UNIVERSIDADE FEDERAL DE PERNAMBUCO

GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA

2008.1

UM ESTUDO SOBRE PROCESSAMENTO E ANÁLISE DE IMAGENS DE MICROARRANJOS DE DNA

TRABALHO DE GRADUAÇÃO

Aluno: Rodrigo Silva Campos (rsc3@cin.ufpe.br)

Orientador: Tsang Ing Ren (tir@cin.ufpe.br)

Recife, julho de 2008

Agradecimentos

Agradeço ao meus pais, os quais me ensinaram importantes valores que carregarei sempre. Obrigado por todo o carinho, por sempre quererem o melhor para os seus filhos e por dar total apoio na busca pelos meus objetivos.

Agradeço às minhas irmãs por serem compreensivas com a minha ausência e ainda sim terem boas palavras para se dizer nos momentos de encontros da família.

Agradeço à Mila, que soube durante esse período ser compreensiva e paciente, que sempre esteve pronta para dar carinho nas horas de sufoco e disposta a ajudar. Não posso esquecer também de sua família que não deixaram de me apoiar além de intercederem por mim nos momentos de crise, obrigado a todos.

Agradeço à Tsang, pelo apoio e incentivo nesse trabalho, por estar presente nas horas em que precisei além de dar boas orientações.

Agradeço à todos os meus companheiros de turma no Cin, principalmente àqueles que preferiram sempre fazer os projetos durante a madrugada, sempre com bom humor e competência. Obrigado à todos que tentaram e conseguiram me carregar para longe do computador quando eu não reconhecia que estava precisando esquecer o peso das responsabilidades.

Resumo

Microarranjos de DNA consistem num conjunto ordenado de milhares de moléculas de DNA cuja seqüência é conhecida. Dessa forma é criada uma matriz de fragmentos genéticos distintos e posicionados numa ordem pré-definida, a qual pode ter sua imagem capturada bem como digitalizada. Isso permite avaliar a expressão de milhares de genes simultaneamente através de métodos de processamento computacional de imagens. A técnica que utiliza microarranjos de DNA tem se revelado uma poderosa ferramenta para a análise de experimentos genéticos contribuindo, por exemplo, na busca de tratamento e descoberta de novos tipos de doenças além de predizer ou diagnosticar aquelas cuja expressão genética é conhecida. Este trabalho propõe a pesquisa das principais técnicas utilizadas no processamento e análise de imagens de microarranjos de DNA com o intuito principal de obter o arcabouço necessário à comparação de resultados e sugestão de melhorias.

Palavras-chave: microarranjos de DNA, expressão gênica, processamento de imagens, algoritmos de clusterização.

Abstract

Microarrays of DNA consist of an ordered set of thousands of molecules of DNA whose sequence is known. Thus is created a matrix of different genetic fragments and placed in a pre-defined order, which can have their image captured and digitized. This allows evaluating the expression of thousands of genes simultaneously through computational methods for processing of images. The technique that uses microarrays DNA has revealed a powerful tool for the analysis of genetic experiments contributing, for example, in the search for treatment and discovery of new types of diseases in addition to predict or diagnose those whose gene expression is unknown. This paper proposes a search of the main techniques used in the processing and analysis of images from microarrays DNA with the primary purpose of getting the framework necessary for the comparison of results and suggestions for improvements.

Keywords: DNA microarrays, gene expression, image process, algorithms of clustering.

Sumário

Índice de Figuras

Índice de Equações

1. Introdução

O rápido crescimento do volume de dados gerados pelo seqüenciamento de genomas assim como a quantidade de dados sobre expressão gênica vêm tornando cada vez mais complexo o entendimento das funções dos genes nos organismos. Dessa forma, há sempre a demanda por métodos que possibilitem o processamento e a análise dos dados não somente de maneira mais eficiente, mas que também mantenham um alto grau de confiabilidade. A técnica que utiliza microarranjos de DNA surge nesse cenário, possibilitando o estudo da expressão gênica sob diversas condições e a um custo de tempo cada vez menor. Nos experimentos de microarranjos de DNA são produzidas imagens de expressões gênicas de um organismo. A partir do processamento dessas imagens é possível identificar e quantificar os dados biológicos cuja análise tem levado à várias descobertas e implicações. Por exemplo, os dados obtidos de experimentos com microarranjos vêm auxiliando a busca, diagnóstico e tratamento de diversos tipos de doenças, assim como na previsão da reação de organismos sob diversas condições de interesse.

O processamento de imagens de microarranjos de DNA tem de lidar com vários desafios que envolvem desde a correta identificação dos sinais de expressão gênica à necessidade de detectar e eliminar as expressões inválidas. Devido à existência de diferentes modos de se preparar os microarranjos, é comum o desenvolvimento de ferramentas orientadas a um determinado padrão de imagens, o que dificulta a comparação de resultados. Assim, é necessário o conhecimento das diversas técnicas empregadas atualmente, para poder decidir qual a melhor forma de se conduzir os experimentos na busca por conclusões biológicas mais precisas. Esse trabalho traz um resumo das principais técnicas de processamento e análise de imagens de microarranjos de DNA com o intuito principal de fornecer o arcabouço necessário à comparação e sugestão de melhorias. Após a abordagem dos tópicos teóricos são realizados experimentos que reproduzem os principais passos necessários ao tratamento das imagens.

Os capítulos a seguir estão organizados de modo a apresentar as etapas consecutivas no processamento das imagens. Antes, no segundo capítulo, procura-se entender os principais conceitos biológicos por trás da tecnologia de microarranjos. No terceiro capítulo é explorada a importância dos diversos procedimentos e as possíveis influências externas durante a aquisição das imagens. O quarto capítulo envolve a busca e definição das áreas de interesse através de diversos métodos. Fechando o fluxo de eventos, segue o quinto capítulo introduzindo o problema da quantificação dos dados sob variadas exigências. O sétimo capítulo diz respeito a realização de experimentos utilizando algumas das técnicas abordadas modificadas a fim de aumentar o desempenho. Finalizando, são expostas algumas conclusões retiradas a partir do que foi pesquisado.

2. A tecnologia de Microarranjos

Nesse capítulo são introduzidos alguns conceitos fundamentais da Biologia Molecular e como eles são aplicados na construção dos microarranjos de DNA.

2.1 Fundamentos Biológicos

A biologia molecular retrata o estudo das células e moléculas. Uma célula é definida como a unidade fundamental dos seres vivos, ou a menor unidade capaz de manifestar as propriedades de um ser vivo. Estruturalmente, uma célula é classificada como procariótica caso seja de um organismo unicelular ou eucariótica quando faz parte de um organismo pluricelular. Elas estão envolvidas numa membrana chamada citoplasma. O citoplasma compreende todo o volume da célula, com exceção do núcleo. O núcleo das células eucarióticas controla todas as suas atividades (sintetizar seus componentes, crescer, se multiplicar, etc.) [1].

2.1.1 Ácidos Nucléicos

Todo organismo vivo armazena informação biológica na forma de moléculas de ácidos nucléicos, formadas por nucleotídeos. Cada nucleotídeo por sua vez, consiste de: uma molécula de açúcar (desoxirribose ou ribose), um grupo fosfato e, uma base nitrogenada. Os ácidos nucléicos são, portanto, classificados como Desoxirribonucléico (DNA) ou Ribonucléico (RNA) [2]. No DNA são encontrados quatro tipos de bases nitrogenadas: adenina (A), citosina (C), guanina (G) e timina (T). É representado por uma fita dupla (emparelhamentos de nucleotídeos) complementar e antiparalela. No esquema de emparelhamento, o nucleotídeo A sempre se liga ao nucleotídeo T e o nucleotídeo C sempre se liga ao G através de pontes de hidrogênio [2].

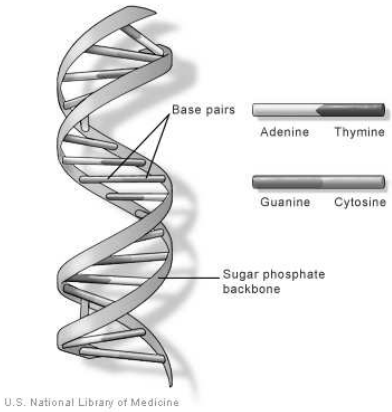

Figura 1: Estrutura do DNA [34]

O RNA também engloba cinco tipos de bases trocando-se timina pela uracila (U). Apresenta-se como uma cadeia única de tamanho menor que o DNA e com grande diversidade de estruturas secundárias relacionadas às suas funções desempenhadas na célula. São reconhecidos três tipos:

1. mRNA (mensageiro), que contém a informação para a codificação de proteínas;
2. tRNA (transportador), que é responsável pelo transporte de aminoácidos;
3. rRNA (ribossomal), que possui papel estrutural.

Quando duas seqüências complementares de ácido nucléicos se combinam, esse processo bioquímico é chamado de hibridização [2].

2.1.2 Expressão Gênica

A seqüência de milhões de bases emparelhadas que formam o DNA é subdividida em fragmentos de diversos tamanhos denominados genes. Neles estão contidas as informações que especificam a estrutura das proteínas, macromoléculas que de fato realizam as principais ações nos organismos. Por expressão gênica, entende-se o processo em que um gene é utilizado na construção de uma proteína ou para controlar a expressão de outros genes. A síntese das proteínas pode ser resumida em dois passos:

1. O DNA é transcrito ao ribonucléico mensageiro (mRNA) em um processo chamado transcrição;
2. O mRNA é transformado em proteína ou em parte dela (aminoácido), em um processo chamado tradução.

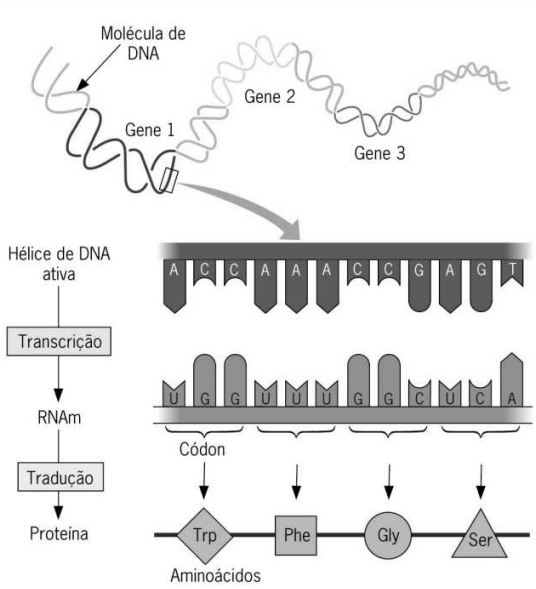

Figura 2: Síntese de proteínas [35]

A transcrição é a reprodução de uma fita de DNA em uma seqüência de RNA complementar. O modelo onde o DNA serve como molde para a síntese de moléculas de RNA, que por sua vez, coordenam a síntese de proteínas, é conhecido como o Dogma Central da Biologia Molecular. Outro mecanismo importante é a replicação dos genes que permite a hereditariedade das informações carregadas no DNA. O conjunto de toda a informação codificada no DNA de um organismo é chamado de genoma.

Numa célula característica, cerca de 10.000 a 20.000 genes são expressos simultaneamente. O nível de expressão gênica é um número que mede a quantidade de mRNA associado a um gene particular e está relacionado com a quantidade de proteína que ele produz [1]. As técnicas de análise de expressão gênica consistem principalmente em interceptar a etapa de transcrição. Desta forma, é possível estudar os "comandos" ativos nos organismos em diferentes estados biológicos. Uma das principais técnicas modernas utilizadas para a análise em larga-escala é o microarranjo [3].

2.2 Microarranjos de DNA

Os microarranjos de DNA (tradução do termo em inglês *microarray*), ou chips de DNA, consistem num conjunto ordenado de milhares de moléculas de DNA cujas seqüências são conhecidas. São utilizados em experimentos de análise de expressão gênica em larga escala. Essa técnica foi impulsionada pela necessidade de se analisar a grande quantidade de informação gerada pelo seqüenciamento de genomas. Os *spots* (sondas) representam as amostras microscópicas depositadas na superfície para atuar como detectores dos genes expressos. O material detector pode ser composto de:

1. Oligonucleóticos - pequenas moléculas de DNA com pouca quantidade de nucleotídeos (bases), constituídas por segmentos não repetidos e que hibridizam apenas com um dos mRNA a ser utilizado;
2. cDNA - molécula sintética de DNA mais estável obtida a partir da transcrição reversa do mRNA que se deseja observar.

2.3 Metodologia de Preparação

A tecnologia de microarranjos é um processo baseado em hibridização que possibilita observar a concentração de mRNA de uma amostra de células analisando a luminosidade de sinais fluorescentes [1]. A preparação de um experimento envolve a escolha dos genes que se deseja estudar e a síntese dos seus detectores correspondentes. Em resumo, o processo pode ser dividido nas seguintes fases:

1. Células cujo DNA possui os genes de interesse são cultivadas em duas soluções distintas: uma correspondendo à situação biológica normal (padrão), outra correspondendo à situação a ser estudada;

2. Recolhe-se o mRNA das duas soluções, “marcando” os mRNA de cada solução com uma substância fluorescente. Normalmente são utilizados corantes cy3 (verde) e cy5 (vermelho);
3. Os mRNA marcados são misturados e aplicados nos microarranjos de DNA;
4. Ocorre a hibridização dos microarranjos com a mistura de mRNA.

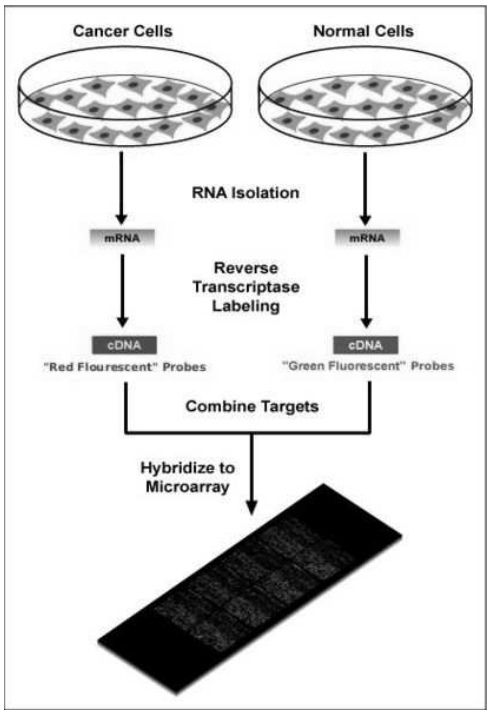

Figura 3: Construção dos Microarranjos [36]

A base de um experimento envolvendo microarranjos de DNA é o processo de hibridização. Somente os genes que possuírem seqüências complementares aos mRNA marcados deverão apresentar algum nível de expressão.

A partir da utilização de luz adequada é possível excitar o material fluorescente contido nas amostras de mRNA que hibridizaram com os detectores genéticos dos microarranjos. A intensidade de fluorescência é proporcional ao mRNA original, porém a constante de proporcionalidade é desconhecida, inviabilizando a quantificação absoluta. Nesse modelo, assume-se que a razão de expressão gênica entre as duas condições biológicas de interesse é estimada pela razão dos níveis de hibridização entre seus respectivos mRNA. Assim, pode-se procurar por diferenças importantes entre as condições, ou como é conhecido na área de análise de microarranjos, descobrir a expressão gênica diferencial entre as duas amostras [3]. Essa questão volta a ser abordada no Capítulo 5, quando os dados das intensidades são quantificados.

As imagens são geradas por dispositivos de varredura (escaner) especiais que utilizam lasers microscópicos. Cada imagem é uma representação do microarranjo varrido com

várias matrizes em duas dimensões. Como resultado, as imagens digitalizadas apresentam a reação de fluorescência de todos os *spots* contidos na lâmina varrida pelo laser. A intensidade da reação representa o nível de expressão diferencial de cada gene e está relacionada com a abundância do respectivo mRNA na solução. Os *spots* da que contêm amostras marcadas com o fluoróforo cy3 devem aparecer na imagem como círculos verdes intensos, aqueles com amostras marcadas com o fluoróforo cy5 aparecem como círculos vermelhos, no caso de quantidades iguais dos dois corantes os círculos devem aparecer amarelos.

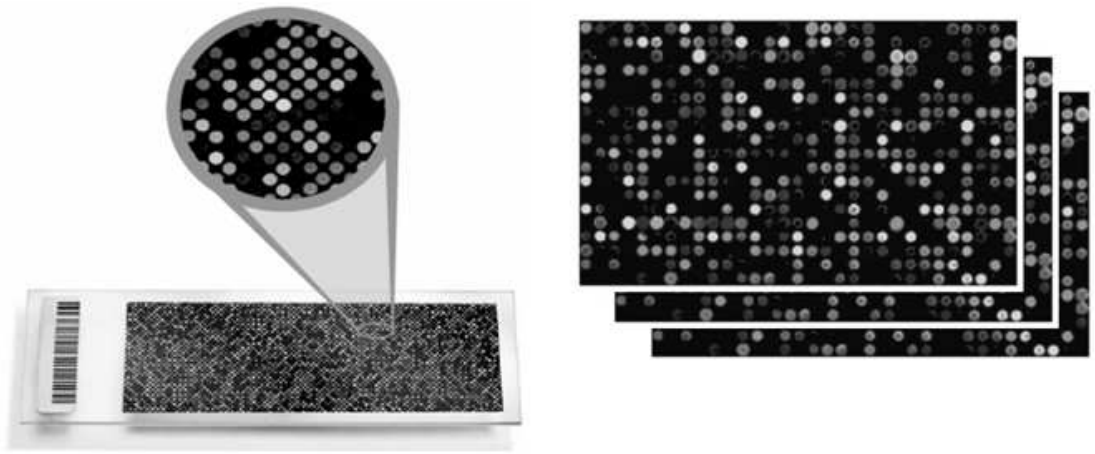

Figura 4: Imagem digitalizada dos microarranjos de DNA

O método básico para extração de dados de uma imagem de microarranjo envolve a identificação e medição da intensidade de fluorescência de cada elemento individual sobre a seqüência da matriz através de sistemas computacionais. O software de aquisição de dados precisa identificar o formato dos microarranjos, incluindo o esquema de distribuição, tamanho, forma e intensidade do *spots*, distância entre *spots*, resolução da imagem, além da fluorescência do plano de fundo (mais conhecido como *background*).

2.3 Fluxo do Processamento de Dados

Desde a invenção da tecnologia de microarranjos em 1995 [4], foram desenvolvidos vários métodos de processamento de imagens, modelos estatísticos e técnicas de mineração de dados específicos para análise de microarranjos de DNA [5]. Essa análise faz parte de um fluxo de dados comum durante o processamento de imagens de microarranjos que inclui:

1. Alinhamento da grade;
2. Segmentação dos *spots*;
3. Avaliação da qualidade;
4. Quantificação e normalização dos dados;
5. Identificação dos genes diferencialmente expressos;

6. Mineração dos dados (interpretação dos resultados).

O processo é ilustrado na Figura 5 logo abaixo. O subconjunto de passos realtivos ao processamento de imagem está dedelimitado por uma linha tracejada. O principal passo da análise de microarranjos é a extração dos descritores de intensidade de cada *spot* que representam os níveis de expressão gênica. Tais valores são posteriormente utilizados para análises mais aprofundadas. Deste modo são obtidas conclusões biológicas baseadas nos resultados da mineração de dados e análise estatística de todas as características extraídas.

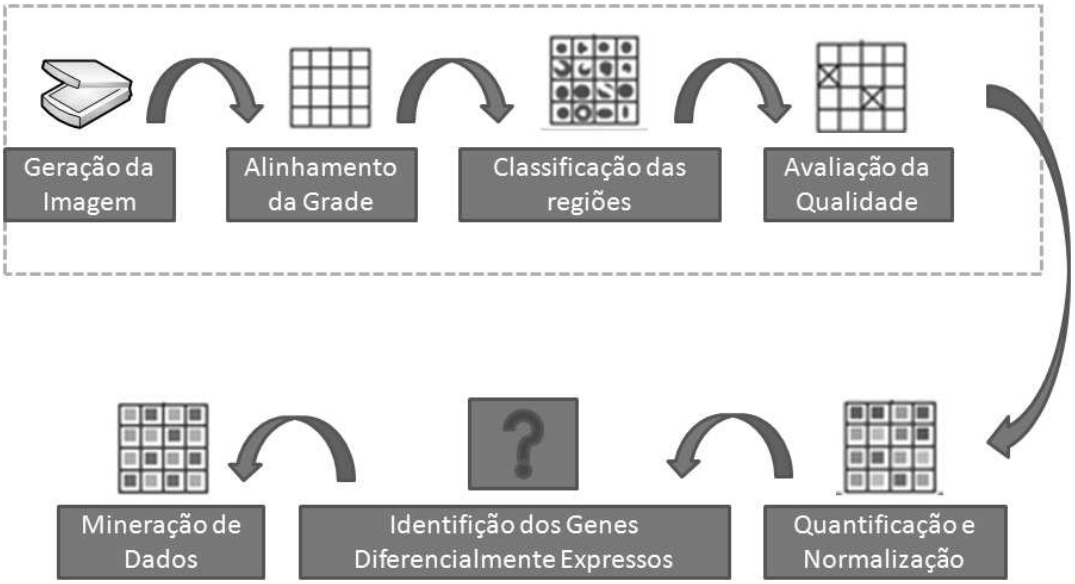

Figura 5: Fluxo do processamento de dados de microarranjos.

3. Geração das Imagens

Esse capítulo reúne informações sobre como são geradas as imagens dos microarranjos. Mesmo seguindo todos os procedimentos, existem diversos fatores incontroláveis que contribuem para a inserção de ruídos nos experimentos. Por isso também são abordadas as causas e formas comuns dos problemas de variação nas imagens.

3.1 Layout do microarranjo

O layout de uma imagem de microarranjo refere-se à forma como são dispostos os *spots* onde ocorre cada expressão gênica. Em quase todos os esquemas encontrados, os *spots* são organizados em grupos maiores, limitados por um *grid* bidimensional (2D) que define a localização de cada *spot* do grupo pela linha e coluna ou pelas coordenadas absolutas (x, y). O termo comumente aceito para denominar cada conjunto de *spots* dentro em uma imagem de microarranjos seria *grid*. Abaixo é demonstrado um exemplo de organização dos microarranjos em *grid*s.

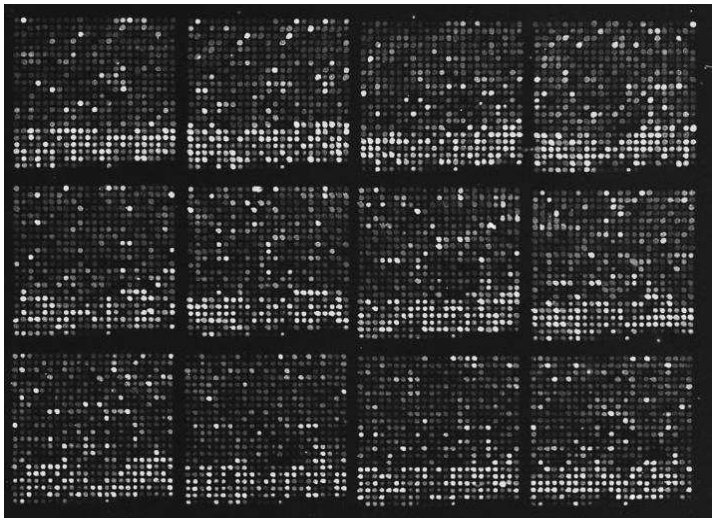

Figura 6: Imagem de microarranjo subdividido em 3x4 *grid*s

O planejamento da organização dos *grid*s influencia diretamente no modo como será tratada a imagem após a digitalização. Basicamente a definição do layout de qualquer imagem de microarranjo é dependente de fatores como o tipo de equipamento a ser utilizado para sintetizar o microarranjo, e de considerações para a posterior análise da imagem. Muitas tecnologias comerciais mantêm um layout fixo com mecanismos de análise das imagens otimizados para um esquema particular. Desse modo, os sistemas de processamento de imagens de microarranjos funcionam com bastante eficiência quando tratam de chips de DNA produzidos pelo mesmo centro de pesquisas onde foram desenvolvidos.

3.2 Impressão do microarranjo

Embora existam diferentes abordagens na impressão dos microarranjos, nesse trabalho é considerado o procedimento mais conhecido por *Stanford/Pat Brow* [33]. Essa abordagem além de ser mais barata quando comparada a outras tecnologias, é a mais utilizada nos centros acadêmicos para a produção personalizada dos microarranjos cujos exemplos serão considerados mais adiante. Ha também uma maior flexibilidade para o estudo das seqüências de DNA, a partir da síntese de longas seqüência de cDNA. A construção dos microarranjos segue exatamente o padrão descrito no Capítulo 2, sendo que os *spots* são impressos nos chips mecanicamente através de pinos microscópicos que depositam o material genético controlados por um robô.

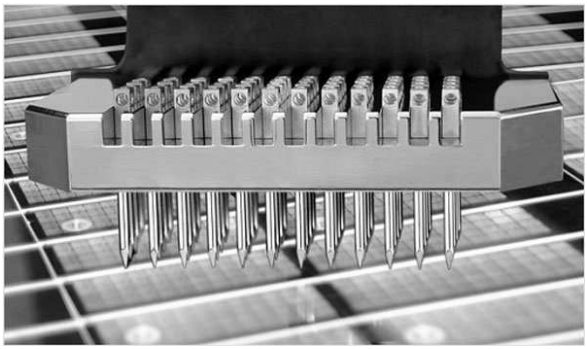

Figura 7: Impressão de microarranjos

Cada unidade de impressão, ou pino, cria um bloco individual de *spot*. O número de pinos pode ser modificado de maneira a produzir um novo layout para a imagem do microarranjo. A distância entre os *spots* e número de linhas e colunas dentro de cada *grid* é controlada pelo sistema de impressão. Um arquivo de análise é criado contendo informações relativas à quantidade de blocos, linhas, colunas, distância entre blocos, diâmetro aproximado do *spot*, além da anotação dos genes (ou fragmento de genes representados por cada bloco).

Uma abordagem alternativa é utilizada pela empresa Affymetrix [6] na produção de chips comerciais. Os *spots* são definidos coma tecnologia *Nimblegem* que utiliza repetitivos processos de fotolitografia. A abordagem é mais custosa e são utilizadas seqüências menores de códigos genéticos (oligonucleotídeos) que podem ser sintetizadas com maior facilidade que o cDNA [5].

3.3 Formato do arquivo

Tipicamente, a varredura a laser de um cDNA gera dois arquivos TIFF (*Tagded Image Format File*). Estes dois arquivos contêm informações das florescências detectadas durante a excitação pelo laser. No caso dos experimentos com microarranjos, os arquivos são gerados com 16 bites para cada canal de cor RGB. A escolha é baseada na faixa

alcance das intensidades de fluorescência e sensibilidade do escaner. Os valores de fluorescência detectados após a amplificação e conversão de analógico-digital devem ficar dentro do intervalo [0, 65.535] (2^16-1 = 65.535). Caso contrário os valores mais elevados são truncado para o máximo (pixel saturado) [5].

O formato TIFF também inclui opções de compressão de imagens (com ou sem perdas), porém não é recomendado utilizá-las a fim de prevenir a perda de informações dos *spots* e evitar a queda de precisão na extração das características. Outro formato utilizado para representar imagens de microarrajos é o JPG. O JPG ocupa menos espaço que o TIFF, mas além de possuir menor qualidade, utiliza algoritmos de compressão que levam a perda de dados.

3.4 Imagem ideal

Idealmente o conteúdo da imagem gerada deveria ser caracterizado por uma geometria determinística do *grid*, formas pré-definidas dos *spots*, intensidades constantes e proporcionais ao fenômeno biológico (tanto de *background* quanto de *foreground*). A Figura abaixo ilustra um exemplo de uma imagem ideal de microarranjo.

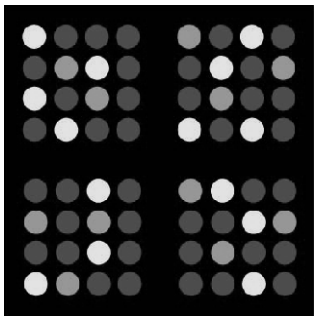

Figura 8: Imagem ideal de um microarranjo

Do ponto de vista estatístico, a aquisição de uma imagem ideal está diretamente relacionada à quantidade de pixels destinada a cada *spot* (resolução da imagem). Logo, o custo dos experimentos, as limitações do escaner e as dificuldades para se armazenar resoluções extremamente altas são os reais impedimentos à geração de imagens mais próximas da ideal.

Embora na prática seja impossível encontrar uma imagem ideal de microarranjos, é um bom ponto inicial para entender as variações da imagem e possivelmente simulá-las [7]. Simulações são úteis na geração de conjunto de dados para testes dos algoritmos de processamento das imagens. Além disso, também podem prover inferências sobre os impactos causados por diferentes fatores na construção dos microarranjos e na precisão das conclusões biológicas finais.

3.5 Fontes de variações na imagem

A tecnologia de microarranjos de DNA é um complexo processo que envolve várias etapas. Portanto, existem vários fatores aleatórios atuando em cada fase numa proporção difícil de ser estimada. A seguir são abordadas algumas fontes de variações de imagem comumente observadas na área dos *spots* (*foreground*), no plano de fundo dos *spots* (*background*) e na informação de intensidade.

3.5.1 Variações nos canais da imagem

Dependendo do tipo de marcação do cDNA durante a preparação do microarranjo (hibridização), pode-se obter: únicas, duplas ou múltiplas fluorescências numa mesma imagem. É mais comum encontrar fontes de dados que representam imagens duplamente fluorescentes produzidas por dispositivos que operam em dois comprimentos de onda. Na Figura abaixo são observadas as fluorescências detectadas com o escaner operando com comprimentos de onda 532 nm (vermelho) e 632 nm (verde) [5].

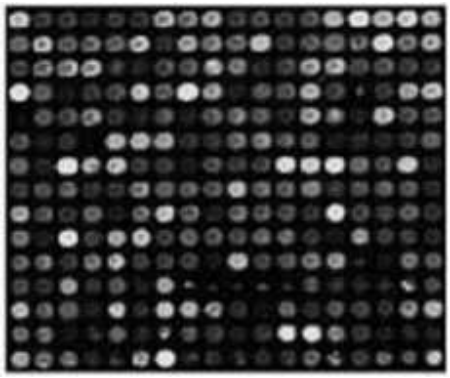

Figura 9: Imagem com *spots* verdes e vermelhos.

Em geral, os dados da imagem de microarranjo podem consistir em um número qualquer de canais. É possível prever a utilização de mais de dois ou três canais, no futuro, através de imagens hiperespectrais [5].

Outra variação consiste na forma como o arquivo é armazenado, se houve compressão dos dados e qual foi a precisão utilizada (número de bytes por pixel). Por exemplo, um arquivo armazenado num formato com perda de dados introduz borramento espacial dos *spots* e a análise da imagem torna-se menos precisa. Similarmente, o número de bytes por pixel precisa acomodar a faixa do sinal analógico produzido pela excitação dos corantes fluorescentes. Essa faixa corresponde a máxima menos a mínima medida de amplitude, e qualquer valor fora do intervalo [*min*, *max*] é mapeado para um dos extremos. Para um número fixo de bytes, o aumento da faixa conseqüentemente diminui a precisão de cada medida de intensidade.

Os algoritmos de processamento de imagens devem ser capazes de tratar qualquer número de canais de entrada, formato de arquivos e precisão dos dados. De qualquer maneira a análise dos resultados da imagem irá conter alguma incerteza devido ao processo de armazenagem do arquivo e as restrições da imprecisão dos dados.

3.5.2 Variações da geometria dos grids

A própria preparação das lâminas de microarranjos é considerada uma fonte de variação na geometria do *grid*. Por exemplo, uma máquina que deposita amostras genéticas através de pinos de imersão imprime múltiplas matrizes de *spots*. Ao longo do tempo a imersão dos pinos pode se alterar, causando irregularidades na disposição dos *spots* impressos [8]. Similarmente, qualquer deslocamento rotacional de uma lâmina ou dos pinos causará uma rotação do *grid* na imagem do microarranjo em relação à borda da imagem. A Figura 10 demonstra um exemplo de um *grid* rotacionado com linhas e colunas espaçadas irregularmente.

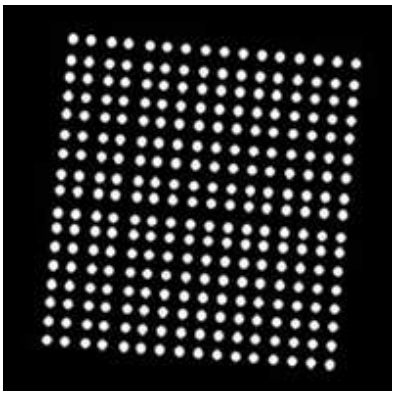

Figura 10: Espaçamento e geometria irregular do *grid* de *spots* do microarranjo

A localização dos *spots* também é afetada pelo material da lâmina (vidro, nylon) e os tipos de sondas utilizados (marcação com elementos radioativos ou químicos fluorescentes) [9]. As variações têm diferentes causas:

- Esforços mecânicos (nylon);
- Baixa potência de discriminação (vidro);
- Forte sinal de *background* (marcação fluorescente);
- Interferência do sinal pelos sinais dos *spots* vizinhos (marcação radioativa).

A baixa potência de discriminação merece atenção especial porque devido a ela, muitos *spots* deixam de ser detectados [8]. A Figura 11 ilustra que muitos *spots* estão faltando na matriz porque seus sinais não foram distinguidos do *background*. A ausência de *spots* introduz um desafio para o alinhamento automático do *grid*. Por exemplo, um método de alinhamento totalmente automático de *grid*s deverá falhar em detectar corretamente um *grid* se uma linha de *spots* longo da borda estiver faltando completamente (nenhuma evidência de existência da linha).

Figura 11: Lâmina de microarranjo (o *grid* embaixo à direita possui uma linha a menos que os outros).

3.5.3 Variações de background

As diferenças de *background* ocorrem devido a:

a. Preparação da lâmina do microarranjo;
b. Procedimentos inapropriados de aquisição (presença de poeira ou sujeira);
c. Instrumentos de aquisição.

Os tipos (a) e (b) de variações de *background* devem ser detectadas pela avaliação da qualidade do microarranjo. A variação devido aos instrumentos de aquisição pode ser removida por um usuário.

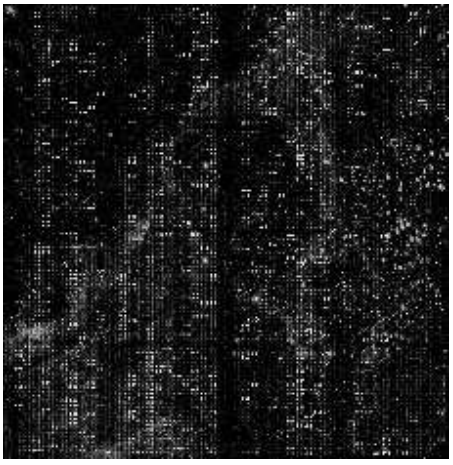

Figura 12: Variação de *background* (contaminação do fundo com um dos corantes).

Muitos algoritmos de processamento de imagens compensam as variações de *background* através da modelagem via funções de distribuição de probabilidade (FDP). O modelo mais utilizado é a FDP Gaussiana (também chamada de Normal) [7]. Outros modelos estatísticos considerados são a distribuição Uniforme e a distribuição Funcional, dependendo das propriedades observadas nas imagens adquiridas. A Figura 13 mostra exemplos de *background* modelados por distribuições de probabilidade.

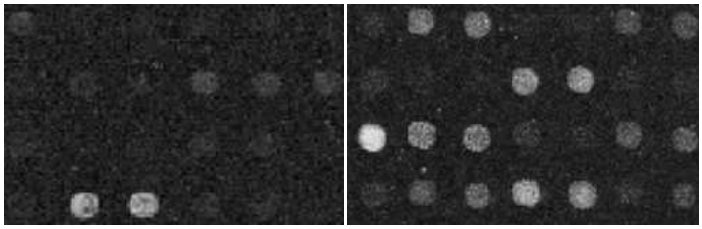

Figura 13: Ruídos de *background* modelados (FDP Normal à esquerda e FDP t-Student à direita).

Vale ressaltar que, embora todos os canais das imagens dos microarranjos possam seguir a mesma FDP, cada canal precisa de seus próprios parâmetros para o modelo de distribuição escolhido.

3.5.4 Variações da formas dos *spots*

Também é preciso considerar a forma dos elementos dos *grid*s no microarranjo (ou formas primitivas do *grid*). Apesar de a maioria dos atuais microarranjos de cDNA serem produzidos com *spots* circulares, pode-se encontrar o uso de outras formas, como linhas ou retângulos. Para os *spots* circulares, existe um grande número de desvios a serem modelados. A Figura 14 mostra algumas classes de desvios morfológico encontradas em imagens de microarranjos.

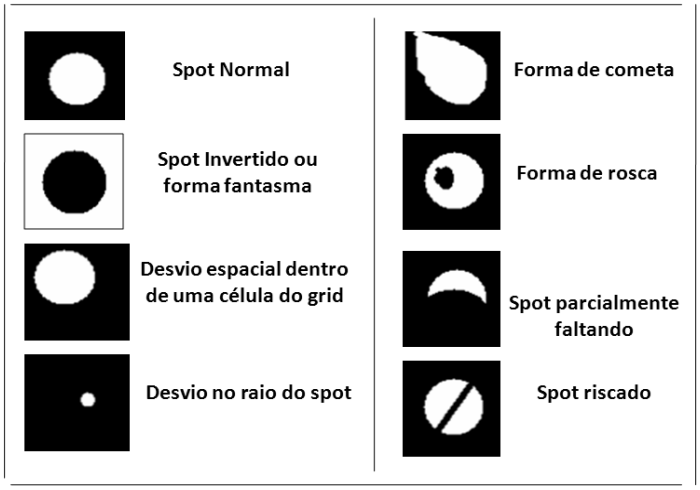

Figura 14: Variações espaciais e morfológicas dos *spots*.

Existem muito mais variações de forma que precisam ser analisados durante a avaliação da qualidade dos *spots* a fim de determinar a validade da medida de intensidade. A análise dos desvios dos *spots* ajuda a identificar o sucesso ou a falha da um experimento particular.

3.5.5 Variações das intensidades de foreground e background

Variações nas intensidades do *foreground* e do *background* estão também presentes na análise das imagens dos microarranjos devido ao material da lâmina e as variadas técnicas de marcação. Enquanto o tipo de marcação de fluorescência conduz a imagens com fundos escuros e *spots* brilhantes (*foreground* claro contrastando com o *background* escuro), outro tipo de marcação com radio-isótopo leva a imagens com fundo claro e *spots* escuros.

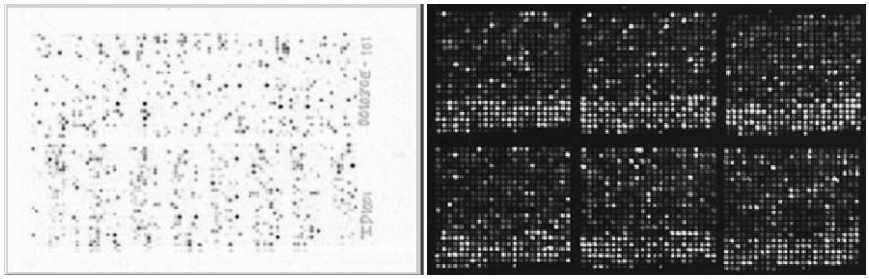

Figura 15: Microarranjos com diferentes marcadores (radioativo à esquerda e fluorescente à direita).

A diferença de intensidade do *background* e do *foreground* é bastante relevante para o significado biológico como será demonstrado posteriormente. Por isso é de vital importância a separação precisa dessas duas classes. Entretanto, a faixa de diferença de intensidade (*max* - *min*) e a amplitude das variações afetam a discriminação das classes, influenciando diretamente na definição das áreas de interesse.

4. Processamento das Imagens

Nesse capítulo são encontrados os principais métodos de processamento das imagens de microarranjos. Basicamente, essa etapa no tratamento de microarranjos de DNA consiste no alinhamento dos *grid*s seguida e na definição das regiões de *background* e *foreground*.

Diante das variações das imagens dos microarranjos, é desejável desenvolver algoritmos de processamento automático de imagens que sejam robustos à todas ou à maioria delas. A robustez deve incluir:

1. Qualquer número de canais;
2. Qualquer armazenamento e representação computacional;
3. Localizações variáveis dos *grid*s e *spots*;
4. Ruído de *background* desconhecido;
5. Esquemas variáveis de *background* e *foreground*;
6. Desvios nas formas dos *spots*;
7. Desvios dos perfis de intensidades esperados dos *spots*.

Além disso, os algoritmos devem reconhecer aqueles casos em que faltam *spots* para desabilitar a automação devido à falta de evidências da forma real do *grid*.

Para qualquer pesquisador que realiza experimentos com a tecnologia microarranjo, é importante para garantir o determinismo do processamento das imagens. Assumindo que um algoritmo é executado com os mesmos dados, é esperada a obtenção dos mesmos resultados após a execução. A fim de conseguir essa meta, os algoritmos devem ser o máximo possível livres de parâmetros, para que seus resultados possam facilmente ser repetidos sem tanta dependência do usuário. Tomando como exemplo o posicionamento manual de um padrão de *grid*. Além de ser tedioso e de consumir bastante tempo, também é indesejável, já que o passo de alinhamento do *grid* não pode ser repetido facilmente. Um exemplo concreto da questão de repetição é apresentado em [10].

4.1 Alinhamento de grids

O alinhamento de *grid* (também conhecido como endereçamento ou procura de *spot* consiste no passo de processamento de imagens de microarranjos que registra um conjunto de linhas e colunas paralelas desigualmente espaçadas. Os padrões encontrados representam partes do conteúdo da imagem como uma matriz bi-dimensional de *spots* [11]. Como dito anteriormente, essas matrizes são conhecidas como *grid*s (grades). O objetivo é encontrar todos os padrões na imagem que definem um conjunto de *spots*, ou seja, as coordenadas das linhas e suas orientações, de modo que os pares de linhas perpendiculares definam localizações aproximadas dos *spots* (células do *grid*).

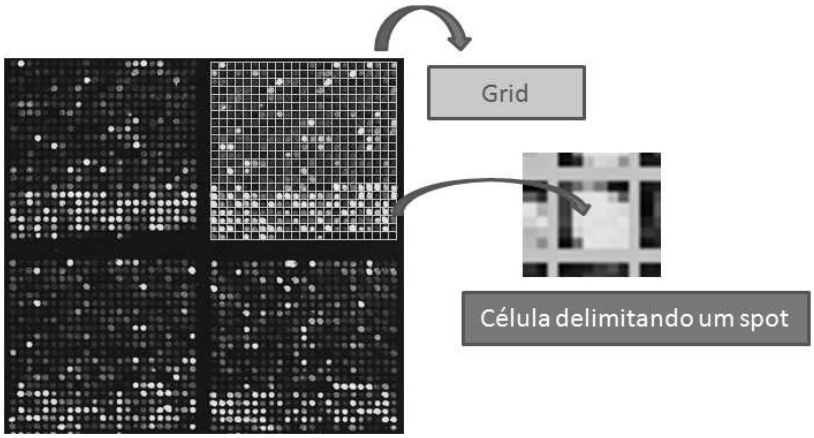

Figura 16.: Imagem de um *grid*.

4.1.1 Alinhamento do ponto de vista da automação

Alinhamento manual

Considerando o fato de que a geometria do *spot* seja muito similar a de um *grid* (ou de um conjunto de *subgrid*s), esse método de alinhamento é baseado em um modelo de *grid* definido manualmente. Um usuário especifica as dimensões do *grid* e um raio para o *spot* criando assim o modelo. De posse dessas informações um *grid* é construído na imagem sem a preocupação inicial com a localização correta dos *spots*. Em seguida, através dos mecanismos de iteração com o usuário oferecidos pelo sistema (ex: mouse), a posição do *grid* pode ser ajustada de modo a corresponder ao layout dos *spots* no microarranjo.

É possível obter um alinhamento de *grid* perfeito assumindo que o software de interação com o usuário dê suporte ao ajuste individual da forma e da localização de cada *spot*. No entanto, é evidente que essa abordagem, além de consumir muito tempo, exige paciência e dificilmente pode ser repetida de modo a obter os mesmos resultados. Logo, não é um método interessante quando se visa um alto rendimento na análise de imagens de microarranjos.

Alinhamento semi-automático

Em geral, há algumas etapas do alinhamento de *grid* que podem ser executadas pelos computadores de forma confiável, enquanto outras que são dependentes das entradas do usuário. Assim, pode-se definir um método de alinhamento que utiliza alguns dados de entrada definidos pelo usuário para construir um modelo do *grid*. Além da redução do esforço do usuário, o alinhamento semi-automático do *grid* também contribui para a geração de resultados determinísticos. Todavia, o método pode não ser adequado quando levadas em conta as exigências de alto rendimento da análise de imagens de microarranjos.

Como exemplo dessa abordagem basta considerar uma inicialização manual (seleção dos *spots* das bordas, especificação das dimensões do *grid*) seguida de uma procura automática das linhas do *grid* [12]. O componente automatizado pode se basear nas propriedades dos dados observadas durante a varredura da imagem ou utilizar técnicas de correlação de imagens com um modelo de *grid* previamente definido.

Alinhamento totalmente automático

Esse método consiste em identificar todos os *spots* sem qualquer intervenção humana, baseado em uma ação única. Essa ação única serve para opcionalmente incorporar qualquer conhecimento prévio sobre um layout da imagem no algoritmo de alinhamento a fim de reduzir seu espaço de busca. Muitas vezes, o desafio no desenvolvimento de métodos totalmente automáticos consiste em identificar e calcular todos os parâmetros que representam o conhecimento a priori além de quantificar restrições para todos eles. Tipicamente, esses métodos são orientados aos valores de intensidade dos pixels e precisam otimizar internamente vários parâmetros do algoritmo durante a busca espacial para compensar as variações de imagem descritas anteriormente.

Essa abordagem depende inteiramente do conteúdo dos dados. Por exemplo, no caso de uma linha de *spots* faltando na imagem (a cor do *spot* não é distinguível do *background*), um algoritmo automático na deve ser capaz de encontrar qualquer evidência da linha do *grid*. *Grid*s de baixa confiabilidade devem ser definidos à parte através da inspeção humana. Outra alternativa é definir na imagem algum *spot* de confiança que possa servir de guia (modelo padrão) durante o processamento e assim prover um mecanismo de alto-correção.

4.1.2 Alinhamento do ponto de vista de análise da imagem

Abordagem baseada em modelos

Nesse método, são definidos modelos que se aproximam do layout dos *grid*s. O mais comum envolve a utilização de *spots* circulares com tamanhos e quantidades pré-definidas. A partir do *grid* gerado busca-se o casamento de padrão com uma subimagem do microarranjo. A abordagem baseada em modelo é a mais comum na literatura, estando presente em ferramentas conhecidas: GenePix Pro [13], ScanAnalyse [14], entre outras.

A maioria dos softwares disponíveis permite o ajuste da correspondência de padrões manualmente (tamanho do *spot*, espaçamento entre *spots*, localização dos *grid*s). Outros já incorporam um refinamento automático da localização do *grid* a partir do tamanho e do espaçamento dos *spots* [13]. O refinamento é executado pela maximização da correlação com uma imagem padrão formada a partir das entradas do usuário ou com uma imagem do microarranjo processada sobre um conjunto de possíveis modelos de localizações (ex: translação e rotação da posição inicial definida pelo usuário). Também é possível

empregar modelos deformáveis [15] para alcançar certas flexibilidades no alinhamento do *grid*.

Essa abordagem é considerada apropriada quando a geometria do *grid* observado não se desvia muito do modelo definido como padrão. Se as medidas dos *spots* são imprevisivelmente irregulares então os resultados tornam-se imprecisos. Um exemplo de alinhamento impreciso é demonstrado a seguir.

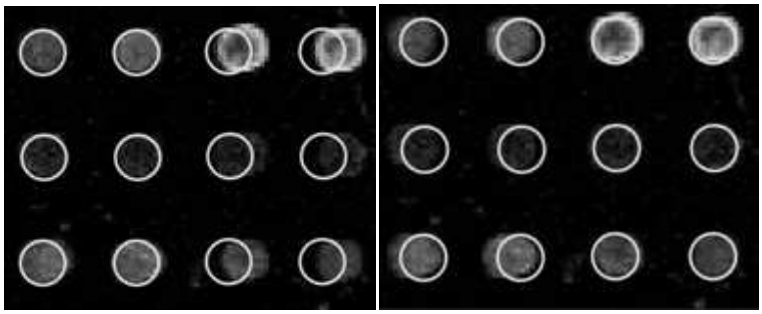

Figura 17: Alinhamento baseado em modelo de *grid*s com a ferramenta ScanAlyse.

Na imagem à esquerda, o modelo de *grid* apresenta *spots* mal definidos nas últimas colunas. Na tentativa de se melhorar o casamento de padrões, translada-se o modelo para a direita, porém o espaçamento irregular dos *spots* leva a desvios nas primeiras colunas.

Abordagem orientada a dados

Essa abordagem é equivalente ao alinhamento automático de *grid*s. Os algoritmos que seguem esse método podem ser divididos em vários componentes, cada um responsável por resolver uma parte do quebra-cabeça do alinhamento.

Definição das linhas

A partir da análise das projeções unidimensionais da imagem é possível descobrir as prováveis localizações das linhas que definirão os *grid*s. Para isso são executados seguintes passos:

- Um somatório de todas as intensidades em uma direção é computado e denotado como um vetor de projeção (das linhas se a direção for vertical ou das colunas se a direção for horizontal);
- Os extremos locais são detectados entres os vetores das projeções. Eles representam uma aproximação dos centros dos *spots*;
- Um conjunto de linhas é determinado a partir dos extremos locais levando-se em conta alguns parâmetros de entrada (por exemplo, número de linhas) ou pela procura de inconsistências no espaçamento entre os extremos locais;
- Todas as intersecções de linhas perpendiculares são calculadas para estimar as localizações dos *spots*.

A outra abordagem algorítmica para encontrar linhas de *grid* é baseada na segmentação da imagem [16] utilizando limiarização adaptativa e operações morfológicas para detectar *spots* guias. *Spots* guias são definidos como sendo aqueles de boa qualidade (forma circular, tamanho apropriado, intensidade constante e maior que o *background*). Com a ajuda dos *spots* guias e da informação sobre o layout do microarranjo, o *grid* final pode ser estimado automaticamente. A limitação dessa abordagem é a suposição da existência dos *spots* guias livres de contaminação.

Processamento de múltiplos canais

O segundo componente aborda o problema da fusão de múltiplos canais da imagem (também chamados de bandas). Como cada canal da imagem é adquirido em tempos diferentes, pode ocorrer um deslocamento espacial entre as aquisições, resultando no registro de canais cruzados. A existência de múltiplas bandas é tratada com a fusão dos canais através de operações lógicas como OR ou AND [11].

A fusão de todos os canais com a operação OR irá propagar as variações de intensidades do *foreground* e do *background* aumentando a robustez do algoritmo de alinhamento. Além disso, a opção de fundir os canais antecipadamente reduz a quantidade de processamento e evita o problema de fundir os múltiplos *grid*s detectados em cada canal.

Definição da rotação do grid

O terceiro está relacionado ao problema da rotação do *grid*. Uma abordagem deste problema é uma busca exaustiva por todos os ângulos de rotação possíveis [11]. Essa abordagem é motivada pelo fato de que a faixa de rotação do *grid* pode ser construída analisando as quatro bordas da matriz 2D. A desvantagem é que um pequeno ângulo de rotação da imagem introduz distorções nos pixels, pois as novas posições não-inteiras são arredondadas para a posição inteira mais próxima.

Definição de múltiplos grids

O quarto componente do problema de múltiplos *grid*s (matrizes 2D de *spots*). *Grid*s distintos no microarranjo também são arrumados em uma matriz, dessa forma o número de *grid*s pode determinado pelas quantidades observadas ao longo dos eixos horizontal e vertical. Essas quantidades podem ser especificadas como parâmetros de entrada e são utilizadas pelo algoritmo para particionar a imagem original em sub-áreas contendo *grid*s individuais.

Se os parâmetros de entrada não estão disponíveis, então o problema pode ser abordado pelo tratamento da imagem inteira, buscando por todas as linhas irregulares e analisando o espaçamento entre todas elas. Toda descontinuidade grande no espaçamento entre as linhas indicará o fim de um *grid* e o início de outro.

4.2 Definição dos spots

O resultado do alinhamento de *grid* é uma aproximação das localizações dos *spots*. Uma localização de *spot* é normalmente definida como uma área retangular da imagem que encapsula uma expressão gênica (também chamada de célula do *grid*). O próximo passo é identificar os pixels que fazem parte do *foreground* (sinal propriamente dito) e do *background*. A extração dos pixels de interesse envolve a segmentação e a clusterização da imagem.

A segmentação da imagem está associada ao problema de particionar uma imagem em regiões espacialmente contíguas com propriedades similares (exemplo: cor ou textura), enquanto que a clusterização se refere ao problema do particionamento de uma imagem em conjuntos de pixels com propriedades similares (exemplo: intensidade, cor ou textura) mas não necessariamente conectados. O objetivo da segmentação dentro de uma célula do *grid* é encontrar uma região que contenha a informação do *foreground* (área de interesse). Se um *spot* é formado por um conjunto de regiões/pixels não contíguos, então a clusterização pode ser aplicada. Embora a segmentação e a clusterização da imagem resultem em agrupamentos de pixels baseados nas similaridades da intensidade, também é freqüente utilizar uma extração baseada em um modelo espacial, onde o modelo segue uma forma padrão de *spot*.

4.2.1 Modelos espaciais

Este tipo de separação assume que um *spot* está centralizado dentro da célula do *grid* e ele aproximadamente corresponde à morfologia esperada para um *spot*. Um modelo espacial consiste tipicamente de dois círculos concêntricos, onde os pixels dentro do círculo menor são marcados como *foreground* (sinal da expressão gênica) e os pixels fora do círculo maior são marcados com *background*.

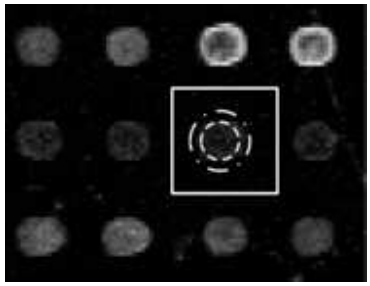

Figura 18: Ilustração de uma célula de *grid* e a separação utilizando modelo espacial com círculos concêntricos.

Todos os pixels entre os dois círculos concêntricos são considerados de transição e não são utilizados. Claramente, esse tipo de marcação do *foreground* falhará para *spots* com raio variável ou com deslocamento espacial do centro dentro da célula. Além disso, pixels

inválidos (contaminados por poeira ou sujeita) podem ser incluídos na área de *foreground* sem nenhuma verificação.

4.2.2 Agrupamento baseado na intensidade

Nessa abordagem a definição da área de interesse se resume ao problema do agrupamento de duas classes [17]. Esse é um problema comum na área de processamento de imagens existindo, portanto, várias formas de tratá-lo.

Um modo simples envolve a técnica de limiarização. A limiarização da imagem é executada pela escolha de um valor de intensidade limite (limiar) e marcação de todos os pixels cujas intensidades estão abaixo ou acima desse valor, dependendo do esquema de cor para *foreground* e *background* adotado no microarranjo. O valor do limiar pode ser escolhido levando-se em conta a porcentagem esperada de pixels do *spot* dentro de uma célula. Mas é necessário o conhecimento prévio da resolução da imagem e do raio do *spot*. A abordagem de limiarização neste caso é vista como um agrupamento por determinar a fronteira de separação entre os grupos.

Uma abordagem mais completa através de algoritmos de clusterização propriamente ditos utiliza intensidades representantes de cada grupo como o k-Means ou k-Medoids. Os pixels da imagem são agrupados de acordo com a similaridade entre suas intensidades e os representantes dos grupos. A cada definição dos grupos, os representantes são atualizados levando-se em conta os novos valores agrupados. Até que os representantes dos grupos não se modifiquem ou tenham mudanças mínimas, o processo de reagrupamento prossegue. Normalmente a distância euclidiana é utilizada como função para cálculo da similaridade e a média ou mediana são utilizadas para o cálculo dos representantes. Detalhes dos algoritmos k-Means são encontrados em [18] e do k-Medoids em [19].

4.2.3 Segmentação baseada na intensidade

Existem muitos métodos de segmentação disponíveis na literatura do processamento de imagens [20]. No caso de imagens de microarranjos são conhecidas a segmentação com crescimento de região e a segmentação em cascata.

A segmentação com crescimento de região começa com um conjunto inicial de posições de pixels (sementes) [20]. O algoritmo agrupa simultaneamente os pixels com intensidades similares às das sementes de modo a formar um conjunto de pixels contíguos (regiões). O agrupamento é executado incrementalmente ao mesmo tempo em que o limiar de similaridade decresce. A segmentação é completada quando todos os pixels são marcados para uma das regiões crescidas a partir das sementes iniciais. Em imagens de microarranjos, a semente do *foreground* pode ser escolhida como sendo a posição do centro de uma célula de *grid* (centro mais provável do *spot*) ou o pixel de máxima intensidade dentro da célula. Similarmente, a semente do *background* pode ser

selecionada como sendo o ponto médio entre dois *spots* ou o pixel de intensidade mínima dentro da célula de *grid*.

A segmentação via transformações em cascata é realizada com operadores de imagem derivados de morfologia matemática [21]. São utilizados dois operadores básicos, dilatação e erosão, e dois operadores compostos, fechamento e abertura. Esses operadores favorecem a filtragem de estruturas claras ou escuras de imagens de acordo com tamanho e forma pré-definidos. No caso das imagens de microarranjos, operadores morfológicos podem filtrar grupos de pixels que se desviam muito da forma e tamanho esperado para um *spot*. O resultado desse tipo de segmentação é a região mais provável que corresponde ao sinal do *spot* (*foreground*).

4.2.4 Métodos híbridos

Vários métodos de separação de *foreground* tentam integrar o conhecimento prévio sobre a morfologia do *spot*, localização do *spot* e distribuição de intensidade esperada. Na abordagem híbrida, as técnicas são combinadas a fim de se refinar a extração dos pixels de interesse.

Segmentação e Agrupamento espacialmente restritos

Por exemplo, a separação de *foreground* usando segmentação leva a uma região conectada que é ajustada para um modelo espacial [16]. Se o melhor círculo de ajuste desvia muito do modelo então o *spot* é marcado como inválido. Outro exemplo seria a separação de *foreground* utilizando agrupamento com minimização restringida pela dispersão do grupo [22].

Ajuste espacial e de intensidade

A partir da análise da distribuição de intensidade dos pixels do *foreground* e do *background* definidos por um modelo espacial, são descartando aqueles pixels classificados como fora do padrão (outliers) da distribuição [23]. O ajuste espacial é alcançado pela marcação inicial do *foreground* e *background* sobre um modelo de *spot*, enquanto que o ajuste de intensidade é conseguido pela remoção dos pixels com intensidade fora do padrão em relação às distribuições do *foreground* e do *background*. Assim, é esperado que se remova os pixels considerados inválidos (alterados por poeira ou sujeira) nas regiões de *foreground* e *background* e que contribuem para o desvio da forma normal do *spot*.

5. Análise das Imagens

Nesse capítulo são definidas formas de se decidir sobe a validade ou não das regiões de *foreground* definidas para cada *spot*. Após o controle de qualidade, as medições enfim podem ser realizadas. Para isso são demonstradas algumas técnicas comuns na literatura pesquisada.

5.1 Avaliação da qualidade dos spots

De posse da forma mais aproximada do *spot* real após a definição dos pixels de *foreground* e *background*, é necessário também identificar e eliminar as células dos *grid*s que contendo *spots* inválidos. Esse passo é importante porque na quantificação dos dados é assumido que todos os *spots* disponíveis são válidos e, portanto, contribuem para uma análise precisa das expressões gênicas.

A fim de detectar *spots* inválidos ou defeituosos, são definidos critérios de validade (métricas) e valores limites de desvio para classificar os *spots* como válidos ou inválidos. Em geral, o critério de avaliação da validade de um *spot* é dividido em dois tipos:

1. Avaliação a partir das intensidades do *foreground* e do *background*, que inclui a determinação de:
    a. Níveis absolutos do *background* e *foreground*;
    b. Variação do *background*;
    c. Saturação do *foreground*;
    d. Razão entre as intensidades do *foreground* e do *background* (ou razão do sinal pelo ruído).
2. Avaliação a partir das propriedades morfológicas do *foreground*, tais como:
    a. Forma do *spot*;
    b. Irregularidades no tamanho do *spot*;
    c. Localização do *spot*.

Além disso, é preciso entender o relacionamento entre os defeitos detectados de *spots* inválidos e as fontes daqueles defeitos no experimento de microarranjo. Dessa forma considera-se esse tipo de análise como controle de qualidade do *spot*. A avaliação da qualidade do *spot* é necessária para a geração de dados confiáveis, sendo considerada a última etapa onde os defeitos dos *spots* podem se detectados. Nas seções seguintes, será focada a avaliação da qualidade baseada na imagem do *spot*.

5.1.1 Critérios para avaliação das intensidades

Variações da intensidade do background

Existem dois tipos de critérios de avaliação das variações do *background*. No primeiro, métricas variabilidades local e global do *background* são modeladas para avaliar o ruído

da região. Elas são indiretamente proporcionais à variação do *background*, ou seja, definidas como uma multiplicação das estimativas do desvio padrão do *background* [24]. Embora as métricas locais possam detectar a presença de ruídos numa célula de *grid*, as métricas globais fornecem indicações sobre a variação em toda a lâmina de microarranjo.

No segundo critério, as métricas são modeladas baseadas na observação de que algumas células de *grid* devem ter médias de ruído de *background* maiores que a lâmina geralmente. Por exemplo, de acordo com as fórmulas abaixo [13], a métrica de qualidade ***q*** deveria se aproximar de um para *spots* válidos e zero *spots* inválidos.

Equação 1: q é a métrica de qualidade, m é a mediana, μ é a média. A notação BKG se refere ao *background*.

$$q_{BKG}^{LOC\&GLOB\,1} = \frac{\mu_{BKG}^{GLOBAL}}{\mu_{BKG}^{LOCAL} + \mu_{BKG}^{GLOBAL}};$$

Equação 2: q é a métrica de qualidade, m é a mediana, μ é a média. A notação BKG se refere ao *background*.

$$q_{BKG}^{LOC\&GLOB\,2} = \frac{m_{BKG}^{GLOBAL}}{m_{BKG}^{LOCAL} + m_{BKG}^{GLOBAL}}$$

Uniformidade na intensidade de foreground e background

Neste caso assume-se que o *foreground* e o *background* possuem uma distribuição de intensidade uniforme. Dessa forma, uma grande variação de intensidade no *foreground* indica um *spot* menos confiável, enquanto que uma grande variação de intensidade no *background* significa que o sinal foi corrompido durante a preparação da lâmina do microarranjo (*spot* com ruído). Para detectar os defeitos de *foreground*, utilizam-se métricas estatísticas como a Equação 3 [23]. A métrica se aproxima de um para *spots* válidos (variância zero) e compensa o fato de que *spots* com maiores magnitudes de intensidade possam ter maiores variações (divisão pela média amostral do *foreground*).

Equação 3: q é a métrica de qualidade estatística, μ é a média e σ é o desvio padrão dos pixels de *foreground* (FRG).

$$q_{FRG}^{STAT} = 1 - \frac{\sigma_{FRG}}{\mu_{FRG}}$$

Outro par de métricas para o *foreground* e o *background* relaciona os valores absolutos das intensidades [13]:

Equação 4: q é a métrica de qualidade utilizando valores de intensidade absolutos, I é a intesidade máxima ou mínima do *foreground* (FRG) e RANGE é uma faixa de intensidade.

$$q_{FRG}^{ABS} = 1 - {(I_{FRG,\max} - I_{FRG,\min})}/{Range};$$

Equação 5: q é a métrica de qualidade utilizando valores de intensidade absolutos, I é a intesidade máxima ou mínima do *background* (BKG), e RANGE é uma faixa de intensidade.

$$q_{BKG}^{ABS} = 1 - {(I_{BKG,\max} - I_{BKG,\min})}/{Range}$$

Saturação da intensidade do foreground

Anteriormente foi visto que a saturação da intensidade ocorre quando as intensidades dos pixels excedem a faixa de detecção do dispositivo de varredura, acarretando na gravação de um valor truncado. Em decorrência da saturação, as estimativas das expressões gênicas são corrompidas [25]. Ainda que não seja possível diferenciar os pixels saturados de genes altamente expressos daqueles saturados por variações externas (contaminações). Porém, uma forma de minimizar o impacto do problema envolve a aplicação de métricas de saturação para todos os tipos de pixels saturados seguida da aplicação de métricas da forma do *spot* para um posterior refinamento dos resultados.

A fim de detectar a saturação, métricas contínuas e categóricas têm sido propostas. Uma métrica contínua utiliza a proporção de pixels saturados no *spot* [13]:

Equação 6: q é a métrica de saturação contínua, count indica a quantidade de pixels total ou saturados do *spot*.

$$q_{SATURATION}^{CONT} = 1 - {count_{saturated}}/{count_{all}}$$

Uma métrica discreta (ou melhor, binária) classifica um *spot* como válido ou inválido é baseada numa quantidade limite de pixels saturados do *spot* [24].

Equação 7: q é a métrica de saturação discreta, count é a porcentagem de pixels saturados na imagem do *spot*, T é a porcentagem que limita a validade do *spot*.

$$q_{SATURATION}^{CATEG} = \begin{cases} 1; & if \quad count_{saturated} < T\% \\ 0; & if \quad count_{saturated} \geq T\% \end{cases}$$

Razão do sinal pelo ruído

Esse valor representa a propriedade mais explorada (SNR) na avaliação de qualidade do *spot* [13]. O critério SNR elimina *spots* com sinal muito fraco (1 < SNR < limiar), nenhum sinal (SNR ≈ 1), ou *spots* fantasmas (SNR < 1). Ela é baseada na informação de intensidade e definida com valores da média e da mediana conforma a fórmula abaixo.

Equação 8: q é a métrica de qualidade baseada em valores médios ou medianos, m é a mediana, μ é a média. A notação FRG se refere ao *foreground*.

$$q_{SNR}^{MEAN} = \mu_{FRG} /(\mu_{FRG} + \mu_{BKG});$$

Equação 9: q é a métrica de qualidade baseada em valores médios ou medianos, m é a mediana, μ é a média. A notação BKG se refere ao *background*.

$$q_{SNR}^{MEDIAN} = m_{FRG} /(m_{FRG} + m_{BKG})$$

5.1.2 Critérios para avaliação de propriedades morfológicas

Forma do *spot*

Quando se leva em consideração a forma do *spot*, várias métricas são propostas. As mais comuns basicamente utilizam as informações espaciais dos *spots* como: área, perímetro e diâmetro.

As métricas de qualidade baseadas na área do *spot* podem ser computadas de acordo com as seguintes fórmulas [23]:

Equação 10: q é a métrica de qualidade baseada no formato da área, A é a área dos pixels marcados como *foreground* e A_0 é a área esperada para o *spot*.

$$q_{SHAPE}^{AREA1} = \frac{|A - A_0|}{A_0};$$

Equação 11: q é a métrica de qualidade baseada no formato da área, A é a área dos pixels marcados como *foreground* e A_0 é a área esperada para o *spot*.

$$q_{SHAPE}^{AREA2} = \exp(-\frac{|A - A_0|}{A_0})$$

Métricas da qualidade baseadas no perímetro do *spot* são computadas de acordo com as fórmulas 10 e 11, com **A** e $\mathbf{A_0}$ substituídos pelo perímetro da área marcada como *foreground* e da circunferência de um *spot* respectivamente. Entretanto, para pequenos perímetros de *spot* estimados a métrica se torna muito imprecisa devido à natureza das imagens digitais. Assim, outra forma de se avaliar a qualidade do *spot* com relação ao perímetro é demonstrada a seguir [13]:

Equação 12: q é a métrica de qualidade baseada no perímetro da forma, A é a área estimada e C é a circunferência esperada para o *spot*.

$$q_{SHAPE}^{PERIM} = 4\pi A / C^2$$

As métricas de qualidade baseada no diâmetro do *spot* avaliam o desvio da forma circular esperada pela estimativa do diâmetro da área ou pela medição dos tamanhos das seções transversais que passam pelo centro do *spot* em múltiplas direções angulares [28]. Se o diâmetro estimado ou o tamanho da seção transversal desvia do valor esperado por mais que uma porcentagem especificada então o *spot* é considerado inválido.

Equação 13: q é a métrica de qualidade baseada no tamanho do diâmetro, L é o tamanho da seção transversal dos pixels marcados como *foreground* e L_0 é o tamanho esperado.

$$q_{SHAPE}^{X-SECTION1} = \frac{|L - L_0|}{L_0},$$

Equação 14: q é a métrica de qualidade baseada no tamanho do diâmetro, L é o tamanho da seção transversal dos pixels marcados como *foreground* e L_0 é o tamanho esperado.

$$q_{SHAPE}^{X-SECTION2} = \exp(-\frac{|L - L_0|}{L_0})$$

Localização do spot

A métrica de localização do *spot* é definida como uma distância euclidiana entre o centróide de todos os pixels marcados como *foreground* e o centro esperado do *spot*. Neste caso supõe-se que o algoritmo de alinhamento do *grid* tenha uma precisão muito boa. Dessa forma é possível considerar o centro de cada célula do *grid* como sendo o centro esperado do *spot*.

5.1.3 Aplicação de critérios de qualidade

A aplicação dos critérios de qualidade é realizada através da combinação de métricas de qualidade. Para isso, cada uma tem de ser normalizada dependendo da faixa de seus valores. Em seguida, uma composição do escore de qualidade é formada pela aplicação de operadores para o conjunto de métricas selecionadas. Normalmente são utilizados os operadores de multiplicação (para métricas contínuas) e de lógica booleana como o AND (para métricas categóricas) conforme as Equações 15 e 16 respectivamente. A escolha desses operadores se deve ao fato de que todos os critérios precisam ser aplicados simultaneamente durante a avaliação da qualidade do *spot*. Porém, um tratamento especial é normalmente dado para a incorporação de métricas de saturação [25].

Equação 15: q é a métrica de qualidade resultante da multiplicação de critérios.

$$q_{COMPOSITE}^{CONT} = \prod_{i=1}^{m} q_i ;$$

Equação 16: q é a métrica de qualidade resultante da aplicação da operação lógica AND de todos os critérios.

$$q_{COMPOSITE}^{CATEG} = \bigcap_{i=1}^{m} q_i$$

Em se tratando de imagens com múltiplos canais, geralmente cada canal tem sua qualidade avaliada separadamente. A decisão final sobre a validade de cada *spot* é definida por um mecanismo de votação, ou seja, se as avaliações de um canal definido como majoritário levam a uma classificação inválida então o *spot* é marcado como inválido. Também é possível criar composições dos escores de qualidade do *spot* pela combinação de métricas para todos os canais.

A maioria das ferramentas comerciais de análise de expressões gênicas utiliza as métricas de avaliação mais comuns. Por exemplo, todas as métricas definidas aqui são encontradas nos sistemas GenePix e QuantArray.

5.2 Quantificação dos dados

Dado um conjunto válido de *spots* com suas respectivas regiões de *foreground* e *background* delimitadas, a próxima etapa na análise de imagens de microarranjos consiste em extrair as informações de cada *spot* e, baseado nelas, tirar conclusões a respeito da regulação dos genes. Esse processo é chamado de quantificação de dados, refere-se à extração de valores descritivos dos pixels de *foreground* e *background* de cada *spot*. Idealmente, as informações obtidas (também chamados de características ou atributos) devem ser diretamente proporcional à quantidade de mRNA na solução que foi depositada num *spot*, representando assim o nível de expressão do gene depositado.

5.2.1 Extração dos atributos do spot

Em geral, os atributos do *spot* se encaixam em duas categorias:

1. Características absolutas e relativas;
2. Características estatísticas e determinísticas.

Conforme visto no Capítulo 3, as intensidades puras do microarranjo não podem ser interpretadas como medidas absolutas devido à variabilidade aleatória e sistemática na preparação dos dados da imagem do microarranjo. Por essa razão, em experimentos com microarranjos de DNA o interesse está nas diferenças estatísticas entre os níveis de

expressão de genes na amostras de referência e teste (hibridização da mistura de mRNA). Portanto, as considerações a seguir são focadas nos atributos estatísticos relativos.

Cálculo de descritores

As características relativas de *spots* das seqüências de cDNA são computadas em termos de razões (simples, logarítmica, regressão) dos valores de intensidades puras derivados dos canais vermelho e verde [25]. Enquanto que as características estatísticas são obtidas considerando-se os conjuntos de intensidades como aproximações de alguma função de probabilidade. Os descritores estatísticos mais comuns dos conjuntos de pixels do *foreground* e do *background* são suas médias, medianas e modas [5]. Abaixo, são demonstradas formas de se quantizar os descritores de *spot* a partir de valores relativos estatísticos:

Equação 17: des é o descritor estatístico utilizando razão simples, X é o símbolo para a média, mediana ou moda, o sobrescrito FRG refere-se ao *foreground* e o sobrescrito CHANNEL aos canais de cor da imagem.

$$des_{RATIO}^{X} = \frac{X_{FRG}^{\text{CHANNEL 0}}}{X_{FRG}^{\text{CHANNEL 1}}}$$

Equação 18: des é o descritor relativo estatístico utilizando razão logarítmica, X é o símbolo para a média, mediana ou moda, os sobrescritos FRG e BGK referem-se ao *foreground* e ao *background* e o sobrescrito CHANNEL aos canais de cor.

$$des_{LOG\ RATIO}^{X\ WRT\ BKG} = \log_2\left(\frac{X_{FRG}^{\text{CHANNEL 0}} - X_{BKG}^{\text{CHANNEL 0}}}{X_{FRG}^{\text{CHANNEL 1}} - X_{BKG}^{\text{CHANNEL 1}}}\right)$$

Enquanto a Equação 17 representa uma razão direta de valores absolutos, a Equação 18 é uma razão logarítmica de diferenças relativas. Com a utilização de diferenças relativas reduz-se os efeitos da fluorescência não específica (por exemplo, auto-fluorescência em lâminas de vidro), porém é preciso verificar os casos de *spots* fantasmas (quando as intensidades de *foreground* são menores que as de *background*). Além disso, parâmetros estatísticos podem ser calculados a fim de medir o formato da distribuição de intensidade (espalhamento, inclinação, simetria) e indicar intervalos de confiança para os descritores extraídos. Por exemplo, um alto desvio padrão observado entre as médias computadas de diferentes *spots*, significa uma grande variação dos valores. Conseqüentemente, a confiança em obter repetidamente o descritor exato é baixa (alta incerteza de valores absolutos para experimentos repetidos) [5].

Existe também a razão de regressão, que supõe a existência de uma relação de dependência linear entre as intensidades dos pixels dos canis verde e vermelho da

imagem [26]. A razão de regressão significa uma estimativa dessa relação e pode ser calculada através de análises de correlação. Dessa forma, espera-se que as características extraídas se ajustem às diferentes eficiências dos marcadores fluorescentes quando detectados (corantes vermelhos têm maior eficiência que os verdes) e às diferentes quantidades de mRNA das amostras. A razão é computada ajustando-se uma linha reta com nenhuma interseção no gráfico de dispersão formado pelas intensidades vermelho e verde dos pixels de *foreground* e *background*. Se o valor β é utilizado no ajuste da reta Y=β*X + ε para Y'=X' então a análise é também chamada de regressão linear dos canais [5]. O mecanismo de regressão é ilustrado a seguir:

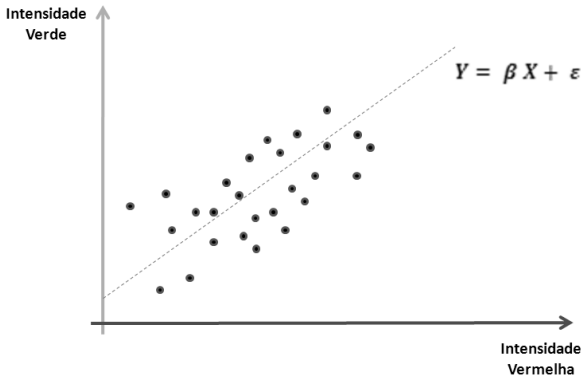

Figura 19: Ilustração da calibração linear dos canais verde e vermelho.

Visualização dos descritores

O jeito mais comum de se visualizar os descritores extraídos dos *spots* consiste na inspeção de uma tabela. Porém, como o número de *spots* vem aumentando nos experimentos de microarranjos, uma tabela com milhares de linhas equivalentes a cada *spot* analisado não fornece um mecanismo muito eficiente de visualização. Considerando o de que os *spots* nos microarranjos são organizados sob uma forma padrão de *grid*s regulares, torna-se natural apresentar os descritores de *spot* extraídos da mesma forma. Essa abordagem de visualização preserva a localização espacial relativa dos *spots*, além de possibilitar a inspeção diretamente através de imagens características. Um exemplo desse tipo de visualização é demostrado na Figura 20 abaixo:

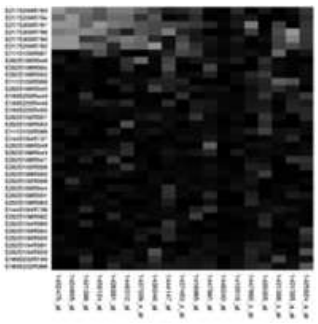

Figura 20: Visualização dos descritores de *spots*

O objetivo a longo prazo é poder combinar descritores de expressões gênicas com informações extras sobre os genes. Por exemplo, seria possível demonstrar em várias escalas a informação da expressão gênica do microarranjo aliada à estrutura 3D das seqüências [27].

Seleção de Descritores de Spot

É importante entender os prós e contras de cada tipo de descritor de *spot*, a fim de escolher os mais apropriados para a análise dos dados de microarranjos. Os descritores determinísticos absolutos, como a soma das intensidades são dependentes do tamanho dos *spots* e sensíveis à contaminação além da saturação dos pixels. Similarmente, os descritores absolutos são inapropriados para *spots* de microarranjos de cDNA já que as medidas da intensidade fluorescente dependem dos marcadores de referência.

Em relação aos descritores estatísticos, o uso da média amostral reduz a variação de intensidade do *spot*, mas é sensível a intensidades fora do padrão (outliers). A mediana por sua vez é mais resistente a outliers, mas também é mais cara computacionalmente. O descritor moda é definido como a intensidade mais freqüente ocorrendo num conjunto de intensidades de *foreground* ou *background*. É resistente aos outliers e simples de se computar. No entanto, é difícil de estimar a confiabilidade quando a freqüência de ocorrências de intensidade (histograma de intensidade) contém múltiplos picos (distribuição de intensidade multimodal) [5]. Quando a distribuição de intensidade é unimodal e simétrica, as estimativas da média, mediana e moda são todas iguais.

No caso da computação de descritores relativos, aborda-se o problema sob os pontos de vista da modelagem estatística e da análise de correlação. Do ponto de vista de uma modelagem estatística, é preferível usar operadores (transformações) que levem a uma variável aleatória segundo uma distribuição Gaussiana devido à facilidade de manipulação matemática. Do ponto de vista da análise de correlação, a razão de regressão não descreve bem as características da maioria das imagens, visto que na prática a relação entre os pixels das diferentes bandas da imagem não deve ser linear. Assim, a razão de regressão é um descritor mais apropriado para imagens microarranjos com contrastes de alta intensidade entre *foreground* e *background* [5].

5.3 Normalização dos Dados

As medidas da intensidade da fluorescência em cada canal de cor da imagem podem ser distorcidas durante os passos de preparação de dados. Logo, para garantir a confiabilidade nas comparações entre os resultados obtidos de diferentes lâminas de microarranjos, são utilizadas técnicas de normalização dos dados.

A dificuldade de realizar comparações significativas decorre das diferenças entre as preparações das lâminas de microarranjos com relação a: quantidades de mRNA, conFigurações do scanner, protocolos de microarranjos ou marcações específicas. O

propósito da normalização é o ajuste dessas variações, principalmente da eficiência da marcação e da hibridização das amostras. Dessa forma espera-se descobrir as verdadeiras variações biológicas resultantes da análise dos níveis de expressões gênicas.

Nas próximas seções são discutidas as principais abordagens para normalização dos dados.

5.3.1 Normalização utilizando descritores estatísticos

Descritores estatísticos incluem média, mediana, moda ou percentil da distribuição de intensidade das amostras. A normalização pode ser realizada pela divisão ou subtração dos descritores estatísticos. A Equação 5.18 abaixo representa a transformada-Z [26].

Equação 19: Normalização pela transformada-Z, onde I é o valor de intensidade de um pixel numa determinada posição (x, y), μ é a média e σ é o desvio padrão das intensidades da imagem.

$$I^{NORM\ STAT}_{Z-TRANSFORM}(row, col) = \frac{I(row, col) - \mu}{\sigma}$$

5.3.2 Normalização utilizando spots de controle

Nessa abordagem, são inseridos *spots* de intensidades conhecidas ou genes de nível de expressão conhecidos na lâmina de microarranjo. As intensidades desses *spots* são utilizadas como referência (controle) para a normalização das intensidades de todos os outros *spots*. Nesse caso, a fim de se obter uma maior precisão, inclusive nas variações locais, os *spots* de controle são espalhados ao longo da lâmina do microarranjo.

5.3.3 Normalização utilizando análise de regressão

As razões de regressão são freqüentemente usadas como partes da normalização dos canais da imagem de microarranjos. Entre os métodos de normalização que utilizam análise de regressão o mais utilizado é o Intra-lâmina (*withinslide*) que consiste na subtração de um fator de normalização *c* das razões logarítmicas individuais das intensidades [28].

Equação 20: Modelo de normalização por análise de regressão.

$$I(row, col) = log_2 R/G - c$$

O fator de normalização é calculado separadamente para cada lâmina de microarranjo, utilizando somente os dados das seqüências que hibridizaram. Como c é uma função, há diversas maneiras de defini-la sendo a mais comum a técnica de normalização global, onde é assumido que as intensidades verde e vermelha se relacionam por um fator constante (relação linear, R = β*G). Assim, assumindo um deslocamento nulo para a dependência linear entre os canais:

Equação 21: Normalização global.

$$I(row, col) = log_2 R/G - c = log_2 R/G - log_2 \beta = log_2 R/(\beta * G)$$

6. Experimentos

Esse capítulo é direcionado à demonstração da realização de alguns experimentos utilizando os conceitos propostos. Em virtude da existência de várias ferramentas consolidadas e dos desafios que existem na área de processamento de imagens de microarranjos, preferiu-se direcionar os esforços no desenvolvimento de um método automatizado de definição e análise dos *spots* de microarranjos de DNA.

6.1 Método das estimativas para alinhamento do grid

Conforme o Capítulo 4, para alcançar altos rendimentos no processamento de imagens de microarranjos, além dos requisitos de robustez e confiabilidade, os algoritmos precisam ser o máximo possível livre de parâmetros determinados por um usuário. Além disso, métodos automatizados precisam identificar, calcular e restringir os parâmetros que representariam o conhecimento à priori, ou seja, parâmetros que necessitariam da intervenção de um usuário.

Basicamente a proposta se enquadra na classe de abordagem totalmente automática do ponto de vista da automação e na classe da abordagem orientada aos valores dos dados do ponto de vista de análise.

6.1.1 Algoritmo das estimativas

O passo único de interação com o usuário incorpora apenas dois parâmetros de inicialização:

1. Quantidade de *spots* por microarranjo na horizontal;
2. Quantidade de *spots* por microarranjo na vertical.

Todos os demais valores necessários são inferidos a partir desses dois parâmetros. Uma visão básica das técnicas utilizadas para estimativas é explicada na seção seguinte. A implementação foi realizada na ferramenta Matlab.

Basicamente, o algoritmo de processamento da imagem, é dividido em três fases:

1. Detecção da quantidade e das fronteiras dos elementos;
    a. Somatório das intensidades dos pixels na direção horizontal;
        i. Estimativa da quantidade de *grid*s na direção horizontal;
        ii. Estimativa da quantidade de *spots* em cada *grid* na direção horizontal.
    b. Somatório das intensidades dos pixels na direção vertical;
        i. Estimativa da quantidade de *grid*s na direção horizontal;
        ii. Estimativa da quantidade de *spots* em cada *grid* na direção horizontal.
2. Definição das localizações a partir das estimativas da fase 1;

a. Estimativa da localização da área de cada *grid;*
b. Estimativa da localização de cada *spot* dentro de um *grid.*

3. Refinamento das localizações dos *spots* aplicando-se as fases 1 e 2 em cada subimagem formada por um *grid*;
4. Refinamento das localizações dos *spots* a partir de cada subimagem formada por uma célula de *grid*.

6.1.2 Descrição de passo a passo

Detecção da quantidade e das fronteiras dos elementos

Utilizando o somatório das intensidades dos pixels numa determinada direção, é possível inferir características importantes da imagem, dentre elas quantos e como estão distribuídos os *grid*s nos microarranjos [29]. Para auxiliar na visualização, basta observar o padrão gráfico dos valores apresentados conforme a Figura abaixo:

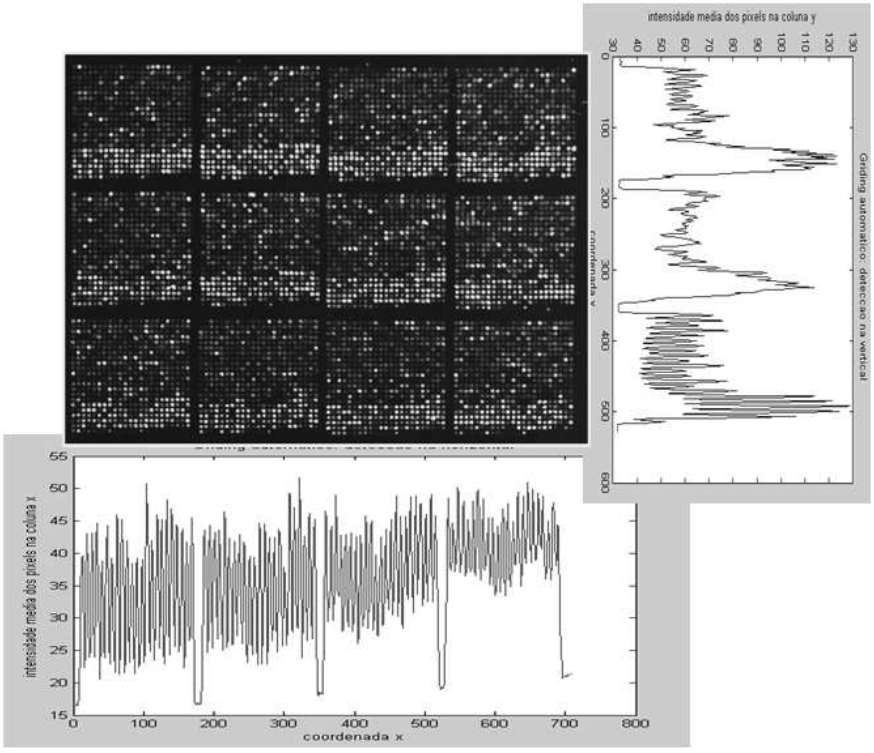

Figura 21: Trecho de experimento de microarranjo de DNA com os respectivos gráficos das intensidades médias dos pixels nas direções horizontal e vertical.

Percebe-se que tanto na direção horizontal quanto na vertical, existe uma queda brusca nos valores das intensidades que caracteriza o espaçamento entre os *grid*s. Além disso, cada pico observado é interpretado como o centro aproximado de um *spot*. Dessa forma é possível estimar além das quantidades de *grid*s, o tamanho médio dos *spots* em cada direção.

Primeiramente o tamanho dos *spots* é calculado observando as transições entre os picos de intensidade segundo a rotina:

1. *Dado o vetor AvgIntensity[] contendo o somatório médio dos pixels em uma determinada direção;*
2. *Detecte todos os picos de intensidade observando as transições entre os pixels;*
    a. *Para cada pixel p(i) = AvgIntensity[i]:*
        i. *Se p(i-1) < p(i) > p(i+1) o pixel é marcado como pico positivo*
        ii. *Se p(i-1) > p(i) < p(i+1) o pixel é marcado como pico negativo*
        iii. *Caso contrário o pixel não é marcado*
3. *Calcule o tamanho de cada spot observado a quantidade de pixels entre dois picos negativos consecutivos;*
4. *Estime o tamanho do spot a partir da mediana dos valores calculados anteriormente.*

Nesse caso, a escolha da mediana ao invés da média se deve ao fato dessa métrica ser menos sensível à influência de prováveis *spots* de baixa confiança (contaminação, ruídos) presentes no vetor de intensidade.

Definição das localizações

Dado o tamanho escolhido para o *spot*, com o conhecimento à priori da quantidade desse elemento em determinada direção, é possível estimar o tamanho médio dos *grid*s. Para isso, uma janela de pixels é criada com o tamanho do *grid* calculado (quantidade de *spots* vezes o tamanho mediano do *spot*) e percorre-se todo vetor de intensidades a fim de se delimitar as fronteiras dos *grid*s em cada direção. As fronteiras são marcadas de modo a maximizar uma função que determina o peso de cada delimitação de fronteira possível para o *grid*. No algoritmo, o peso foi definido pelo somatório de todas as intensidades englobadas pelas fronteiras do *grid*. Durante a varredura do vetor de intensidades na direção horizontal ou vertical, é verificado se houve aumento no peso do *grid*, se sim, continua-se a varredura.

A busca pára quando após cinco iterações, não ocorre melhora no valor da soma das intensidades dos pixels delimitados pela janela. A estratégia de cinco iterações é baseada no fato de que o espaçamento entre os *grid*s é maior que cinco pixels (empírico). Outra forma de se estimar o valor seria através do padrão gráfico de espaçamento entre os *grid*s demonstrado na Figura 20. O ideal seria o valor padrão do layout de espaçamento entre os *grid*s, porém como foi visto no Capítulo 2, não há uma padronização que regule o layout de preparação de lâminas com as variadas tecnologias de microarranjos. Abaixo se encontra uma demonstração da rotina utilizada para a delimitação das fronteiras dos *grid*s:

1. *Dado o vetor AvgIntensity[] contendo o somatório dos pixels em uma determinada direção;*
2. *Dado o tamanho em pixels da janela do microarray;*
3. *Percorre-se o vetor AvgIntensity[] com a janela window de pixels*

   a. *s(i) = somatório das intensidades dos pixels cobertos pela janela window(i) com pixel inicial na posição i de AvgIntensity[]*
   b. *Se s(i) > s(i - 1), então houve melhora, a window(i) é marcada como fronteira do microarray e a varredura continua*
   c. *Se não houver melhora após 5 iterações, passe para a definição do próximo microarray*
4. *Continua até que toda a quantidade de pixels não varridos do vetor de intensidades não seja suficiente para delimitar um novo microarray com o tamanho calculado*

Para a os *grid*s na imagem, é realizado um casamento da posição de cada fronteira marcada na direção horizontal com a posição da fronteira na direção vertical.

1. *Dado o vetor Columns[] contendo as fronteiras de cada microarray na direção horizontal;*
2. *Dado o vetor Rows[] contendo as fronteiras de cada microrarray na direção vertical;*
3. *Para cada par $c_{(i)}$ = Columns[i], e $r_{(i)}$ = Rows[i]*
   a. *Define as coordenadas que localizam as fronteiras de um microarray*
      i. $(x1, y1) = (r_{(i)}, c_{(i)})$
      ii. $(x2, y2) = (r_{(i)}, c_{(i+1)})$
      iii. $(x3, y3) = (r_{(i+1)}, c_{(i)})$
      iv. $(x4, y4) = (r_{(i+1)}, c_{(i+1)})$

Similarmente, a posição de cada *spot* também é calculada dentro de cada *grid*. Para a de microarranjo representa pela Figura 20 acima, obtém-se o seguinte resultado na fase 2 do algoritmo:

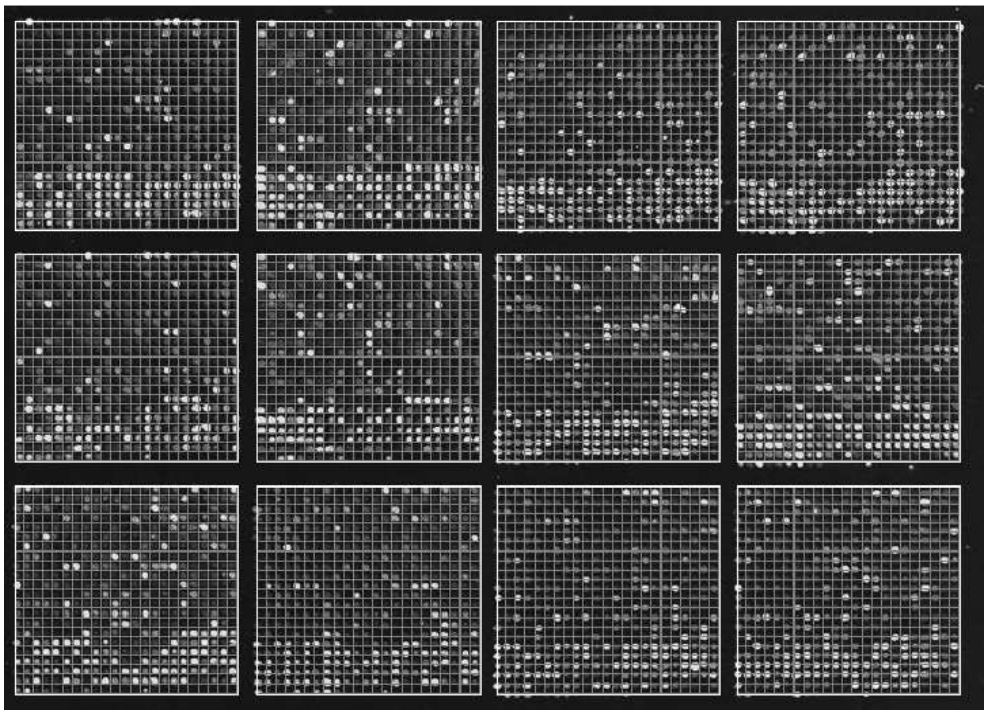

Figura 22: Estimativa de localização fornecida na fase 2 do algoritmo.

Refinamento das Localizações

A primeira estimativa de localização na imagem apresenta *spots* sobrepostos pelas divisões das áreas, ou seja, baixa precisão na localização. Caso fosse aplicado algum dos critérios de avaliação de *spot* definidos no Capítulo 5, as regiões de *foreground* e *background* seriam de baixa confiança para as subimagens formadas para esses *spots*. Uma forma encontrada para diminuir os erros, foi submeter as primeiras estimativas de localização a um refinamento.

No algoritmo proposto, tal refinamento é realizado utilizando os mesmos passos descritos. Porém, uma vez que já se possui uma localização razoável dos *grid*s, não é necessário usar toda a imagem no refinamento. Por isso são utilizadas apenas a subimagem correspondente a cada *grid*, extrapolando o tamanho real, para permitir uma possível correção na posição. A Figura abaixo ilustra uma subimagem representante de um *grid*, obtida a partir de sua primeira estimativa de posição:

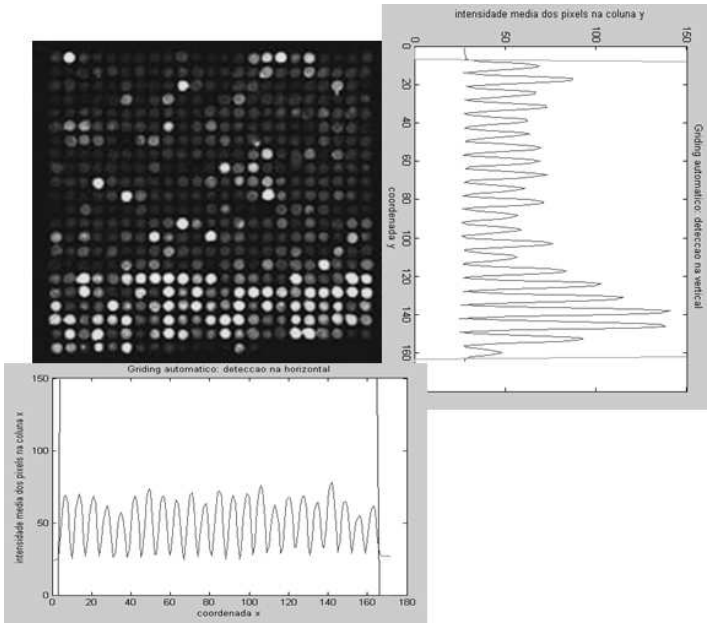

Figura 23: Subimagem correspondente a um *grid* submetido ao refinamento de posição.

Dessa vez os gráficos na direção horizontal e vertical apresentam-se mais suaves e com picos bem definidos. Como somente a subimagem do próprio *grid* está sendo utilizada, elimina-se a influência das outras distribuições de intensidades. As linhas verticais representam a delimitação das fronteiras do *grid*. Como resultado do refinamento das localizações, cada *grid* torna-se independente dos demais e por isso há uma melhora na definição dos *spots*. A seguir, é demonstrado o refinamento das posições dos *grid*s para a imagem de microarranjo da Figura 21.

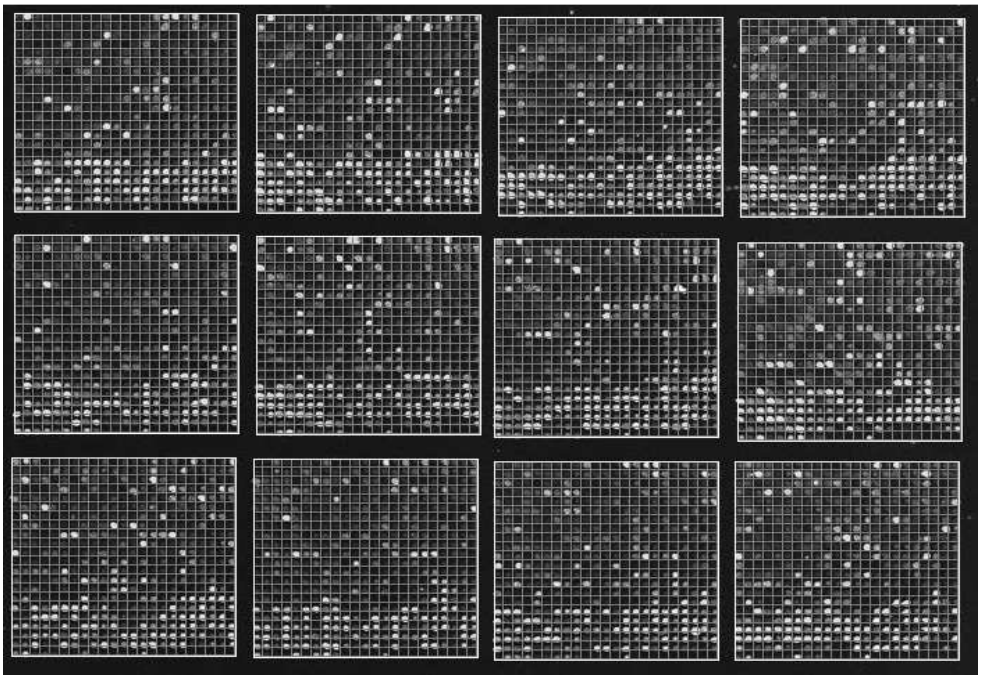

Figura 24: Novas posições dos *grid*s após o refinamento.

Enquanto as primeiras estimativas são obtidas a partir de toda imagem (global), o refinamento atua localmente para cada subimagem aumentando a precisão da localização. Embora seja constatada a melhoria na precisão dos *spots* definidos, ainda pode não ser suficiente para os critérios de controle de qualidade. Para alguns *grid*s, essa melhora ainda não é o suficiente de acordo a Figura 24. Numa inspeção mais detalhada são encontradas células mal delimitadas, cuja área abrange mais de um *spot*.

Uma alternativa para remover esse tipo de erro é aplicar um novo refinamento, dessa vez utilizando somente a subimagem do próprio *spot*. O ajuste novamente é realizado a partir da análise do padrão de distribuição de intensidade nas direções horizontal e vertical.

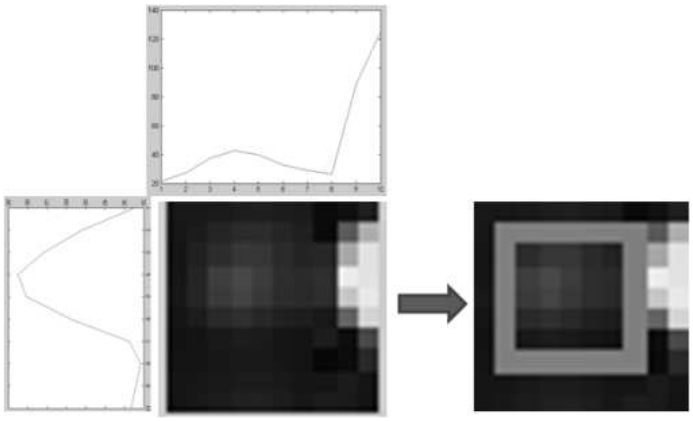

Figura 25: Subimagem correspondente a uma célula de *grid* submetida ao refinamento de posição.

Idealmente, o padrão de intensidade média de um spot é caracterizado pela presença de um pico central, dividindo a área do gráfico em duas partes simétricas. Então, neste novo refinamento, as fronteiras de cada *spot* são delimitadas de modo a obter uma subimagem que contenha um padrão de distribuição de intensidade o mais próximo possível do ideal. Espera-se que dessa forma a influência das intensidades de outros *spots* seja eliminada o que aumentaria as chances da classificação precisa das regiões de *foreground* e *background*.

O último passo de ajuste é aplicado então para todas as células de cada *grid*, contribuindo para uma melhor definição dos *spots*. A Figura 25 reflete o resultado final do alinhamento dos *grids* pelo método das estimativas, com cada spot individualmente localizado.

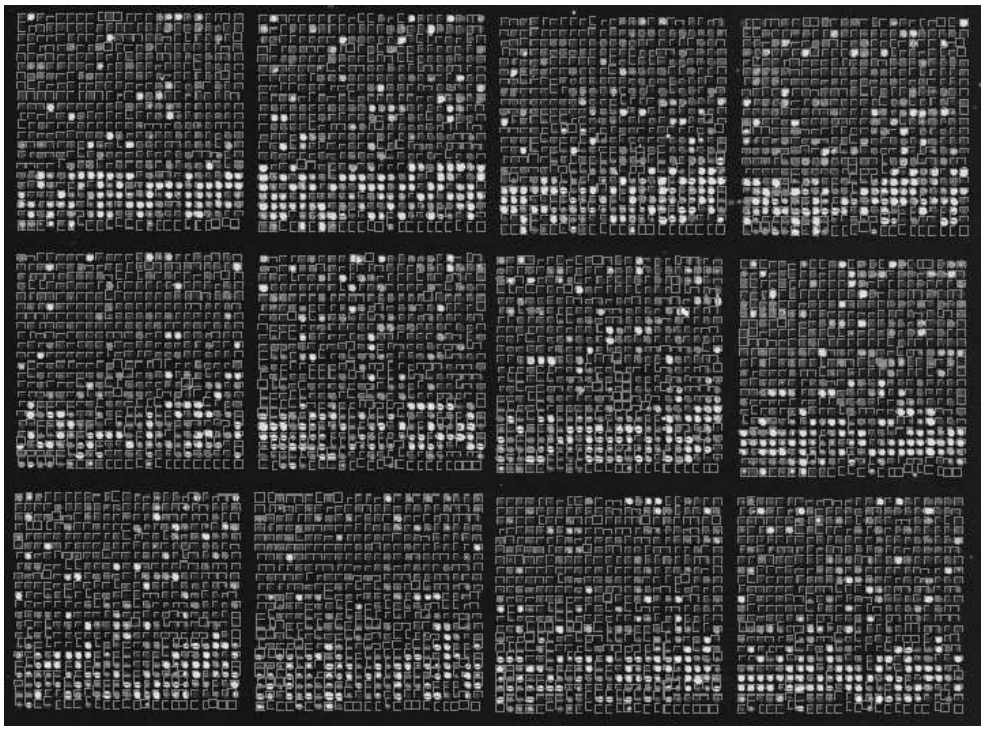

Figura 26: Novas posições das células após o último passo de refinamento de localização.

6.2 Classificação das regiões de foreground e background

Como parte do fluxo de dados (ver Capítulo 2) do processamento de imagens de microarranjos de DNA, após o alinhamento dos *grid*s, segue-se o processo de definição das regiões de *foreground* e *background* nos *spots* localizados. A fim de se observar as regiões geradas a parir das localizações estimadas pelo algoritmo de alinhamento de *grid* proposto, foi aplicado um algoritmo de clusterização nas subimagens representantes de cada *spot* definido. O algoritmo escolhido foi o k-Means, o qual é utilizado com bastante freqüência na segmentação de imagens de microarranjos [31].

Para a utilização do k-Means foram utilizadas as seguintes considerações:

- Base de dados: subimagem definida pela localização aproximada de um *spot*, sendo os pixels as amostras a serem agrupadas;
- Atributos da amostra: coordenadas relativas (x, y) e intensidade do pixel;
- Função de similaridade: distância euclidiana às intensidades dos pixels.

Embora seja prático, esse tipo de agrupamento apresenta a limitação de gerar resultados imprevisíveis. A fim de eliminar o não-determinismo das regiões geradas (característica indesejável nos algoritmos de processamento de imagens de microarranjos (ver Capítulo 4), foi utilizada uma regra de inicialização do k-Means. Ao invés de escolher representantes iniciais para os grupos de forma aleatória, passou-se a adotar os pixels mais distantes em termos de intensidade, ou seja, aqueles com o maior e o menor valor de intensidade. Segue o resultado da aplicação do k-Means nos *grid*s representados na Figura 26:

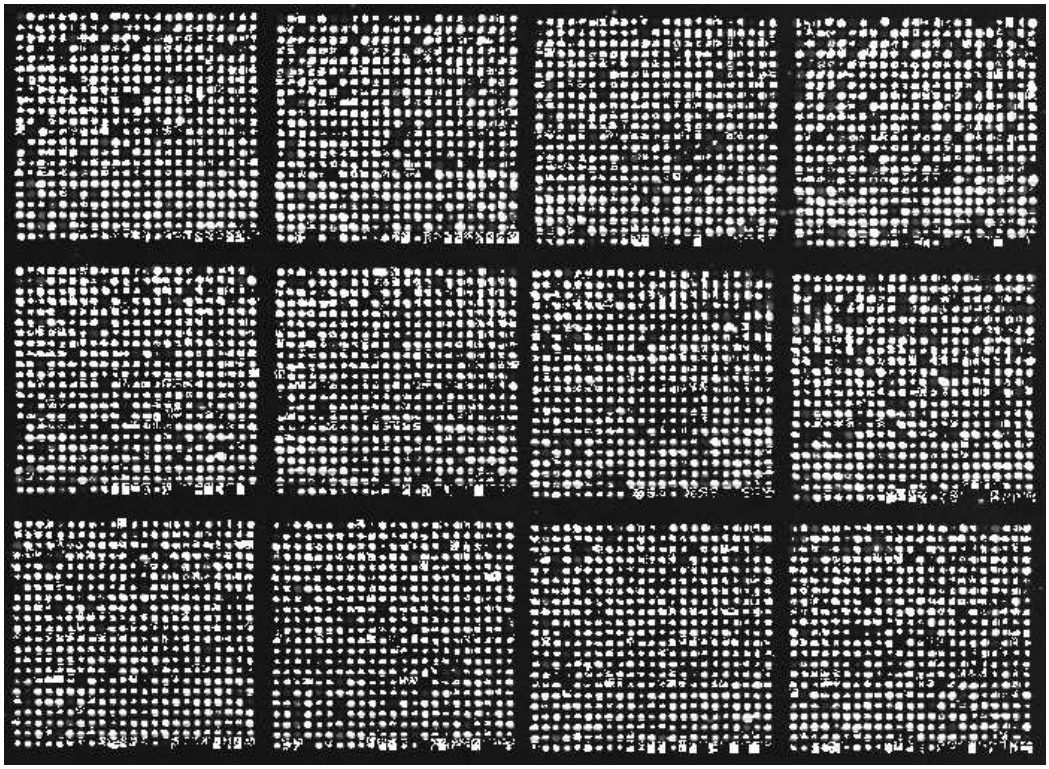

Figura 27: *Spots* segmentados pelo k-Means após o refinamento das posições (os pixels classificados como *foreground* foram marcados com a cor branca).

6.3 Desenvolvimento de ferramenta para interação com usuário

Também em caráter experimental, foi desenvolvida uma ferramenta de alinhamento manual de *grid* com funcionalidades que melhorariam a precisão das subimagens representantes dos *spots*. Vista a necessidade de interagir com um usuário, optou-se pela implementação em linguagem de programação Java, pelas facilidades de se criar interfaces.

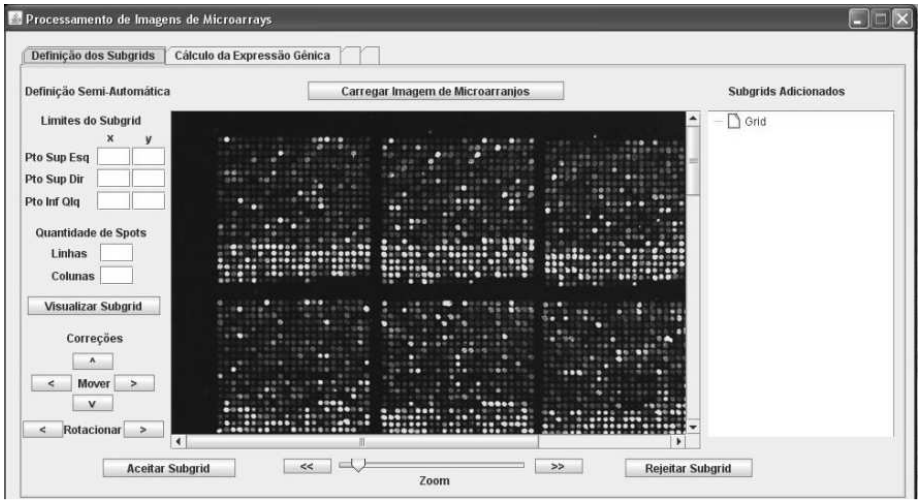

Figura 28: Ferramenta desenvolvida para testes com o alinhamento manual de *grid*.

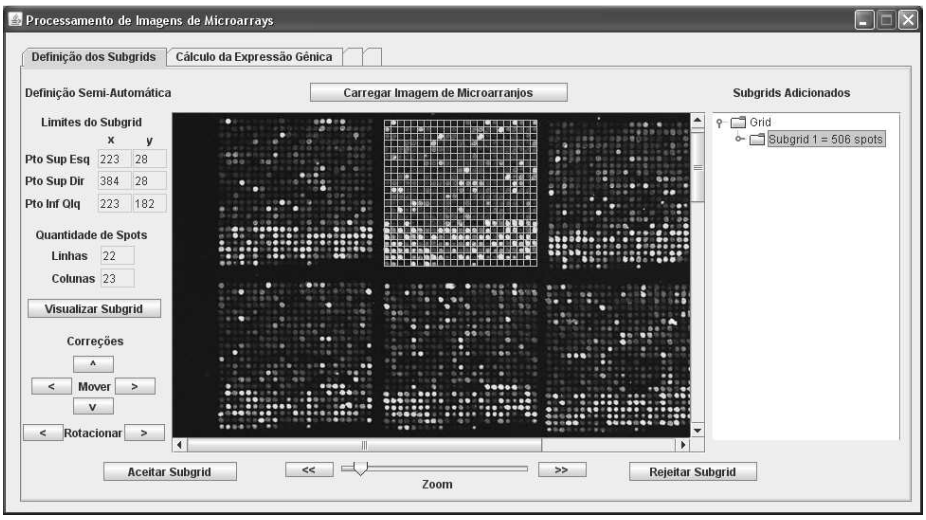

Figura 29: Exemplos de *grid*s criados

O intuito é posteriormente integrar a agilidade do algoritmo das estimativas ao ajuste manual das posições dos *grid*s baseado na visão do usuário. Como o algoritmo fornecendo as primeiras estimativas das localizações, o aumento no rendimento do processamento das imagens de microarranjos nesse caso é evidente.

7. Considerações Finais

7.1 Conclusões

Em busca do entendimento da área de processamento de imagens de microarranjos de DNA, descobriu-se que existe um complexo fluxo de eventos que divide o experimento com microarranjos em várias etapas:

1. Preparação
2. Aquisição
3. Processamento
4. Análise
5. Interpretação

A primeira, relacionada à preparação do experimento, utiliza-se o conhecimento de diversos mecanismos da biologia molecular para obter as chamadas expressões gênicas.

A segunda diz respeito a forma como são organizados os microarranjos, onde há uma preocupação especial em evitar e conhecer as diversas formas de variações que diminuem a qualidade das imagens geradas pelos detectores especiais de fluorescência.

A terceira envolve a aplicação de algoritmos para definir as regiões de interesse na imagem, ou seja, escolher os pixels que serão utilizados na quantificação da expressão gênica. Sendo um grande desafio o desenvolvimento de técnicas automatizadas que diminuam o esforço humano ao mesmo tempo em que mantenham as exigências de desempenho e de precisão.

A quarta, exige uma seleção das imagens processadas a fim de se manter os níveis de qualidade que garantam uma boa precisão durante a transformação das características em valores comparáveis e interpretáveis.

Uma revisão das principais técnicas de processamento assim como os desafios a serem superados foram reunidos nesse trabalho. Pretendeu-se com isso, prover um conhecimento mínimo para auxiliar na decisão quanto às abordagens a serem seguidas durante a execução de um experimento envolvendo imagens de microarranjos. Ainda sim, percebeu-se a escassez de formas de se comparar os resultados obtidos de diferentes ferramentas computacionais que tratam diferentes tecnologias de microarranjos de DNA

Por fim, foi possível pelo que foi exposto, desenvolver uma nova abordagem automática para a definição das células dos *grid*s que delimitam as imagens dos microarranjos. Essa abordagem necessita de apenas dois parâmetros que devem ser fornecidos por um usuário e permite a obtenção de resultados determinísticos. Os resultados experimentais demonstram que a modelagem é factível, sendo possível ainda a realização de várias melhorias que levem a um aumento de precisão assim como a eliminação de parâmetros.

7.2 Trabalhos Futuros

As técnicas de análise foram abordadas de forma superficial, portanto encontra-se aí, uma oportunidade de expandir o trabalho, com um detalhamento matemático mais formal assim como a pesquisa de outras abordagens. Além disso, como o foco esteve nas ações aplicadas às imagens, não foi pesquisada a parte de interpretação de dados que leva a construção das conclusões biológicas. Esta é outra área a ser considerada para a continuidade do trabalho.

É esperado que, com o conhecimento mínimo de todas as etapas do processo, seja possível a construção de uma ferramenta genérica, capaz de lidar com diversos formatos e padrões de imagens e permitir a obtenção de resultados confiáveis com a mínima dependência de interação com o usuário. Antes disso, contudo, é necessário encontrar uma forma confiável de se comparar os métodos pesquisados.

Referências Bibliográficas

[1] RUIVO, H. M. *Análise Integrada de Dados Ambientais Utilizando Técnicas de Classificação e Agrupamento de Microarranjos de DNA*: **Dissertação de mestrado em Computação Aplicada**. São José dos Campos: Instituto Nacional de Pesquisas Espaciais. 2007.

[2] MENA-CHALCO, J. P. *Indentificação de regiões codificantes de proteína através da transformada modificada de Morlet*. **Dissertação de mestrado em Ciências da Computação**. São Paulo: Universidade de São Paulo. 2005.

[3] VÊNCIO, R. Z. N. *Análise Estatística na Interpretação de Imagens: microarranjos de dna e ressonância magnética funcional*: **Dissertação de doutorado em Bioinformática**. São Paulo: Universidade de São Paulo. 2006.

[4] SCHENA, M.; SHALON, D.; DAVIS, R.W.; BROWN, P.O. *Quantitative monitoring of gene expression patterns with complementary DNA microarray*. **Science**, n. 270, p. 467-470. 1995

[5] KAMBEROVA, G.; SHAH, S. **DNA Array Image Analysis**: Nuts & Bolts, DNA Press. 2007. 280p. (Nuts & Bolts)

[6] Affymetrix. **GENECHIP® ARRAYS**. [on line]. Disponível em: http://www.affymetrix.com/index.affx. Último acesso: 10, junho, 2008.

[7] BALAGURUNATHAN, Y. et al. *Simulation of cDNA Microarrays via a Parameterized Random Signal Model*, **Journal of Biomedical Optics**. 2002.

[8] BUHLER, J.; IDEKER, T.; HAYNOR, D. *Dapple: Improved Techniques for Finding Spots on DNA Microarrays.* **UV CSE Technical Report UWTR**. 2000.

[9] STEINFATH, M. et al. *Automated image analysis for array hybridization experiments*, **Bioinformatics**, v. 17, p. 634-641. 2001.

[10] LAWRENCE, N. D. et al. *Reducing the variability in cDNA microarray image processing by Bayesian inference*, **Bioinformatics**, v. 20, n. 4, p. 518-526. 2004.

[11] BAJCSY P. *Gridline: Automatic Grid Alignment in DNA Microarray Scans.* **IEEE Transactions on Image Processing**, v. 13, n. 1, p. 15-25. 2004.

[12] BRANDLE, N.; BISCHOF, H.; LAPP, H. *Robust DNA Microarray Image Analysis.* **Machine Vision and Applications**, v. 15, p. 11-28. 2003.

[13] Molecular Devices. **GenePix Pro Microarray Image Analysis**. Disponível em: http://www.moleculardevices.com/pages/software/gn_genepix_pro.html. Último acesso: 10, junho, 2008.

[14] Eisen M. **ScanAlyze**. [on line]. Disponível em: http://rana.lbl.gov/EisenSoftware.htm. Último acesso: 08, maio, 2008.

[15] HARTELIUS, K.; CARTSTENSEN, J. M. *Bayesian Grid matching*. **IEEE Transactions on Pattern Analysis and Machine Intelligence**. v. 25, n. 2, p.162-173. 2003.

[16] LIEW, A.; YAN, H.; YANG, M. *Robust Adapttive Spot Segmentation of DNA Microarray Images.* **Pattern Recognition**. v. 36, p. 1251-1254. 2003.

[17] STEINFATH, M. et al. *Automated image analysis for array hybridization experiments*. **Bioinformatics**. v. 17, p. 634-641. 2001.

[18] HARTIGAN, J. A.; WONG, M. A. *A K-Means Clustering Algorithm.* **Applied Statistics**. v. 28, n. 1, p. 100-108. 1*979*

[19] BELACEL, N.; WANG, Q.; CUPERLOVIC-CULF, M. *Clustering Methods for Microarray Gene Expression Data.* **OMICS: A Journal of Integrative Biology**. v.10, n.4, p. 507-531. 2006

[20] GONZALEZ, R. C.; WOODS, R. E. **Processamento de Imagens Digitais.** 1. ed. São Paulo: E. Blücher, 2005. 509 p.

[21] ANGULO, J.; SERRA, J. *Automatic Analysis of DNA Microarray Images Using Mathematical Morphology.* **Bioinformatics.** v. 19, n. 5, p. 553-562. 2003.

[22] BOZINOV, D.; RAHNENFUHRER, *Unsupervised Technique for Robust Target Separation and Analysis of DNA Microarray Spots Through Adaptive Pixel Clustering.* **Bioinformatics**. v. 18, n. 5, p. 747-756. 2002.

[23] DRAGHICI, S. **Data Analysis Tools for DNA Microarrays**. Chapman & Hall, 2003. 512p. (CRC Mathematical Biology and Medicine Series).

[24] WANG, X.; GOSH, S.; GUO, S-W. *Quantitative quality control in microarray image processing and data acquisition.* **Nucleic Acids Research**, v. 29. 2001.

[25] DODD, L. E. et al. *Correcting Log Ratios for Signal Saturation in cDNA Microarrays.* **Bioinformatics**. v. 20, n. 16, p. 2685-2693. 2004.

[26] QUACKENBUSH, J. *Computational analysis of microarray data*. **Nature Reviews Geneticst**.v. 2, p. 418-427. 2001.

[27] ADAMS, R. M. et al. *Case Study: A Virtual Environment for Genomic Data Visualization.* **IEEE Transactions on Visualization**. 2002.

[28] KNUDSEN, S. **Guide to Analysis of DNA Microarray Data.** Wiley-Liss, 2004. 184 p.

[29] JAIN A. N. et al. *Fully Automated Quantification of Microarray Image Data* **Genome Research**, v. 12, n. 2, p. 325-332. 2002

[30] KOIDE, T. *Análise global da expressão gênica de Xyella fastidiosa submetida a estresses ambientais*. **Dissertação de doutorado em Ciências Biológicas**: Universidade de São Paulo, 2006.

[31] PEREIRA, O. *Análise de Dados de Microarrays de DNA*. **Genômica Funcional**. Universidade de Aveiro, 2003.

[32] TAN. P. K. et al. *Evaluation of gene expression measurements from commercial microarray platforms*. **Nucleic Acids Research**. v. 31, p. 5676-5684. 2003.

[33] GRADINER-GARDEN, M.; LITTLEJOHN, G. *A comparison of microarray databases Margaret*. **Briefings In Bioinformatics**. v. 2, n. 2, p. 143-158. 2001.

[34] FREUDENRIC, C. **How DNA Works**. [on line]. Disponível em: http://science.howstuffworks.com/dna1.htm. Último acesso: 27, abril, 2008.

[35] JÚNIOR, C. S.; SASSON, S. **Biologia: O Metabolismo Celular**. 7. ed. São Paulo: Saraiva, 2002. 400 p.

[36] WIKIPEDIA. **DNA Microarray**.[on line]. Disponível em: http://en.wikipedia.org/wiki/DNA_microarray. Último acesso: 27, abril, 2008.