], sendo mesmo superada pela correlação de Pearson (apenas seqüências de não-manutenção) em um ponto percentual (Tabela 4-13). Tal melhoria discreta não invalida a utilização da taxa de conservação da ordenação MESs

como métrica de similaridade, já que o *clustering* é uma forma de aprendizagem eminentemente exploratória e a taxa de conservação provê informações interessantes relacionando as amostras por si só. Além disso, agrupar réplicas de amostras por meio da conservação da ordenação permite avaliar de modo mais claro se a tecnologia empregada na geração dos dados de expressão é reprodutiva no sentido de que os experimentos podem ser reproduzidos de forma confiável. Isso é possível, pois a métrica de similaridade de conservação da ordenação MESs informa o número de pares de seqüências mantém sua ordem relativa de expressão. Em duas réplicas do mesmo experimento, espera-se idealmente que a taxa de conservação seja total (100%). Caso isso não ocorra, então imprecisões ocorreram em um dos dois pontos listados acima.

É comum na análise de dados de expressão gênica que seja feita a remoção de *outliers* de modo que quaisquer seqüências que estejam acima ou abaixo de três vezes o desvio-padrão da média sejam descartadas, assumindo uma distribuição normal dos dados (Glaura Franco, UFMG, comunicação pessoal, 2006/2007; Alon *et al*., 1999). Já Ge *et al*. (2005) simplesmente descartaram 5% das seqüências mais e menos expressas das amostras. Nesse caso, as amostras são de tecidos "normais" de *H. sapiens*. Entretanto, a eliminação de 5% das seqüências mais expressas representa cerca de 50% das MESs, de acordo com os dados da Tabela 6 para a série gse2361, considerando-se as classes de "seqüências de não-manutenção" e "todas", mas tais seqüências foram utilizadas no cálculo da métrica de similaridade MESs.

As duas métricas de distância (ou similaridade) entre amostras apresentadas aqui são prontamente aplicáveis a quaisquer outros tipos de medidas de expressão gênica, pois os sinais são convertidos em ranqueamento. Este tratamento torna os procedimentos aqui desenvolvidos mais facilmente aplicáveis. Todo o algoritmo é prontamente programável em qualquer linguagem, sendo que os *bytecodes* Java já foram utilizados por outros membros de nosso grupo de pesquisa. Estudos envolvendo amostras de SAGE e freqüências de ESTs estão sendo desenvolvidos com as métricas apresentadas neste trabalho e os programas desenvolvidos.

5 Conclusões e Perspectivas

Eu, Jesus, enviei o meu anjo, para vos testificar estas coisas nas igrejas.
Eu sou a Raiz e a Geração de Davi, a resplandecente Estrela da manhã.
E o Espírito e a esposa dizem: Vem!
E quem ouve diga: Vem
E quem tem sede venha; e quem quiser tome de graça da água da vida.
Apocalipse 22.16-17

Eis que estou à porta e bato;
se alguém ouvir a minha voz e abrir a porta
entrarei em sua casa e com ele cearei, e ele, comigo.
Apocalipse 3.20

O *clustering* de amostras de dados de expressão gênica é uma técnica exploratória que provê informações importantes sobre os fenótipos sendo analisados. Tradicionalmente, métricas de similaridade já utilizadas em outras áreas de conhecimento, como medidas de distância e de correlação entre os dados são utilizadas para agrupar amostras inteiras de dados de expressão. Este trabalho representa um passo inicial de uma abordagem que busca inspiração biológica para novas métricas de similaridade baseadas na expressão gênica de tecidos, órgãos e mesmo organismos como um todo.

As duas métricas de similaridade apresentadas são inspiradas em duas premissas simples: (i) os genes mais expressos em uma amostra devem ser capazes de caracterizá-la e (ii) a ordem de expressão dos genes mais expressos deve ser mais conservada em amostras mais parecidas. O presente trabalho representa apenas duas abordagens simples que exploram essas duas premissas. Os resultados apresentados são satisfatórios no sentido de que eles superam aqueles produzidos pelas métricas tradicionais e, para a métrica de similaridade MESs, eles são similares aos de outras abordagens que exploram técnicas estatísticas e matemáticas avançadas, com a vantagem de prover significado biológico para a similaridade.

Correntemente, dados de ESTs estão sendo analisados por meio do *clustering* utilizando a similaridade MESs. Espera-se aplicar e analisar dados de SAGE. Além disso, uma aplicação interessante seria aplicar o *clustering* MESs em amostras de dados de expressão de diferentes organismos, o que permitiria compará-los em relação à expressão gênica. Uma maneira de se alcançar tal objetivo seria explorar a homologia das seqüências presentes nas diversas plataformas a fim de relacionar as seqüências e permitir a aplicação das métricas.

A Definição dos *Clusters* de Amostras nos Dados Utilizados

As tabelas abaixo listam as séries de dados utilizadas neste trabalho bem como os *clusters* de suas amostras. Para cada série é informado o seu código de acesso ao banco de dados GEO do NCBI, o organismo correspondente. Em seguida é informado o número de *clusters* esperados e, finalmente, cada *cluster* é discriminado por um número seguido do nome da amostra como disponível nos arquivos em formato "soft" disponibilizados publicamente pelo NCBI.

Clusters Severos

gse607	ath	GSE1912	mmu
3		8	
0	Leaf_GC2	0	Postnatal day 1, mouse 1
0	Leaf_GH1	0	Postnatal day 1, mouse 2
0	Leaf_GH2	0	Postnatal day 1, mouse 3
1	STEM_GC7	1	Postnatal day 6, mouse 1
1	STEM_GC8	1	Postnatal day 6, mouse 2
1	STEM_GH7	1	Postnatal day 6, mouse 3
1	STEM_GH8	2	Postnatal day 14, mouse 1
2	FLOWER_GC5	2	Postnatal day 14, mouse 2
2	FLOWER_GC6	2	Postnatal day 14, mouse 3
2	FLOWER_GH5	3	Postnatal day 17, mouse 1
2	FLOWER_GH6	3	Postnatal day 17, mouse 2
END;		3	Postnatal day 17, mouse 3
gse3416	ath	4	Postnatal day 23, mouse 1
6		4	Postnatal day 23, mouse 2
0	00h Col-0 replicate A	4	Postnatal day 23, mouse 3
0	00h Col-0 replicate B	5	Postnatal 9-week, mouse 1
0	00h Col-0 replicate C	5	Postnatal 9-week, mouse 2
1	04h Col-0 replicate A	5	Postnatal 9-week, mouse 3
1	04h Col-0 replicate B	6	Postnatal 5-month, mouse 1
1	04h Col-0 replicate C	6	Postnatal 5-month, mouse 2
2	08h Col-0 replicate A	6	Postnatal 5-month, mouse 3
2	08h Col-0 replicate B	7	Postnatal 1-year, mouse 1
2	08h Col-0 replicate C	7	Postnatal 1-year, mouse 2
3	12h Col-0 replicate A	7	Postnatal 1-year, mouse 3
3	12h Col-0 replicate B	7	Postnatal 1-year, mouse 4
3	12h Col-0 replicate C	END;	

4 16h Col-0 replicate A
4 16h Col-0 replicate B
4 16h Col-0 replicate C
5 20h Col-0 replicate A
5 20h Col-0 replicate B
5 20h Col-0 replicate C
END;
GSE9311 ath
2
0 Root control rep 1
0 Root control rep 2
0 Root Selenate rep 1
0 Root Selenate rep 2
1 Shoot control rep1
1 Shoot control rep2
1 Shoot Selenate rep1
1 Shoot Selenate rep2
END;
gse775 mmu
6
0 PGA-lv_1h_511
0 PGA-lv_1h_512
0 PGA-lv_1h_514
0 PGA-lv2_1h_514
0 PGA_MI_ilv_1h_392
0 PGA_MI_ilv_1h_394
0 PGA_MI_ilv_1h_395
0 PGA_MI_nilv_1h_392
0 PGA_MI_nilv_1h_394
0 PGA_MI_nilv_1h_395
1 PGA-lv_1w_651
1 PGA-lv_1w_672
1 PGA-lv_1w_674
1 PGA-lv2_1w_674
1 PGA_MI_ilv_1w_662
1 PGA_MI_ilv_1w_663
1 PGA_MI_ilv_1w_670
1 PGA_MI_nilv_1w_662
1 PGA_MI_nilv_1w_663
1 PGA_MI_nilv_1w_670
2 PGA-lv_24h_510
2 PGA-lv_24h_515
2 PGA-lv_24h_517
2 PGA-lv2_24h_517
2 PGA_MI_ilv_24h_360
2 PGA_MI_ilv_24h_361
2 PGA_MI_ilv_24h_363
2 PGA_MI_nilv_24h_360
2 PGA_MI_nilv_24h_361
2 PGA_MI_nilv_24h_362
3 PGA-lv_48h_518
3 PGA-lv_48h_519
3 PGA-lv_48h_520
3 PGA-lv2_48h_519
3 PGA_MI_ilv_48h_351
3 PGA_MI_ilv_48h_354
3 PGA_MI_ilv_48h_355
GSE2195 mmu
7
0 AO 1hr
0 CTL2 AO 1hr
0 TMRI AO 1hr
0 E2 1hr
0 CTL2 E2 1hr
0 TMRI E2 1hr
1 AO 2hr
1 CTL2 AO 2hr
1 TMRI AO 2hr
1 E2 2hr
1 CTL2 E2 2hr
1 TMRI E2 2hr
2 AO 4hr
2 CTL2 AO 4hr
2 TMRI AO 4hr
2 E2 4hr
2 CTL2 E2 4hr
2 TMRI E2 4hr
3 AO 8hr
3 CTL2 AO 8hr
3 TMRI AO 8hr
3 E2 8hr
3 CTL2 E2 8hr
3 TMRI E2 8hr
4 AO 24hr
4 CTL2 AO 24hr
4 TMRI AO 24hr
4 E2 24hr
4 CTL2 E2 24hr
4 TMRI E2 24hr
5 AO 48hr
5 CTL2 AO 48hr
5 TMRI AO 48hr
5 E2 48hr
5 CTL2 E2 48hr
5 TMRI E2 48hr
6 AO 72hr
6 CTL2 AO 72hr
6 TMRI AO 72hr
6 E2 72hr
6 CTL2 E2 72hr
6 TMRI E2 72hr
END;
GSE2719 hsa
23
0 brain
1 stomach
2 colon
3 pancreas
4 prostate
5 skin
6 small intestine
7 adrenal
8 connective tissue
9 heart

3 PGA_MI_nilv_48h_351
3 PGA_MI_nilv_48h_354
3 PGA_MI_nilv_48h_355
4 PGA-lv_4h_506
4 PGA-lv_4h_507
4 PGA-lv_4h_509
4 PGA-lv2_4h_506
4 PGA_MI_ilv_4h_389
4 PGA_MI_ilv_4h_390
4 PGA_MI_ilv_4h_391
4 PGA_MI_nilv_4h_389
4 PGA_MI_nilv_4h_390
4 PGA_MI_nilv_4h_391
5 PGA-lv_8w_329
5 PGA-lv_8w_339
5 PGA-lv_8w_340
5 PGA_MI_ilv_8w_311
5 PGA_MI_ilv_8w_326
5 PGA_MI_ilv_8w_332
5 PGA_MI_nilv_8w_311
5 PGA_MI_nilv_8w_326
5 PGA_MI_nilv_8w_332
END;
GSE1001 rno
6
0 Normal Rat Retina
0 Normal Rat Retina 2
0 Normal Rat Retina 3
1 Injured 4h Rat Retina
1 Injured 4h Rat Retina 2
1 Injured 4h Rat Retina 3
2 Injured 1 day Rat Retina 1
2 Injured 1 day Rat Retina 2
2 Injured 1 day Rat Retina 3
3 Injured 3 day Rat Retina 1
3 Injured 3 day Rat Retina 2
3 Injured 3 day Rat Retina 3
4 Injured 7 day Rat Retina 1
4 Injured 7 day Rat Retina 2
4 Injured 7 day Rat Retina 3
5 Injured 30 day Rat Retina 1
5 Injured 30 day Rat Retina 2
5 Injured 30 day Rat Retina 3
END;
GSE1036 hsa
6
0 HeminTimecourse_0hA
0 HeminTimecourse_0hB
1 HeminTimecourse_6hA
1 HeminTimecourse_6hB
2 HeminTimecourse_12hA
2 HeminTimecourse_12hB
3 HeminTimecourse_24hA
3 HeminTimecourse_24hB
4 HeminTimecourse_48hA
4 HeminTimecourse_48hB
5 HeminTimecourse_72hA
10 kidney
11 liver
12 lung
13 skeletal muscle
14 spleen
15 fibrosarcoma 1
15 fibrosarcoma 2
15 fibrosarcoma 3
15 fibrosarcoma 4
15 fibrosarcoma 5
15 fibrosarcoma 6
15 fibrosarcoma 7
16 GIST 1
16 GIST 2
17 Leiomyosarcoma 1
17 Leiomyosarcoma 2
17 Leiomyosarcoma 3
17 Leiomyosarcoma 4
17 Leiomyosarcoma 5
17 Leiomyosarcoma 6
18 Lipo dediff 1
18 Lipo dediff 2
18 Lipo dediff 3
18 Lipo dediff 4
19 Lipo pleo 1
19 Lipo pleo 2
19 Lipo pleo 3
20 MFH 1
20 MFH 2
20 MFH 3
20 MFH 4
20 MFH 5
20 MFH 6
20 MFH 7
20 MFH 8
20 MFH 9
21 Round cell 1
21 Round cell 2
21 Round cell 3
21 Round cell 4
22 Synovial sarcoma 1
22 Synovial sarcoma 2
22 Synovial sarcoma 3
22 Synovial sarcoma 4
END;
GSE952 rno
27
0 frontal cortex wistar3658
0 frontal cortex wistar3659
0 frontal cortex wistar3660
0 frontal cortex wistar3662
0 frontal cortex wistar 3920
0 frontal cortex wistar 3921
0 frontal cortex wistar 3922
0 frontal cortex wistar 3923
0 frontal cortex wistar 3924
0 frontal cortex wistar 3925

5	HeminTimecourse_72hB	0	frontal cortex wistar ky3926
END;		0	frontal cortex wistar ky3927
GSE1156	rno	0	frontal cortex wistar ky3928
5		0	frontal cortex wistar ky3929
0	NINDS-RHS-Ctr_1h-1aUA-s2	0	frontal cortex wistar ky3930
0	NINDS-RHS-Ctr_1h-2aUA-s2	0	frontal cortex wistar ky3931
0	NINDS-RHS-Ctr_1h-3aUA-s2	0	frontal cortex sprague3932
0	NINDS-RHS-Kainate_1h-1aUA-s2	0	frontal cortex sprague3933
0	NINDS-RHS-Kainate_1h-2aUA-s2	0	frontal cortex sprague3934
0	NINDS-RHS-Kainate_1h-3aUA-s2	0	frontal cortex sprague3935
1	NINDS-RHS-Ctr_240h-1aUA-s2	0	frontal cortex sprague3936
1	NINDS-RHS-Ctr_240h-2aUA-s2	0	frontal cortex sprague3937
1	NINDS-RHS-Ctr_240h-3aUA-s2	1	amygdala wistar3722
1	NINDS-RHS-Kainate_240h-1aUA-s2	1	amygdala wistar3723
1	NINDS-RHS-Kainate_240h-2aUA-s2	1	amygdala wistar3724
1	NINDS-RHS-Kainate_240h-3aUA-s2	1	amygdala wistar3942
2	NINDS-RHS-Ctr_24h-1aUA-s2	1	amygdala, central nucleus sprague1
2	NINDS-RHS-Ctr_24h-2aUA-s2	1	amygdala sprague3950
2	NINDS-RHS-Ctr_24h-3aUA-s2	1	amygdala wistar ky3946
2	NINDS-RHS-Kainate_24h-1aUA-s2	1	amygdala, central nucleus sprague2
2	NINDS-RHS-Kainate_24h-2aUA-s2	1	amygdala wistar 3721
2	NINDS-RHS-Kainate_24h-3aUA-s2	1	amygdala wistar 3720
3	NINDS-RHS-Ctr_6h-1aUA-s2	2	hippocampus wistar3772
3	NINDS-RHS-Ctr_6h-2aUA-s2	2	hippocampus wistar ky3956
3	NINDS-RHS-Ctr_6h-3aUA-s2	2	hippocampus sprague3957
3	NINDS-RHS-Kainate_6h-1aUA-s2	3	cerebellum wistar3884
3	NINDS-RHS-Kainate_6h-2aUA-s2	3	cerebellum wistar3885
3	NINDS-RHS-Kainate_6h-3aUA-s2	3	cerebellum wistar3886
4	NINDS-RHS-Ctr_72h-1aUA-s2	3	cerebellum wistar3887
4	NINDS-RHS-Ctr_72h-2aUA-s2	3	cerebellum wistar3888
4	NINDS-RHS-Ctr_72h-3aUA-s2	3	cerebellum wistar3889
4	NINDS-RHS-Kainate_72h-1aUA-s2	3	cerebellum wistar ky3890
4	NINDS-RHS-Kainate_72h-2aUA-s2	3	cerebellum wistar ky3891
4	NINDS-RHS-Kainate_72h-3aUA-s2	3	cerebellum wistar ky3892
END;		3	cerebellum wistar ky3893
GSE1432	hsa	3	cerebellum wistar ky3894
4		3	cerebellum sprague3896
0	IFNG Microglia Sample B18 1hr	3	cerebellum sprague3897
0	IFNG Microglia Sample B18 6hrs	3	cerebellum sprague3898
0	IFNG Microglia Sample B18 24hrs	3	cerebellum sprague3899
0	Control Microglia Sample B18 1hr	3	cerebellum sprague3900
0	Control Microglia Sample B18 6hrs	3	cerebellum sprague3901
0	Control Microglia Sample B18 24hrs	4	cerebral cortex wistar 3902
1	IFNG Microglia Sample O 1hr	4	cerebral cortex wistar 3903
1	IFNG Microglia Sample O 6hrs	4	cerebral cortex wistar 3904
1	IFNG Microglia Sample O 24hrs	4	cerebral cortex wistar ky3910
1	Control Microglia Sample O 1hr	4	cerebral cortex wistar ky3911
1	Control Microglia Sample O 6hrs	4	cerebral cortex wistar ky3912
1	Control Microglia Sample O 24hrs	4	cerebral cortex wistar ky3913
2	IFNG Microglia Sample W 1hr	4	cerebral cortex sprague3914
2	IFNG Microglia Sample W 6hrs	4	cerebral cortex sprague3915
2	IFNG Microglia Sample W 24hrs	4	cerebral cortex sprague3916
2	Control Microglia Sample W 1hr	4	cerebral cortex sprague3917
2	Control Microglia Sample W 6hrs	4	cerebral cortex sprague3918
2	Control Microglia Sample W 24hrs	4	cerebral cortex sprague3919
3	IFNG Microglia Sample Y20 1hr	5	ventral striatum wistar 3958
3	IFNG Microglia Sample Y20 6hrs	5	ventral striatum wistar ky3959

3	IFNG Microglia Sample Y20 24hrs
3	Control Microglia Sample Y20 1hr
3	Control Microglia Sample Y20 6hrs
3	Control Microglia Sample Y20 24hrs
END;	
GSE1493	hsa
3	
0	lin+CD34+ I replicate
0	lin+CD34+ II replicate
1	Lin-CD34+ I replicate
1	lin-CD34+ II replicate
2	lin-CD34- I replicate
2	Lin-CD34- II replicate
END;	
GSE1541	hsa
5	
0	mRNA from A549 Static Control 1hr Jan 31 2003
0	RNA from A549 Cells - Static Control 1hr Replicate 1
0	A549 RNA Static Control 4hr Jan 31 2003
0	RNA A549 cells Static Control 4hr Replicate 1
1	A549 LPS 1h Jan 31 2003
1	A549 LPS 1hr Oct 21 2002
1	A549 LPS 4hr Jan 31 2003
1	A549 LPS 4hr Nov 21 2002
2	A549 Stretch 1hr Jan 31 2003
2	A549 cells Stretch 1hr Oct 20 2002
2	A549 Stretch 4hr Nov 21 2002
2	A549 stretch 4hr Jan 31 2003
3	A549 TNF 1hr Jan 31 2003
3	A549 TNF 1hr Oct 20 2002
3	A549 TNF 4hr Jan 31 2003
3	A549 TNF 4hr Nov 21 2002
4	A549 TNF+Stretch 1hr Jan 31 2003
4	A549 TNF+Stretch Oct 20 2002
4	A549 TNF+Stretch 4hr Jan 31 2003
4	A549 TNF+Stretch 4hr Nov 21 2002
END;	
GSE1614	hsa
3	
0	JF2dR1
0	JF2dR2
0	JF2dR3
0	JF2dR4
1	JF8dR1
1	JF8dR2
1	JF8dR3
1	JF8dR4
2	JF15dR1
2	JF15dR2
2	JF15dR3
2	JF15dR4
END;	

5	ventral striatum wistar ky3960
5	ventral striatum sprague3961
5	ventral striatum sprague3962
5	ventral striatum sprague3963
5	ventral striatum wistar 4026
5	ventral striatum wistar 4028
6	dorsal striatum wistar ky3970
6	dorsal striatum wistar ky3972
6	dorsal striatum wistar 3966
6	dorsal striatum-sprague3978
6	dorsal striatum-sprague3981
7	dorsal root ganglion fisher1
7	dorsal root ganglion fisher2
8	pineal sprague1
8	pineal sprague2
9	pituitary sprague1
9	pituitary sprague2
10	nucleus accumbens, whole wistar1
10	nucleus accumbens, whole wistar2
10	nucleus accumbens_core sprague1
10	nucleus accumbens_core sprague2
10	nucleus accumbens_shell sprague1
10	nucleus accumbens_shell sprague2
11	small intestine sprague1
11	small intestine sprague2
12	skeletal muscle sprague1
12	skeletal muscle sprague2
13	bone marrow sprague1
13	bone marrow sprague2
14	kidney sprague1
14	kidney sprague2
15	heart sprague1
15	heart sprague2
16	large intestine sprague1
16	large intestine sprague2
17	spleen sprague1
17	spleen sprague2
18	thymus sprague1
18	thymus sprague2
19	endothelial cells sprague1
19	endothelial cells sprague2
20	ventral tegmental area sprague1
20	ventral tegmental area sprague2
21	locus coeruleus sprague1
21	locus coeruleus sprague2
22	prefrontal cortex sprague1
22	prefrontal cortex sprague2
23	dorsal raphe sprague1
23	dorsal raphe sprague2
24	hypothalamus sprague1
24	hypothalamus sprague2
25	primary cortical neurons sprague1
25	primary cortical neurons sprague2
26	cornea Sprague1
26	cornea Sprague2
END;	

GSE8692	hsa
3	
0	GBM/gliosar-coma_KH_24016_24007
0	gliosarcoma_762_26300_26291
0	gliosarcoma_631_6958_6955
1	GBM_1745_6948_6920
1	GBM_1507_34705_34696
1	GBM_1354_4782_4767
1	GBM_1043_14629_14597
1	GBM_932_15432_15410
1	GBM_931_4776_4764
2	oligodendrogli-oma_IC_26342_26338
2	AMG_1649_34707_34697
2	AMG_1326_34703_34695
END;	

Clusters Não-Severos

gse607	ath	GSE1912	mmu
3		8	
0	Leaf_GC2	0	Postnatal day 1, mouse 1
0	Leaf_GH1	0	Postnatal day 1, mouse 2
0	Leaf_GH2	0	Postnatal day 1, mouse 3
1	STEM_GC7	1	Postnatal day 6, mouse 1
1	STEM_GC8	1	Postnatal day 6, mouse 2
1	STEM_GH7	1	Postnatal day 6, mouse 3
1	STEM_GH8	2	Postnatal day 14, mouse 1
2	FLOWER_GC5	2	Postnatal day 14, mouse 2
2	FLOWER_GC6	2	Postnatal day 14, mouse 3
2	FLOWER_GH5	3	Postnatal day 17, mouse 1
2	FLOWER_GH6	3	Postnatal day 17, mouse 2
END;		3	Postnatal day 17, mouse 3
gse3416	ath	4	Postnatal day 23, mouse 1
6		4	Postnatal day 23, mouse 2
0	00h Col-0 replicate A	4	Postnatal day 23, mouse 3
0	00h Col-0 replicate B	5	Postnatal 9-week, mouse 1
0	00h Col-0 replicate C	5	Postnatal 9-week, mouse 2
1	04h Col-0 replicate A	5	Postnatal 9-week, mouse 3
1	04h Col-0 replicate B	6	Postnatal 5-month, mouse 1
1	04h Col-0 replicate C	6	Postnatal 5-month, mouse 2
2	08h Col-0 replicate A	6	Postnatal 5-month, mouse 3
2	08h Col-0 replicate B	7	Postnatal 1-year, mouse 1
2	08h Col-0 replicate C	7	Postnatal 1-year, mouse 2
3	12h Col-0 replicate A	7	Postnatal 1-year, mouse 3
3	12h Col-0 replicate B	7	Postnatal 1-year, mouse 4
3	12h Col-0 replicate C	END;	
4	16h Col-0 replicate A	GSE2195	mmu
4	16h Col-0 replicate B	2	
4	16h Col-0 replicate C	0	AO 1hr
5	20h Col-0 replicate A	0	CTL2 AO 1hr
5	20h Col-0 replicate B	0	TMRI AO 1hr

5	20h Col-0 replicate C	0	AO 2hr
END;		0	CTL2 AO 2hr
GSE9311	ath	0	TMRI AO 2hr
2		0	AO 4hr
0	Root control rep 1	0	CTL2 AO 4hr
0	Root control rep 2	0	TMRI AO 4hr
0	Root Selenate rep 1	0	AO 8hr
0	Root Selenate rep 2	0	CTL2 AO 8hr
1	Shoot control rep1	0	TMRI AO 8hr
1	Shoot control rep2	0	AO 24hr
1	Shoot Selenate rep1	0	CTL2 AO 24hr
1	Shoot Selenate rep2	0	TMRI AO 24hr
END;		0	AO 48hr
gse775	mmu	0	CTL2 AO 48hr
3		0	TMRI AO 48hr
0	PGA-lv_1h_511	0	AO 72hr
0	PGA-lv_1h_512	0	CTL2 AO 72hr
0	PGA-lv_1h_514	0	TMRI AO 72hr
0	PGA-lv2_1h_514	1	E2 1hr
0	PGA-lv_1w_651	1	CTL2 E2 1hr
0	PGA-lv_1w_672	1	TMRI E2 1hr
0	PGA-lv_1w_674	1	E2 2hr
0	PGA-lv2_1w_674	1	CTL2 E2 2hr
0	PGA-lv_24h_510	1	TMRI E2 2hr
0	PGA-lv_24h_515	1	E2 4hr
0	PGA-lv_24h_517	1	CTL2 E2 4hr
0	PGA-lv2_24h_517	1	TMRI E2 4hr
0	PGA-lv_48h_518	1	E2 8hr
0	PGA-lv_48h_519	1	CTL2 E2 8hr
0	PGA-lv_48h_520	1	TMRI E2 8hr
0	PGA-lv2_48h_519	1	E2 24hr
0	PGA-lv_4h_506	1	CTL2 E2 24hr
0	PGA-lv_4h_507	1	TMRI E2 24hr
0	PGA-lv_4h_509	1	E2 48hr
0	PGA-lv2_4h_506	1	CTL2 E2 48hr
0	PGA-lv_8w_329	1	TMRI E2 48hr
0	PGA-lv_8w_339	1	E2 72hr
0	PGA-lv_8w_340	1	CTL2 E2 72hr
1	PGA_MI_ilv_1h_392	1	TMRI E2 72hr
1	PGA_MI_ilv_1h_394	END;	
1	PGA_MI_ilv_1h_395	GSE2719	hsa
1	PGA_MI_ilv_1w_662	23	
1	PGA_MI_ilv_1w_663	0	brain
1	PGA_MI_ilv_1w_670	1	stomach
1	PGA_MI_ilv_24h_360	2	colon
1	PGA_MI_ilv_24h_361	3	pancreas
1	PGA_MI_ilv_24h_363	4	prostate
1	PGA_MI_ilv_48h_351	5	skin
1	PGA_MI_ilv_48h_354	6	small intestine
1	PGA_MI_ilv_48h_355	7	adrenal
1	PGA_MI_ilv_4h_389	8	connective tissue
1	PGA_MI_ilv_4h_390	9	heart
1	PGA_MI_ilv_4h_391	10	kidney
1	PGA_MI_ilv_8w_311	11	liver
1	PGA_MI_ilv_8w_326	12	lung
1	PGA_MI_ilv_8w_332	13	skeletal muscle
2	PGA_MI_nilv_1h_392	14	spleen

2 PGA_MI_nilv_1h_394
2 PGA_MI_nilv_1h_395
2 PGA_MI_nilv_1w_662
2 PGA_MI_nilv_1w_663
2 PGA_MI_nilv_1w_670
2 PGA_MI_nilv_24h_360
2 PGA_MI_nilv_24h_361
2 PGA_MI_nilv_24h_362
2 PGA_MI_nilv_48h_351
2 PGA_MI_nilv_48h_354
2 PGA_MI_nilv_48h_355
2 PGA_MI_nilv_4h_389
2 PGA_MI_nilv_4h_390
2 PGA_MI_nilv_4h_391
2 PGA_MI_nilv_8w_311
2 PGA_MI_nilv_8w_326
2 PGA_MI_nilv_8w_332
END;
GSE1001 rno
2
0 Normal Rat Retina
0 Normal Rat Retina 2
0 Normal Rat Retina 3
1 Injured 4h Rat Retina
1 Injured 4h Rat Retina 2
1 Injured 4h Rat Retina 3
1 Injured 1 day Rat Retina 1
1 Injured 1 day Rat Retina 2
1 Injured 1 day Rat Retina 3
1 Injured 3 day Rat Retina 1
1 Injured 3 day Rat Retina 2
1 Injured 3 day Rat Retina 3
1 Injured 7 day Rat Retina 1
1 Injured 7 day Rat Retina 2
1 Injured 7 day Rat Retina 3
1 Injured 30 day Rat Retina 1
1 Injured 30 day Rat Retina 2
1 Injured 30 day Rat Retina 3
END;
GSE1036 hsa
6
0 HeminTimecourse_0hA
0 HeminTimecourse_0hB
1 HeminTimecourse_6hA
1 HeminTimecourse_6hB
2 HeminTimecourse_12hA
2 HeminTimecourse_12hB
3 HeminTimecourse_24hA
3 HeminTimecourse_24hB
4 HeminTimecourse_48hA
4 HeminTimecourse_48hB
5 HeminTimecourse_72hA
5 HeminTimecourse_72hB
END;
GSE1156 rno
2
0 NINDS-RHS-Ctr_1h-1aUA-s2
15 fibrosarcoma 1
15 fibrosarcoma 2
15 fibrosarcoma 3
15 fibrosarcoma 4
15 fibrosarcoma 5
15 fibrosarcoma 6
15 fibrosarcoma 7
16 GIST 1
16 GIST 2
17 Leiomyosarcoma 1
17 Leiomyosarcoma 2
17 Leiomyosarcoma 3
17 Leiomyosarcoma 4
17 Leiomyosarcoma 5
17 Leiomyosarcoma 6
18 Lipo dediff 1
18 Lipo dediff 2
18 Lipo dediff 3
18 Lipo dediff 4
19 Lipo pleo 1
19 Lipo pleo 2
19 Lipo pleo 3
20 MFH 1
20 MFH 2
20 MFH 3
20 MFH 4
20 MFH 5
20 MFH 6
20 MFH 7
20 MFH 8
20 MFH 9
21 Round cell 1
21 Round cell 2
21 Round cell 3
21 Round cell 4
22 Synovial sarcoma 1
22 Synovial sarcoma 2
22 Synovial sarcoma 3
22 Synovial sarcoma 4
END;
GSE952 rno
27
0 frontal cortex wistar3658
0 frontal cortex wistar3659
0 frontal cortex wistar3660
0 frontal cortex wistar3662
0 frontal cortex wistar 3920
0 frontal cortex wistar 3921
0 frontal cortex wistar 3922
0 frontal cortex wistar 3923
0 frontal cortex wistar 3924
0 frontal cortex wistar 3925
0 frontal cortex wistar ky3926
0 frontal cortex wistar ky3927
0 frontal cortex wistar ky3928
0 frontal cortex wistar ky3929
0 frontal cortex wistar ky3930

0	NINDS-RHS-Ctr_1h-2aUA-s2	0	frontal cortex wistar ky3931
0	NINDS-RHS-Ctr_1h-3aUA-s2	0	frontal cortex sprague3932
0	NINDS-RHS-Ctr_240h-1aUA-s2	0	frontal cortex sprague3933
0	NINDS-RHS-Ctr_240h-2aUA-s2	0	frontal cortex sprague3934
0	NINDS-RHS-Ctr_240h-3aUA-s2	0	frontal cortex sprague3935
0	NINDS-RHS-Ctr_24h-1aUA-s2	0	frontal cortex sprague3936
0	NINDS-RHS-Ctr_24h-2aUA-s2	0	frontal cortex sprague3937
0	NINDS-RHS-Ctr_24h-3aUA-s2	1	amygdala wistar3722
0	NINDS-RHS-Ctr_6h-1aUA-s2	1	amygdala wistar3723
0	NINDS-RHS-Ctr_6h-2aUA-s2	1	amygdala wistar3724
0	NINDS-RHS-Ctr_6h-3aUA-s2	1	amygdala wistar3942
0	NINDS-RHS-Ctr_72h-1aUA-s2	1	amygdala, central nucleus sprague1
0	NINDS-RHS-Ctr_72h-2aUA-s2	1	amygdala sprague3950
0	NINDS-RHS-Ctr_72h-3aUA-s2	1	amygdala wistar ky3946
1	NINDS-RHS-Kainate_1h-1aUA-s2	1	amygdala, central nucleus sprague2
1	NINDS-RHS-Kainate_1h-2aUA-s2	1	amygdala wistar 3721
1	NINDS-RHS-Kainate_1h-3aUA-s2	1	amygdala wistar 3720
1	NINDS-RHS-Kainate_240h-1aUA-s2	2	hippocampus wistar3772
1	NINDS-RHS-Kainate_240h-2aUA-s2	2	hippocampus wistar ky3956
1	NINDS-RHS-Kainate_240h-3aUA-s2	2	hippocampus sprague3957
1	NINDS-RHS-Kainate_24h-1aUA-s2	3	cerebellum wistar3884
1	NINDS-RHS-Kainate_24h-2aUA-s2	3	cerebellum wistar3885
1	NINDS-RHS-Kainate_24h-3aUA-s2	3	cerebellum wistar3886
1	NINDS-RHS-Kainate_6h-1aUA-s2	3	cerebellum wistar3887
1	NINDS-RHS-Kainate_6h-2aUA-s2	3	cerebellum wistar3888
1	NINDS-RHS-Kainate_6h-3aUA-s2	3	cerebellum wistar3889
1	NINDS-RHS-Kainate_72h-1aUA-s2	3	cerebellum wistar ky3890
1	NINDS-RHS-Kainate_72h-2aUA-s2	3	cerebellum wistar ky3891
1	NINDS-RHS-Kainate_72h-3aUA-s2	3	cerebellum wistar ky3892
END;		3	cerebellum wistar ky3893
GSE1432	hsa	3	cerebellum wistar ky3894
2		3	cerebellum sprague3896
0	IFNG Microglia Sample B18 1hr	3	cerebellum sprague3897
0	IFNG Microglia Sample O 1hr	3	cerebellum sprague3898
0	IFNG Microglia Sample W 1hr	3	cerebellum sprague3899
0	IFNG Microglia Sample Y20 1hr	3	cerebellum sprague3900
0	IFNG Microglia Sample B18 6hrs	3	cerebellum sprague3901
0	IFNG Microglia Sample O 6hrs	4	cerebral cortex wistar 3902
0	IFNG Microglia Sample W 6hrs	4	cerebral cortex wistar 3903
0	IFNG Microglia Sample Y20 6hrs	4	cerebral cortex wistar 3904
0	IFNG Microglia Sample B18 24hrs	4	cerebral cortex wistar ky3910
0	IFNG Microglia Sample O 24hrs	4	cerebral cortex wistar ky3911
0	IFNG Microglia Sample W 24hrs	4	cerebral cortex wistar ky3912
0	IFNG Microglia Sample Y20 24hrs	4	cerebral cortex wistar ky3913
1	Control Microglia Sample O 1hr	4	cerebral cortex sprague3914
1	Control Microglia Sample B18 1hr	4	cerebral cortex sprague3915
1	Control Microglia Sample W 1hr	4	cerebral cortex sprague3916
1	Control Microglia Sample Y20 1hr	4	cerebral cortex sprague3917
1	Control Microglia Sample B18 6hrs	4	cerebral cortex sprague3918
1	Control Microglia Sample W 6hrs	4	cerebral cortex sprague3919
1	Control Microglia Sample Y20 6hrs	5	ventral striatum wistar 3958
1	Control Microglia Sample O 6hrs	5	ventral striatum wistar ky3959
1	Control Microglia Sample B18 24hrs	5	ventral striatum wistar ky3960
1	Control Microglia Sample O 24hrs	5	ventral striatum sprague3961
1	Control Microglia Sample W 24hrs	5	ventral striatum sprague3962
1	Control Microglia Sample Y20 24hrs	5	ventral striatum sprague3963
END;		5	ventral striatum wistar 4026

GSE1493	hsa	5	ventral striatum wistar 4028
3		6	dorsal striatum wistar ky3970
0	lin+CD34+ I replicate	6	dorsal striatum wistar ky3972
0	lin+CD34+ II replicate	6	dorsal striatum wistar 3966
1	Lin-CD34+ I replicate	6	dorsal striatum-sprague3978
1	lin-CD34+ II replicate	6	dorsal striatum-sprague3981
2	lin-CD34- I replicate	7	dorsal root ganglion fisher1
2	Lin-CD34- II replicate	7	dorsal root ganglion fisher2
END;		8	pineal sprague1
GSE1541	hsa	8	pineal sprague2
2		9	pituitary sprague1
0	mRNA from A549 Static Control 1hr Jan 31 2003	9	pituitary sprague2
0	RNA from A549 Cells - Static Control 1hr Replicate 1	10	nucleus accumbens, whole wistar1
0	A549 LPS 1h Jan 31 2003	10	nucleus accumbens, whole wistar2
0	A549 LPS 1hr Oct 21 2002	10	nucleus accumbens_core sprague1
0	A549 Stretch 1hr Jan 31 2003	10	nucleus accumbens_core sprague2
0	A549 cells Stretch 1hr Oct 20 2002	10	nucleus accumbens_shell sprague1
0	A549 TNF 1hr Jan 31 2003	10	nucleus accumbens_shell sprague2
0	A549 TNF 1hr Oct 20 2002	11	small intestine sprague1
0	A549 TNF+Stretch 1hr Jan 31 2003	11	small intestine sprague2
0	A549 TNF+Stretch Oct 20 2002	12	skeletal muscle sprague1
1	A549 RNA Static Control 4hr Jan 31 2003	12	skeletal muscle sprague2
1	RNA A549 cells Static Control 4hr Replicate 1	13	bone marrow sprague1
1	A549 LPS 4hr Jan 31 2003	13	bone marrow sprague2
1	A549 LPS 4hr Nov 21 2002	14	kidney sprague1
1	A549 Stretch 4hr Nov 21 2002	14	kidney sprague2
1	A549 stretch 4hr Jan 31 2003	15	heart sprague1
1	A549 TNF 4hr Jan 31 2003	15	heart sprague2
1	A549 TNF 4hr Nov 21 2002	16	large intestine sprague1
1	A549 TNF+Stretch 4hr Jan 31 2003	16	large intestine sprague2
1	A549 TNF+Stretch 4hr Nov 21 2002	17	spleen sprague1
END;		17	spleen sprague2
GSE1614	hsa	18	thymus sprague1
3		18	thymus sprague2
0	JF2dR1	19	endothelial cells sprague1
0	JF2dR2	19	endothelial cells sprague2
0	JF2dR3	20	ventral tegmental area sprague1
0	JF2dR4	20	ventral tegmental area sprague2
1	JF8dR1	21	locus coeruleus sprague1
1	JF8dR2	21	locus coeruleus sprague2
1	JF8dR3	22	prefrontal cortex sprague1
1	JF8dR4	22	prefrontal cortex sprague2
2	JF15dR1	23	dorsal raphe sprague1
2	JF15dR2	23	dorsal raphe sprague2
2	JF15dR3	24	hypothalamus sprague1
2	JF15dR4	24	hypothalamus sprague2
END;		25	primary cortical neurons sprague1
		25	primary cortical neurons sprague2
		26	cornea Sprague1
		26	cornea Sprague2
		END;	
		GSE8692	hsa
		3	
		0	GBM/gliosarcoma_KH_24016_24007
		0	gliosarcoma_762_26300_26291
		0	gliosarcoma_631_6958_6955

```
1       GBM_1745_6948_6920
1       GBM_1507_34705_34696
1       GBM_1354_4782_4767
1       GBM_1043_14629_14597
1       GBM_932_15432_15410
1       GBM_931_4776_4764
2       oligodendroglioma_IC_26342_26338
2       AMG_1649_34707_34697
2       AMG_1326_34703_34695
END;
```

Bibliografia

(Adams *et al*. 1999) Adams, M.D. et al., "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project", *Science,* 252, 1651-1656, 1991.

(Adams *et al.* 1993) Adams, M.D., Soares, M.B., Kerlavage, A.R., Fields, C., Venter, J.C., "Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library.", Nat Genet.4(4):373-80, Aug, 1993.

(Addya *et al.* 2004) Addya S, Keller MA, Delgrosso K, Ponte C M, *et al.* (2004). Erythroid-induced commitment of K562 cells results in clusters of differentially expressed genes enriched for specific transcription regulatory elements, Physiol Genomics 19:117-130. 1994.

(Affymetrix 2002) *Statistical Algorithms Description Document*, 2002.

(Akutsu *et al.* 1998) Akutsu, T., Miyano, S., Kuhara, S., "Identification Of Genetic Networks from a Small Number of Gene Expression Patterns under the Boolean Network Model", *Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithms*, 695-702, 1998.

(Alon *et al.* 1999) Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. J., "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays", PNAS, vol. 96, pp 6745-6750, june, 1999.

(Altman 1998) Altman R. B., "Bioinformatics in support of molecular medicine", *AMIA Annual Symposium*, Orlando, FL, 1998.

(Barker *et al.* 1998) Barker W. C., Garavelli J. S., Haft D. H., Hunt L. T., Marzec C. R., Orcutt B. C., Srinivasarao G. Y., Yeh L. S., Ledley R. S., Mewes H. W. , Pfeiffer F., Tsugita A., "The PIR — Protein Information Resource", *Nucleic Acids Research.*, 26(1):27-32, January, 1998.

(Barrett *et al*., 2005) Barrett T, Suzek TO, Troup DB, Wilhite SE, *et al*. "NCBI GEO: mining millions of expression profiles — database and tools", Nucleic Acids Res.1;33 Database Issue:D562-6.

(Bay 2004) Bay S. D., Chrisman L., Pohorille A., Shrager J., "Temporal aggregation bias and inference of causal regulatory networks", *Journal of Computational Biology*, 11(5), 971-85, 2004

(Ben-Dor, Shamir & Yakhini, 1999) Ben-Dor A., Shamir R. and Yakhini Z. "Clustering gene expression patterns". Journal of Computational Biology, 6(3/4):281–297, 1999.

(Ben-Dor *et al*., 2000) Ben-Dor, A., ZBruhn L., Friedman, N. Nachman I, Schummer M, Yakhini Z. "Tissue classification with gene expression profiles". J Comput Biol. 2000;7(3-4):559-83.

(Ben-Dor, Friedman & Yakhini, 2001) Ben-Dor, A., Friedman, N. and Yakhini, Z. "Class discovery in gene expression data", In *Proc. Fifth Annual Inter. Conf. on Computational Molecular Biology (RECOMB 2001)*, pages 31–38. ACM Press, 2001.

(Berg 2002) Berg, J. M., Tymoczko, J. L., Stryer, L., *Biochemistry*, 5th edition, W. H. Freeman and Company, 2002.

Bergmann DC, Lukowitz W, Somerville CR (2004). Stomatal Development and Pattern Controlled by a MAPKK Kinase, Science 304, 1494.

(Bläsing *et al*., 2005) Bläsing OE, Gibon Y, Günther M, Höhne M *et al*. "Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in *Arabidopsis*". Plant Cell 2005 Dec; 17(12):3257-81.

(Boni *et al*. 2005) Boni JP, Leister C, Bender G, Fitzpatrick V, *et al*. Population pharmacokinetics of CCI-779: Correlations to safety and pharmacogenomic responses in patients with advanced renal cancer, CLINICAL PHARMACOLOGY & THERAPEUTICS, 77(1):76-89.

(Brazma & Vilo, 2000) Brazma, A., and Vilo, J. *Gene Expression Data Analysis*. FEBS Letters 480, 17-24, 2000.

(Chu 2003) Chu, T., *Learning from SAGE Data*, Ph.D. Dissertation. Philosophy Dept., Carnegie Mellon University, Jan., 2003 (disponível em http://www.phil.cmu.edu/projects/genegroup/papers.html)

(D'haeseleer 2005) D'haeseleer, Patrik, "How does gene expression clustering work?", *Nature Biotechnology*, vol. 23, n.12, December, 2005.

(Ding, 2002) Ding, C. "Analysis of gene expression profiles: class discovery and leaf ordering". In Proc. of International Conference on Computational Molecular Biology (RECOMB), pp. 127–136,Washington, DC., April 2002.

(dos Santos *et al*. 2004) dos Santos CC, Han B, Andrade CF, Bai X, *et al*. (2004). DNA microarray analysis of gene expression in alveolar epithelial cells in response to TNF-γ, LPS, and cyclic stretch. Physiol Genomics 19:331-342.

(Eisen *et al.* 1998) Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D., "Cluster analysis and display of genome-wide expression patterns" *Proc. Natl. Acad. Sci. USA* Vol. 95, pp. 14863–14868, December 1998

(Faria-Campos *et al*. 2003) Faria-Campos, A.C., "Mining Microorganism EST Databases in The Quest for New Proteins", *Genet Mol Res*, 2, 169-77.

(Fitch & Sokhansanj 2000) Fitch J. P., Sokhansanj B., "Genomic Engineering: Moving Beyond DNA Sequence to Function", *Proceeding of IEEE*, December 2000. Disponível em http://www.ece.ogi.edu/~strom/research/cnrg.htm ou

http://www.ece.ogi.edu/~strom/research/papers_talks/fitch_dna_ieee_proc.pdf

(Fleet *et al*. 2003) Fleet JC, Wang L, Vitek O, Craig BA, *et al*. Gene expression profiling of Caco-2 BBe cells suggests a role for specific signaling pathways during intestinal differentiation, Physiol Genomics 13:57-68.

(Franco *et al.* 1997) Franco, G.R., Rabelo, E.M.L., Azevedo, V., Pena, H.B., Ortega, J.M., Santos, T.M., Meira, W.S.F., Rodrigues, N.A., Dias, C.M.M., Harrop, R. *et al*. "Evaluation of cDNA libraries from different developmental stages of Schistosoma mansoni for production of expressed sequence tags (ESTs)", *DNA Res.* 4: 231-240, 1997.

(Ge *et al.* 2005) Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S.M., Aburatani, H.: Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. Genomics 86 (2005) 127-141.

(Golub *et al*. 1999) Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, D.D., Lander E.S., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", *Science*, Vol. 286(15):531–537, October 1999.

(gse775, 2008) A Mouse Model of Myocardial Infarction [http://cardiogenomics.med.harvard.edu/groups/proj1/pages/mi_home.html]. Acessado em 10/10/2008.

(gse1156, 2008) http://pepr.cnmcresearch.org/jsp/background.jsp. Acessado em 10/10/2008.

(Hunter 1995) Hunter L., *Molecular Biology for Computer Scientists*, in *Artificial Intelligence & Molecular Biology*. Disponível em http://www.biosino.org/mirror/www.aaai.org/Press/Books/Hunter/hunter-contents.html [acessado em 26/09/2008]

(Ibrahim *et al.* 2005) Ibrahim, A. F. M., Hedley, P. E., Cardle, L., Kruger, W., Marshall, D. F., Muehlbauer, G. J., Waugh R., "A comparative analysis of transcript abundance using SAGE and Affymetrix arrays", Funct Integr Genomics 5: 163–174, 2005.

(Ishii *et al*. 2000) Ishii, M., Hashimoto, S., Tsutsumi, S.,Wada,Y., Matsushima, K., Kodama, T., Aburatani, H., "Direct Comparison of GeneChip and SAGE on the Quantitative Accuracy in Transcript Profiling Analysis", *Genomics*, 68, 136-143, 2000.

(Jain, Murty & Flynn, 1999) Jain, A. K., Murty, M. N., Flynn, P. J., "Data Clustering: A Review", ACM Computing Surveys, vol. 31, nº 3, setembro, 1999.

(Jiang *et al.* 2004) Jiang, D., Tang, C., Zhang, A., "Cluster Analysis for Gene Expression Data: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370-1386, Nov., 2004.

(Kim & Nam, 2006) Kim V. N., Nam J. W. "Genomics of miRNA". Trends Genet. 22:165–173. 2006.

(Kohonen 1984) Kohonen, T., *Self-Organization and Associative Memory*, Spring-Verlag, Berlin, 1984.

(Kuo *et al.* 2002) Kuo W., Jenssen T., Butte A., Ohno-Machado L., Kohane I., "Analysis of matched mRNA measurements from two different microarray technologies", *Bioinformatics*, March,18(3):405-12, 2002.

(Lin *et al*. 2004) Lin KK, Chudova D, Hatfield GW, Smyth P, Andersen B: Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance. Proc Natl Acad Sci USA 2004, 9;101(45):15955-60.

(Lipshutz *et al.*, 1999) Lipshutz R. J., Fodor S. P. A., Gingeras T. R., Lockhart D. J., "High density synthetic oligonucleotide arrays", Nature Genetics, supplement. Vol. 21, 1999.

(Liu *et al*. 2007) Liu T, Papagiannakopoulos T, Puskar K, Qi S, *et al*. (2007). Detection of a microRNA signal in an in vivo expression set of mRNAs. PLoS ONE 29;2(8):e804.

(Lockhart *et al.* 1996) Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., *et al*., "Expression monitoring by hybridization to high-density oligonucleotide arrays", *Nature Biotechnology* 14: 1675−1680, 1996.

(Lodish *et al.* 2003) Lodish, H., Scott, M. P., Matsudaira, P., Darnell, J., Zipursky, L., Kaiser, C. A., Berk, A, Krieger M., *Molecular Cell Biology*, *5th edition*, W. H. Freeman, 2003.

(Lu *et al.* 2004) Lu J, Lal A, Merriman B, Nelson S, Riggins G., "A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips.", Genomics, Oct;84(4):631-6, 2004.

(Luger, 2004) Luger, G. F., *Inteligência Artificial: Estruturas e Estratégias para a Solução de Problemas Complexos*, 4ª edição, Editora Bookman, 2004.

(Luscombe *et al*. 2001) Luscombe N. M., Greenbaum D., Gerstein M., "What is Bioinformatics: a Proposed Definition and Overview of the Field", *Methods of Information in Medicine*, 40(4):346-58, 2001.

(Manfredini *et al*. 2005) Manfredini R, Zini R, Salati S, Siena M, *et al*.. (2005). The Kinetic Status of Hematopoietic Stem Cell Subpopulations Underlies a Differential Expression of Genes Involved in Self-Renewal, Commitment, and Engraftment, Stem Cells 2005;23:496-506. 2005.

(MaQC Consortium 2006) "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements", *Nature Biotechnology*, 24(9), pp. 1151 – 1161, September, 2006.

(McQueen 1967) McQueen, J.B., "Some methods for classification and analysis of multivariate observations", In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, University of California, Berkeley, 1967.

(Mei *et al.*, 2003) Mei R., Hubbell E., Bekiranov S., Mittmann M, *et al.*, "Probe selection for high-density oligonucleotide arrays", PNAS 2003;100;11237-11242; originally published online Sep 19, 2003;doi:10.1073/pnas.1534744100

(Menck & van Sluys, 2004) "Fundamentos da Biologia Molecular — A Construção do Conhecimento" *in Genômica*, Editora Atheneu, 2004.

(Mudado & Ortega, 2006) Mudado, M. A. e Ortega, J. M., "A picture of gene sampling/expression in model organisms using ESTs and KOG proteins", *Genet. Mol. Res.* 5 (1): 242-253, 2006.

(Pinto & Ortega, 2007) Pinto, S., e Ortega, J. M. "Finding Normalizers Genes by Means of Homology Searches on Expressed Sequence Tags and Oligonucleotide Array Data". Proceedings of the Brazilian Symposium on Bioinformatics (BSB2007). pp 160-171. Angra dos Reis. Brazil. 2007.

(Pontius, Wagner & Schuler, 2003) Pontius, J. U, Wagner, L, Schuler, G. D. UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information; 2003.

(Rock *et al.,* 2005) Rock R.B., Hu S., Deshpande A., Munir S., *et al.*, "Transcriptional response of human microglial cells to interferon-γ", Genes and Immunity 6, 712–719, 2005.

(Russell & Norvig, 2003). Artificial Intelligence: A Modern Approach (2nd Ed.), Upper Saddle River, NJ. Prentice-Hall. (http://aima.cs.berkeley.edu/).

(Schena *et al.* 1995) Schena, M., Shalon, D., Davis, R.W., and Brown, P.O.. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", *Science*, 1995 .

(Shamir & Sharan 2000) Shamir, R. and Sharan, R., "Click: A clustering algorithm for gene expression analysis", In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00). AAAI Press.*, 2000.

(Sharan *et al.* 2002) Sharan, R., Elkon, R., and Shamir, R., "Cluster Analysis and its Applications to Gene Expression Data, *Ernst Schering workshop on Bioinformatics and Genome Analysis*, pp. 83--108, Springer Verlag, Berlin 2002.

(Slonim 2002) Slonim, D. K., "From patterns to pathways: gene expression data analysis comes of age", *Nature Genetics Suplement*, vol 32., pp 502-508, December, 2002.

(Staunton *et al*. 2001) Staunton, J. E., Slonim, D. K., Coller, H. A., Tamayo, P., Angelo, M. J., Park, J., Scherf, U., Lee, J. K., Reinhold, W. O., Weinstein, J. N., Mesirov, J. P., Lander, E. S.,and Golub, T. R., "Chemosensitivity prediction by transcriptional profiling", PNAS, July, 2001.

(Stekel *et al.* 2000) Stekel D.J., Git Y., Falciani F.,"The Comparison of Gene Expression from Multiple cDNA Libraries", *Gen. Res.* 10:2055-2061, 2000.

(Su *et al*., 2002) Su AI, Cooke MP, Ching KA, Hakak Y, *et al.,* "Large-scale analysis of the human and mouse transcriptomes", PNAS, vol. 99, no. 7, 4465–4470.

(Tibshirani *et al*., 2002) Tibshirani R., Hastie T., Narasimhan B., Chu G. "Diagnosis of multiple cancer types by shrunken centroids of gene expression". PNAS , vol. 99, num. 10, pp 6567-6572, 2002.

(Van Hoewyk *et al*. 2008) Van Hoewyk D, Takahashi H, Inoue E, Hess A, Tamaoki M, Pilon-Smits E: Transcriptome analyses give insights into selenium-stress responses and selenium tolerance mechanisms in Arabidopsis.Physiol Plant. 2008 Feb;132(2):236-53.

(Vázquez-Chona *et al.*, 2004) Vázquez-Chona F, Song BK, Geisert EE Jr. "Temporal changes in gene expression after injury in the rat retina". Invest Ophthalmol Vis Sci 2004 Aug;45(8):2737-46.

(Velculescu 1995) Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W., "Serial Analysis Of Gene Expression", *Science* 270, 484-487, 1995.

(Walker *et al*., 2004) Walker J. R., Su A. I., Self D. W., Hogenesch J. B. *et al*. "Applications of a rat multiple tissue gene expression data set". Genome Research 2004 Apr;14(4):742-9.

(Witten & Frank, 2000) Witten IH, Frank E. Data mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann Publishers. (WEKA site: www.cs.**waikato**.ac.nz/ml/**weka**/)

(Xing & Karp, 2001) Xing, E. P. e Karp, R. M., "CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts". Bioinformatics. Vol 17. suppl. 1. pp S306-S315. 2001.

(Xu & Zhang, 2005) Xu, X. e Zhang, A., "Virtual Gene: A Gene Selection Algorithm for Sample Classification on Microarray Datasets". International Conference on Computational Science (2) 2005: 1038-1045. Atlanta. GA. USA.

(Yeung & Ruzzo, 2000) Yeung, Ka Yee e Ruzzo, Walter L. "An empirical study on principal component analysis for clustering gene expression data". Technical Report UW-CSE-2000-11-03, Department of Computer Science & Engineering, University of Washington, 2000.

(Yoon *et al.*, 2006) Yoon SS, Segal NH, Park PJ, Detwiller KY *et al*. "Angiogenic profile of soft tissue sarcomas based on analysis of circulating factors and microarray gene expression". J Surg Res 2006 Oct;135(2):282-90.