UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Análise de Dados de Altas Dimensões

Autor: Gilberto Müller Beuren
Orientadora: Profa. Dra. Jandyra Maria Guimarães Fachel

Porto Alegre, Julho de 2010.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Análise de Dados de Altas Dimensões

Autor: Gilberto Müller Beuren

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Orientadora:
Professora Dra. Jandyra Maria Guimarães Fachel

Porto Alegre, Julho de 2010.

BANCA EXAMINADORA

Professora Dra. Jandyra Maria Guimarães Fachel

Professora Dra. Luciana Neves Nunes

Dedico este trabalho aos meus pais Tarcisio Beuren e Lia Maria Müller Beuren e ao meu irmão Marcelo Müller Beuren, que além de serem minha família sempre souberam ser os meus melhores amigos.

“You didn't see me on the floor weeping
You didn't see me lying by the door
You didn't see me swallowing my tablets
You can't look inside my eyes no more
Alone on the floor”

Genesis P-Orridge

Agradecimentos

Agradeço em primeiro lugar a toda minha família, que sempre me apoiou e incentivou em todas as decisões que eu tomei. Agradeço também a minha namorada Laura Santaló Rebello, que soube sempre transformar os meus “dias de chuva” em “dias de sol” e foi uma grande companheira acima de tudo; aos meus colegas de curso e turma, em especial Maicom Frozza, Rodrigo Coster e Renan Xavier Cortes, que foram fundamentais na minha formação e, acima de tudo, foram ótimos amigos e companheiros de estudo; ao meu grande amigo Ruben Ladwig, que me deu alguns dos conselhos mais valiosos que levo na minha vida; aos meus amigos Diego Barbosa de Souza e Paulo Fernando Nericke Motula; aos meus colegas de trabalho da Siqueira Campos Associados, com os quais aprendi e continuo aprendendo não apenas sobre trabalho e estatística, mas também sobre a vida, em especial a Caroline Legramanti Rodrigues; aos funcionários do Núcleo de Assessoria Estatística da UFRGS, onde trabalhei dois anos como bolsista e desenvolvi minha paixão pela estatística e por fim, mas não por isso menos importante, à professora e orientadora Jandyra Maria Guimarães Fachel, que foi a minha “mãe adotiva” no curso.

Resumo

O avanço tecnológico vem permitindo que novas áreas da ciência se desenvolvam e, com isso, novas técnicas começam a surgir. Problemas antes considerados sem solução passam a ser resolvidos computacionalmente. E isto não é diferente na estatística. Bancos de dados cada vez maiores vem surgindo e, com isso, surgem dados com altas dimensões. Áreas como química, genética e biociências têm um crescente interesse em analisar este tipo de dado. As técnicas multivariadas utilizadas atualmente não se aplicam neste tipo de caso e, para isso, novos métodos devem ser desenvolvidos. Este trabalho tem como principal foco fazer uma revisão da literatura existente sobre o tema de dados com altas dimensões. São mostradas técnicas como análise de similaridade, redução de dimensões, análise de *cluster* e medidas de distância entre sequências de DNA. É apresentado o *software* ImageMaster™ 2D Platinum, que é uma alternativa viável para a realização da análise de similaridade entre imagens e o *software* PAST e, por fim, três exemplos práticos ilustram a utilização dos métodos abordados em dados com altas dimensões.

Palavras chave: Dados com Altas Dimensões, Eletroforese Bidimensional, Sequências de DNA, Comparação de Imagens.

Abstract

Technological progress is enabling new science areas to develop and, thus, new techniques are beginning to emerge. Problems previously considered unsolvable now have a computational solution. And this is not different in statistics. Increasing databases has emerged and, with this, high dimensional data arise. Areas such as chemistry, genetics and life sciences have an increasing interest in analyzing this kind of data. The current multivariate techniques do not apply in this case and, therefore, new methods must be developed. This monograph has as its main focus to review the existing literature on the topic of high dimensional data. It is shown techniques such as similarity analysis, dimensionality reduction, cluster analysis and measures of distance between DNA sequences. We present the software ImageMaster™ 2D Platinum, which is a viable alternative to the analysis of similarity between images and the software PAST and, finally, three examples illustrate the use of the methods with high dimensional data.

Keywords: High Dimensional Data, Bidimensional Electrophoresis, DNA sequences, Image Comparison.

Sumário

1. Introdução

O avanço cada vez mais crescente da tecnologia vem abrindo portas antes inacessíveis em todos os ramos da ciência. Problemas antes considerados impossíveis de se solucionar, devido a sua alta complexidade, são hoje resolvidos de forma computacional em tempo cada vez mais hábil. E para acompanhar esta "revolução" tecnológica, novas técnicas e áreas vêm surgindo e evoluindo muito rapidamente.

Na estatística isso não é diferente. Na última década os avanços computacionais e tecnológicos fizeram com que a resolução de muitos problemas considerados urgentes na estatística pudessem ser resolvidos. Estes problemas envolvem quantidades enormes de dados ou então dados onde as dimensões das observações são muito grandes, chegando até a ser maior do que o tamanho da amostra.

As técnicas multivariadas conhecidas até então não geram bons resultados para este tipo de dados, mesmo sob fortes suposições. Este problema, juntamente com a necessidade crescente de uma análise de uma quantidade enorme de informação genética fez com que novas técnicas, como análises de *cluster* e de similaridade para dados com muitas dimensões, especialmente criadas para dados com altas dimensões, começassem a se desenvolver e a serem discutidas em diversos *workshops* voltados ao tema, como por exemplo, na vigésima quinta Conferência Internacional de Biometria (XXVth International Biometric Conference), a ser realizada em Florianópolis no ano de 2010, onde diversas palestras estão relacionadas ao tema de alta dimensionalidade de dados.

Em muitas áreas de ciência, o problema da dimensionalidade alta de dados começa a surgir de forma mais recorrente. Apresentamos alguns dos exemplos mais comuns, que são ilustrados na página seguinte: química (utilizando espectrografias – Figura 1), astronomia (utilizando dados automáticos de telescópios – Figura 2), genética (dados de micro arranjos de DNA – Figura 3) e biociências (utilizando eletroforese bidimensional para detecção de proteínas, representadas por manchas – Figura 4).

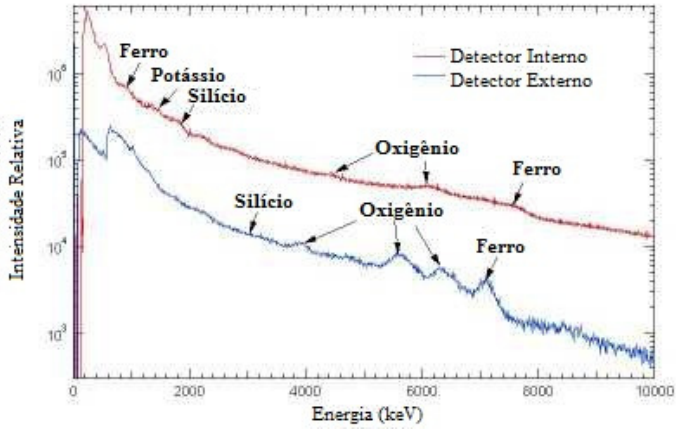


Figura 1: Espectrografia utilizada para detecção de elementos químicos

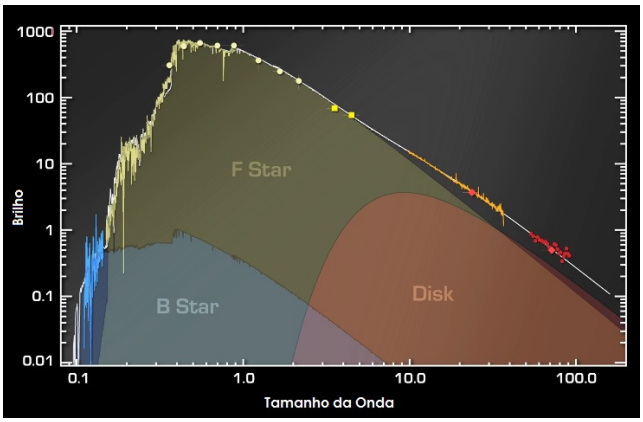


Figura 2: Gráfico de dados coletado por telescópio

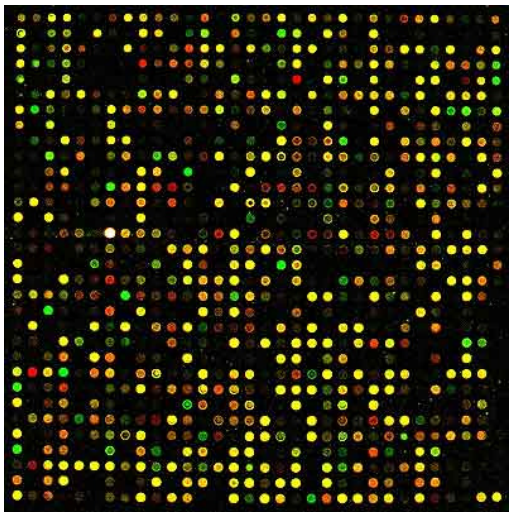


Figura 3: Micro arranjo de DNA

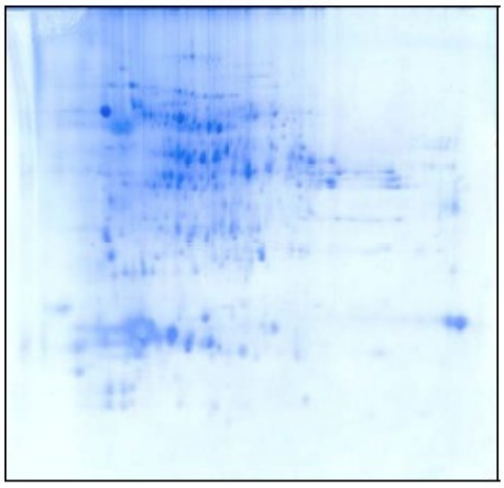


Figura 4: Eletroforese bidimensional de proteínas de raízes de algodoeiro

O interesse em resolver de uma forma mais precisa este tipo de dado já vem de muito tempo atrás. A indústria química, por exemplo, tem interesse em analisar dados de composição química de tintas e solventes desde que estes começaram a ser comercializados. O mesmo vale para as farmácias e indústrias cosméticas, com seus perfumes e xampus. É interessante estudarmos a composição destes produtos para podermos aperfeiçoá-los e reduzirmos o seu custo de produção. Veterinários e biocientistas trabalham com muitos dados que envolvem composição de proteínas em determinados tecidos de animais. Geneticistas e médicos estão na busca de curas para doenças associadas com arranjos de DNA.

Surgiu então uma necessidade de analisar os dados com altas dimensões que novas tecnologias geram. É de interesse destas áreas o desenvolvimento de técnicas

capazes de analisar satisfatoriamente este tipo de informação, utilizando a imagem inteira, sem perder nenhum tipo de informação, que é muito valiosa nestas situações.

Este trabalho tem como objetivo abordar algumas destas técnicas que vem sendo discutidas ao longo da última década, como análise de similaridade e distâncias entre sequências de DNA. Também apresentaremos informação sobre o *software* ImageMaster™ 2D Platinum (GE Healthcare), planejado para resolver problemas nesta área e sobre o *software* PAST, utilizado para cálculos de distância entre sequências de DNA e T^2 de Hotelling. Serão também mostrados três exemplos práticos onde dados com altas dimensões são analisados.

No próximo capítulo o leitor será introduzido a alguns conceitos básicos que serão fundamentais para a compreensão do resto do trabalho.

2. Conceitos Básicos

Um tipo específico de dados, onde existe um número muito grande de dimensões, que chega a ser comparável com o tamanho de amostra, ou então, algumas vezes até maior do que ele, é conhecido como dados com altas dimensões. Um exemplo prático disto é a presença de muitos genes, mas poucos pacientes com uma determinada doença. Este tipo de dado não pode ser tratado pelas técnicas tradicionais existentes, que não suportam essa alta dimensionalidade dos dados. Veremos a seguir, os problemas que este tipo de dado pode causar.

Em termos gerais, problemas com alta dimensionalidade resultam do fato de que um número fixo de pontos torna-se cada vez mais "esparso" quando o número de dimensões vai aumentando. Para tornar a visualização deste problema mais compreensível, vamos considerar 100 pontos distribuídos aleatoriamente segundo uma distribuição uniforme no intervalo [0,1]. Se este intervalo é dividido em 10 células, é altamente provável que cada célula contenha algum ponto. Entretanto, se mantivermos fixo este número de pontos, porém, distribuídos em um quadrado unitário (onde, logicamente cada ponto passa a ser bidimensional), mantendo a divisão proposta anteriormente (discretização de 0,1 para cada dimensão), desta vez teremos 100 células bidimensionais e é razoável propor que alguma das células ficará vazia. Partindo para um exemplo tridimensional, teremos 1000 células, resultando numa quantidade muito maior de células vazias do que células com pontos, visto que o número de células é muito maior do que de pontos. Os dados começam a se "perder no espaço" à medida que aumentamos as dimensões.

No caso do agrupamento de dados (análise de *cluster*) com altas dimensões, o problema da dimensionalidade afeta principalmente a medida de distância ou de similaridade. A maioria das técnicas de análise de *cluster* depende desta medida e, em geral, agrupam objetos mais próximos em grupos separados. O mesmo problema será encontrado quando utilizamos medidas de similaridade para detectar padrões semelhantes em imagens multidimensionais.

O comportamento das distâncias em dados com altas dimensões vem sendo estudado há alguns anos. É mostrado por Beyer et al (1998) que para alguns tipos de distribuições de dados, a distância relativa entre o ponto mais próximo e o ponto mais

distante de um determinado ponto escolhido ao acaso vai a zero quando o número de dimensões aumenta. É comum se dizer então que as distâncias entre pontos se tornam relativamente uniformes em dados com altas dimensões. O mesmo problema é encontrado ao se utilizar a distância absoluta ao invés da relativa, onde, segundo Hinneburg et al (2000), para dados com mais de duas dimensões, o uso da distância entre pontos é insignificante para análise de *cluster*.

Em resumo, isto tudo mostra que dados com altas dimensões definitivamente não podem receber o mesmo tratamento que dados com poucas dimensões e, portanto, necessitam de diferentes abordagens.

O conhecimento deste tipo de dificuldade apresentada é de grande interesse para uma grande gama de áreas da ciência. Para tal entendimento, é importante que alguns conceitos destas áreas sejam definidos também.

Na área genética, em 1990, se inicia o Projeto Genoma Humano, com um financiamento inicial de 50 bilhões de dólares e duração prevista de 15 anos. Entre os seus principais objetivos, encontram-se sequenciar e decodificar todo o DNA do genoma humano. Genoma é uma sequência completa de DNA. Esta abordagem foi pioneira na obtenção de dados com altas dimensões. A partir deste projeto, inúmeras outras pesquisas no ramo genético começaram a se desenvolver.

Um exemplo do que foi citado acima é a criação de um termo muito utilizado na área veterinária, que é o proteoma. Ele vem da união das palavras PROTEína e genOMA e é o conjunto de proteínas expressas por algum genoma. Segundo definição de Wilkins et at (1995), em termos gerais, proteoma é o equivalente protéico ao genoma. Porém, o genoma de um indivíduo é praticamente constante, independente de qual célula está sendo analisada, enquanto que o seu proteoma varia bastante de célula para célula (neurônio e linfócitos, por exemplo).

Para a análise de proteoma, muito útil nas áreas de genética, medicina, biociências e veterinária, é utilizada uma técnica conhecida como eletroforese, que se baseia na migração das moléculas carregadas, numa solução, em função da aplicação de um campo elétrico. Esta técnica foi primeiramente utilizada no ano de 1937. Porém, uma nova técnica, em 1975 começa a se desenvolver, gerando imagens bidimensionais similares à Figura 4. Esta técnica se chama eletroforese bidimensional e tem como principal objetivo a detecção de proteínas, assim como a sua quantidade, em determinadas células estudadas. Com o surgimento desta técnica, começou a surgir a necessidade de algum tipo de análise que comparasse diferentes imagens (chamadas de géis) para se verificar a eficiência de algum novo tratamento celular, já que este tipo de

dado apresenta altas dimensões. No entanto, estes géis são uma representação bidimensional de uma imagem originalmente composta por três dimensões, como pode ser visto na Figura 5 abaixo:

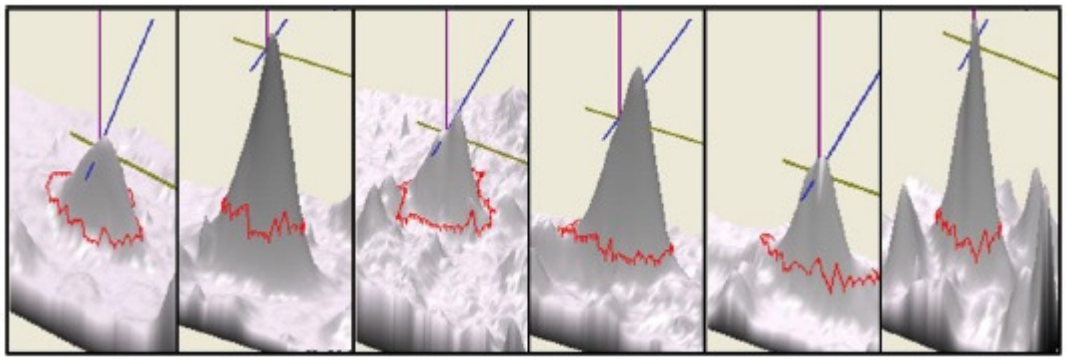


Figura 5: Imagem tridimensional de seis eletroforeses

Os picos em eletroforeses tridimensionais semelhantes aos apresentados na Figura 5 irão resultar nas manchas mais intensas presentes nos géis gerados através da eletroforese bidimensional semelhante à Figura 4.

Ainda na área genética, a análise da sequência de DNA vem se desenvolvendo de forma crescente. Cada sequência de DNA apresenta nucleotídeos, que são compostos ricos em energia e que auxiliam os processos metabólicos. O DNA apresenta nucleotídeos compostos por duas bases púricas (adenina e guanina, representadas pelas letras A e G, respectivamente) e duas bases pirimídicas (citosina e timina, representadas pelas letras C e T, respectivamente). Para as análises realizadas em enormes micro arranjos do DNA, são realizados cortes em sequências alvo específicas, chamados de sítios. Estes sítios então serão estudados, deixando o trabalho mais objetivo e prático.

Já na área química, para a detecção de determinados tipos de elementos presentes em algum objeto, são utilizados gráficos com espectros, chamados de espectrogramas. Determinados picos no gráfico podem determinar a presença de algum tipo de elemento. A mesma necessidade de comparação de imagens surge aqui, onde temos diversos espectrogramas e queremos verificar se seguem o mesmo padrão de comportamento com relação aos seus elementos químico.

Por fim, na área de astronomia, com o avanço tecnológico, milhões de dados surgem de telescópios. São, portanto, tiradas diversas fotos e gerados milhares de gráficos que precisam de uma análise mais sofisticada devido a sua alta dimensionalidade.

Nos próximos capítulos serão abordadas técnicas estatísticas que vem sendo desenvolvidas com o intuito de comparar grupos de imagens ou então agrupar informações de dados com altas dimensões.

3. Análise de Similaridade

A técnica de eletroforese bidimensional teve as suas primeiras citações feitas por O'Farrell (1975). Com esta técnica, é possível separarmos as proteínas primeiramente utilizando os seus pontos isoelétricos e depois no seu peso molecular. Isto gera uma imagem bidimensional (chamada de gel), onde, como pode ser observado na Figura 4, manchas escuras indicam a presença de uma proteína. Um dos maiores interesses deste tipo de técnica é detectar diferentes expressões protéicas.

Para isto ser feito, é necessário algum tipo de medida de similaridade robusta e precisa entre as manchas. Devido à complexidade biológica, física e química do processo, a localização da mesma mancha protéica pode diferir de maneira tanto global quanto local, o que torna quase impossível o registro perfeito das imagens. Isso torna ainda mais fácil a detecção de diferenças entre duas imagens. Entretanto, este tipo de análise é de extremo interesse, já que traz informações muito pertinentes para biólogos.

A partir destas ideias, Xin e Zhu (2009) propuseram um método simples e preciso para medir a similaridade entre manchas baseado em múltiplas informações. Este método proposto tem inspiração no princípio de atração e explora simultaneamente a distância entre as manchas, a intensidade das manchas e a informação de padrão da mancha para, de forma precisa e automática, relacionar manchas em duas imagens bidimensionais.

O princípio de atração da lei universal de gravidade da física é utilizado, este princípio diz que cada ponto de massa atrai cada outro ponto de massa por um ponto de força ao longo da linha que intersecta os dois pontos; e que a força é proporcional ao produto das duas massas e inversamente proporcional ao quadrado da distância entre os pontos de massa. O método de similaridade de manchas proposto é baseado nestas ideias encontradas no princípio de atração.

Suponha que I_r e I_f são duas imagens bidimensionais de entrada, representando, respectivamente, as imagens de referência e de flutuação; suponha que são dados dois conjuntos de manchas bidimensionais $\phi_r = \{s_{ri} | i=1,2,..., N\}$ e $\phi_f = \{s_{fj} | j=1,2,..., M\}$, onde N e M representam o número de manchas em I_r e I_f, respectivamente. Seja c um operador que dá as coordenadas (x, y) para o centróide da mancha s, ou seja, $c(s_{ri}) = (x_{ri}, y_{ri})$, são as coordenadas do centróide para a i-ésima mancha s_{ri} da imagem de referência,

e $c(s_{fj}) = (x_{fj}, y_{fj})$ as coordenadas do centróide para a *j*-ésima mancha s_{fj} na imagem de flutuação. De mesmo modo, $g(c(s_{ri}))$ e $g(c(s_{fj}))$ denotam as intensidades dos centróides (chamados de níveis cinza) em I_r e c, respectivamente.

Sabendo disto, podemos achar a correspondência entre as manchas nas imagens bidimensionais de referência e de flutuação. Assuma que cada mancha s_{fj} em I_f atrai cada mancha s_{ri} em I_r por um ponto de força **f** ao longo da linha entre as duas manchas da seguinte maneira

$$\mathbf{f}(s_{ri},\ s_{fj}) = \frac{K(s_{ri},\ s_{fj})}{D(s_{ri},\ s_{fj})}\mathbf{r}, \tag{3.1}$$

onde $K(s_{ri}, s_{fj})$ é um critério de similaridade, $D(s_{ri}, s_{fj})$ é a função de distância e **r** é o vetor unitário de s_{fj} a s_{ri}.

I_f se deformará de acordo com a força **f** que então poderá ser escrita da forma

$$\mathbf{f}(s_{ri},\ s_{fj}) = \frac{w_g w_p K_g K_p}{w_d D_d + \lambda}\ .\frac{\mathbf{d}}{\|\mathbf{d}\|}. \tag{3.2}$$

Os parâmetros em (3.2) são descritos a seguir:

O parâmetro K_g representa a intensidade da similaridade (nível cinza) e é definido por

$$K_g = \frac{\bar{g}(c(s_{ri}))\bar{g}\left(c(s_{fj})\right)}{\left(\left|\bar{g}(c(s_{ri})) - \bar{g}\left(c(s_{fj})\right)\right| + 1\right)\max\left(\bar{g}(c(s_{ri})),\ \bar{g}\left(c(s_{fj})\right)\right)}, \tag{3.3}$$

com $\bar{g}(c(s_{ri})) = 255 - g(c(s_{ri}))$ e $\bar{g}(c(s_{fj})) = 255 - g(c(s_{fj}))$.

O parâmetro K_p mede o padrão de similaridade dado por

$$K_p = \exp\left(-\ \text{diff}\ (P_{fj},\ P_{ri})\right), \tag{3.4}$$

onde P_{ri} e P_{fj} codificam o padrão de informação das manchas em I_r e I_f, respectivamente. O termo diff(P_{fj}, P_{ri}) expressa a diferença em padrão entre a imagem de flutuação e a imagem de referência. Quando a diferença é zero, K_p é máximo. Quanto maior a diferença, menor será o parâmetro de padrão de similaridade K_p (isto é, os dois padrões são dissimilares). Isso nos leva a definir K_p dentro da vizinhança circular Ω das manchas s_{ri} e s_{fj} como

$$K_p = \exp\left(-\frac{\left\|\sum_{n\neq j\in\Omega}\boldsymbol{v}_{fj,fn} - \sum_{m\neq i\in\Omega}\boldsymbol{v}_{ri,rm}\right\|}{\delta}\right)$$
$$\times\ \exp\left(-\frac{1}{2\pi}arctg\frac{\sum_{n\neq j\in\Omega}y_{fj,fn} - \sum_{m\neq i\in\Omega}y_{ri,rm}}{\sum_{n\neq j\in\Omega}x_{fj,fn} - \sum_{m\neq i\in\Omega}x_{ri,rm}}\right), \tag{3.5}$$

na qual δ denota o raio de vizinhança das manchas s_{ri} e s_{fj}, $\mathbf{v}_{ri,rm}$ o vetor de deslocamento da mancha s_{fj} para a mancha s_{fn} (que pode ser observado nas Figuras 6A e 6B que é apresentada a seguir). Em (3.5), o padrão de informação é modelado como a soma dos vetores de deslocamento, e o padrão de similaridade é modelado como um número real igual ao produto do módulo (com o fator de escala δ) e fase do vetor residual (com fator de escala 2π) resultante da subtração entre a soma de $\mathbf{v}_{fj,fn}$ na imagem de flutuação e a soma de $\mathbf{v}_{ri,rm}$ na imagem de referência.

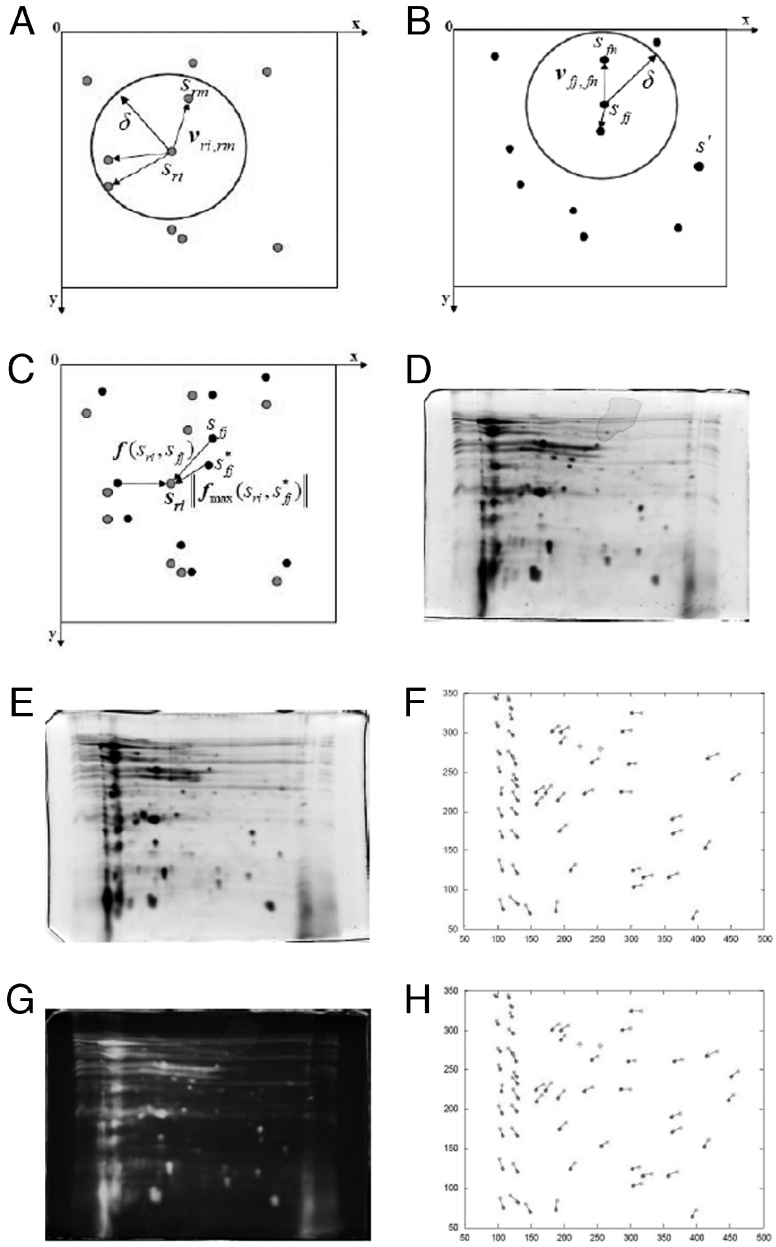


Figura 6: Princípio da análise de similaridade de manchas e alguns resultados. Dados da análise de O'Farrell (1975) sobre as proteínas *Escherichia coli*. (A) informação padrão das manchas em I_r. (B) informação padrão das manchas em I_j. (C) correspondência de s_{ri} e s_{fj} através do valor máximo da força atrativa f. (D) imagem de referência original. (E) imagem de flutuação original. (F) campo de vetores apontando de uma mancha para a outra (ou seja, se deslocando de uma macha para outra; da imagem de referência para a imagem de flutuação). (G) imagem da diferença, ilustrando as manchas que corresponderam nas duas imagens. (H) campo de vetores representando os pares de manchas correspondidas mostradas em (G).

A medida de distância D_d é obtida de

$$D_d = \frac{\|\mathbf{d}\|^2}{\delta^2}, \tag{3.6}$$

onde **d** designa o vetor de distância euclidiana entra as duas manchas s_{ri} e s_{fj}.

Os coeficientes de ponderação w_d, w_g e w_p têm como fim ajustar a força da distância, intensidade e padrão de informação. O coeficiente $\lambda \in (0,1]$ previne situações instáveis quando $\|\mathbf{d}\|^2$ está próximo de zero.

Para cada mancha $s_{ri} \in \Phi_r$ é computada a força atrativa entre s_{ri} e cada mancha $s_{fi} \in \Phi_f$. Se s_{fj}^* resulta no valor máximo de $\|\mathbf{f}(s_{ri}, s_{fj})\|$, então s_{ri} e s_{fj}^* irão ser consideradas manchas correspondentes, como pode ser visto na Figura 6C. Obtemos, desta maneira, dois conjuntos correspondentes de manchas $\Phi_r^* = \{s_{ri} | i = 1, 2, ..., q\}$ e $\Phi_f^* = \{s_{fj}^* | i = 1, 2, ..., q\}$ em I_r e I_f, respectivamente, onde q representa o número de manchas combinadas. Outras manchas sem correspondência (como a mancha *s'* mostrada na Figura 6B) podem ser separadas em seguida para uma análise posterior.

Apesar do enfoque na área de biociências com a utilização de imagens de eletroforese bidimensional dado ao método descrito neste capítulo, este é aplicável em qualquer uma das áreas em que são apresentadas duas imagens bidimensionais com o objetivo de encontrarmos similaridades entre eles.

4. Redução de Dimensões e Análise de Cluster

4.1 Redução de Dimensões

Uma alternativa bastante razoável ao nos depararmos com dados com altas dimensões consiste em encontrarmos uma maneira de reduzirmos a dimensionalidade destes dados, para então realizarmos alguma análise estatística. Para tal, existem dois tipos de técnicas utilizadas: transformação das características (*feature transformation*) e seleção das características (*feature selection*).

Técnicas de transformação das características têm como objetivo reduzir um conjunto de dados em menos dimensões através da combinação dos atributos originais. Este tipo de técnica é muito eficiente em encobrir estruturas latentes em conjuntos de dados. Entretanto, são menos eficientes quando há uma quantidade elevada de atributos irrelevantes "escondidos", visto que preservam a distância relativa entre objetos. Além disso, as novas características serão formadas pelas combinações das originais e isso pode causar certo transtorno na hora de interpretá-las.

Técnicas de seleção das características selecionam apenas as dimensões mais relevantes de um conjunto de dados. No caso particular de uma possível análise de *cluster* posteriormente, uma limitação deste tipo de técnica ocorre quando os grupos são formados em diferentes subespaços. Este tipo de problema motivou a criação de algoritmos que usam ideias das técnicas de transformação das características e posteriormente selecionam apenas os subespaços relevantes para cada *cluster* separadamente.

4.1.1 Transformação das Características

Transformações das características são muito utilizadas em dados com altas dimensões. Estes métodos incluem técnicas como análise de componentes principais e decomposição em valores singulares. As transformações geralmente preservam as distâncias relativas originais dos objetos. Desta maneira, o conjunto de dados é resumido através da combinação linear dos atributos, mostrando estruturas latentes.

Geralmente é utilizado antes de alguma outra análise dos dados, permitindo o uso das novas características criadas.

Apesar de na maioria das vezes ser útil, este tipo de técnica não deixa de levar nenhum dos atributos originais em consideração. Pelo contrário, a informação das dimensões irrelevantes é preservada, fazendo com que estas técnicas não sejam eficientes para revelar *clusters*, por exemplo, quando há um grande número de atributos irrelevantes.

Outra desvantagem de se usar combinações dos atributos é a dificuldade de interpretação. Por causa disto, transformações das características são mais adequadas para conjuntos de dados que apresentem a maioria das dimensões sendo relevantes para uma análise posterior, mas muitas delas sendo redundantes ou altamente correlacionadas.

4.1.2 Seleção das Características

Técnicas de seleção das características têm como objetivo descobrir os atributos de um conjunto de dados que são mais relevantes para uma análise posterior. São técnicas poderosas e altamente utilizadas para a redução da dimensionalidade para níveis mais manejáveis.

A seleção das características consiste em procurar através de vários subconjuntos de características e avaliar cada um destes subconjuntos utilizando algum critério, como os de Pena et al (2001) ou de Yu e Liu (2003).

As estratégias mais comuns de procura são as buscas sequenciais através do espaço de características podendo ser feita tanto para frente (*forward*) quanto para trás (*backward*).

4.2 Análise de Cluster em Imagens

Uma boa alternativa quando trabalhamos com dados com altas dimensões e necessitamos de alguma forma agrupar este tipo de dados é a Análise de C*luster* em imagens, que pode ser realizada tanto com os dados brutos ou padronizados quanto com os dados após a realização de uma redução de dimensões. Neste trabalho serão apresentadas apenas algumas técnicas de *cluster* (agrupamento) baseadas em *grids*.

Em sua forma mais básica, agrupamento baseado em *grids* é relativamente simples:

a) Divida o espaço de amplitude dos dados em (hiper) células retangulares, por exemplo, particionando todo o intervalo de dados de cada dimensão em células de tamanhos iguais. Na Figura 7, é possível visualizar um exemplo bidimensional deste tipo de *grid*:

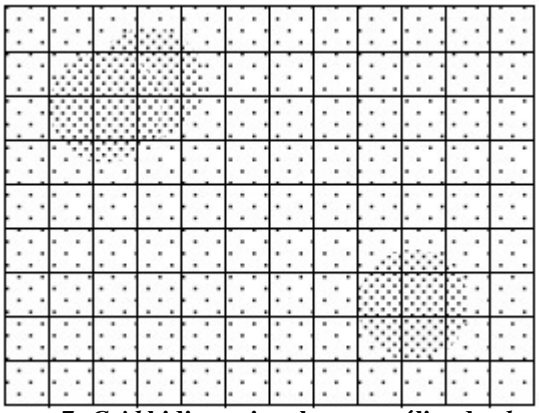


Figura 7: *Grid* bidimensional para análise de *cluster*

b) Descarte as células menos densas. Isto resulta em uma definição de densidade baseada em *clusters*, isto é, regiões altamente densas representam *clusters*, enquanto que regiões menos densas representam ruído. Geralmente esta é uma boa suposição, entretanto, quando os *clusters* apresentam densidades muito diferentes, podemos ter problemas com este tipo de abordagem.
c) Combine células adjacentes com alta densidade para formar *clusters*. Se as regiões mais densas são adjacentes, então elas podem ser juntadas para formar um único *cluster*.

Existem algumas preocupações óbvias com relação aos métodos de agrupamento baseados em *grids*. Como os *grids* criados são quadrados ou retangulares, em muitos casos eles não irão se encaixar perfeitamente no formato do *cluster*. Para resolver este tipo de problema, pode-se aumentar o número de *grids*, de forma a deixá-los menores e assim aproximar-se da forma real do *cluster*. Entretanto, há um preço a se pagar por isto, e neste caso, será o aumento do tempo de trabalho computacional. Outro problema que pode surgir a partir deste aumento de *grids* é a aparição de "buracos" dentro dos *clusters*, devido ao tamanho muito reduzido que estes irão assumir, ainda mais se estivermos trabalhando com um tamanho de amostra não tão grande (resultando em uma menor quantidade de pontos).

Além do alerta descrito anteriormente a respeito deste tipo de técnica, existem sérios problemas quando a dimensionalidade dos dados aumenta muito. Para se perceber isto basta imaginar o caso em que cada dimensão é dividida em apenas 2 *grids*. Sendo *d* o número de dimensões presentes, teremos 2^d células. Dados com 30 dimensões iriam, então, utilizar no mínimo 1 bilhão de células. Inclusive para grandes conjuntos de dados, a maioria das células ficaria vazia.

Outro problema é encontrar *clusters* entre as dimensões. Para compreender isto, imagine que cada ponto em um dos *clusters* da Figura 7 é aumentado com muitas

variáveis adicionais e que os valores atribuídos aos pontos nestas dimensões são uniformemente e aleatoriamente distribuídos. Então quase todos os pontos irão "cair" em células separadas no novo espaço alto dimensional. Assim, percebemos que grupos de pontos podem estar presentes em apenas alguns subespaços do modelo com altas dimensões.

A seguir, serão mostradas quatro técnicas mais comuns de agrupamento de dados baseado em *grids* que estão disponíveis: CLIQUE, MAFIA, DENCLUE e OptiGrid.

4.2.1 CLIQUE

CLIQUE, como pode ser visto em Agrawal et al (1998), é um algoritmo para agrupamento que tenta lidar com os problemas citados anteriormente e que tem a sua abordagem baseada na seguinte observação: uma região que é densa em um determinado subespaço deve criar regiões densas quando projetado em dimensões menores. Um exemplo disto pode ser obtido através da Figura 8:

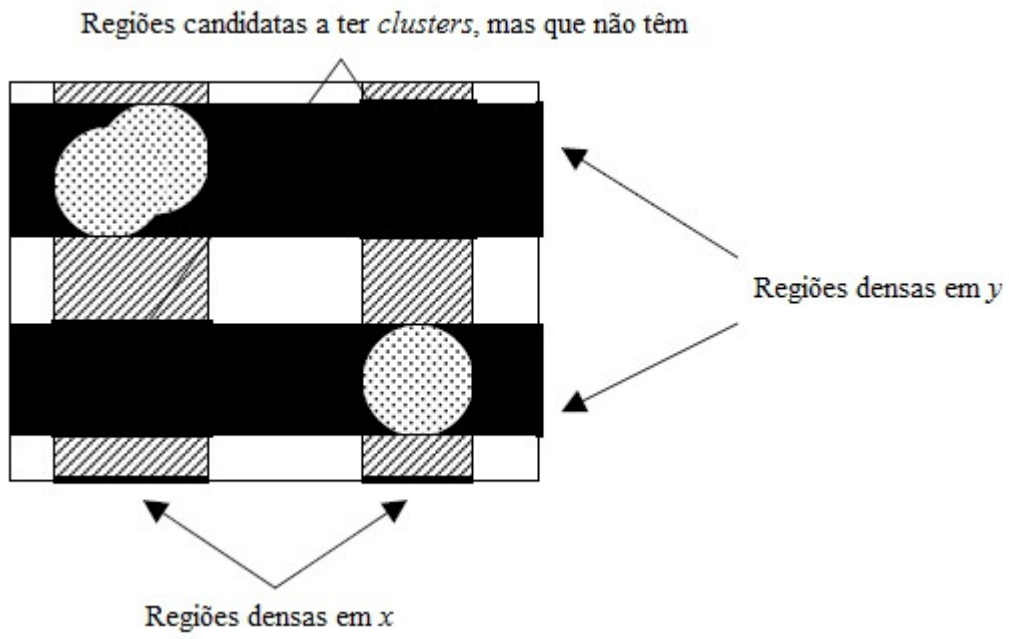


Figura 8: Ilustração da ideia de que alta densidade em altas dimensões implica em alta densidade em baixas dimensões, mas não vice-versa

Na Figura 8, se examinamos as coordenadas *x* (horizontal) e *y* (vertical) de cada um dos pontos, nós vemos a presença de regiões densas nas distribuições unidimensionais, que refletem a existência de regiões densas bidimensionais. As colunas pretas horizontais e as colunas hachuradas verticais indicam as projeções dos *clusters* nos eixos vertical e horizontal, respectivamente. A Figura 8 também nos mostra que alta densidade em baixas dimensões pode apenas sugerir possíveis localizações de *clusters* em dimensões mais altas, já que as regiões das dimensões mais altas são

formadas por intersecções de duas regiões densas de menor dimensão, e isso nem sempre corresponde a um *cluster*.

Entretanto, ao iniciar com intervalos unidimensionais densos, é possível encontrar potenciais intervalos bidimensionais densos e, ao inspecionar estes, encontrar os verdadeiros intervalos. Este procedimento pode ser estendido para se encontrar regiões densas em qualquer subespaço de forma muito mais eficiente do que formar células correspondentes a todos os possíveis subespaços de dimensões e então procurar pelas unidades densas nestas células. Porém, o CLIQUE ainda necessita de habilidade de descobrir regiões densas para reduzir os subespaços investigados. Além disso, sua complexidade computacional, mesmo sendo linear no número de pontos de dados, é não-linear no número de dimensões.

4.2.2 MAFIA

MAFIA (Merging Adaptative Finite Intervals And is more than a clique), como pode ser visto em Harasha et al (1999), é um refinamento da abordagem CLIQUE. Esta técnica encontra melhores *clusters* e alcança maior eficiência ao utilizar *grids* não-uniformes, conforme pode ser visto na Figura 9:

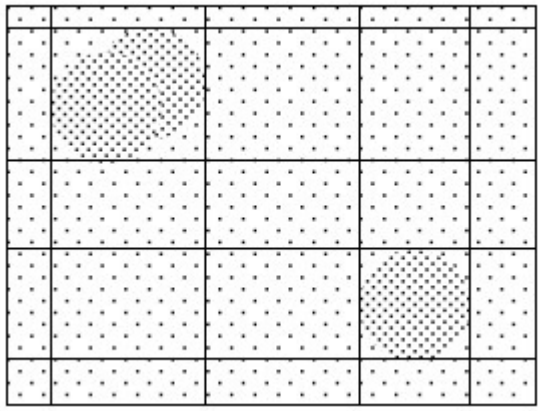


Figura 9: *Grid* utilizado na abordagem MAFIA

Mais especificamente, ao invés de arbitrariamente dividir os dados em intervalos pré-determinados e pré-espaçados, MAFIA particiona cada dimensão usando um número variável de intervalos que se adaptam e que melhor refletem a distribuição dos dados naquela dimensão.

Para melhor ilustrar, CLIQUE utiliza um *grid* similar ao encontrado na Figura 7, e assim, quebra cada um dos intervalos densos unidimensionais em um determinado número de subintervalos, incluindo alguns que apresentam menor densidade, já que ele inclui parte da região não-densa.

Conceitualmente, MAFIA começa com um grande número de pequenos intervalos para cada dimensão e então combina intervalos adjacentes de densidades similares para terminar com um menor número de intervalos maiores. Assim, um *grid* utilizando a abordagem MAFIA é bem representado pelo *grid* visto na Figura 9.

4.2.3 DENCLUE

Uma diferente abordagem do mesmo problema é fornecida pelo método DENCLUE (DENsity CLUstEring), em estudo feito por Hinnenburg e Keim (1998). O DENCLUE é um agrupamento por densidade que leva em conta uma abordagem mais formal ao método baseado pelas densidades para modelar a densidade total de um conjunto de pontos como a soma de funções de "influência" associadas a cada ponto. A função de densidade total terá picos locais, como por exemplo, local de densidade máxima e, a partir destes picos locais, chegaremos aos *clusters* de forma simples.

Especificamente, para cada ponto de dado, encontraremos o pico mais próximo associado a ele. O conjunto de todos os pontos associados a um particular pico (chamado de densidade atratora local) se torna um *cluster*. Entretanto, se a densidade em um pico local é muito baixa, então os pontos no *cluster* associado a este pico são considerados ruídos e então descartados. Além disso, se um pico local puder ser conectado a um segundo pico local por um trajeto de pontos e a densidade em cada ponto deste trajeto é maior do que um limiar mínimo, então os *clusters* associados a estes pontos se fundem. Com isso, *clusters* com qualquer formato podem ser encontrados. Na Figura 10, temos um exemplo de função de densidade total utilizado para a formação de *clusters* na abordagem DENCLUE:

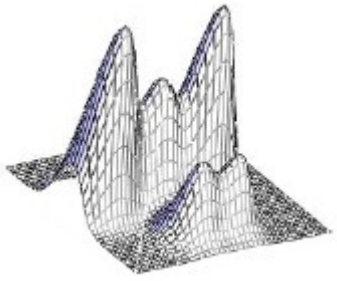


Figura 10: Exemplo de função de densidade total

4.2.4 OptiGrid

Apesar das características atraentes do DENCLUE em espaços com poucas dimensões, esta abordagem não lida tão bem com os dados a medida que a dimensão

aumenta ou quando há presença de ruído. Por isso, os mesmos criadores do DENCLUE, Hinnenburg e Keim (1999) desenvolveram o OptiGrid.

O algoritmo descrito pelos autores segue seis passos, que serão resumidamente citados a seguir:

1) Para cada dimensão:
    a) Faça um histograma dos dados. Note que isto é equivalente a contar os pontos em um *grid* uniforme unidimensional impostos nos valores;
    b) Determine o nível de ruído. Isto pode ser feito através de uma inspeção manual do histograma, se a dimensionalidade não é muito alta. Em caso contrário, este processo necessita ser automatizado;
    c) Encontre o ponto máximo mais à direita e mais à esquerda e o q-1 máximo entre eles (onde q é o número de partições dos dados que nós procuramos e todas estas partições podem estar em uma dimensão);
    d) Escolha q, o mínimo entre os máximos encontrado no passo anterior. Estes pontos representam localizações para possíveis cortes, isto é, localizações onde o hiperplano pode ser posicionado para particionar os dados. Escolher células pouco densas minimiza a chance de cortes através de um *cluster*;
    e) Dê um escore para cada corte potencial, por exemplo, pela sua densidade.
2) De todas as dimensões, selecione os melhores q cortes, isto é, aqueles cortes com menores densidades.
3) Utilizando estes cortes, crie um *grid* que particiona os dados.
4) Encontre as células mais densas e adicione elas para a lista de *clusters*.
5) Refine a lista de *clusters*.
6) Repita as etapas 1-5 utilizando cada *cluster*.

Em resumo, a abordagem OptiGrid é semelhante à MAFIA no sentido de que cria um *grid* utilizando uma partição dependente dos dados. Entretanto, o OptiGrid não se preocupa em localizar o melhor subespaço para usar esta partição. Ele localiza potenciais *clusters* entre o conjunto de células formadas através do seu plano de cortes. Com relação à eficiência, este tipo de abordagem é muito melhor.

Porém, também existem alguns problemas neste tipo de técnica como, por exemplo, o fato do número de cortes necessários ser bastante vago. Com o objetivo de solucionar este tipo de problema, novos tipos de abordagens estão sendo desenvolvidas, como o PDDP (Power-Delay-Direction Profile).

5. Medidas de Distância entre Sequências de DNA

Em muitos casos, quando temos dados moleculares, temos a necessidade de realizar algum tipo de análise para comparar diferentes indivíduos. No caso de sequências de DNA, algum tipo de medida de distância entre duas sequências poderá nos dar uma boa ideia da similaridade entre dois indivíduos e inclusive ajudar a detectar graus de parentesco. A seguir, serão abordados três diferentes tipos de medidas de distância entre sequencias de DNA, que vem sendo mais utilizadas: distâncias baseadas em modelos, distância log determinante e as distâncias de Hamming.

5.1 Distâncias Baseadas em Modelos

Considere duas sequências de DNA semelhantes, que divergiram há *t* unidades de tempo. A probabilidade de que duas bases de nucleotídeos em sítios correspondentes sejam idênticas no tempo *t* é dada por *I(t)*. Em genética, existe a chamada taxa de substituição, que neste caso será considerada constante e igual a α. Então, para um dado sítio, a probabilidade de que dois nucleotídeos homólogos permaneçam idênticos em *t + 1* quando eles são idênticos no tempo *t* é igual a $[(1-3\alpha)^2 + 3\alpha^2]I(t)$.

A probabilidade mostrada acima envolve dois eventos mutuamente excludentes: ambas as bases mudam para duas bases idênticas com probabilidade $3\alpha^2$ e, ambos os nucleotídeos permanecem inalterados com probabilidade $(1-3\alpha)^2$. E, a probabilidade de que dois sítios de nucleotídeos tornem-se idênticos no tempo no tempo *t+1* quando eles são diferentes em *t* é igual a $[2\alpha(1-3\alpha) + 2\alpha^2][1-I(t)]$, o que também consiste em dois eventos mutuamente excludentes: uma mudança ocorre em um dos sítios e o outro sítio permanece inalterado com probabilidade $2\alpha(1-3\alpha)$ e, ambas bases mudam simultaneamente para outras duas bases idênticas com probabilidade $2\alpha^2$. Chegamos assim na seguinte probabilidade:

$$I(t+1) = [(1-3\alpha)^2 + 3\alpha^2]I(t) + [2\alpha(1-3\alpha) + 2\alpha^2][1-I(t)]. \qquad (5.1)$$

Com a condição inicial em (5.1) de que $I(0) = 1$ temos que

$$I(t) = \frac{1}{4}[1 + 3(1 - 8\alpha + 16\alpha^2)^t]. \quad (5.2)$$

Entretanto, α é, em geral, muito pequeno, fazendo com que os termos α^2 se tornem desprezíveis em (5.2), resultando finalmente em

$$I(t) \approx \frac{1}{4}[1 + 3(1 - 8\alpha)^t] \approx 1 - \frac{3}{4}(1 - e^{-8\alpha t}). \quad (5.3)$$

Uma abordagem feita por Gojobori et al (1990) diz que a distância evolutiva *K*, que representa o número médio de substituições de nucleotídeos acumulados por sítio no tempo *t*, pode ser calculada da seguinte maneira:

$$K = 2 \times 3\alpha t = -\frac{3}{4}\ln\left(1 - \frac{4}{3}F_D\right), \quad (5.4)$$

onde $F_D = 1 - I(t)$. O erro padrão de *K*, é dado por

$$\sigma_K = \frac{\sqrt{\frac{1}{n}F_D 1(1-F_D)}}{1-\frac{4}{3}F_D}, \quad (5.5)$$

onde *n* é o número total de sítios comparados.

Quando comparamos várias sequências de dois grupos (geralmente chamados de *espécies*), devemos considerar a possibilidade de que quaisquer duas sequências podem ser descendentes de diferentes sequências na população ancestral. Seja *S* uma medida de similaridade intra-espécie, que depende do tamanho da população e da taxa de mutação. Com a população ancestral em equilíbrio, *S* é esperado permanecer constante com o tempo a algum valor $\hat{S}$. Levando em consideração a variação intra-espécies, chega-se em

$$K_W = \frac{3}{4}\ln\left(\frac{4\hat{S}-1}{4I-1}\right), \quad (5.6)$$

que mede a distância entre as populações recente e ancestral e é chamada de distância de Jukes-Cantor. Conforme Weir e Basten (1990), *S* é estimada a partir da variação observada em cada uma das duas espécies recentes. Suponha que é feita uma

amostragem na população *i*, onde n_i sequências são amostradas. Define-se como $r_{ijj'}$ a proporção de bases homólogas que são iguais nas sequências *j* e *j'*. A similaridade amostral dentro da população *i* é dada por

$$\tilde{S}_i = \frac{1}{n_i(n_i-1)} \sum_{i=1}^{n_i} \sum_{j \neq j'} r_{ijj'} \,. \tag{5.7}$$

Se $s_{jj'}$ é o número de bases idênticas entre a sequência *j* na população 1 sequência *j'* na população 2, a similaridade entre populações é estimada por

$$\tilde{I} = \frac{1}{n_1 n_2} \sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} s_{jj'} \,. \tag{5.8}$$

5.2 Distância Log Determinante

Com o objetivo de calcular distâncias entre sequências de DNA com diferentes composições de nucleotídeos, é introduzida por Lockhart et al (1994) a medida de distância Log Determinante (LogDet). Ela é baseada na matriz de divergência F_{xy}, que é encontrada da seguinte maneira: para sequências *x* e *y*, o elemento (*i*, *j*) da matriz é dado pela proporção de sítios em que *x* está no nucleotídeo *i* enquanto *y* é *j*. Com isso, a soma de todos os elementos da matriz é igual a 1. A distância LogDet entre *x* e *y* é definida como

$$d_{xy} \equiv \ln\left(\det F_{xy}\right), \tag{5.9}$$

onde det é o determinante da matriz F_{xy}. Para assinalar uma distância de zero entre uma sequência e ela mesma, a fórmula (5.9) é modificada, tornando-se a seguinte:

$$d'_{xy} \equiv -\frac{1}{4} \ln\left(\frac{\det F_{xy}}{\sqrt{(\det F_{xx})(\det F_{yy})}}\right). \tag{5.10}$$

Note que quando as freqüências das bases são iguais, $\det F_{xx} = \det F_{yy} = \left(\frac{1}{4}\right)^4$ e o valor esperado da Distância LogDet é então chamado de número médio de substituições por sítio.

5.3 Distância de Hamming

A distância de Hamming, conforme pode ser visto em Seillier-Moiseiwitsch et al (1994), é muito utilizada como análise descritiva. Definimos $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{iK})'$ como sendo o vetor que representa a sequência de DNA i de tamanho K. X_{iK} é então o nucleotídeo presente na posição k. Considere as duas sequências $\mathbf{X}_i$ e $\mathbf{X}_{i'}$ de mesmo tamanho. A distância $H_{ii'}$ de Hamming é dada por

$$H_{ii'} = \frac{1}{K} \sum_{k=1}^{K} \delta(X_{ik} \neq X_{i'k}), \tag{5.11}$$

onde δ denota a função indicadora, que é igual a 1 caso o evento $(X_{ik} \neq X_{i'k})$ ocorra e 0 em caso contrário. Ou seja, δ representa o número de posições onde $\mathbf{X}_i$ e $\mathbf{X}_{i'}$ diferem. Quanto menor o valor da distância de Hamming, maior é a semelhança entre as duas sequências de DNA. Enquanto esta distância só deve ser tratada como uma estatística descritiva, em muitas situações ela nos dá uma estimativa razoável da atual distância.

A distância de Hamming nos permite reescrever a distância de Jukes-Cantor, vista em (5.6) da seguinte forma:

$$K_W = -\frac{3}{4} \ln\left(1 - \frac{4}{3} H_{ii'}\right), \tag{5.12}$$

onde $H_{ii'}$ representa a distância de Hamming calculada para a sequência de DNA. A mesma interpretação dada à distância de Hamming pode ser dada à distância de Jukes-Cantor.

6. Software ImageMaster™ 2D Platinum

Como visto no capítulo 3, a análise de similaridade além de não ser simples manualmente, leva um tempo razoável para ser completamente executada. No entanto, existem alternativas computacionais para este tipo de análise, tornando a tarefa de comparação de imagens bidimensionais muito mais fácil e rápida. A partir destas ideias, foi proposto o *software* ImageMaster™ 2D Platinum, pelo Instituto Suíço de Bioinformática (Swiss Institute of Bioinformatics) em colaboração com a GeneBio™ e a Amersham Biosciences. Atualmente é fabricado pela GE Healthcare e sua versão atual é a 7.0.

O objetivo principal deste programa é a análise automática de géis bidimensionais na área de biociências. Na Figura 11 apresentada abaixo, é possível visualizarmos a interface do *software*:

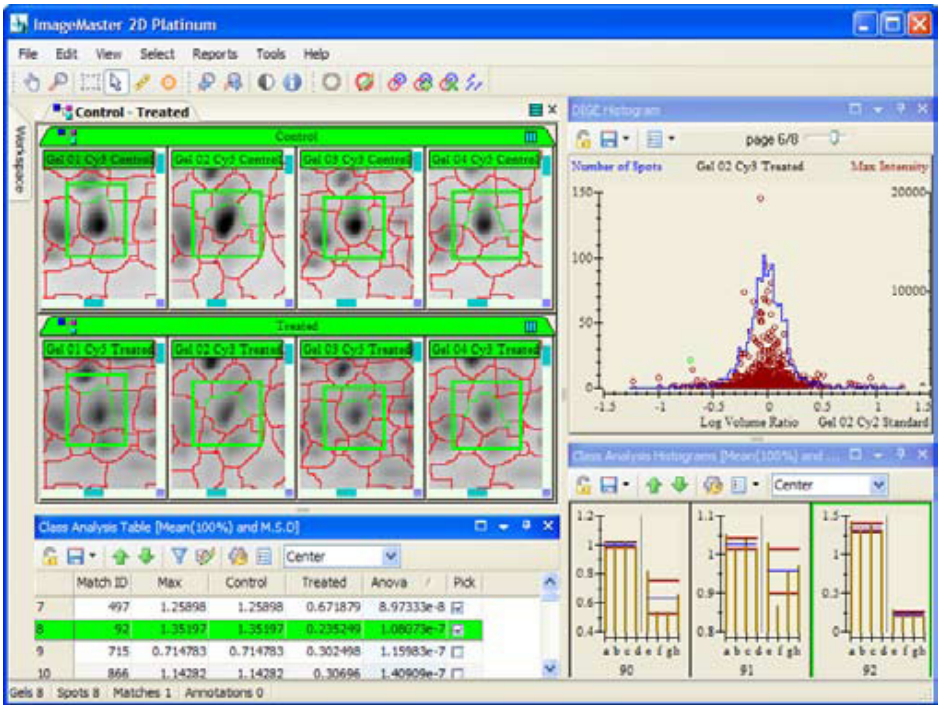


Figura 11: Interface de usuário do *software* ImageMaster™ 2D Platinum, ilustrando a análise de géis bidimensionais

Algumas das funcionalidades presentes nas versões mais recentes do programa incluem a visualização 3D de múltiplos géis por vez, a identificação direta de similaridades ou diferenças entre imagens bidimensionais e o alinhamento destas

imagens para a identificação visual nas expressões de proteínas, além de realizar uma filtragem do gel antes das análises para evitar a influência de ruídos indesejados.

O *software* permite uma fácil organização de géis e experimentos dentro do espaço de trabalho e a sua interface permite o acesso a vários tipos de dados. Após a análise realizada, as imagens podem ser coloridas para obter-se uma melhor visualização de combinações entre géis e cada uma das saídas pode ser comentada.

A detecção de manchas pode ser feita de maneira automática, semi-automática ou manual. Com a escolha da primeira opção, o usuário irá intervir o mínimo possível, deixando o próprio programa definir quais são as manchas mais apropriadas para posterior análise. São utilizados parâmetros robustos de detecção e parâmetros de saliência que tem como objetivo distinguir manchas de interesse de ruídos indesejáveis. O *software* identifica cada uma das manchas (que representa determinado tipo de proteína, por exemplo) com uma linha vermelha ou verde ao seu redor, conforme pode ser visto na Figura 12 a seguir. É então realizada uma identificação automática de posição, forma e tamanho das manchas. O usuário pode optar então pela divisão ou união de manchas propostas pelo programa, caso seja de seu interesse. Além disso, podem ser selecionadas áreas de interesse da imagem para a análise ao invés de se analisar o gel por inteiro.

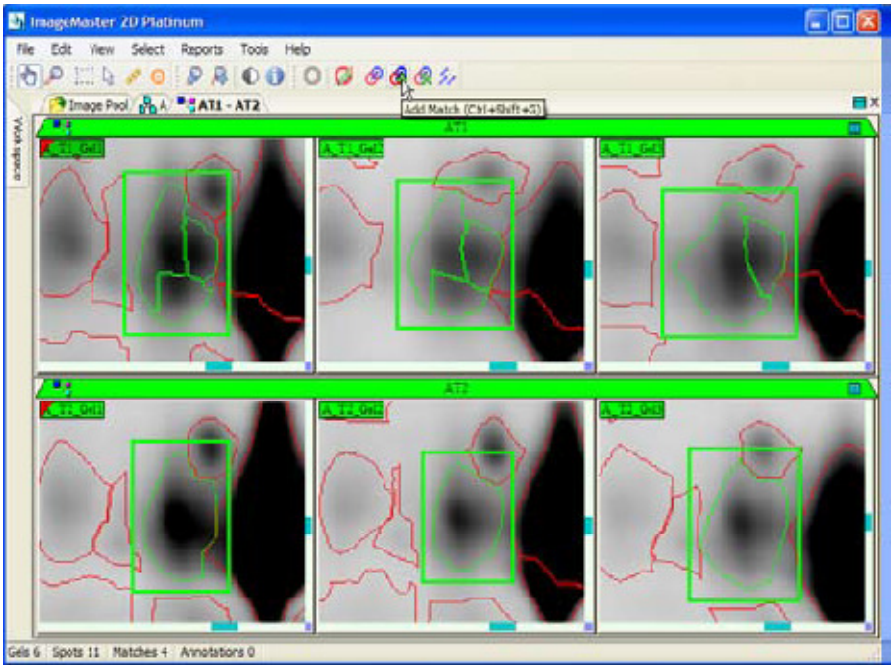


Figura 12: Janela de detecção de manchas do *software* ImageMaster™ 2D Platinum

Depois de detectadas as manchas, a análise de similaridade entre as imagens é realizada. No *software*, existem as opções de correspondência de manchas automática ou semi-automática. Desta vez, o usuário não precisa controlar nenhum parâmetro e a análise é feita levando-se em consideração a localização, a forma e a posição das

manchas. As comparações entre as imagens são sempre pareadas e é possível escolhermos uma imagem de referência e compararmos ela outras imagens de flutuação, resultando em ilimitadas correspondências entre manchas. Depois de realizada a análise podemos utilizar uma ferramenta de transparência para sobrepor um gel sobre o outro e identificarmos semelhanças e diferenças através de pinturas diferentes das manchas. Na Figura 13 temos um exemplo de como é feita a análise de similaridade entre os géis:

Figura 13: Análise de similaridade utilizando o *software* ImageMaster™ 2D Platinum

Além destas análises, o programa também realiza a chamada normalização dos géis, aonde as variações vindas do escaneamento das imagens ou a presença de diferentes tipos de intensidade de manchas são removidas, padronizando todas as imagens. São removidos também *outliers* e ruídos que possam interferir no resultado.

Diversos processos de filtragem podem ser realizados nas imagens, utilizando critérios baseados em tamanho ou forma das manchas e tem como objetivo a visualização e eliminação destas manchas do gel. As expressões de variação entre imagens ou classes de imagens podem ser investigadas utilizando-se um *ranking* automático das manchas. Os resultados então são apresentados em relatórios temporários ou permanentes.

Outra funcionalidade do programa é a criação de géis artificiais através dos géis originais. São utilizadas as médias de posição, formato e intensidade das manchas correspondentes para a formação do novo gel. Com isto, é possível se ter uma ideia do padrão protéico de determinada célula e é possível a visualização de diferentes

expressões de proteínas. Além disto, as manchas que não foram combinadas podem ser vistas também em um gel artificial se o usuário assim desejar.

Após cada ação realizada, relatórios são gerados para o usuário. Um exemplo são os relatórios sobre o grupo de manchas combinadas, sobre as classes de géis ou então anotações de categorias. O *software* também permite a edição destes relatórios e a geração de gráficos, como por exemplo, histogramas, para a melhor visualização dos resultados da análise. É possível também adicionar vários tipos de anotações nas imagens, para melhor compreensão das similaridades e diferenças entre géis, conforme pode ser visto na Figura 14:

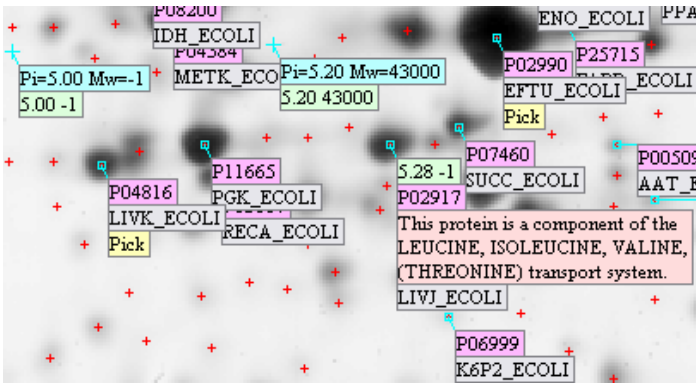


Figura 14: Anotações feitas em gel após análise de similaridade no *software* ImageMaster™ 2D Platinum

Análises estatísticas e gráficas também podem ser realizadas utilizando-se este programa. Gráficos de dispersão, por exemplo, podem ser usados para analisarmos similaridades ou variações experimentais nas imagens. Estatísticas descritivas, como medidas de tendência central ou de dispersão podem ser calculadas tanto nos relatórios para cada grupo de géis, quanto na geração de histogramas. Análise fatorial para identificação de manchas similares dentro de géis pode ser realizada antes ou depois da análise de similaridade de imagens. Além disto, existe a possibilidade de comparação de médias, através do teste *t* para duas amostras, *one-way* ANOVA e testes não paramétricos como os de Mann-Whitney ou de Wilcoxon. Por fim, o *software* também realiza o teste de Kolmogorov-Smirnov para verificar se a distribuição de duas amostras vem de uma mesma distribuição de probabilidade.

Outra característica interessante do *software* com relação à realização de análises estatísticas é a análise de *cluster* realizada para a identificação de conjuntos de imagens comuns. Os géis são classificados automaticamente baseados em suas semelhanças de manchas. Na Figura 15, na página seguinte, é possível termos uma visualização da análise de *cluster* sendo realizada:

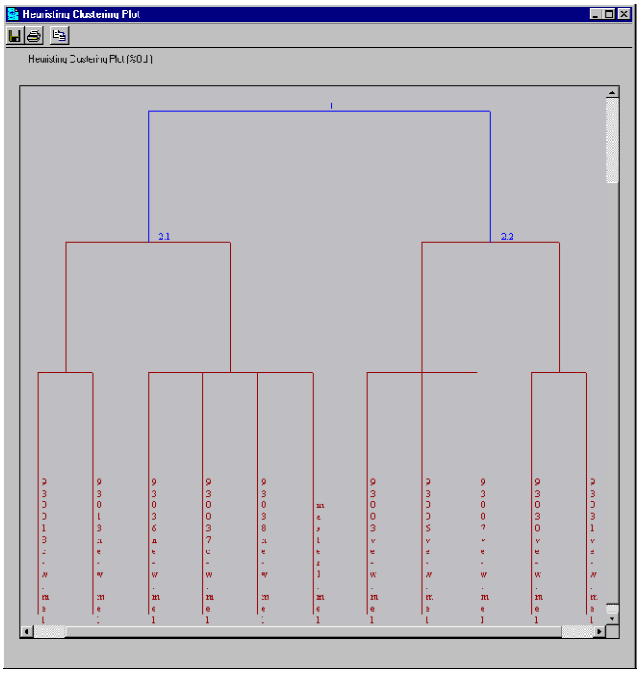


Figura 15: Análise de *cluster* realizada no *software* ImageMaster™ 2D Platinum

Todas as análises realizadas no programa podem ser exportadas para os formatos XML e Microsoft Office Excel e todos os relatórios gerados podem ser salvos para uma futura utilização. As imagens, os gráficos e as janelas de dados podem ser copiadas e inseridas em outros programas; todas as tabelas de dados são compatíveis com a plataforma Windows.

O *software*, entretanto, não é gratuito e sua aquisição pode ser feita diretamente no *site* da empresa fabricante GE Healthcare (http://www.gelifesciences.com). Uma licença básica da versão 7.0 custa atualmente R$ 30.780,00 (aproximadamente US$15000). Uma versão de avaliação de 14 dias está disponível gratuitamente no *site* do fabricante para que o usuário possa testar o programa. Como alternativa, existe a possibilidade da compra de uma licença da versão anterior do *software* (6.0) pelo preço de R$5.544,00 (aproximadamente US$2772).

7. Software PAST

Como visto no capítulo 5, existem diversas técnicas desenvolvidas com o objetivo de se calcular uma medida de distância entre sequências de DNA para vermos o quão distante duas sequências estão uma da outra. Um *software*, chamado PAST, desenvolvido para resolver diversos problemas estatísticos, tem uma seção própria para a análise de sequências genéticas. O presente capítulo tem como objetivo mostrar aos leitores esta seção citada, além de outra seção do *software* que aborda a análise descriminante de Hotelling.

O *software* PAST (PAlaeontological STatistics) foi desenvolvido pelo professor Øyvind Hammer da Universidade de Oslo em conjunto com o supervisor David A.T. Harper do museu geológico de Copenhagen e do programador P.D. Ryan. Pode ser obtido gratuitamente na página do professor Øyvind Hammer (http://folk.uio.no/ohammer/past/). É de simples uso e apresenta apenas uma planilha para escrevermos os dados e um menu onde escolhemos a análise a ser realizada com estes dados, conforme pode ser visto na Figura 17:

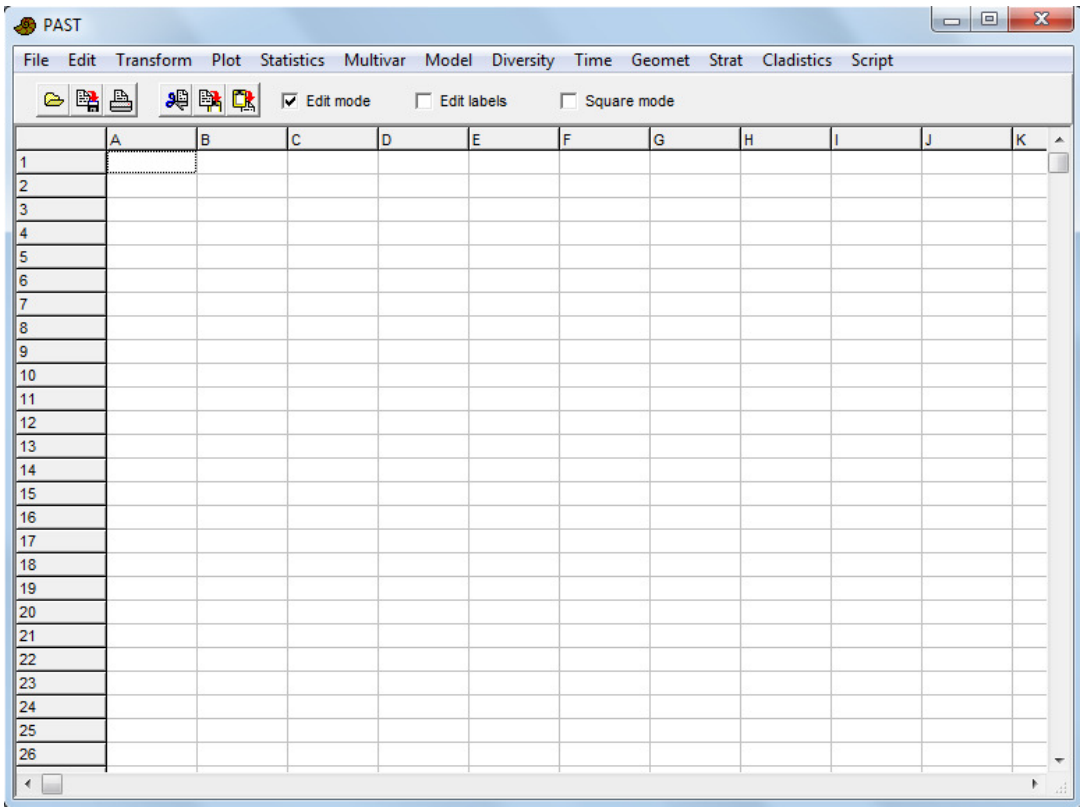
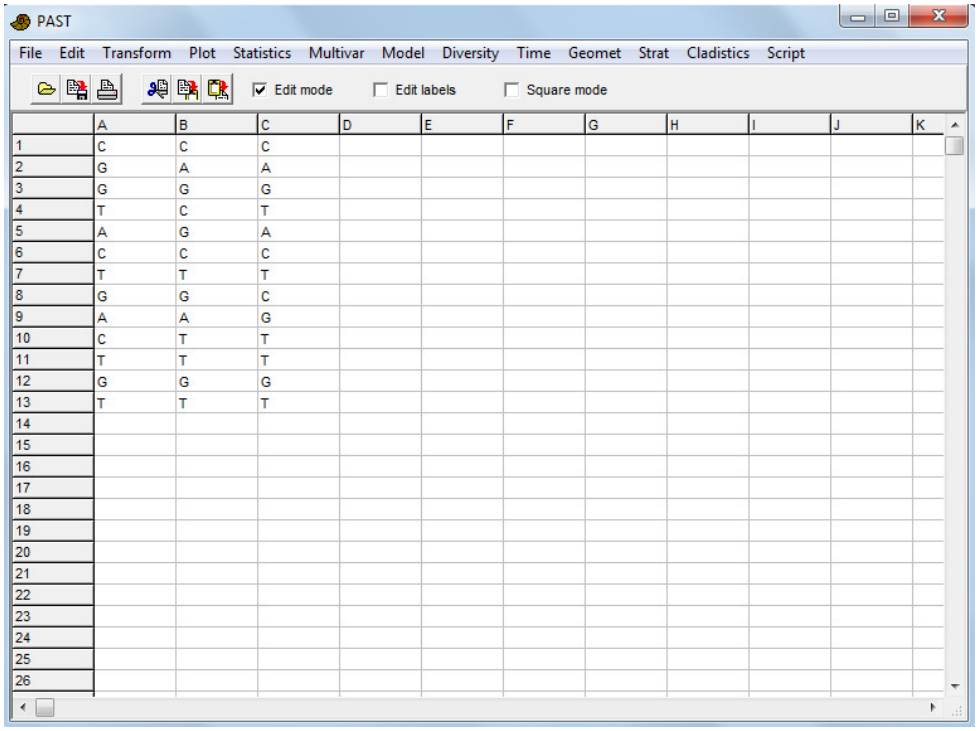


Figura 17: Interface gráfica do *software* PAST

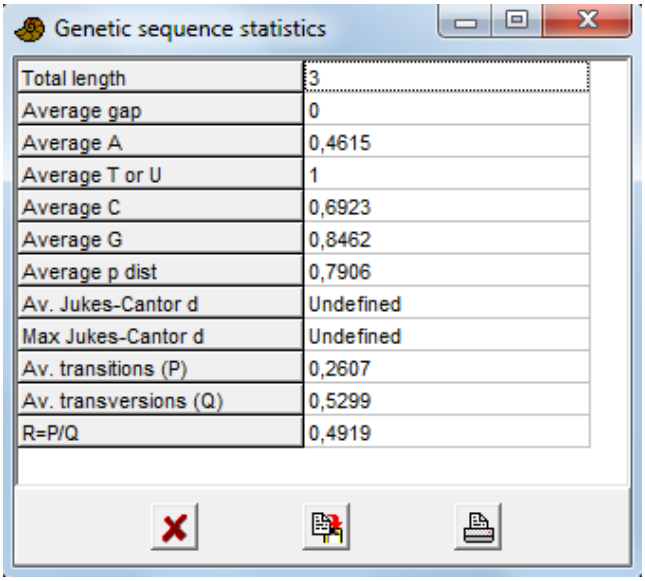
O PAST permite que o banco de dados seja importado apenas dos formatos *.dat* (próprio do *software*) e *.txt* (bloco de notas) e permite que os dados sejam salvos nos formatos *.dat* (PAST) e *.nex* (NEXUS) e *.xls* (Excel).

Uma das seções que irá nos interessar neste trabalho é a *Genetic sequence stat*, contida dentro do menu *Statistics*. Após preenchermos a planilha de dados com os nucleotídeos presentes nas sequências de DNA, onde cada coluna representa uma sequência, conforme pode ser visto na Figura 18, a seção mencionada acima irá fornecer uma análise estatística dos dados similar a apresentada na Figura 19:



	A	B	C
1	C	C	C
2	G	A	A
3	G	G	G
4	T	C	T
5	A	G	A
6	C	C	C
7	T	T	T
8	G	G	C
9	A	A	G
10	C	T	T
11	T	T	T
12	G	G	G
13	T	T	T

Figura 18: Disposição dos dados no *software* PAST para a realização da medida de distância entre sequências de DNA



Total length	3
Average gap	0
Average A	0,4615
Average T or U	1
Average C	0,6923
Average G	0,8462
Average p dist	0,7906
Av. Jukes-Cantor d	Undefined
Max Jukes-Cantor d	Undefined
Av. transitions (P)	0,2607
Av. transversions (Q)	0,5299
R=P/Q	0,4919

Figura 19: Exemplo de resultados da análise estatística de sequências de DNA utilizando o *software* PAST

Além de calcular a proporção média de cada um dos nucleotídeos, são calculadas as distâncias médias de Hamming (*Averege p dist*) e de Jukes-Cantor (*Av. Jukes-Cantor d*). No caso de apenas duas sequências de DNA serem comparadas, as distâncias médias representam simplesmente a distância bruta entre as duas sequências, conforme visto no capítulo 5. Os resultados podem ser impressos diretamente através do PAST.

Esta seção do *software* PAST será utilizada no próximo capítulo para confirmação dos cálculos realizados em um exemplo prático.

Outra seção do PAST que será utilizada no próximo capítulo é a *Discriminant/Hotelling*, contida dentro do menu *Multivar*. Este comando realiza o teste T^2 multivariado de Hotelling e tem como objetivo detectar diferenças significativas entre dois conjuntos de dados. O teste funciona da seguinte forma: dados dois conjuntos de dados, um eixo que maximiza a diferença entre os conjuntos é construída. Os dois conjuntos de dados são então plotados ao longo deste eixo através de um histograma.

Para a realização desta análise no *software*, os dois conjuntos de dados devem estar separados por cores diferentes nas linhas, conforme pode ser visto na Figura 20:

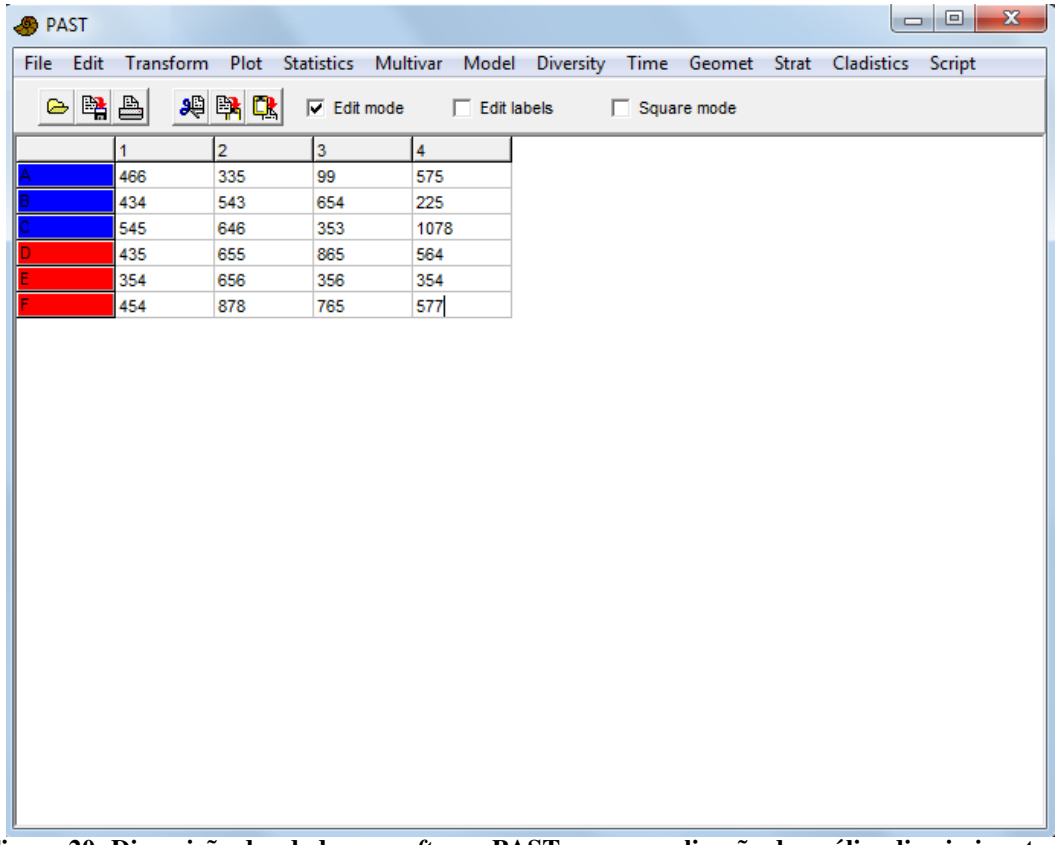


Figura 20: Disposição dos dados no *software* PAST para a realização da análise discriminante de Hotelling

Além disto, o número de colunas deve ser no máximo a metade do número de linhas para a análise ser realizada, ou seja, o número de caso deve ser no mínimo duas

vezes maior do que o número de variáveis. Os resultados da análise, conforme pode ser observado na Figura 21, serão apresentados da seguinte maneira:

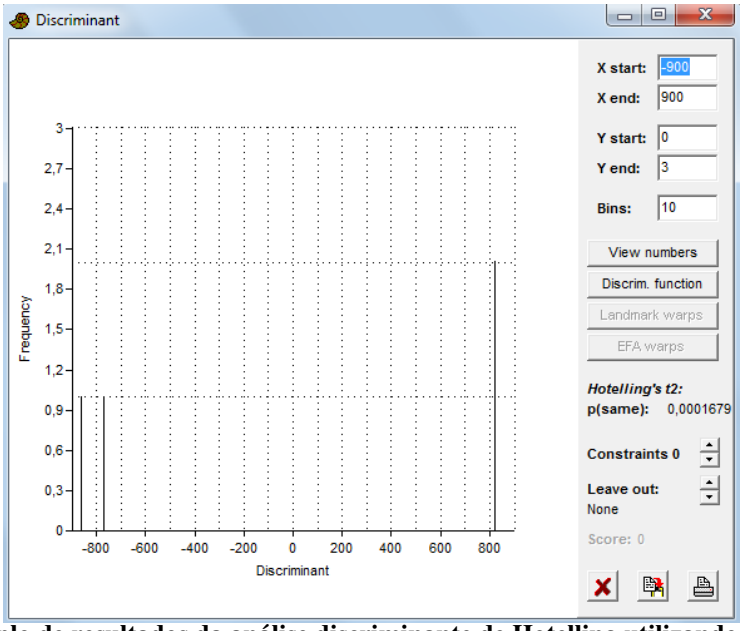


Figura 21: Exemplo de resultados da análise discriminante de Hotelling utilizando o *software* PAST

A hipótese nula do teste é de que as médias dos dois conjuntos são iguais, ou seja, que os conjuntos provêm da mesma população. O valor-p do teste é apresentado junto com o gráfico de histogramas citado anteriormente. Os resultados podem ser impressos diretamente através do PAST.

8. Exemplos Práticos

Neste capítulo serão abordados três exemplos práticos em que dados com altas dimensões estão presentes e foram necessárias técnicas estatísticas específicas para dados com esta característica para a sua resolução. O primeiro exemplo é sobre um estudo do perfil de expressão proteica do mosquito causador da malária, o segundo trata de uma comparação da distância entre duas sequências de DNA e o terceiro aborda um estudo de comparação entre um medicamento original e um falso.

8.1 Exemplo 1

O primeiro exemplo a ser abordado neste trabalho vem de um projeto realizado pela Universidade Federal do Amazonas (UFAM) em parceria com o Instituto Nacional de Pesquisas da Amazônia (INPA) e com a Universidade de Brasília (UnB), coordenado pelo Dr. Edmar Vaz de Andrade, nomeado "Proteoma do Intestino de *Anopheles darlingi*: Principal Vetor da Malária no Brasil" (Revista FAPEAM, número 5). O projeto tem como principal objetivo analisar o perfil de expressão proteica do intestino médio de fêmeas e machos de *Anopheles darlingi* em diferentes condições alimentares. A espécie *Anopheles darlingi* é representada pelo mosquito hospedeiro e transmissor da malária, que vive em regiões tropicais e subtropicais da América Central e do Sul.

Para a realização da pesquisa, os machos do mosquito são alimentados com glicose e as fêmeas com glicose e com sangue sadio, formando três grupos. Após a coleta destes mosquitos, estes tem os seus intestinos dissecados em laboratório. É então realizada a análise proteômica.

A primeira análise a ser feita, utilizando-se o gel eletroforético (Figura 22), foi a do rendimento médio de proteínas intestinais solúveis.

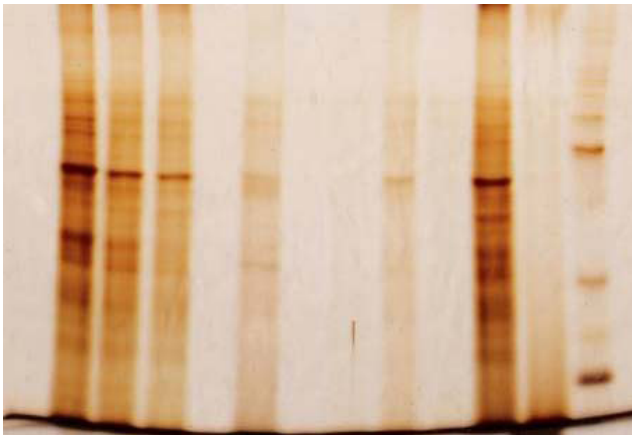


Figura 22: Géis eletroforéticos unidimensionais apresentando estrutura proteômica de mosquitos da espécie *Anopheles darlingi*

A partir do estudo feito na figura acima, os géis unidimensionais apresentados na Figura 22 são "transformados" em informação numérica descrita na tabela 1:

Tabela 1: Rendimento médio de proteínas intestinais solúveis (µg) por mosquito em diferentes condições alimentares

Sexo/Tipo de Alimentação		
Fêmeas/Glicose	Fêmeas/Sangue	Machos/Glicose
1,92	1,21	1,21
2,75	1,98	0,56
2,55	1,25	0,43
2,02	0,71	0,09
2,57	0,78	0,37
1,00	1,51	0,32
12,81	7,44	2,98

Este primeira medição é de caráter apenas descritivo e não há interesse em analisarmos diferenças entre os sexos nem entre os tipos de alimentação. A segunda análise, de cunho comparativo, verifica através das eletroforeses bidimensionais dos três grupos estudados, o perfil da expressão proteica encontrada, como pode ser visto na Figura 23.

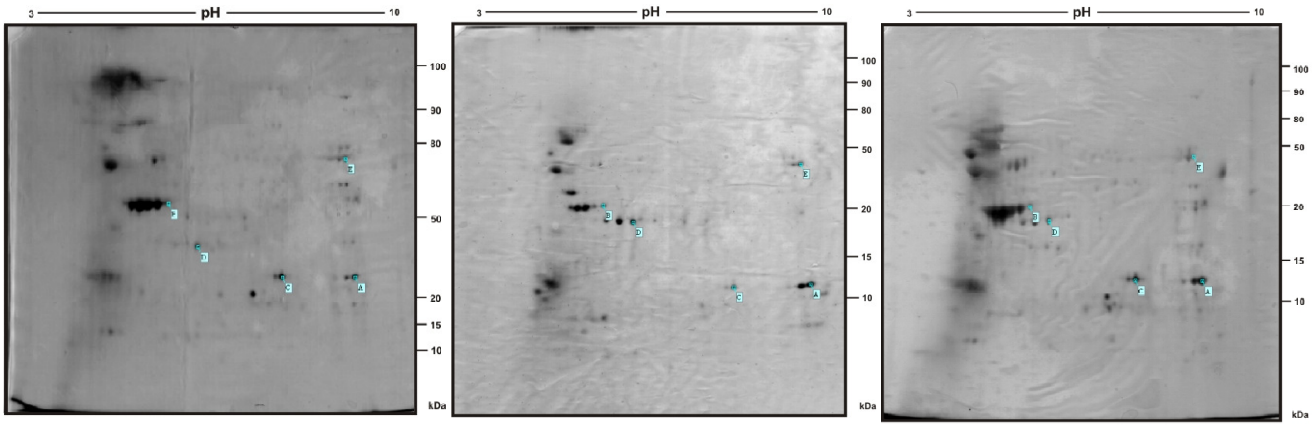


Figura 23: Eletroforeses bidimensionais dos machos alimentados com glicose, das fêmeas alimentadas com glicose e das fêmeas alimentadas com sangue, respectivamente

Com os géis em mãos, através de um aparelho, são medidos os pontos isoelétricos e as massas moleculares de cada um dos pontos protéicos encontrados nas três imagens. Foram observados 39 pontos protéicos nos machos alimentados com glicose, 38 pontos protéicos nas fêmeas alimentadas com glicose e 48 pontos protéicos nas fêmeas alimentadas com sangue. Com estes pontos medidos, a tabela 2 é montada:

Tabela 2: Variação da intensidade de pH e massa molecular das proteínas intestinais de *Anopheles darlingi* em diferentes condições alimentares

Sexo	Tipo de alimentação	Número de manchas	pI		MM	
			$\bar{x}$ ± EP	Variação	$\bar{x}$ ± EP	Variação
Fêmea	Glicose	38	6,6 ± 0,30	4,3 – 9,7	22,7 ± 2,3	7,8 – 55,2
	Sangue	48	6,5 ± 0,23	4,3 – 9,5	20,8 ± 1,7	7,1 – 61,0
Macho	Glicose	37	6,5 ± 0,29	4,3 – 9,7	34,8 ± 4,3	7,7 – 90,5

PI = ponto isoelétrico; MM = massa molecular, $\bar{x}$ = média amostral, EP = erro padrão

Conforme se pode perceber na tabela 2, não há diferenças significativas entre os três grupos, tanto em relação a ponto isoelétrico quanto em relação a massa molecular, já que a variação, que é calculada através de um intervalo de confiança de 95%, apresenta intersecções entre os grupos.

Para aprofundamento dos resultados encontrados, são realizadas análises de similaridade, primeiramente entre as imagens de machos e fêmeas alimentados com glicose e a seguir entre imagens de fêmeas. Por fim, comparam-se os resultados obtidos nas imagens com o intuito de se traçar um perfil de proteínas desenvolvidas devido aos dois tipos de alimentações e sexo. A Figura 24 nos mostra as análises de similaridade realizadas entre machos e fêmeas alimentados com glicose e entre fêmeas alimentadas com glicose e sangue.

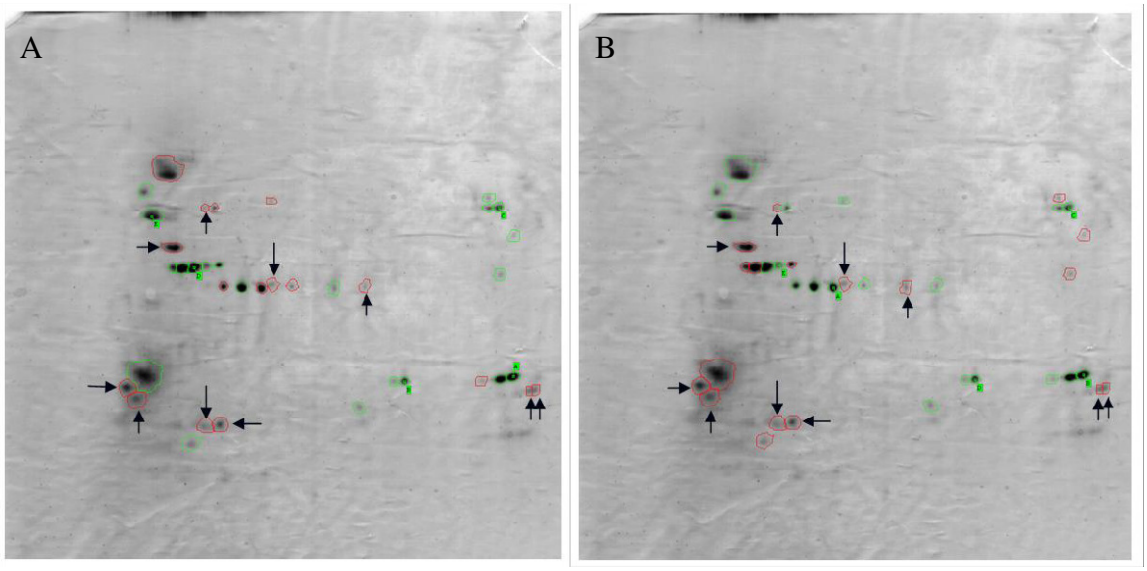


Figura 24: (A) comparação entre fêmeas e machos alimentados com glicose (B) comparação entre fêmeas alimentadas com glicose e sangue

Pontos circulados em verde na Figura 24 representam manchas comuns. Na Figura 24A os pontos circulados em vermelho representam manchas específicas em fêmeas, e as setas pretas indicam as manchas presentes também em machos. Já na Figura 24B os pontos circulados em vermelho representam manchas específicas em fêmeas alimentadas com glicose e as setas pretas indicam as manchas presentes também em fêmeas alimentadas com sangue.

Outro resultado final, oriundo da análise de similaridade, é uma visualização geral do perfil das fêmeas alimentadas com glicose, apresentada na Figura 25:

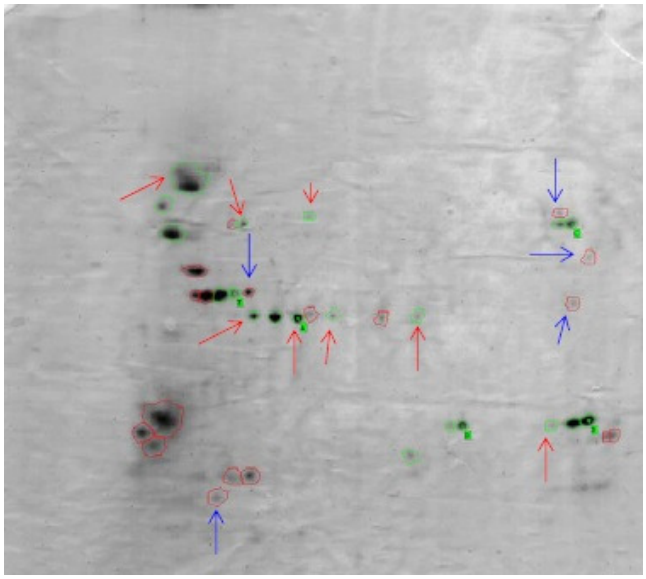


Figura 25: Gel de fêmeas alimentadas com glicose após análise de similaridade

As setas vermelhas da Figura 25 indicam as proteínas características das fêmeas (independente da sua condição alimentar) e a seta azul indica aquelas proteínas que estão presentes em fêmeas alimentadas apenas com glicose.

Ao final do estudo, chega-se à conclusão de que, apesar do padrão de proteínas entre os três grupos não ser significativamente diferente, existem alguns tipos de proteínas que só estão presentes em fêmeas, ou então que só estão presentes em alimentações baseadas em glicose. Concluí-se que o perfil de proteínas dos grupos apresenta tamanho e localidades semelhantes.

8.2 Exemplo 2

O segundo exemplo a ser apresentado neste trabalho é fictício e tem como objetivo apresentar algumas das técnicas de medidas de distância entre sequências de DNA abordadas no capítulo 5. Na tabela abaixo, são apresentadas as duas sequências de DNA que serão utilizadas ao longo deste exemplo:

Tabela 3: Sequências de DNA

Posição	Sequência 1	Sequência 2
1	C	C
2	C	C
3	T	G
4	G	G
5	A	A
6	A	G
7	A	A
8	G	G
9	G	G
10	T	T
11	C	C
12	A	A
13	C	T
14	A	A
15	T	T

Para iniciarmos o nosso estudo, faremos uma análise descritiva de ambas as sequências de DNA apresentadas acima. Na tabela 4 são apresentadas freqüências de cada um dos nucleotídeos em ambas as sequências.

Tabela 4: Tabela de Frequências

Sequência 1		**Sequência 2**	
Nucleotídeo	Frequência	Nucleotídeo	Frequência
A	5	A	4
C	4	C	3
G	3	G	5
T	3	T	3

Como visto anteriormente, no capítulo 5, outra análise descritiva pode ser feita através da distância de Hamming. Para isto, devemos verificar o número de posições em que o nucleotídeo difere nas duas sequências. Ao olharmos a tabela 3, vemos que apenas nas posições 3, 6 e 13 houve uma mudança de nucleotídeos de uma sequência para a outra. Temos então que $\sum_{k=1}^{K} \delta(X_{ik} \neq X_{i\prime k}) = 3$ e que K = 15. Ao aplicarmos esta informação na fórmula (5.11), chegamos na distância de Hamming:

$$H_{ii\prime} = \frac{1}{K} \sum_{k=1}^{K} \delta(X_{ik} \neq X_{i\prime k}) = \frac{1}{15}.3 = 0{,}2.$$

A distância de Hamming, portanto, é igual a 0,2, ou seja, este número indica que 20% das posições apresentam nucleotídeos diferentes nas duas sequências.

Para complementarmos esta primeira medida de distância, calcularemos agora a distância de Jukes-Cantor, que faz uma correção na distância de Hamming. Utilizando a fórmula (5.12), obtemos o seguinte:

$$K_W = -\frac{3}{4}\ln\left(1 - \frac{4}{3}H_{ii'}\right) = -\frac{3}{4}\ln\left(1 - \frac{4}{3}0{,}2\right) \cong -\frac{3}{4}\ln(0{,}73) \cong 0{,}2326\,.$$

Após a realização destas duas distâncias de caráter mais descritivo, partiremos para a distância log determinante. Para isso, devemos montar as matrizes de divergência F_{xy}, F_{xx} e F_{yy} e então calcularmos os seus determinantes.

Para a montagem das matrizes de divergência a seguir, são consideradas as linhas e colunas seguindo a respectiva ordem: A, C, G e T. Assim, temos que:

$$F_{xy} = \begin{pmatrix} 4/15 & 0 & 1/15 & 0 \\ 0 & 3/15 & 0 & 1/15 \\ 0 & 0 & 3/15 & 0 \\ 0 & 0 & 1/15 & 2/15 \end{pmatrix},$$

$$F_{xx} = \begin{pmatrix} 5/15 & 0 & 0 & 0 \\ 0 & 4/15 & 0 & 0 \\ 0 & 0 & 3/15 & 0 \\ 0 & 0 & 0 & 3/15 \end{pmatrix} \text{e}$$

$$F_{yy} = \begin{pmatrix} 4/15 & 0 & 0 & 0 \\ 0 & 3/15 & 0 & 0 \\ 0 & 0 & 5/15 & 0 \\ 0 & 0 & 0 & 3/15 \end{pmatrix}.$$

Calculando os determinantes destas três matrizes, obtemos o seguinte:

$$\det F_{xy} = 0{,}0014 \text{ e } \det F_{xx} = \det F_{yy} = 0{,}0035$$

Aplicando estas informações na fórmula (5.10), obtemos a distância log determinante:

$$d'_{xy} \equiv -\frac{1}{4}\ln\left(\frac{\det F_{xy}}{\sqrt{(\det F_{xx})(\det F_{yy})}}\right) \equiv -\frac{1}{4}\ln\left(\frac{0{,}0014}{\sqrt{0{,}0035^2}}\right) \equiv -\frac{1}{4}\ln\left(\frac{0{,}0014}{\sqrt{0{,}0035^2}}\right) \equiv 0{,}229.$$

Como se pode perceber, o resultado encontrado para a distância log determinante (0,229) é bastante similar aos resultados encontrados utilizando-se a distância de Hamming (0,2) e a distância de Jukes-Cantor (0,2326).

Os cálculos realizados no exemplo 2 podem ser feitos através do *software* PAST apresentado no capítulo 7.

8.3 Exemplo 3

O terceiro exemplo apresentado neste trabalho surgiu através de uma assessoria do NAE (Núcleo de Assessoria Estatística) do Instituto de Matemática da UFRGS. Um pesquisador tem interesse em identificar diferenças entre medicamentos de uma marca usual do mercado recolhidos em diversos pontos para identificar se os mesmos são originais ou falsificados.

Para isto, os dados são apresentados em uma forma espectrocópica, ou seja, gráficos de linha que apresentam um pico ao longo de um eixo, que varia de 400 a 1000 nanômetros, onde a altura é medida. Para a análise, foram escolhidos oito medicamentos que são supostamente originais (medicamentos 1 a 8) e dois que são supostamente falsificados (medicamentos 9 e 10) e é de interesse verificar se esta suposição é verdadeira baseando-se na altura dos picos de cada curva.

A área selecionada para a análise engloba a faixa de 400 a 413 nanômetros (com cinco medições), local de maior pico para ambos os grupos. Com isto, chegamos aos seguintes dados apresentados na tabela 5:

Tabela 5: Altura das Curvas

Medicamento/Nanômetros	400	403	407	410	413
Medicamento 1	418134	418493	419669	420209	418040
Medicamento 2	329109	331034	336477	342347	343882
Medicamento 3	438977	438713	437946	436769	435001
Medicamento 4	416736	417032	417886	417310	415077
Medicamento 5	392990	393721	395836	397036	394993
Medicamento 6	408499	408807	409684	409691	407159
Medicamento 7	453098	452549	451081	449737	447524
Medicamento 8	418858	419080	419809	419066	415493
Medicamento 9	81621	82745	85624	93210	96115
Medicamento 10	65382	66730	69211	75819	83587

A análise a ser realizada é a análise descriminante de Hotelling, com o auxilio do *software* PAST, conforme visto no capítulo 7.

Chegamos assim em um valor-p aproximadamente igual a 0,00000355, que é muito próximo de zero ($p<0,001$). Com isto, até mesmo para um nível de significância de 0,1%, rejeitamos a hipótese nula de que os dois conjuntos de medicamentos apresentam médias iguais, ou seja, os dados dos dois grupos diferem, nos dando fortes indícios de que as medidas obtidas para os medicamentos do grupo 2 marcado em vermelho na Figura 26 (medicamentos 9 e 10) não são iguais aos do outro grupo que vem de medicamentos originais. Isto também pode ser visto através da Figura 27, onde é apresentado o histograma das variáveis no eixo que maximiza a distância entre os conjuntos. Podemos ver que são discriminados de forma bastante evidente os dois grupos, que representam os medicamentos originais e os falsificados.

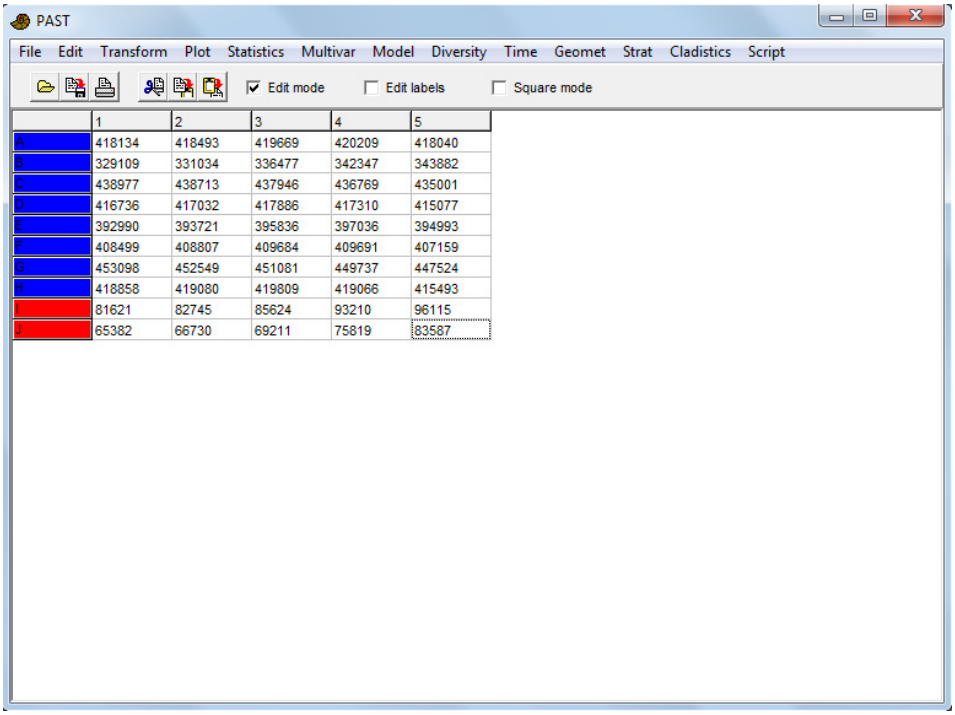


Figura 26: Disposição dos dados do Exemplo 3 no *software* PAST para a realização da análise discriminante de Hotelling

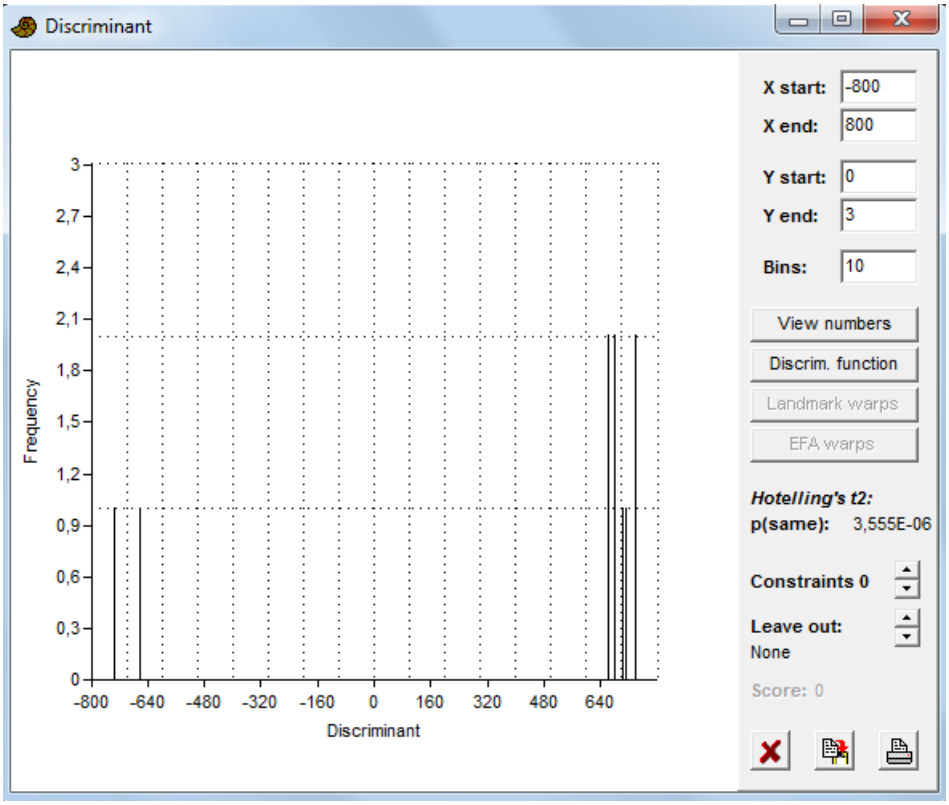


Figura 27: Resultados da análise discriminante de Hotelling no *software* PAST

Com isto, podemos concluir dizendo que, com um nível de significância de 0,1%, temos indícios de que os medicamentos 9 e 10 são realmente produtos diferentes dos medicamentos originais, visto que apresentam uma diferença significativa entre o conjunto de medidas no intervalo de 400 a 413 nanômetros.

9. Conclusões e Considerações Finais

O futuro da análise de dados com altas dimensões é bastante promissor. Não é por acaso que este é um dos ramos da estatística que mais vem crescendo comparado com outras áreas mais antigas. A medida que a tecnologia avança e novos métodos computacionais se desenvolvem, o trabalho com este tipo de dado tende a se intensificar. Dentro da estatística, ainda existe certa escassez de informação sobre este tema, devido a sua complexidade e difícil utilização sem métodos computacionais disponíveis. Porém, novos problemas surgem todos os dias, e cada vez em maior número, resultando em uma necessidade de maior conhecimento nesta área. O estatístico do futuro deverá ter bons conhecimentos em análise de dados com altas dimensões.

Como podemos ver ao longo deste trabalho, existem diversas aplicações para este tipo de dado, em várias áreas do conhecimento. Geneticistas sempre irão necessitar de análises em suas sequências de DNA. Químicos estarão interessados nas composições de seus produtos. Astrônomos desejam identificar novas estrelas e galáxias. A importância de estudos mais aprofundados dos dados com altas dimensões é enorme, já que as técnicas existentes não produzem resultados satisfatórios.

Entretanto, como se trata de uma área relativamente nova, as informações a respeito de técnicas e métodos para a análise dos dados com altas dimensões ainda permanecem muito dispersas. Organizar a informação já obtida, destacar ligações entre as técnicas utilizadas nas diferentes áreas e exemplificar de forma clara o seu uso são tarefas árduas e de grande valor. Para isto, um trabalho de revisão da literatura e apresentação de métodos é importante. Esta é uma das principais contribuições deste trabalho, assim como o estímulo deixado para outros alunos ou pesquisadores para continuarem a desenvolver esta área.

Para concluir, podemos ver o quão importante este estudo é a partir das áreas da ciência que a utilizam. Sem um estudo mais completo de dados com altas dimensões, não seria possível, por exemplo, a identificação de padrões de DNA, ou então o desenvolvimento de novas tintas ou de novos cosméticos. E mais: a identificação de sequências de DNA ou de genomas pode determinar algum tipo de doença e então

salvar vidas; a descoberta da composição de algum tipo de vegetal ou fungo, através da análise dos proteomas, pode ser determinante para a criação de remédios mais eficazes.

A alta dimensionalidade dos dados já vem ganhando destaque nos últimos anos e reconhecer a sua importância para o desenvolvimento da estatística é importante. Em muitos congressos, palestras e *workshops* sobre este assunto já vem sendo abordados e espera-se que este número cresça ainda mais.

O desconhecimento do estatístico com relação a este tipo de análises não poderá mais passar despercebido por muito tempo. O mundo já começou a voltar os seus olhos para o problema dos dados com altas dimensões.

Referências Bibliográficas

[1] AGRAWAL, R., GEHRKE, J., GUNOPULOS D. & RAGHAVAN, P. (1998). Automatic subspace clustering of high-dimensional data for data mining applications. Em: *ACM SIGMOD Conference on Management of Data*. Seattle, WA, EUA. pp. 94-105.

[2] ANDRADE, M. & PINHEIRO, H.P. (2002). Métodos Estatísticos Aplicados em Genética Humana. Em: *15º Simpósio Nacional de Probabilidade e Estatística*. São Paulo, SP, Brasil.

[3] BEYER, K., GOLDSTEIN, J., RAMAKRISHNAN, R. & SHAFT, U. (1998). When is 'nearest neighbor' meaningful? Em: *Proceedings of 7th International Conference on Database Theory*. Jerusalem, Israel. pp. 217-235.

[4] CHURCH, S. (2004). Advances in two-dimensional gel matching technology. *Biochem Soc Trans*. **Vol. 32, No. 3**. pp. 511-516.

[5] DOMENICONI, C., PAPADOPOULOS, D. & GUNOPULOS, D. *Subspace Clustering of High Dimensional Data*. Disponível em http://www.siam.org/proceedings/datamining/2004/dm04_058domeniconic.pdf. Acesso em: 17/03/2010.

[6] GOJOBORI, T., MORYIANA, E.N. & KIMURA, M. (1990). Statistical Methods for Estimating Sequence Divergence. *Methods in Enzimology*. **Vol. 183**. pp. 531-550.

[7] HINNENBURG, A. & KEIM, D.A. (1998). An Efficient Approach to Clustering in Large Multimedia Databases with Noise. Em: *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. Nova Iorque, NY, EUA. pp. 58-65.

[8] HINNENBURG, A. & KEIM, D.A. (1999). Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering. Em: *Proceedings of 25th International Conference on Very Large Data Bases*. Edimburgo, Escócia. pp. 506-517.

[9] HINNENBURG, A., AGGARWAL, C. & KIEM, D.A. (2000). What is the nearest neighbor in high dimensional spaces? Em: *Proceedings 26th International Conference on Very Large Data Bases*. Cairo, Egito. pp. 506-515.

[10] KACZMAREK, K., WALCZAK, B., DE JONG, S. & VANDEGINSTE, B.G. (2003). Matching 2D gel electrophoresis images. *J Chem Inf Comput Sci*. **Vol. 43, No. 3**. pp. 978-986.

[11] LAFETÁ, B.N., SANTOS, S. SILVA, V.L., CARVALHO, M.A.R., DINIZ, C.G., SILVA, N. (2008). Determinação do perfil protéico da membrana externa da Leptospira interrogans sorovariedade Hardjoprajitno. *Arq. Bras. Med. Vet. Zootec.* **Vol. 60, No. 6**. pp. 1301-1306.

[12] LOCKHART, P.J., STEEL, M.A., HENDY, M.D. & PENNY, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*. **Vol. 11**. pp. 605-612.

[13] NAGESH H., GOIL, S. & CHOUDHARY, A. (1999). MAFIA: Efficient and Scalable Subspace Clustering for Very Large Data Sets. Em: *Center for Parallel and Distributed Computing*. Northwestern University.

[14] O'FARRELL, P.H. & BIOL, J. (1975). High Resolution Two-dimensional Electrophoresis of Proteins. *J. Biol. Chem*. **Vol. 250**. pp. 4007-4021.

[15] PARSONS, L., HAQUE, E. & LIU, H. (2004). Subspace Clustering for High Dimensional Data: A Review. *SIGKDD Explor. Newsl.*, **Vol. 6, No. 1**. pp. 90-105.

[16] PENA, J.M., LOZANO, J.A., LARRANGA, P. & INZA, I. (2001). Dimensionality reduction in unsupervised learning of conditional gaussian networks. *IEEE Transactions*. **Vol. 23, No. 6**. pp. 590-603.

[17] SEILLIER-MOISEIWITSCH, F., MARGOLIN, B.H. & SWANSTROM, R. (1994). Genetic Variability of the Human Immunodeficiency Virus: Statistical and Biological Issues. *Anual Review of Genetics*. **Vol. 28**. pp. 559-596.

[18] SERDOBOLSKII, V.I. (2000). *Multivariate Statistical Analysis: A High-Dimensional Approach*. 1ª ed., Kluwer Academic Publishers, Londres, Inglaterra.

[19] STEINBACH, M., ERTÖZ, L. & KUMAR, V. (2004). The Challenges of Clustering High Dimensional Data. Em: *NEW DIRECTIONS in Statistical Physics*. Luc T.Willie, Boca Raton, FL, EUA. pp. 273-307.

[20] VAN DE WIEL, M. *High-Dimensional Data Analysis: Microarrays and Multiple Testing*. Disponível em www.math.vu.nl/sto/onderwijs/statlearn/Hdda_microarray_mt1.ppt. Acesso em: 07/04/2010.

[21] WAAIJENBORG, S. & ZWINDERMAN, A.H. (2010). Association of repeatedly measured intermediate risk factors for complex diseases with high dimensional SNP data. *Algorithms for Molecular Biology*. **Vol. 5, No. 1**. pp. 17.

[22] WATANABE, Y., TAKAHASHI, K. & NAKAZAWA, M. (1997). Automated detection and matching of spots in autoradiogram images of two-dimensional electrophoresis for high-speed genome scanning. Em: *Proceedings of the 1997 International Conference on Image Processing*. Washington, DC, EUA. **Vol. 3**. pp. 496-499.

[23] WEIR, B.S. & BASTEN, C.J. (1990). Sampling Strategies for DNA Sequence Distances. *Biometrics*. **Vol. 46**. pp. 551-582.

[24] WILKINS, M.R., WASINGER, V.C., CORDWELL, S.J., CERPA-POLJAK, A., YAN, J.X., GOOLEY, A.A., DUNCAN, M.W., HARRIS, R., WILLIAMS, K.L. & HUMPHERY- SMITH, I. (1995). Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis*. **Vol. 7**. pp. 1090-1094.

[25] XIN, H.M. & ZHU, H. (2009). Multiple information-based spot matching method for 2-DE images. *Electrophoresis*. **Vol. 30, No. 6**. pp. 2477-2480.

[26] YU, L. & LIU, H. (2003). Feature selection for high-dimensionaldata: a fast correlation-based solution. Em: *Proceedings of the twentieth International Conference on Machine Learning*. Washington, DC, EUA. pp. 856-863.

Anexos

Anexo 1 – Páginas da Internet com as informações importantes tratadas na monografia

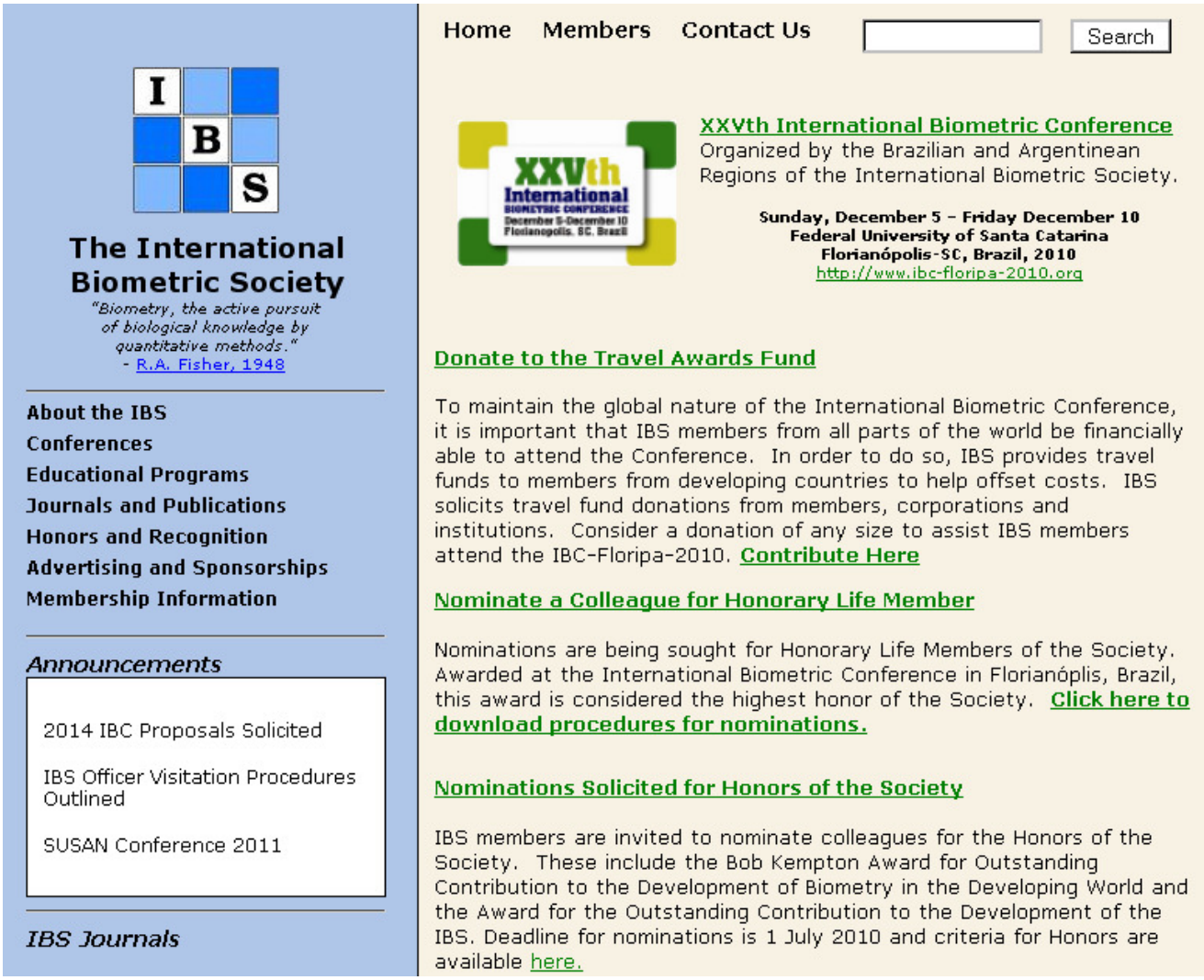


Figura A1: Página oficial da XXVth International Biometric Conference
http://www.rbras.org.br/~ibcfloripa2010/

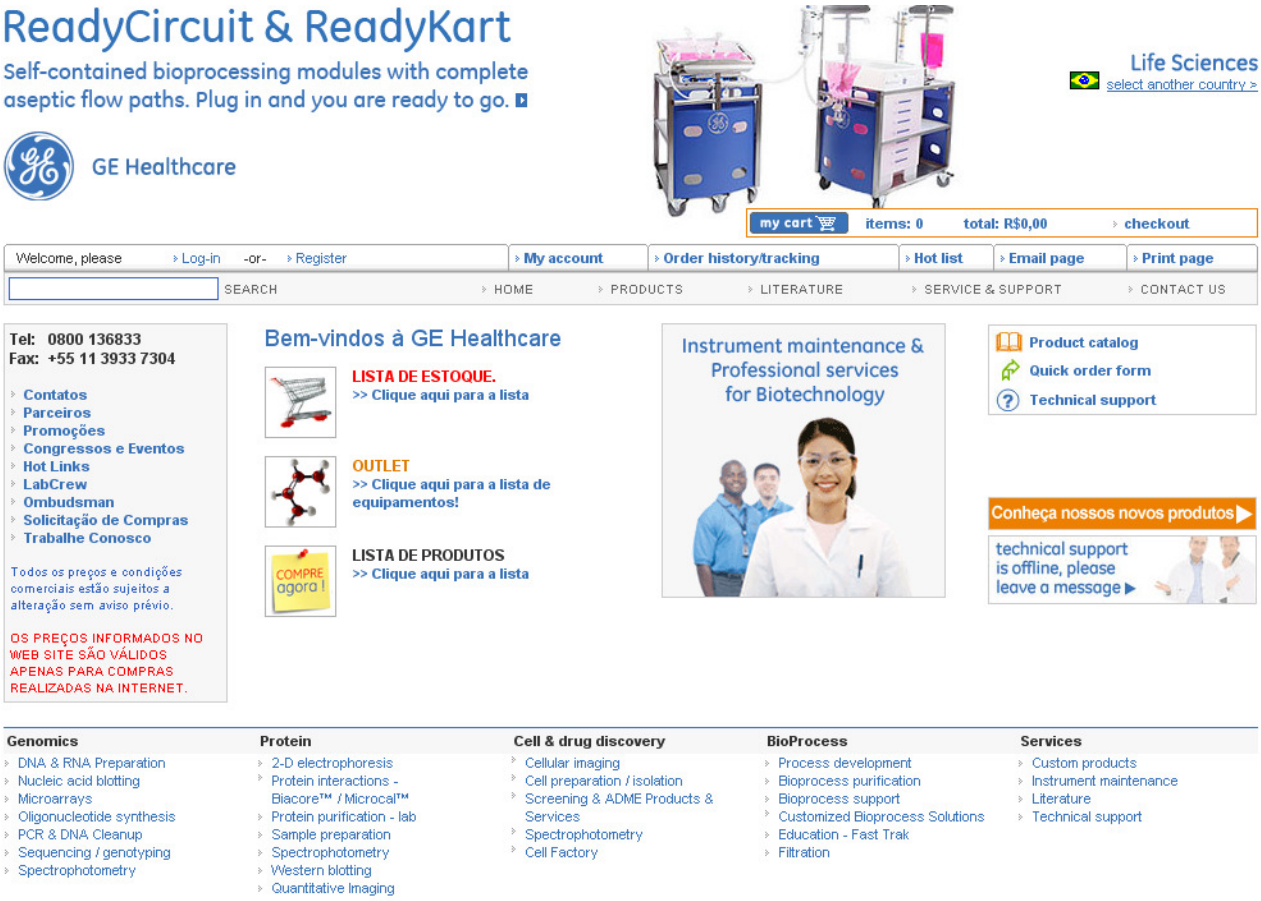


Figura A2: Página no Brasil da GE Healthcare, empresa fabricante do *software* ImageMaster™ 2D Platinum
http://www.gehealthcare.com/

Figura A3: Página do professor Øyvind Hammer, desenvolvedor do *software* PAST
http://folk.uio.no/ohammer/past/