

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

Análise de Dados de Expressão Gênica

Katti Faceli

André C P L F de Carvalho

Marcilio Carlos Pereira de Souto

Nº 250

RELATÓRIOS TÉCNICOS



São Carlos – SP
Jan./2005

SYSNO	1418701
DATA	1 / 1
ICMC - SBAB	

Análise de Dados de Expressão Gênica

Katti Faceli
André C.P.L.F. de Carvalho

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação e Estatística
Laboratório de Inteligência Computacional
Caixa Postal 668, 13560-970 - São Carlos, SP, Brasil
e-mail: {katti, andre}@icmc.usp.br

Marcílio Carlos Pereira de Souto

Universidade Federal do Rio Grande do Norte
Departamento de Informática e Matemática Aplicada - DIMAp
Campus Universitario
59072-970 - Natal, RN, Brazil
e-mail: marcelio@dimap.ufrn.br

Resumo

A análise de dados de expressão gênica é bastante útil para diversas aplicações na biologia. Por exemplo, esse tipo de análise pode fornecer informações importantes sobre as funções de uma célula, os mecanismos de regulação dos genes e as diferenças moleculares entre vários tipos de células ou tecido.

A expressão de um gene consiste da conversão das instruções genéticas presentes no gene, em uma proteína, responsável por uma função específica dentro de uma célula.

A mensuração da expressão gênica a partir dos transcritos (mRNAs) de um organismo pode ser feita por meio de diversas técnicas, que geram dados em larga escala, tais como SAGE (*Serial Analysis of Gene Expression*) e *microarray*.

A partir desses dados de expressão dos genes, faz-se necessária uma análise criteriosa para a extração de informações úteis. Essa análise pode ser realizada em diferentes níveis e de diferentes formas, variando de análises de baixo nível, como filtragens e normalizações, até análises estatísticas para identificação de expressão diferente em diversas condições, identificação de padrões e mineração de dados.

Palavras-Chave: Expressão gênica, Microarray, SAGE, RT-PCR, MPSS.

Janeiro 2005
Draft impresso em 11 de janeiro de 2005

Sumário

1	Introdução	1
2	Processo de Expressão Gênica	2
3	Técnicas para Medir a Expressão Gênica	6
3.1	<i>Microarray de DNA</i>	7
3.2	<i>SAGE - Serial Analysis of Gene Expression</i>	9
3.3	<i>MPSS - Massively Parallel Signature Sequencing</i>	10
3.4	<i>RT-PCR - Reverse Transcription - Polymerase Chain Reaction</i>	11
3.5	<i>Comparação das Técnicas de Medição da Expressão Gênica</i>	12
4	Processamento e Análise Computacional dos Dados de Expressão Gênica	14
5	Conclusão	17
	Referências	20

Lista de Figuras

1	Elementos, estrutura e representação do DNA.	3
2	Genes em um cromossomo.	4
3	Processo de produção de uma proteína em eucariotos.	5

Lista de Tabelas

1	Comparação entre as técnicas de medição da expressão gênica.	12
2	Exemplo de uma matriz de expressão.	14

1 Introdução

A bioinformática ou biologia computacional, área de pesquisa em grande expansão, diz respeito à utilização de técnicas e ferramentas da computação para a resolução de problemas da biologia (Baldi & Brunak 1998). Uma área importante da biologia, que necessita da aplicação da computação é a biologia molecular (Setúbal & Meidanis 1997), englobando a resolução de problemas em tópicos que vão desde a comparação de seqüências e montagem de fragmentos até a identificação e análise da expressão de genes e determinação da estrutura de proteínas.

A análise de dados de expressão gênica é bastante útil para diversas aplicações na biologia. Por exemplo, esse tipo de análise pode fornecer informações importantes sobre as funções de uma célula, uma vez que as mudanças na fisiologia de um organismo são acompanhadas por mudanças nos padrões de expressão de seus genes (Chan, Hontzeas, & Park 2000), os mecanismos de regulação dos genes e as diferenças moleculares entre vários tipos de células ou tecido.

A expressão de um gene consiste da conversão das instruções genéticas presentes no gene, em uma proteína, responsável por uma função específica dentro de uma célula. Existem duas abordagens para a avaliação da expressão gênica: a análise do transcriptoma e a análise do proteoma. A primeira se baseia na análise dos produtos da transcrição, mRNA (RNA mensageiro), intermediários no processo de produção de uma proteína, ou expressão gênica. A segunda se refere à análise direta das proteínas produzida no processo de expressão. Uma proteína é produzida a partir de um mRNA, porém, a quantidade de mRNA presente em uma célula, apesar de relacionada, nem sempre corresponde à quantidade de proteína produzida.

Apesar de analisar diretamente o produto final da expressão de um gene, que são as proteínas, a análise do proteoma é muito mais trabalhosa do que a análise do transcriptoma. Com os avanços nas tecnologias relacionadas às análises de proteína, as análises baseadas no proteoma vêm crescendo rapidamente. Apesar disso, a análise da expressão gênica tem sido feita, em sua grande maioria, pela análise do transcriptoma.

A mensuração da expressão gênica a partir dos transcritos (mRNAs) de um organismo pode ser feita por meio de diversas técnicas, que geram dados em larga escala, tais como SAGE (*Serial Analysis of Gene Expression*) (Velculescu, Zhang, Vogelstein, & Kinzler 1995) e *microarray* (Schena, Shalon, Davis, & Brown 1995).

A partir dos dados sobre expressão de genes, faz-se necessária uma análise criteriosa para a extração de informações úteis. Essa análise pode ser realizada em diferentes níveis e de diferentes formas.

Este relatório contém uma descrição de vários aspectos relacionados à análise de expressão gênica. Na Seção 2 é descrito o processo de expressão gênica, de forma bastante simplificada. A Seção 3 contém uma descrição resumida de algumas das técnicas existentes atualmente para a medição da expressão gênica. Essas técnicas também são brevemente comparadas e a forma dos dados gerados por elas é apresentada. O processamento necessário e algumas possíveis análises dos dados de expressão gênica são brevemente discutidos na Seção 4.

2 *Processo de Expressão Gênica*

Colocado de maneira simplificada, o funcionamento de um organismo é baseado na produção de proteínas, que ocorre em suas células. A maioria das substâncias presentes em um organismo são proteínas. Proteínas são moléculas grandes compostas de uma ou mais cadeias de moléculas menores chamadas aminoácidos. Existem diferentes tipos de proteínas, com tamanho variando de dezenas a centenas de aminoácidos, desempenhando papel estrutural ou funcional no organismo. O organismo humano possui cerca de 30.000 tipos de proteínas, desempenhando papéis específicos. Por exemplo, a hemoglobina é uma proteína que transporta oxigênio no sangue, o colágeno é uma proteína estrutural presente nos tendões e nos lóbulos das orelhas, a actina e a miosina são proteínas que interagem para promover a contração muscular e a insulina é um hormônio protéico que controla o transporte de açúcar do sangue para o interior das células.

Toda a variedade de proteínas que um organismo pode produzir, bem como outras informações necessárias para a produção dessas proteínas, está codificada em moléculas chamadas ácidos nucleicos. As regiões dos ácidos nucleicos que contêm os códigos necessários para a produção das proteínas são chamadas genes. Os genes e todas as demais informações contidas nessas moléculas constituem a informação genética. Existem dois tipos de ácidos nucleicos, o DNA (ácido desoxirribonucleico) e o RNA (ácido ribonucleico).

O DNA é uma molécula composta de duas fitas estruturadas na forma de uma dupla-hélice. Cada fita consiste de uma cadeia de nucleotídeos. Um nucleotídeo é formado por um grupo fosfato, um açúcar (desoxirribose) e uma base nitrogenada. A base nitrogenada pode ser uma adenina (*A*), uma guanina (*G*), uma citosina (*C*) ou uma timina (*T*), dando origem a quatro nucleotídeos diferentes. As duas cadeias de nucleotídeos são unidas por pontes de hidrogênio entre as bases nitrogenadas. As reações de ponte de hidrogênio que ocorrem entre as bases nitrogenadas são chamadas de pareamento de bases. Esse pareamento só ocorre entre bases específicas, que são ditas complementares. A base *A* deve parear com a base *T* e a base *G* deve parear com a base *C*. Embora nucleotídeos e bases sejam coisas distintas, por simplicidade, o nucleotídeo é identificado pela base que ele possui, e o termo base é utilizado como sinônimo de nucleotídeo, dizendo-se, por exemplo, que uma molécula possui 200 bases ou 200 nucleotídeos. Uma molécula de DNA costuma ser representada simplesmente pela sequência de nucleotídeos de uma fita específica. A Figura 1 ilustra os elementos, a estrutura dupla-hélice e a representação de uma molécula de DNA.

O RNA também é uma molécula composta por uma cadeia de nucleotídeos, assim como o DNA. Porém, o RNA possui uma única fita, não apresentando a forma de dupla-hélice. O açúcar dos nucleotídeos do RNA é a ribose e não a desoxirribose, como no DNA. Além disso, o RNA não contém a base nitrogenada timina, que é substituída pela base uracila (*U*), que também parecia com a base adenina.

Na maioria dos organismos, a informação genética está codificada em moléculas de DNA, embora em alguns vírus, por exemplo, esta informação esteja codificada em moléculas de RNA. Neste trabalho, as descrições se referem ao primeiro caso. Cada molécula de DNA presente em uma célula constitui um cromossomo. O conjunto dos cromossomos de um organismo é chamado de genoma e está presente em todas as células do organismo.

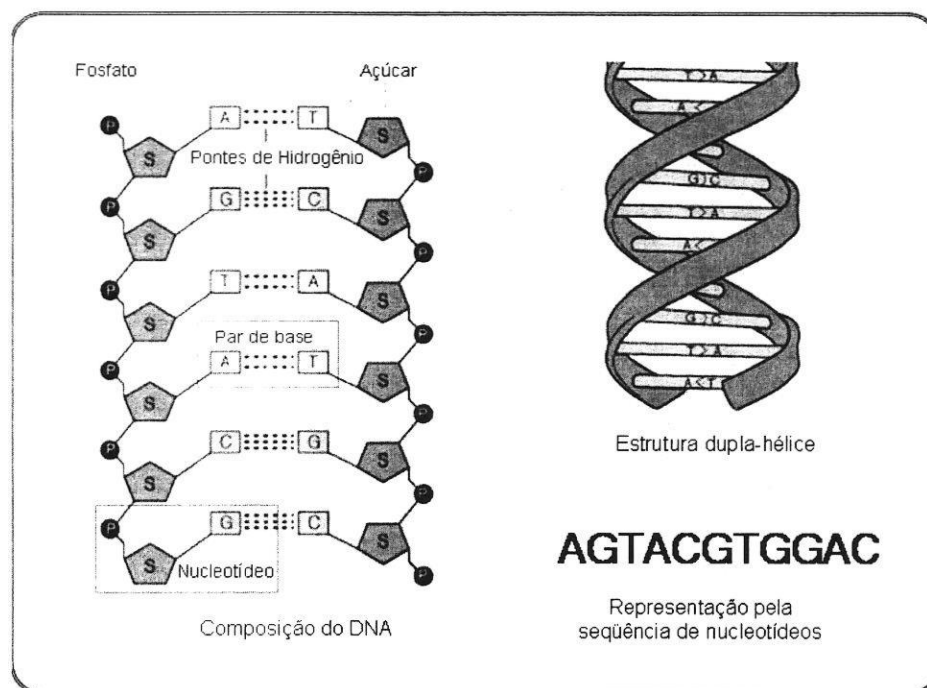


Figura 1: Elementos, estrutura e representação do DNA.

A Figura 2 mostra uma representação dos genes em um cromossomo.

Em resumo, “o DNA codifica todas as proteínas que a(s) célula(s) do organismo deve(m) sintetizar e as proteínas, por sua vez, são responsáveis (direta ou indiretamente) pelas funções necessárias para a sobrevivência” Lewin (2001). A seguir é descrito o processo de síntese de uma proteína a partir do código presente no DNA, ou seja, o processo de expressão de um gene.

A produção de uma proteína a partir de um gene também é chamada de expressão do gene. A expressão de um gene é mais complexa nos organismos eucariotos* do que nos procariotos†, mas o processo básico é o mesmo. A informação genética é representada no DNA pela sequência das bases (ou nucleotídeos) nas suas fitas. A relação entre a sequência de bases do DNA e a sequência de aminoácidos da proteína correspondente é chamada de código genético. Cada sequência de 3 bases ao longo da fita de DNA é um código químico para um dos 20 aminoácidos que compõem as proteínas. Com raríssimas exceções, o código genético é universal, sendo basicamente o mesmo para todos os animais, plantas e microrganismos. A síntese de uma proteína se baseia no código genético e se divide em duas etapas: transcrição e tradução. A Figura 3 apresenta as etapas do processo de expressão de um gene.

Simplificadamente, a transcrição é a etapa de formação de uma fita de RNA complementar a uma região do DNA, ou seja, uma fita de RNA cujas bases nitrogenadas da

*Organismos cujas células possuem núcleo. O genoma desses organismos é composto de várias moléculas de DNA e está situado no núcleo.

†Organismos cujas células não possuem núcleo. Seu genoma consiste de uma única molécula de DNA livre na célula.

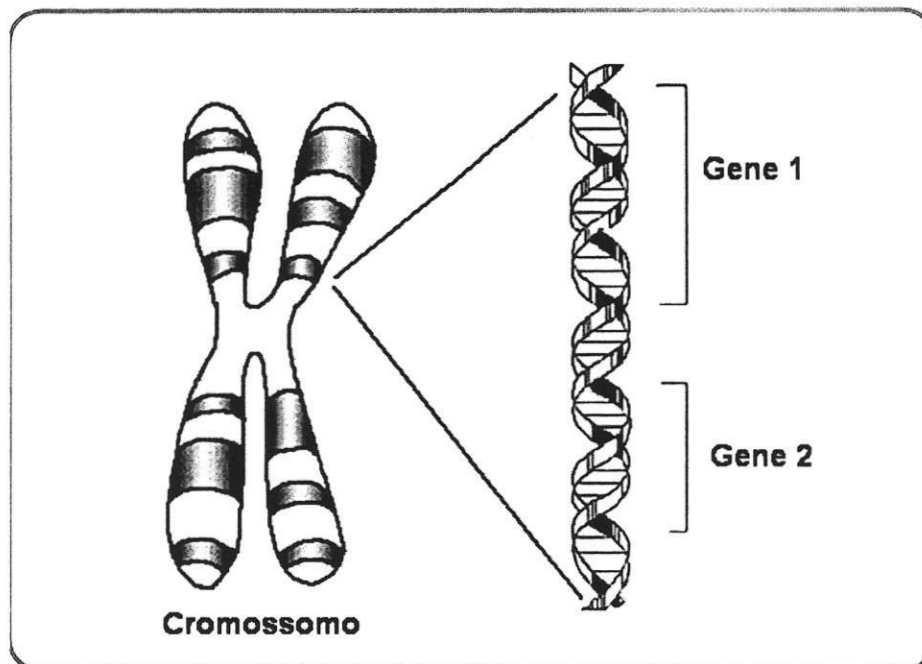


Figura 2: Genes em um cromossomo.

seqüência são complementares às bases da seqüência do DNA. Três tipos de RNA podem ser formados durante a transcrição: o RNA mensageiro (mRNA), que contém a seqüência que codifica uma proteína; o RNA transportador (tRNA), que transporta os aminoácidos até uma estrutura celular chamada ribossomo e possibilita a leitura da informação contida no mRNA durante a tradução e o RNA ribossômico (rRNA), que faz parte da estrutura dos ribossomos. Na produção do mRNA, o início do gene que codifica a proteína é reconhecido na molécula de DNA. Essa região do DNA é desenrolada e suas fitas são separadas. Em seguida, a seqüência do gene é utilizada para gerar uma fita de RNA, complementar à seqüência do gene. Nos eucariotos, esse RNA precisa passar por algumas transformações para que represente a seqüência exata para a produção da proteína. Nesses organismos, essa etapa ocorre no interior do núcleo. Esse RNA “pronto” é o mRNA.

A tradução é a etapa em que o mRNA é utilizado como molde para a fabricação de proteína. A tradução é realizada nos ribossomos. Nos eucariotos, os ribossomos estão localizadas no citoplasma. Neste caso, o mRNA sai do núcleo para participar da tradução. O código do mRNA é lido conforme ele passa pelo interior do ribossomo. Os aminoácidos são adicionados à proteína, um a um, conforme são trazidos aos ribossomos por tRNAs. Os tRNAs fazem a conexão entre uma trinca de nucleotídeos e um aminoácido, permitindo que a seqüência do RNA seja traduzida na seqüência da proteína, de acordo com o código genético. A cada três nucleotídeos do mRNA que passam pelo ribossomo, um tRNA, com uma trinca de nucleotídeos com código complementar ao mRNA, se liga ao ribossomo trazendo o aminoácido adequado. Esse aminoácido é ligado à cadeia de aminoácidos que está sendo construída e que corresponderá à proteína. Após a tradução, as proteínas, geralmente, ainda passam por algumas modificações para que possam exercer adequadamente suas funções.

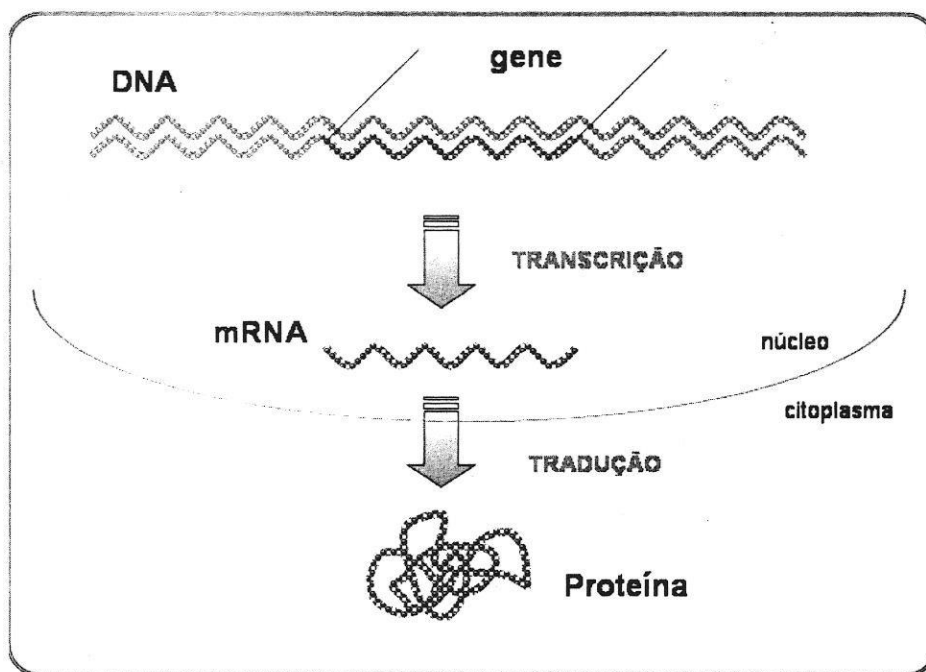


Figura 3: Processo de produção de uma proteína em eucariotos.

Uma visão resumida do processo de expressão gênica pode ser encontrada em diversos livros e artigos sobre bioinformática ou que abordem algum aspecto da análise de expressão gênica, como Setúbal & Meidanis (1997) e Valafar (2002). Uma visão detalhada pode ser encontrada em Lewin (2001).

Como já foi dito, o genoma de um organismo está presente em todas as suas células. Em outras palavras, toda a informação genética necessária ao desenvolvimento e manutenção do organismo sob uma gama enorme de condições ambientais distintas está presente em todas as suas células. Mas, a cada momento, uma célula emprega apenas uma fração consideravelmente reduzida de toda esta informação genética. Isso significa que cada célula produz diferentes tipos de proteínas em diferentes quantidades, em um dado instante.

Todas as fases do processo de expressão de um gene são controladas por uma série de mecanismos que são ativados/desativados de acordo com informações presentes no DNA, com as proteínas e outras substâncias presentes na célula e com fatores ambientais como temperatura. Esses mecanismos de controle determinam quando, onde e em que quantidade as proteínas são sintetizadas. Dentre os fatores que influenciam a atuação desses mecanismos estão: mudanças no ambiente da célula, alterações na forma ou comportamento da célula, idade da célula ou infecção.

Além disso, o tipo e quantidade de proteínas produzidas também dependem do tipo da célula. Células presentes no sangue produzem proteínas diferentes daquelas produzidas pelas células de músculos, por exemplo. Atualmente, estima-se que o DNA de uma célula humana contenha aproximadamente 30.000 genes. Apenas cerca de 20% desses genes estão ativos (produzindo proteínas) em uma célula, em um dado momento. Assim, observando a

quantidade de proteínas presentes em uma célula em um dado momento, pode-se analisar o comportamento da célula sob todos esses fatores e entender melhor o seu comportamento em diversas condições.

Embora o produto final da expressão de um gene seja a proteína, as técnicas mais baratas e utilizadas hoje em dia analisam a expressão de um gene a partir da quantidade de mRNA correspondente à proteína, presente na célula. Embora a relação entre a quantidade de um mRNA e a respectiva proteína em uma célula não seja direta, a abundância de mRNA nas células constitui uma ferramenta importante para os estudos de expressão gênica. A quantidade de mRNA em uma célula, ou conjunto de células, pode ser chamada de nível de expressão do gene Slonim, Tamayo, Mesirov, Golub, & Lander (2000).

A análise do nível de expressão dos genes em uma dada amostra pode fornecer informações importantes para diversas aplicações na área de biologia. Por exemplo, a análise da expressão gênica pode fornecer informações importantes sobre as funções de uma célula, uma vez que as mudanças na fisiologia do organismo são acompanhadas por mudanças nos padrões de expressão dos genes Chan, Hontzeas, & Park (2000). Algumas das aplicações da análise da expressão gênica são:

- Descoberta de genes.
- Determinação de quais genes estão sendo expressos em um determinado tipo de célula em um dado momento e sob certas condições.
- Comparação da expressão dos genes em dois tipos de células diferentes ou duas amostras de tecido diferentes (por exemplo, tecido saudável x tecido doente).
- Observação das mudanças na expressão dos genes em diferentes estágios no ciclo da célula ou durante o desenvolvimento do embrião.
- Estudo dos mecanismos de regulação dos genes.
- Identificação de doenças genéticas complexas.
- Descoberta de medicamentos e estudos de toxicologia.
- Detecção de mutação/polimorfismo nos genes.
- Análise de elemento patogênico.

3 Técnicas para Medir a Expressão Gênica

Atualmente, existem diversas técnicas que permitem identificar o nível de expressão de um grande número de genes simultaneamente, gerando a chamada análise de expressão em larga escala. Dentre essas técnicas podem ser citadas SAGE (*Serial Analysis of Gene Expression*) Velculescu, Zhang, Vogelstein, & Kinzler (1995), MPSS (*Massively Parallel Signature Sequencing Technology*) Brenner, Johnson, Bridgham, Golda, Lloyd, Johnson, Luo, McCurdy, Foy, Ewan, Roth, George, Eletr, Albrecht, Vermaas, Williams, Moon, Burcham, Pallas, DuBridge, Kirchner, Fearon, i Mao, & Corcoran (2000), vários tipos

de *microarray* de DNA (*microarrays* de oligonucleotídeos curtos, *microarrays* de cDNA Schena, Shalon, Davis, & Brown (1995), *microarrays* de oligonucleotídeos longos, *microarrays* de fibra ótica) e *Real-time RT-PCR* (*Reverse Transcription - Polymerase Chain Reaction*). Stanton (2001) faz uma breve revisão de algumas dessas técnicas. Algumas delas são brevemente discutidas a seguir. As pesquisas dos autores estão relacionadas a dados obtidos com as técnicas de *microarray* e SAGE. Assim, algumas informações extras sobre os dados obtidos com essas técnicas são inseridas nas seções seguintes.

3.1 *Microarray de DNA*

O princípio básico por trás da tecnologia de *microarray* de DNA (Harrington, Rosenow, & Retief 2000; Murphy 2002) é a hibridização. Um *microarray* consiste de um conjunto de seqüências de DNA conhecidas (sondas) imobilizadas de forma organizada (com uma configuração conhecida) em uma superfície sólida, em geral, de nylon ou vidro. Conjuntos de sondas correspondentes a cada gene que faz parte do *microarray* são fixos em pontos específicos da superfície, chamados *spots*. Assim, cada *spot* representa um gene conhecido, que é identificado pela sua localização no *microarray*.

A escolha das seqüências de DNA a serem colocadas no *microarray* determina quais genes podem ser detectados. Para organismos com genoma completamente sequenciados pode-se colocar no *microarray* todos os seus genes conhecidos ou ORF (*Open Reading Frames*) suspeitos.

A utilização do *microarray* se dá da seguinte maneira. Inicialmente, o mRNA da amostra de interesse é extraído. Em seguida, cada molécula de mRNA é transformada em uma molécula de DNA complementar (cDNA) marcada de forma radioativa ou fluorescente. Um cDNA é uma molécula de DNA construída em laboratório por meio de transcrição reversa a partir de uma molécula de mRNA. Essa molécula é utilizada no lugar do mRNA por ser mais estável. O cDNA marcado é chamado de alvo. A amostra com os alvos é então hibridizada no *microarray*. Cada alvo se liga a uma sonda complementar no *spot* correspondente a um gene específico. A abundância de cada mRNA na amostra é capturada digitalmente, de acordo com a quantidade de alvo que hibridizou em cada *spot* do *microarray*. A partir do *microarray* com os alvos hibridizados é gerada uma imagem, que é então analisada por meio de um dos vários softwares disponíveis, para quantificar as informações dessa imagem. Os níveis de expressão são derivados de sinais analógicos de absorvência ou fluorescência.

Os dois tipos principais de *microarray* são os de cDNA (Schena, Shalon, Davis, & Brown 1995) e os de oligonucleotídeos. Cada um deles oferece diferentes possibilidades de análise e também diferentes vantagens e desvantagens. Claverie (1999), Stanton (2001), Murphy (2002) e Nguyen, Arpat, Wang, & Carroll (2002) comparam alguns aspectos desses dois tipos de *microarray*.

O *microarray* de cDNA é construído pela impressão de material genético pré-sintetizado na superfície sólida. Neste caso, o DNA sonda vem de organismos naturais e podem ser feitos de DNA genômico ou de cDNA.

Os *microarrays* de cDNA construídos em superfícies de vidro permitem que sejam realizados experimentos competitivos, em que alvos de duas amostras marcadas de forma

diferente, com corantes fluorescentes, sejam hibridizados simultaneamente em um mesmo *microarray*. Nesse caso, os níveis de expressão são dados como a razão entre os níveis de expressão nas duas amostras (Claverie 1999). Yang & Speed (2002) discutem os vários aspectos do projeto de experimentos com *microarrays* de cDNA. Algumas das vantagens e desvantagens do *microarray* de cDNA são:

Vantagens:

- Muito utilizado para quantificação relativa dos níveis de expressão, com a utilização de experimentos competitivos.
- As sondas representam genes de identidade conhecida ou segmentos de DNA funcional, conhecidos como ESTs (*Expressed Sequence Tags*). Pode também ser aplicada para a descoberta de genes, uma vez que não há a necessidade de se conhecer a sequência dos cDNAs sonda.
- Mede a expressão de dezenas de milhares de genes simultaneamente.
- Capacidade de detectar um grande intervalo de níveis de expressão.
- Tem alto custo inicial, tanto em termos financeiros, quanto de trabalho. Porém, depois de instalada é bastante produtiva.

Desvantagens:

- Necessita de grande quantidade de RNA total para preparar o alvo (50 - 200 μ g de RNA total).
- Não é capaz de detectar genes com baixo nível de expressão.
- Apresenta risco de fragmentos do alvo hibridizarem com sondas complementares similares, produzindo detecções falsas.

O *microarray* de oligonucleotídeos é construído pela síntese das sequências de interesse (oligonucleotídeos) *in situ*, no próprio substrato, sendo cada ponto sintetizado, nucleotídeo por nucleotídeo, utilizando uma máscara de luz. Algumas das vantagens e desvantagens do *microarray* de oligonucleotídeos são:

Vantagens:

- Quantifica níveis absolutos de expressão, permitindo comparações confiáveis dos valores obtidos em experimentos separados.
- Mede a expressão de dezenas de milhares de genes simultaneamente.
- Exige uma quantidade menor de RNA total do que o *microarray* de cDNA (5 μ g de RNA total).
- Tem a capacidade de detectar um intervalo de níveis de expressão maior do que o *microarray* de cDNA.
- Apresenta menor risco de fragmentos do alvo hibridizarem com sondas complementares similares do que o *microarray* de cDNA.

Desvantagens:

- É uma tecnologia proprietária, desenvolvida e comercializada pela Affymetrix.
- É menos flexível que o *microarray* de cDNA, pois o usuário não pode selecionar as sondas para compor o *microarray*. Os *microarrays* são comprados prontos. Existe a possibilidade de se especificar os oligonucleotídeos (sondas) desejados, mas o custo é muito alto.
- As sondas são projetadas especificamente para cada gene de interesse, sendo necessário o conhecimento da sequência dos genes.
- Não possibilita a descoberta de genes novos.
- Tem alto custo.

3.2 SAGE - Serial Analysis of Gene Expression

O método SAGE (Velculescu, Zhang, Vogelstein, & Kinzler 1995) se baseia no fato de que uma sequência curta, com 10 a 14 pares de base, chamada *tag*, contém informações suficientes para identificar unicamente um transcrito (mRNA), desde que a *tag* seja obtida de uma única posição dentro de cada mRNA. O método SAGE captura os mRNA da amostra de interesse, reescreve-os na forma de cDNA (DNA complementar), e corta uma pequena *tag* de cada um. Em seguida, essas *tags* são unidas em moléculas longas, chamadas *concatemers*, que são sequenciadas.

O resultado do sequenciamento é analisado por um software que conta o número de *tags* e obtém uma lista dos genes a que elas pertencem. O nível de expressão de um gene, neste caso, é dado pelo número de *tags* associadas ao gene que foram encontradas na amostra. Algumas das vantagens e desvantagens desta técnica são:

Vantagens:

- É capaz de detectar e quantificar o nível de expressão absoluto de genes conhecidos e desconhecidos. Não é necessário o conhecimento da sequência dos genes. Isso permite a descoberta de novos genes.
- Pode ser aplicada em qualquer laboratório que possui equipamento de sequenciamento, o que é muito comum atualmente.
- Não requer repetição e padronização do experimento com uma mesma amostra, uma vez que estatísticas para dados de contagem são bem modeladas pela distribuição de Poisson.

Desvantagens:

- Precisa de uma quantidade grande de mRNA (2,5 - 5 μ g de mRNA). Porém, já existe a técnica microSAGE que utiliza menos mRNA (5ng de mRNA).
- É uma técnica complexa, que requer muito sequenciamento.
- Precisa de um sequenciamento de alta qualidade para a identificação correta dos mRNAs.

- Não é muito adequada para detectar transcritos raros. Transcritos identificados como tendo menos de 20 cópias por célula não são estatisticamente confiáveis. Isso significa que a técnica tem baixa sensibilidade.

3.3 MPSS - *Massively Parallel Signature Sequencing*

MPSS (Brenner, Johnson, Bridgham, Golda, Lloyd, Johnson, Luo, McCurdy, Foy, Ewan, Roth, George, Eletr, Albrecht, Vermaas, Williams, Moon, Burcham, Pallas, DuBridge, Kirchner, Fearon, i Mao, & Corcoran 2000; Lynx 2004) é uma tecnologia proprietária da Lynx Therapeutics, que é utilizada em conjunto com outra tecnologia proprietária, Megacclone, para quantificar o nível de expressão gênica.

Para a aplicação dessa técnica são utilizadas moléculas de cDNA obtidas a partir da amostra de interesse, composta por moléculas de mRNA. Inicialmente, a tecnologia Megacclone é empregada para transformar a amostra de cDNA em micro-contas, cada uma representando uma molécula de cDNA e contendo cerca de 100.000 cópias desta molécula. Em seguida, em um processo paralelo, a MPSS identifica uma sequência curta com 17 a 20 pares de base, a partir de uma posição definida dentro da molécula representada em cada micro-conta. Essa sequência é chamada assinatura. Cada assinatura identifica um transcrito (mRNA) diferente. Num próximo passo, as assinaturas são contadas, determinando-se o número de moléculas de mRNA de cada gene presente na amostra. Isso é feito utilizando ferramentas de bioinformática. O nível de expressão é representado pelo número de transcritos presentes por milhão de moléculas (tpm). Algumas das vantagens e desvantagens desta técnica são:

Vantagens:

- Quantifica o nível de expressão absoluto identificando e contando quase todas as espécies de mRNA de uma amostra (tipicamente mais de 20.000). Geralmente conta-se 1 milhão de mRNAs em um experimento.
- Não requer o conhecimento da sequência dos genes, sendo adequada para a descoberta de novos genes.
- Tem uma alta sensibilidade, sendo capaz de detectar mRNAs com menos de 5 cópias por célula.
- Da mesma forma que a SAGE, não requer repetição e padronização dos experimentos.
- Possui uma excelente escala dinâmica de quantificação. É capaz de detectar desde 5 tpm até mais de 50.000 tpm.
- Pode detectar quase todos os genes presentes na amostra. Exceções são genes que não possuem a sequência empregada por essa tecnologia para criar as assinaturas, genes com assinaturas difíceis de mapear no genoma, genes cuja assinatura possui algum artefato que interfere no sequenciamento, genes cuja assinatura apresenta erros de sequenciamento e genes cujo transcrito possui menos de 100 pares de base. Todas essas exceções são bastante raras.

Desvantagens:

- Precisa de 100 μ g de RNA total, de preferência, ou 2 μ g de mRNA.
- É uma tecnologia proprietária da Lynx Therapeutics.

3.4 RT-PCR - Reverse Transcription - Polymerase Chain Reaction

Existem vários métodos para quantificar RNA através de PCR, permitindo a avaliação dos níveis de mRNA a partir de amostras pequenas (até uma única célula) (Freeman, Walker, & Vrana 1999; Bustin 2002). A análise dos níveis de expressão gênica é apenas uma das aplicações dessas técnicas. O método básico é o RT-PCR (*Reverse Transcription - Polymerase Chain Reaction*).

Inicialmente, cDNA é produzido a partir da amostra de mRNA, por transcrição reversa. Em seguida, o cDNA é amplificado por PCR. Finalmente, o produto da amplificação é detectado e quantificado. Existem duas classes principais de técnicas para a detecção do produto da amplificação: medidas *end-point*, que analisam a reação depois de completa (RT-PCR convencional), e monitoramento em tempo real da formação do produto da PCR, que monitora a reação conforme ela vai ocorrendo (*real time* RT-PCR). A quantificação pode ser relativa ou absoluta. Uma quantificação relativa determina as mudanças na expressão de um gene. Uma quantificação absoluta tenta determinar o número de cópias de um mRNA específico por célula. Este tipo de quantificação requer condições e tratamentos especiais, não necessários para a quantificação relativa.

Os experimentos com os métodos que empregam RT-PCR são muito comuns para a caracterização e confirmação de padrões de expressão gênica obtidos com outras técnicas, sendo muito comum sua utilização para validação dos resultados de *microarrays*. Algumas das vantagens e desvantagens desta técnica são:

Vantagens:

- É um método naturalmente adequado para uma quantificação indireta (relativa) dos níveis de expressão, mas com condições e tratamentos especiais pode-se obter uma quantificação absoluta.
- Tem uma alta sensibilidade, sendo capaz de detectar mRNAs com menos de 5 cópias por célula.
- Trabalha muito bem com pequenas quantidades de RNA total, podendo ser aplicada ao conteúdo de uma única célula.
- É sensível e preciso em uma escala dinâmica ampla.
- Tem sido utilizada como complemento à análise de expressão realizada com outras técnicas.

Desvantagens:

- É apropriada para manipular apenas poucas dúzias de genes por experimento.
- Não é apropriada para a descoberta de novos genes.

3.5 Comparação das Técnicas de Medição da Expressão Gênica

A Tabela 1 apresenta um resumo comparando algumas características das técnicas de medição da expressão gênica descritas. Os números na tabela correspondem às seguintes características:

1 - Descoberta de genes. Indica se a técnica é apropriada para a identificação de novos genes.

2 - Escala de quantificação. Se refere ao intervalo de valores dos níveis de expressão que pode ser detectado.

3 - Sensibilidade. Indica a capacidade de detecção de mRNAs raros, com poucas cópias por célula.

4 - Tamanho da amostra. Indica a quantidade de mRNA ou RNA total necessária, inicialmente, para se realizar a medição.

5 - Trabalho exigido para a realização de um experimento.

6 - Custo do *set-up* inicial.

Tabela 1: Comparação entre as técnicas de medição da expressão gênica.

Técnica	1	2	3	4	5	6
Microarray de cDNA	sim	amplo	média	50-200 μ g de RNA total	alto	alto
Microarray de oligonucleotídeos	não	> que cDNA	média	5 μ g de RNA total	baixo	alto
SAGE	sim	-	baixa	25-5 μ g de mRNA	médio	baixo
MPSS	sim	excelente	alta	100 μ g de RNA total	baixo	alto
RT-PCR	não	amplo	alta	conteúdo de uma célula	baixo	médio

Algumas informações da tabela merecem comentários. A técnica de *microarray* de cDNA pode detectar novos genes desde que sejam empregadas bibliotecas de cDNA com seqüências desconhecidas. O custo inicial para a técnica SAGE é baixo se houver equipamento de sequenciamento disponível no laboratório. O Trabalho e custo do *microarray* de oligonucleotídeos leva em consideração a compra do *microarray* pronto. No *microarray* de cDNA, o *microarray* deve ser construído. Para a técnica MPSS, pode-se considerar o trabalho necessário pequeno, pois toda a análise é feita pela Lynx, sendo necessário apenas o envio das amostras para a empresa. O seu custo é alto pelo mesmo motivo.

Claverie (1999) resume os métodos de obtenção de expressão gênica em métodos baseados em contagem de *tags* (SAGE, MPSS) e métodos baseados em hibridização (*microarrays* de cDNA e de oligonucleotídeos).

Nos métodos de contagem de *tags*, cada experimento diz respeito a uma amostra representando uma condição de interesse. Esse tipo de método é estatisticamente mais robusto. Não é necessária a repetição nem a padronização dos experimentos, uma vez que estatísticas para contagem são bem modeladas pela distribuição de Poisson (Brenner, Johnson, Bridgham, Golda, Lloyd, Johnson, Luo, McCurdy, Foy, Ewan, Roth, George, Eletr, Albrecht, Vermaas, Williams, Moon, Burcham, Pallas, DuBridge, Kirchner, Fearon,

i Mao, & Corcoran 2000). O resultado de um experimento é uma biblioteca contendo a frequência de cada *tag* na amostra. O número total de *tags* sequenciadas varia de poucas centenas a várias dezenas de milhares.

Nos métodos baseados em hibridização, um experimento pode se referir a uma única amostra da condição de interesse ou à relação entre duas amostras de condições diferentes. Em geral, os valores de expressão têm medidas replicadas (internamente em um *microarray*, ou em mais de um *microarray*) e são normalizados de acordo com um controle interno.

Além da tecnologia empregada, outro aspecto importante é a técnica experimental. Vingron & Hoheisel (1999) fazem uma distinção das técnicas experimentais possíveis de acordo com os dados que elas produzem:

- Técnicas que especificamente comparam duas amostras de mRNA, como o *microarray* de cDNA em experimento competitivo.
- Técnicas que contam o número de moléculas de cada espécie de mRNA presente em uma amostra, como a SAGE.
- Técnicas que lêem o conteúdo de mRNA de uma amostra a partir da intensidade de um sinal de hibridização, como o *microarray* de oligonucleotídeos.

No primeiro caso, os níveis de expressão são relativos, isto é, são dados pela razão dos valores de expressão nas duas condições analisadas, A e B, por exemplo. Genes mais expressos na condição A do que em B são ditos hiper-expressos em A e genes menos expressos são ditos hipo-expressos. Com os dados obtidos pelas duas últimas técnicas também é possível fazer análises comparativas, mas não com um único experimento.

A partir da realização de um conjunto de experimentos com uma das técnicas descritas, pode-se construir perfis de expressão dos genes, caracterizando o funcionamento dinâmico de cada gene (Brazma & Vilo 2000). Cada perfil corresponde aos níveis de expressão de todos os genes em uma amostra. Um perfil de expressão pode ser baseado em:

- Diferentes condições a que se submeta as amostras.
- Diferentes estágios de desenvolvimento do organismo ou células de interesse.
- Diferentes tecidos.

Isso depende das amostras empregadas. As amostras podem ser, por exemplo:

- Tecidos de várias partes do organismo.
- Tecidos com diversas doenças.
- Células submetidas a diferentes tratamentos com medicamentos.

- Células submetidas a várias condições (por exemplo de alimentação ou temperatura).
- Amostras obtidas em vários pontos no tempo (para avaliar, por exemplo, o desenvolvimento normal das células da amostra, seu desenvolvimento sob determinadas condições ou após algum tipo de tratamento).

4 *Processamento e Análise Computacional dos Dados de Expressão Gênica*

Os dados de expressão gênica são geralmente representados em uma matriz, com as linhas representando os genes e as colunas representando as amostras (diferentes tecidos, estágios de desenvolvimento ou tratamentos, por exemplo). Essa matriz é chamada matriz de expressão. Cada posição da matriz contém um número representando o nível de expressão de um gene particular em uma amostra. Como já foi dito, as amostras, em geral, provêm de diferentes experimentos. Devido ao alto custo dos experimentos, uma matriz de expressão é, geralmente, constituída de um grande número de genes (milhares) e poucas amostras (dezenas). A Tabela 2 mostra um exemplo de uma matriz de expressão.

Tabela 2: Exemplo de uma matriz de expressão.

Gene	Amostra 1	Amostra 2	...	Amostra m
gene 1	1009	34		500
gene 2	37	657		45
gene 3	2	456		3512
...				
gene n	0	5		1690

A aplicação de qualquer uma das técnicas de medição da expressão gênica produz grandes quantidades de dados, requerendo métodos computacionais para auxiliar na sua interpretação. A análise desses dados precisa ser cuidadosa porque os sinais biológicos podem ser obscurecidos por ruídos provenientes dos experimentos ou outras influências sistemáticas provenientes da metodologia experimental (Vingron & Hoheisel 1999).

Dependendo de como os níveis de expressão sejam obtidos, podem ser necessárias diferentes transformações para adequar os dados para utilização.

Para os dados obtidos com a técnica SAGE, em geral são necessários ajustes nos dados para remover erros de sequenciamento, adicionar valores ausentes em bibliotecas truncadas e transformar as frequências das *tags* de forma a fazer com que todas as bibliotecas representem o mesmo número total de *tags* (Ng, Sander, & Sleumer 2001).

Vários passos de normalização também são necessários nos dados obtidos com *microarrays*, uma vez que existem muitas fontes de variação sistemática nos experimentos com essas técnicas que afetam os níveis de expressão medidos (Dudoit, Yang, Callow, & Speed 2000; Schuchhardt, Beule, Malik, Wolski, Eickhoff, Lehrach, & Herzog 2000; Yang, Dudoit, Luu, & Speed 2001; Irizarry, Hobbs, Collin, Beazer-Barclay, Antonellis, Scherf, & Speed

2003; Hariharan 2003). Alguns passos se aplicam a cada *microarray*, individualmente, enquanto outros se aplicam à matriz de expressão.

Depois dessas transformações, em geral, os dados estão prontos para serem analisados. Porém, dependendo da análise a ser feita, ainda podem ser necessárias outras manipulações (Quackenbush 2001).

Os níveis de expressão dos genes podem ser convertidos para um intervalo comum, por exemplo, $[0, 1]$ ou $[-1, 1]$ (Dopazo, Zanders, Dragoni, Amphlett, & Falciani 2001). Esse tipo de transformação é útil para identificar genes que mostram um comportamento similar, mas possuem níveis de expressão em diferentes intervalos. Essa transformação é sensível a *outliers*. Alguns algoritmos de aprendizado de máquina fazem uso de dados em determinados intervalos, sendo esta uma transformação necessária, se esses algoritmos forem empregados.

Outra transformação consiste na subtração da média dos valores e divisão pelo seu desvio padrão. Com essa transformação, os dados passam a ter média zero e variância um. Isso reduz o efeito de diferentes escalas nos dados, sendo menos sensível a *outliers* do que a conversão dos dados para um único intervalo. Algumas técnicas de análise dos dados, como PCA (*Principal Component Analysis*), requerem que esse tipo de transformação seja aplicada aos dados.

Algumas técnicas de análise estatísticas dependem da distribuição dos dados. Embora existam técnicas que são independentes da distribuição (testes não-paramétricos), os testes paramétricos apresentam maior poder estatístico. Porém, estes testes exigem que os dados tenham uma distribuição normal (Dopazo, Zanders, Dragoni, Amphlett, & Falciani 2001).

Os dados de expressão geralmente têm uma distribuição normal depois de uma transformação logarítmica. Esse tipo de transformação é muito comum nos dados de *microarray* (Dopazo, Zanders, Dragoni, Amphlett, & Falciani 2001; Murphy 2002; Hariharan 2003). Além da distribuição se tornar aproximadamente normal após esse tipo de transformação, ela auxilia na visualização dos dados e faz com que os níveis de expressão dos genes hiper e hipo-expressos, resultantes de experimentos competitivos, tenham mesma magnitude e sinais opostos (Quackenbush 2001).

Apesar das transformações nos dados serem bastante apropriadas para determinadas situações, elas devem ser aplicadas de forma bastante criteriosa, e somente quando se fizerem realmente necessárias, pois cada tipo de transformação leva à perda de algum tipo de informação contida nos dados (Sarle 2002). Hariharan (2003) discute, entre outros aspectos, a aplicação da transformação logarítmica, incluindo algumas desvantagens de se aplicar essa transformação.

Existem inúmeras formas de análise que podem ser aplicadas a dados de expressão gênica, variando desde análises de baixo nível, como filtragens e normalizações, até análises estatísticas para identificação de expressão diferente em diversas condições, identificação de padrões e mineração de dados. Dependendo do interesse do pesquisador, uma ou mais dessas formas de análise podem ser aplicadas aos dados. Porém, é preciso ter em mente que os métodos utilizados para análise têm profunda influência na interpretação dos resultados. Assim, é necessário um entendimento dos métodos utilizados para se fazer

um projeto experimental adequado e uma análise significativa dos dados.

Segundo Quackenbush (2001), não existe consenso sobre o melhor método para revelar padrões de expressão nos dados e fica cada vez mais claro que não existe uma técnica superior. A aplicação de várias técnicas permite explorar aspectos diferentes dos dados.

Sebastiani, Gussoni, Kohane, & Ramoni (2003) discutem algumas análises que podem ser feitas, sob o ponto de vista experimental. A análise mais simples que pode ser feita com dados de expressão gênica é a identificação de genes expressos de forma diferente em duas condições. As condições podem ser níveis específicos de fatores ambientais controláveis (temperaturas extremas ou falta de alimentação) ou a modificação ou remoção de uma porção específica do genoma.

Questões experimentais mais complexas envolvem o traçado do perfil molecular de várias condições de uma só vez para caracterizar, por exemplo, a “impressão digital” genômica de diferentes tipos de câncer (Alizadeh, Eisen, Davis, Ma, Lossos, Rosenwald, Boldrick, Sabet, Tran, Yu, Powell, Yang, Marti, Moore, Hudson, Lu, Lewis, Tibshirani, Sherlock, Chan, Greiner, Weisenburger, Armitage, Warnke, Levy, Wilson, Grever, Byrd, Botstein, Brown, & Staudt 2000) ou o efeito da mudança simultânea de vários fatores experimentais (Churchill & Oliver 2001). Nestes casos, cada amostra consiste dos níveis de expressão obtidos de células em uma condição particular.

Ainda outra questão se refere ao estudo da evolução temporal de perfis de expressão gênica. Esse tipo de estudo está baseado na intuição de que genes que têm um perfil de expressão similar no decorrer do tempo estão agindo juntos por pertencerem à mesma categoria funcional, ou à categorias similares. Com esse tipo de dado é possível observar os mecanismos regulatórios em ação e tentar dividir o genoma em conjuntos de genes que estão envolvidos nos mesmos processos ou em processos relacionados. Embora necessite de uma análise diferente da realizada para o tipo de experimento anterior, neste caso também são comparados perfis moleculares de células em diferentes condições. Exemplos desse tipo de estudo são o processo de transformação de um tumor que cresce localmente em uma metástase (Clark, Golub, Lander, & Hynes 2000), o ciclo de esporulação de levedura (Spellman, Sherlock, Zhang, Iyer, Anders, Eisen, Brown, Botstein, & Fitcher 1998) e a resposta de fibroblastos humanos ao soro (Iyer, Eisen, Ross, Schuler, Moore, Lee, Trent, Staudt, JR, Boguski, Lashkari, Shalon, Botstein, & Brown 1999).

Outros experimentos se referem ao mecanismo regulatório de células observadas em diferentes condições experimentais. Os caminhos regulatórios em células normais modulam o nível e a duração da expressão dos genes, garantindo que as células respondam apropriadamente aos estímulos fisiológicos e extra-celulares. Porém, quando a ativação de um caminho de regulação de um gene dispara uma sub ou super-produção de certas proteínas, podem surgir diversas doenças. Neste tipo de experimento, questões importantes são a descoberta de novos mecanismos regulatórios e de dependências causais entre expressão gênica (Pilpel, Sudarsanam, & Church 2001).

Quackenbush (2001) faz uma revisão de algumas abordagens computacionais para a análise de dados provenientes de experimentos competitivos em *microarrays* de cDNA. Essas abordagens também são aplicáveis a dados obtidos com outros tipos de experimento com *microarray* ou SAGE, desde que os dados se apresentem no formato adequado.

Claverie (1999) faz uma análise das formas de identificação de genes expressos de forma diferente ou coordenada a partir de dados de expressão gênica. Dudoit, Yang, Callow, & Speed (2000) também discutem métodos estatísticos para identificar genes expressos de forma diferente em experimentos replicados de *microarray*. Outros artigos relacionados à identificação de genes expressos de forma diferente incluem (Pan 2002), (Pepe, Longton, Anderson, & Schummer 2003), (Broberg 2003) e (Jaeger, Sengupta, & Ruzzo 2003).

Lockhart & Winzeler (2000) discutem vários aspectos relacionados à expressão gênica e às informações que podem ser obtidas analisando esse tipo de dado. Molla, Waddell, Page, & Shavlik (2004) fazem uma breve revisão da tecnologia de *microarray* de DNA e descrevem algumas formas de análise que podem ser feitas em dados obtidos com essa tecnologia, utilizando algoritmos de aprendizado de máquina. Essas análises incluem: predição da função biológica de um gene, agrupamento de genes com base em seus níveis de expressão, agrupamento de condições experimentais com base nos níveis de expressão, descoberta de novas classes de doenças, predição de doenças existentes, predição do prognóstico de pacientes e predição da resposta de diferentes pacientes com uma dada doença à diversos medicamentos. Embora essas análises sejam descritas no contexto de dados de *microarray* de DNA, elas podem ser aplicadas a dados de expressão gênica em geral. Brazma & Vilo (2000) fazem uma breve descrição de algumas possibilidades de análise de dados de expressão gênica.

5 Conclusão

A análise de dados de expressão gênica é bastante útil para diversas aplicações na biologia. A extração de informações úteis desse tipo de dado depende de análises criteriosas. Essa análise pode ser realizada em diferentes níveis e de diferentes formas.

Este relatório apresentou os principais aspectos relacionados à análise de expressão gênica, resumindo algumas técnicas de medição dos níveis de expressão gênica que têm sido utilizadas e discutindo algumas formas de manipulação e análises que podem ser realizadas em dados de expressão gênica.

Inicialmente, o processo de expressão gênica foi descrito. As técnicas *microarray*, SAGE, MPSS e RT-PCR foram apresentadas e comparadas. Foram também discutidos o processamento necessário e algumas possíveis análises dos dados de expressão gênica.

Referências

- Alizadeh, A., M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, J. Powell, L. Yang, G. Marti, T. Moore, J. J. Hudson, L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W. Chan, T. Greiner, D. Weisenburger, J. Armitage, R. Warnke, R. Levy, W. Wilson, M. Grever, J. Byrd, D. Botstein, P. Brown, & L. Staudt (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature* 403(6769), 503–511.
- Baldi, P., & S. Brunak (1998). *Bioinformatics: The Machine Learning Approach*. Adaptive Computation and Machine Learning. MIT Press.
- Brazma, A., & J. Vilo (2000). Gene expression data analysis. *FEBS Letters* 480(1), 17–24.
- Brenner, S., M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. i Mao, & K. Corcoran (2000, June). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology* 18(6), 630–634.
- Broberg, P. (2003). Statistical methods for ranking differentially expressed genes. *Genome Biology* 4(6), R41.
- Bustin, S. (2002). Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *J Mol Endocrinol* 29(1), 23–39.
- Chan, V., N. Hontzeas, & V. Park (2000, December). Gene expression. Preprint (Univ Warterloo), <http://www.nslj-genetics.org/microarray/preprint.html> [Acessado em 11/01/2005].
- Churchill, G. A., & B. Oliver (2001). Sex, flies and microarrays. *Nature Genetic* 29, 355–356.
- Clark, E., T. Golub, E. Lander, & R. Hynes (2000). Genomic analysis of metastasis reveals an essential role for RhoC. *Nature* 406, 532–535.
- Claverie, J.-M. (1999). Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics* 8(10), 1821–1832.
- Dopazo, J., E. Zanders, I. Dragoni, G. Amphlett, & F. Falciani (2001). Methods and approaches in the analysis of gene expression data. *Journal of Immunological Methods* 250, 93–112.
- Dudoit, S., H. Yang, M. Callow, & T. Speed (2000). Statistical methods for identifying genes with differential expression in replicated cDNA experiments. *Statistica Sinica* 12(1), 111–139.
- Freeman, W., S. Walker, & K. Vrana (1999). Quantitative RT-PCR: pitfalls and potential. *BioTechniques* 26(1), 112–125.
- Hariharan, R. (2003). The analysis of microarray data. *Pharmacogenomics* 4(4), 477–497.

- Harrington, C., C. Rosenow, & J. Retief (2000). Monitoring gene expression using DNA microarrays. *Curr. Opin. Microbol.* 3, 285–291.
- Irizarry, R., B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, & T. Speed (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2), 249–264.
- Iyer, V. R., M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. H. JR, M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein, & P. Brown (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* 283, 83–87.
- Jaeger, J., R. Sengupta, & W. L. Ruzzo (2003, Jan). Improved gene selection for classification of microarrays. In *Pacific Symposium on Biocomputing*, Kauai, Hawaii, pp. 53–64.
- Lewin, B. (2001). *Genes VII*. Artmed editora.
- Lockhart, D., & E. Winzeler (2000). Genomics, gene expression and DNA arrays. *Nature* 405(6788), 827–836.
- Lynx (2004). Lynx Therapeutics home page. <http://www.lynxgen.com/> [Acessado em 14/01/2004].
- Molla, M., M. Waddell, D. Page, & J. Shavlik (2004). Using machine learning to design and interpret gene-expression microarrays. *AI Magazine (Special Issue on Bioinformatics)* 25(1), 23–44.
- Murphy, D. (2002). Gene expression studies using microarrays: Principles, problems, and prospects. *Advan. Physiol. Educ.* 26(4), 256–270.
- Ng, R. T., J. Sander, & M. C. Sleumer (2001). Hierarchical cluster analysis of SAGE data for cancer profiling. In *Workshop on Data Mining in Bioinformatics (BI-OKDD01)*, pp. 65–72.
- Nguyen, D., A. Arpat, N. Wang, & R. Carroll (2002). DNA microarray experiments: biological and technological aspects. *Biometrics* 58, 701–717.
- Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 18(4), 546–554.
- Pepe, M. S., G. Longton, G. L. Anderson, & M. Schummer (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics* 59(1), 133–142.
- Pilpel, Y., P. Sudarsanam, & G. Church (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* 29, 153–159.
- Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics* 2(6), 418–427.
- Sarle, W. S. (2002). Artificial neural network FAQ - should I normalize/standardize/rescale the data? <ftp://ftp.sas.com/pub/neural/FAQ.html> [Acessado em 09/12/2003].
- Schena, M., D. Shalon, R. W. Davis, & P. O. Brown (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470.

- Schuchhardt, J., D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach, & H. Herzel (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Research* 28(10), e47.
- Sebastiani, P., E. Gussoni, I. Kohane, & M. Ramoni (2003). Statistical challenges in functional genomics (with discussion). *Statistical Science* 18(1), 33–70.
- Setúbal, J. C., & J. Meidanis (1997). *Introduction to Computational Molecular Biology*. PWS Publishing Company.
- Slonim, D. K., P. Tamayo, J. P. Mesirov, T. R. Golub, & E. S. Lander (2000). Class prediction and discovery using gene expression data. In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB)*, Tokyo, Japan, pp. 263–272. Universal Academy Press. <http://citeseer.nj.nec.com/slونim00class.html> [Acessado em 19/07/2002].
- Spellman, P., G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Bostein, & B. Futcher (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* 9, 3273–3297.
- Stanton, L. W. (2001). Methods to profile gene expression. *Trends in Cardiovascular Medicine* 11(2), 49–54.
- Valafar, F. (2002). Pattern recognition techniques in microarray data analysis: A survey. *Special Issue Annals of New York Academy of Sciences, Techniques in Bioinformatics and Medical Informatics* 980, 41–64.
- Velculescu, V. E., L. Zhang, B. Vogelstein, & K. W. Kinzler (1995). Serial analysis of gene expression. *Science* 270, 484–487.
- Vingron, M., & J. Hoheisel (1999). Computational aspects of expression data. *J. Mol. Med.* 77, 3–7.
- Yang, Y., & T. P. Speed (2002). Design issues for cDNA microarray experiments. *Nature Review Genetics* 3(8), 579–588.
- Yang, Y. H., S. Dudoit, P. Luu, & T. P. Speed (2001). Normalization for cDNA microarray data. In *SPIE BiOS 2001*, San Jose, California.