



CENTRO UNIVERSITÁRIO FRANCISCANO

PRÓ-REITORIA DE PÓS-GRADUAÇÃO, PESQUISA E EXTENSÃO

ÁREA DE CIÊNCIAS TECNOLÓGICAS

Curso de Mestrado em Nanociências

SYLVIO ANDRÉ GARCIA VIEIRA

**APLICAÇÃO DE MÁQUINAS DE VETORES DE SUPORTE NA
INVESTIGAÇÃO DA ATIVIDADE GÊNICA DO CÂNCER DE COLO DE
INTESTINO.**

Santa Maria, RS.

2011

SYLVIO ANDRÉ GARCIA VIEIRA

**APLICAÇÃO DE MÁQUINAS DE VETORES DE SUPORTE NA
INVESTIGAÇÃO DA ATIVIDADE GÊNICA DO CÂNCER DE COLO DE
INTESTINO.**

Dissertação apresentada ao Curso de
Mestrado em Nanociências do Centro
Universitário Franciscano de Santa Maria
como requisito parcial para obtenção do
título de Mestre em Nanociências.

Orientadora: Prof^a. Dr^a. **Juliana Kaizer Vizzotto**

Co-orientador: Prof. Dr. **José Carlos Merino Mombach**

Santa Maria, RS.

2011

Ficha Catalográfica

V658a Vieira, Sylvio André Garcia
 Aplicação de máquinas de vetores de suporte na
 investigação da atividade gênica do câncer de colo de
 intestino / Sylvio André Garcia Vieira ; orientadora Juliana
 Kaizer Vizzotto ; co-orientador José Carlos Merino Mombach
 - Santa Maria : Centro Universitário Franciscano, 2011.
 77f. : il.

Dissertação (Mestrado em Nanociências) – Centro
Universitário Franciscano, 2011.

1. Bioinformática 2. Mineração de dados 3. Câncer
4. Adenoma 5. SVM 1. Vizzotto, Juliana Kaizer
II. Mombach, José Carlos Merino III. Título

CDU 004:573:62-181.4

ÁREA DE CIÊNCIAS TECNOLÓGICAS

Mestrado em Nanociências

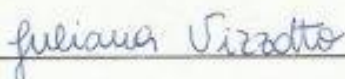
A COMISSÃO EXAMINADORA, ABAIXO-ASSINADA, APROVA A DISSERTAÇÃO:

APLICAÇÃO DE MÁQUINAS DE VETORES DE SUPORTE NA
INVESTIGAÇÃO DA ATIVIDADE GÊNICA DO CÂNCER DE COLO DE
INTESTINO

Elaborada por

SYLVIO ANDRÉ GARCIA VIEIRA

COMISSÃO EXAMINADORA



Profª. Drª. Juliana Kalzer Vizzotto – UFSM



Profª. Drª. Marta Palma Alves – UNIFRA



Prof. Dr. Giovanni Rubert Librelotto – UFSM

Santa Maria, 30 de março de 2011.

RESUMO

A mineração de dados é o processo de descoberta de padrões correlacionados entre os diversos dados existentes em uma base. O GEO é uma base de dados biológicos público, mantido pelo NCBI, onde se buscou as informações referentes a trinta e dois pacientes de Adenoma de colo de intestino, com leituras da expressão de sondas referentes aos genes, extraídas do RNA. Os dados depositados em bancos biológicos, por si só, não produzem informação útil, e por isto, foram selecionados respeitando diversos fatores, como a confiabilidade da informação colhida, a quantidade de informações presentes no maior numero de sondas, e finalmente filtrados pela leitura de maior expressão. Após a base de dados ser tratada e com os genes selecionados, foi então aplicada a ferramenta R com o classificador SVM com o objetivo de identificar, neste pequeno conjunto de genes, a possibilidade de associação deles com a presença do adenoma de colo de intestino.

A partir dos resultados obtidos através da classificação dos dados percebeu-se que as características dos genes são bem distintas e que a atividade varia bastante de gene para gene. Porém, isto ocorre de forma padronizada, o que permitiu que o algoritmo pudesse identificar estes padrões e sugerir sua participação no processo do adenoma.

Palavras-chave: Bioinformática, Mineração de dados, Câncer, Adenoma, SVM.

ABSTRACT

Data mining is the process of discovering patterns correlated with the various existing data in a database. GEO is a public biological database, maintained by NCBI, where they sought information relating to thirty-two patients of colorectal adenoma, with readings from the probes concerning the expression of genes, extracted RNA. The data deposited in biological banks alone do not produce useful information, and therefore, were selected respecting various factors such as the reliability of the information collected, the amount of information present in the greatest number of probes, and finally filtered by reading higher expression. After the database and be treated with the selected genes was then applied to the R tool with the SVM in order to identify this small set of genes, the possibility of their association with the presence of adenoma of the colon of the intestine.

From the results obtained by classifying the data it was noticed that the characteristics of the genes are distinct and that the activity varies greatly from gene to gene. However, this occurs in a standardized manner, allowing the algorithm could identify these patterns and suggest their involvement in the adenoma.

Keywords: Bioinformatics, Data Mining, Cancer, Adenoma, SVM.

SUMÁRIO

LISTA DE FIGURAS	9
LISTA DE TABELAS.....	10
L ISTA DE GRÁFICOS	11
1. INTRODUÇÃO	12
1.1. OBJETIVOS	15
1.2. OBJETIVOS ESPECÍFICOS.....	15
2. REFERENCIAL TEÓRICO	16
2.1. TRABALHOS RELACIONADOS	16
2.2. BIOINFORMÁTICA NO ESTUDO DO CÂNCER	18
2.3. GENE.....	20
2.4. DNA E RNA	22
2.5. ADENOMA	24
2.6. DESCOBERTA DO CONHECIMENTO.....	25
2.7. APRENDIZADO DE MÁQUINA	28
2.8. DATA MINING.....	30
2.9. AMOSTRAGEM	33
2.10. BANCOS DE DADOS	34
2.11. BANCOS DE DADOS BIOLÓGICOS.....	35
2.12. MICROARRANJOS (<i>MICROARRAY</i>).....	38
2.13. AFFYMETRICS	38
2.14. SVM.....	39
2.14.1. <i>FRONTEIRAS LINEARES</i>	40
2.15. TIPOS DE KERNELS SVM	45
2.14.1. <i>SVM e-1071 (Libsvm)</i>	45
2.14.2. <i>SVM – FISCHER</i>	46
2.14.3. <i>MOTIF KERNELS</i>	46
2.15. A FERRAMENTA R	46
2.16. HARDWARE E SOFTWARE.....	47

3. METODOLOGIA.....	48
3.2 DESCRIÇÃO DAS TÉCNICAS APLICADAS.....	49
3.3 RESULTADOS	55
3.4 DIFICULDADES ENCONTRADAS.....	67
4. CONCLUSÕES.....	69
4.1 TRABALHOS FUTUROS.....	70
5. REFERENCIAS BIBLIOGRAFICAS	71

LISTA DE FIGURAS

FIGURA 1 - ESTRUTURA DO DNA [NLM, 2009]	22
FIGURA 2 - VISÃO GERAL DAS FASES DO PROCESSO DE KDD ADAPTADO DE (FAYYAD ET AL., 1996A).....	28
FIGURA 3 - APRENDIZADO DE MÁQUINA SUPERVISIONADO, ADAPTADO DE LORENA ET AL., 2007	30
FIGURA 4 - OBJETIVOS E TAREFAS DE MINERAÇÃO DE DADOS (ADAPTADA DE REZENDE ET AL., 2003).	31
FIGURA 5 - CONJUNTO DE DADOS COM CÍRCULOS E QUADRADOS, ADAPTADO DE TAN ET AL., 2009.	40
FIGURA 6 - CÍRCULOS E QUADRADOS SEPARADOS POR VÁRIAS LINHAS, ADAPTADO DE TAN ET AL. 2009.	41
FIGURA 7- SEPARAÇÃO DOS GRUPOS POR MARGENS, ADAPTADO DE BURGES, 1998.	41
FIGURA 8- VETORES DE SUPORTE, ADAPTADO DE TAN ET AL., 2009.....	42
FIGURA 9- LIMITES DE DECISÃO, ADAPTADO DE LORENA ET AL. 2007.....	43
FIGURA 10 - SVM - ADAPTADO DE LORENA ET AL. 2007 (A) <i>OVERFITTING</i> (B) SVM (C) <i>UNDERFITTING</i>	45

LISTA DE TABELAS

TABELA 1 - MODELO PARCIAL DO ARQUIVO DISPONÍVEL NO GEO	50
TABELA 2 - TABELA GPL570 – FORNECIDA PELA AFFYMETRIX	52
TABELA 3 - TABELA GPL570 – FORNECIDA PELA AFFYMETRIX COM NOMES DUPLOS.....	53
TABELA 4 - ADENOMA1 - PARCIAL	54
TABELA 5 - EXEMPLO DE PREDIÇÃO CORRETA	56
TABELA 6 - EXEMPLO DE PREDIÇÃO INCORRETA.....	56
TABELA 7 - 50 GENES, TREINO COM 32, TESTES COM 32.....	57
TABELA 8 - 50 GENES, TREINO COM 15, TESTES COM 49.....	58
TABELA 9 – 50 GENES, TREINO COM 5, TESTE COM 59	59
TABELA 10 - 20 GENES, TREINO COM 15, TESTE COM 49	60
TABELA 11 - 20 GENES, TREINO COM 32, TESTES COM 32.....	61
TABELA 12 - 15 GENES, TREINO COM 32, TESTE COM 32.....	62
TABELA 13 - 15 GENES, TREINO COM 15, TESTE COM 49	63
TABELA 14- 15 GENES, TREINO COM 5, TESTE COM 59	64
TABELA 15 - 10 GENES, TREINO COM 32, TESTES COM 32.....	65
TABELA 16 - 10 GENES, TREINO COM 15, TESTE COM 49.....	66
TABELA 17 - QUADRO COMPARATIVO DOS RESULTADOS.....	ERRO!
INDICADOR NÃO DEFINIDO.	

L ISTA DE GRÁFICOS

GRÁFICO 1 - 50 GENES, TREINO COM 32, TESTE EM 32.....	57
GRÁFICO 2- 50 GENES, TREINO COM 15, TESTES COM 49	58
GRÁFICO 3 - 50 GENES, TREINO COM 5, TESTE COM 59.....	59
GRÁFICO 4- 20 GENES, TREINO COM 15, TESTE COM 49.....	60
GRÁFICO 5 - 20 GENES, TREINO COM 32, TESTES COM 32	61
GRÁFICO 6 - 15 GENES, TREINO COM 32, TESTE COM 32.....	62
GRÁFICO 7- 15 GENES, TREINO COM 15, TESTE COM 49.....	63
GRÁFICO 8 - 15 GENES, TREINO COM 5, TESTE COM 59.....	64
GRÁFICO 9 - 10 GENES, TREINO COM 32, TESTES COM 32	65
GRÁFICO 10 - 10 GENES, TREINO COM 15, TESTES COM 49	66

1. INTRODUÇÃO

As pesquisas científicas realizadas em laboratórios buscam compreender as razões e as causas de doenças que comprometem a continuidade da vida. Além disso, há o estudo de comportamentos anômalos que comprometem o bom funcionamento das células. Todas estas pesquisas produzem resultados, os quais são centenas de milhares de informações, de dados, que são diariamente e sistematicamente depositados em grandes centros de informações chamados Bancos de Dados Biológicos (LESK, 2008).

Em décadas passadas, o problema residia na capacidade física de armazenamento e processamento dos computadores. Atualmente o foco principal está na maneira como estes dados são analisados, de forma a extrair as informações que sejam realmente úteis. Os dados, na forma natural como são obtidos nas pesquisas e que são armazenados, podem não apresentar conhecimento direto e de fácil visualização, mas se ajustados e manipulados por um processo de extração de conhecimento útil, revelam informações que talvez não fossem possíveis de se obter por meio de técnicas estatísticas normais ou até mesmo com ferramentas analíticas (SCHUCH, 2010).

Já em 1984, John Naisbitt observou, “estamos nos afogando em informação, porém para passar fome em conhecimento”. O problema que existe hoje em dia não é a falta de dados ou informações. Pelo contrário, hoje há excesso de dados, em todas as áreas do conhecimento. Objetivando analisar estes dados oriundos dos laboratórios e depositados nos bancos de dados biológicos, surgiram novas áreas de pesquisa como a Bioinformática e a Biologia Computacional (LAROSE, 2005).

Um banco de dados, segundo Waymire (1999), é uma coleção logicamente coerente de dados que normalmente agrupa registros utilizáveis para o mesmo fim. Um banco de dados contém os dados dispostos numa ordem pré-determinada, em função de um projeto de sistema, sempre para um propósito muito bem definido (SOUZA, 2004). Segundo Silberschatz (2006), um sistema de banco de dados é um método conveniente de ser usado para armazenamento e recuperação de grande quantidade de dados ou informações em aplicações de computador. Algumas de suas aplicações representativas podem ser vistas em organizações financeiras, para informações de clientes, contas, empréstimos, ou ainda em telecomunicações, para manter o registro de chamadas

realizadas, gerar contas mensais, manter saldos de cartões de chamadas pré-pagos, etc. (SILBERSCHATZ, 2006).

A aplicação de banco de dados biológicos, segundo Lesk (2008), inclui, por exemplo: sequências de ácidos nucleicos, inclusive genomas completos; sequências de aminoácidos de proteínas; funções de proteínas, dentre outros.

Segundo Lesk (2005), após a divulgação de que o código genético humano havia sido decifrado, os cientistas permaneceram lado a lado com os políticos. Desta união surgiu o consórcio internacional de pesquisas, que atua hoje em dia (LESK, 2005). A partir deste consórcio, os governos de diversos países passaram a investir em bancos de dados públicos, como por exemplo: o projeto Genoma, que tem a finalidade de armazenar dados oriundos de pesquisas do mapeamento dos genes humanos e o seu sequenciamento, suas funções e interações (DUBITZKY et al., 2007); projeto Proteoma, que se concentrou na pesquisa das proteínas, suas localizações e interações (DUBITZKY et al., 2007); finalmente, o projeto de pesquisas de expressão gênica em tecidos epiteliais (GEO, 2010), objeto de pesquisa deste trabalho.

Para obter conhecimento a partir destes dados, que se acumularam em bancos de dados biológicos públicos, fez-se necessário que as técnicas de análise de informações evoluíssem, o que aconteceu de forma considerável nas últimas décadas. Isto vem tornando as atividades de interpretação das informações mais rápida, prática e confiável. A este conjunto de técnicas de análise de informações foi então dado o nome de Mineração de dados, mais conhecido na expressão em inglês *Data Mining* (TAN et al., 2009).

Segundo Tan et al., (2009), pesquisadores de diferentes disciplinas se engajaram no desenvolvimento de ferramentas computacionais que atendessem a todos, a fim de aumentar a eficiência, permitindo o tratamento de diversos tipos de dados, oriundos de pesquisas das áreas de: engenharia, medicina, biologia, etc.. Esta união permitiu o surgimento da área de mineração de dados. Assim, tornou-se possível construir uma ferramenta que pudesse ser utilizada em todas estas áreas, atendendo as necessidades de todos.

Em especial, a mineração de dados atrai ideias como: amostragem; estimativa e teste de hipóteses a partir de estatísticas e algoritmos de busca; reconhecimento de padrões e aprendizado de máquina (TAN et al., 2009).

Na mineração de dados, há um processo de aprendizagem em que alguns dos dados são oferecidos a uma ferramenta de mineração, como por exemplo, o “R” ou o “Weka”, dentre outros. Durante o processo, a ferramenta aprende os padrões existentes na combinação de características. Assim, pode testar os dados novos e comparar com os padrões criados, classificando os dados em duas classes distintas (HAYKIN, 2001). Neste contexto de extração de informação em grandes quantidades de dados brutos, depositados nas bases públicas, tornou-se possível tratar os microarranjos de dados.

Os microarranjos são tecnologias recentes, que permitem testar simultaneamente a presença de muitas sequências de DNA. Podem ser usados, por exemplo, para determinar padrões de expressão em diferentes proteínas através da detecção de RNAs mensageiros. Além disto, podem genotipar por meio da detecção de variâncias distintas nas sequências de genes (LESK, 2005).

Os experimentos de microarranjos sempre produzem imagens que são analisadas por um software de leitura e tratamento de imagens, no intuito de converter estes dados brutos em expressões gênicas mensuráveis (BOLSTAD, 2007).

Os dados depositados no GEO (GEO, 2010), contêm os valores da expressão gênica de trinta e dois pacientes portadores de adenoma, que segundo Salces et al. (2004) são tumores benignos com um potencial para o desenvolvimento de câncer em um período de 5 a 15 anos, sendo leituras de células doentes e também de células saudáveis. O fato de trabalhar com estas informações, utilizando um algoritmo de classificação eficaz, torna possível a descoberta de padrões entre as leituras gênicas dos pacientes, promovendo a possibilidade de contribuir no diagnóstico de que este paciente poderia estar desenvolvendo o adenoma, através da leitura de sua própria atividade gênica.

1.1. OBJETIVOS

O objetivo deste trabalho é preparar a base de dados oriunda do GEO (GEO, 2010) para que se possam identificar um conjunto reduzido de genes com maior atividade gênica. Aplicar a ferramenta de mineração de dados neste conjunto a fim de contribuir para a identificação dos genes que possam estar envolvidos com o adenoma de colo de intestino.

1.2. OBJETIVOS ESPECÍFICOS

Os objetivos específicos deste trabalho incluem:

- Preparar a base de dados, para que possa ser classificada, isolando somente os genes com informações completas e com atividades dentro de um nível considerado aceitável.
- Identificar dentre os Genes selecionados, os de maior atividade em células com adenoma quando comparadas às células normais;
- Submeter as bases preparadas às ferramentas de mineração para efetuar a classificação;
- Aplicar a ferramenta treinada em novos casos, a fim de realizar a predição da probabilidade do desenvolvimento do adenoma;
- Avaliar os resultados obtidos.

Este texto está estruturado como segue. No Capítulo 2 é abordado o tema desta dissertação, apresentando trabalhos relacionados a este. Já no Capítulo 3, são tratados temas como a bioinformática, genes, DNA e RNA, Adenoma, descoberta do conhecimento, aprendizagem de máquina, mineração de dados e amostragem. No Capítulo 4, são mostradas as ferramentas que foram utilizadas, como bancos de dados, bancos de dados biológicos, descrevendo alguns deles, o classificador SVM e a ferramenta R. Já no Capítulo 5, é abordada a aplicação de microarranjos, o estudo de caso, a descrição das técnicas e apresentados os resultados obtidos. Finalmente no Capítulo 6 são apresentadas as conclusões do trabalho e no sétimo e último Capítulo temos a bibliografia citada no decorrer do texto.

2. REFERENCIAL TEÓRICO

Seguindo a metodologia deste trabalho, foi necessário fazer um levantamento teórico sobre os muitos assuntos, começando a apresentação de alguns trabalhos relacionados ao nosso, logo a seguir trata-se de bioinformática no estudo do câncer, conceitos de genes, DNA e RNA, adenoma, conceitos da área de mineração de dados, das ferramentas utilizadas nesta pesquisa e mais alguns esclarecimentos que se fazem necessários para a compreensão deste trabalho.

2.1. TRABALHOS RELACIONADOS

Esta seção apresenta alguns trabalhos relacionados com a presente dissertação. Essencialmente os trabalhos aqui apresentados tratam de assuntos de mineração de dados na identificação de padrões em bases de dados biológicas com dados de microarranjos e classificação de dados. Estes trabalhos estão relacionados a esta dissertação devido à utilização do mesmo classificador.

Em Guyon et al. (2002), foi abordado o problema da seleção de um pequeno subconjunto de genes a partir de padrões gerais de dados de expressão gênica, oriundos de um microarranjo de DNA, no nosso trabalho, os microarranjos foram de RNA. Foram utilizados exemplos de treinamento disponíveis de câncer e de pacientes normais, da mesma forma que em nosso trabalho, porem, usamos dados do adenoma. Diferente do nosso caso, os autores construíram um classificador adequado para o diagnóstico genético, bem como trabalharam na descoberta de um remédio. Foi demonstrado que existiram tentativas anteriores de resolver este problema de microarranjos com técnicas semelhantes. Os autores propuseram um novo método de seleção genética utilizando *Support Vector Machine* (SVM) baseado na Eliminação de Características Recursivas (*Recursive Feature Elimination* (RFE)). Foi demonstrado experimentalmente que os genes selecionados por esta técnica de produção obteve um melhor desempenho de classificação e dos genes que são biologicamente relevantes para o câncer, assim como nosso trabalho visava identificar os genes possivelmente relevantes na identificação do adenoma. Em contraste com o método de referência, o novo método eliminou a redundância de gene automaticamente e produziu melhores e

mais compactos subconjuntos de genes. No nosso caso, a redundância foi eliminada manualmente. Nos pacientes com leucemia o novo método descobriu 2 genes que produzem zero de erro com a técnica, enquanto que no método de referência, 64 genes são necessários para obtenção do melhor resultado. No Banco de Dados de câncer do cólon, foram utilizados apenas 4 genes no novo método com 98% de precisão, valores semelhantes ou até inferiores aos obtidos em nosso trabalho, enquanto o método de referência do trabalho de Guyon et al. (2002) a precisão ficou com apenas 86%.

Já em Brown et al. (1999), foi desenvolvido um método de classificação funcional de genes utilizando os dados de expressão gênica oriundos de experimentos com a hibridização de microarranjos de DNA, da mesma forma que Guyon, difere de nossa pesquisa em função de termos utilizado amostras de RNA. Este método é baseado nas Máquinas de Vetores de Suporte (SVM), tal qual utilizamos em nosso trabalho. Estas são consideradas como se fosse um método de aprendizagem controlado por computador no intuito de explorar o conhecimento prévio da função do gene, para tentar identificar os genes desconhecidos da função, mas semelhantes aos dados de expressão constantes na base, da mesma forma que utilizamos nesta pesquisa. Segundo os autores, o SVM pode evitar vários problemas associados com os métodos de agrupamento não supervisionado, como por exemplo, o algoritmo de agrupamento hierárquico e o algoritmo de mapas auto-organizáveis. Os SVMs têm muitas características matemáticas que os tornam atraentes para a análise de expressão gênica, incluindo a sua flexibilidade na escolha de uma função de similaridade, há diversas soluções quando trata com conjuntos de dados grandes, a capacidade de lidar com espaços e a capacidade de identificar ruídos. Os autores afirmam que testaram o SVM árduamente utilizando diferentes métricas de similaridade, bem como alguns outros métodos de aprendizagem supervisionada, e identificaram que o SVM foi mais eficaz na identificação de conjuntos de genes com uma função comum, utilizando dados de expressão.

Furey (2000), do departamento de Ciência da Computação da Universidade da Califórnia, assim como os outros trabalhos citados, trabalha com experimentos de *microarray* do DNA gerando milhares de medições de expressão de gene, que estão sendo usados para coletar informações a partir de amostras de tecidos e células sobre

diferenças de expressão gênica que serão úteis no diagnóstico da doença. Foi desenvolvido um novo método para analisar este tipo de dados usando máquinas de vetor de suporte (SVM). Esta análise consiste em ambas as classificações das amostras de tecidos, tal qual foi utilizado neste trabalho, onde analisamos amostras do tecido com adenoma e também do tecido normal, o autor ainda informa que uma exploração dos dados para tecidos mal rotulados ou tecidos com resultados questionáveis foi avaliado, em nosso trabalho estes dados já foram excluídos no pré-processamento.

Nos resultados, os autores demonstraram o método em detalhes sobre amostras que consistem em tecidos de câncer de ovário, tecido normal do ovário e outros tecidos normais. O conjunto de dados consistiu de resultados da expressão de 97.802 cDNAs para cada tecido. Como resultado da análise computacional, uma amostra de tecido é descoberta e confirmada para ser erroneamente rotulada. Após a correção deste erro e a remoção de um caso isolado, a classificação perfeita dos tecidos é conseguida, mas não com a confiança em alta, em nosso trabalho, esta confiança era medida pelo P_value de cada expressão, como pode ser visto em nosso estudo de caso. É então identificado e analisado um subconjunto de genes do conjunto de dados de ovário, cuja expressão é muito diferenciada entre os tipos de tecidos. Para mostrar a robustez do método SVM, dois conjuntos de dados anteriormente publicados a partir de outros tipos de tecidos ou células são analisadas. Os resultados são comparáveis aos obtidos anteriormente. Foi demonstrada também a utilização de outro método de aprendizagem, o Perceptron, para comparar ao SVM. Foram utilizados cinco conjunto de dados para a comparação e o SVM se saiu melhor em todos.

2.2. BIOINFORMÁTICA NO ESTUDO DO CÂNCER

A bioinformática é uma disciplina científica com raízes nas ciências da computação, na estatística e na biologia molecular. A bioinformática desenvolveu-se para enfrentar os resultados das iniciativas de sequenciamento de genes, que produzem uma quantidade cada vez maior de dados sobre proteínas, DNA e RNA. Desse modo, os biólogos moleculares passaram a utilizar métodos estatísticos capazes de analisar grandes quantidades de dados biológicos, a predizer funções dos genes e a demonstrar relações entre genes e proteínas (VOGEL, 2003).

O DNA e o RNA são cadeias de polímeros compostas por uma pequena categoria de substâncias químicas similares. As unidades individuais são denominadas nucleotídeos. (GIBAS e JAMBECK, 2002, Pg. 167).

Esta área de pesquisa multidisciplinar surgiu em meados de 1940, quando o primeiro computador digital foi inventado. O termo bioinformática não existia nesta época, mas já era possível prever que ela aconteceria (SETUBAL, 2003).

Em 1953, quando Watson e Crick descobriram a hélice dupla, evidenciou-se o fato da informação genética também ser armazenada de forma digital, e que poderia ser descrita com um alfabeto, porém não binário, mas quaternário, pois quatro letras são usadas: A, C, G e T. Mais tarde, mas ainda na década de 50, quando se descobriu que os genes também se comportavam de forma binária, ou seja, ligado ou desligado. Ficou claro que futuramente a informática e a biologia estariam andando lado a lado e formariam uma nova área do conhecimento (SETUBAL, 2003).

Após a descoberta do código genético e do fluxo da informação biológica dos ácidos nucléicos para as proteínas, o foco dos estudos passou para a Biologia Molecular. Logo surgiram métodos de sequenciamento destes polímeros (DNA e RNA), principalmente do DNA. Desde então, mais de 18 milhões de sequências já foram produzidas e disponibilizadas em Bancos de Dados públicos (FILHO, 2002). A Biologia Molecular busca encontrar a relação do código genético com o aparecimento de doenças como, por exemplo, o câncer.

Para Christopher Greenman (2007), os cânceres surgem devido a mutações em um subconjunto de genes e conferem a eles, vantagem de crescimento. A disponibilidade da sequência do genoma do ser humano levou a criação de uma proposta de que o sequenciamento sistemático de genomas do câncer de mutações levaria à descoberta de vários genes de cânceres adicionais. Nesta proposta se tem o relato de mais de 1.000 mutações somáticas encontradas em 274 megabases (MB) de DNA correspondentes aos exons que são os codificadores de 518 genes da quinase de proteína em 210 diferentes tipos de câncer humano. Através da pesquisa citada, concluiu-se que houve uma variação significativa no número e no padrão de mutações

em cânceres individualmente, refletindo diferentes exposições, defeitos de reparo do DNA e as origens celulares. A maioria das mutações somáticas é susceptível de serem "passageiros", ou seja, que não possuem contribuição para a oncogênese (criação do câncer). No entanto, houve evidência de mutações "condutoras", que contribuem diretamente para o desenvolvimento dos cânceres estudados, cerca de 120 genes. A sistemática do sequenciamento de genomas do câncer, portanto, revela a diversidade evolutiva de cânceres e implica diretamente na conclusão da existência de um maior repertório de genes do câncer do que era anteriormente previsto (GREENMAN et al., 2007).

2.3. GENE

O gene pode ser definido como a unidade fundamental, física e funcional da hereditariedade. Um gene é uma sequência ordenada de nucleotídeos localizada numa posição particular de um cromossomo, que codifica um produto funcional, uma proteína ou uma molécula de RNA (MAGATÃO e JUNIOR, 2008, Pg. 15).

Em 1859, Charles Darwin afirmou que a evolução e a seleção natural promoviam a adaptação natural das espécies ao meio ambiente, embora não soubesse quais os mecanismos básicos desta adaptação. O processo de transmissão da informação genética ainda era totalmente desconhecido, pois somente no século XX, o padre Gregor Mendel compreendeu que o processo de transmissão das características positivas estava associado a uma unidade básica de informação, o gene (DARWIN, 2004).

O processo que possibilitou a descoberta de como estas características eram armazenadas durou quase cem anos para ser concluído. Somente em 1869 o bioquímico suíço Friedtich Mieschner concluiu que os núcleos celulares possuíam várias substâncias específicas. Estas substâncias poderiam ser separadas em duas categorias, as proteínas e as moléculas ácidas. Estas moléculas ainda eram desconhecidas e receberam o nome de Ácidos Nucléicos (LINDEN, 2008).

Já em 1909, o russo Phoebus A. T. Levene, que também estudava os ácidos nucleicos, identificou a ribose como açúcar de um dos dois tipos de ácidos nucleicos, e chamou de ácido ribonucleico, e certos componentes do outro, o ácido

desoxirribonucleico. Levene também identificou corretamente a estrutura do DNA, dada por fosfato-açúcar-base (LINDEN, 2008).

Levene e muitos de seus colegas estavam convencidos de que, com ácidos nucleicos e proteínas no núcleo, as complexas e abundantes moléculas de proteínas armazenavam todas as informações genéticas nos cromossomos. Esta teoria sobre o propósito do DNA, de meramente manter unidas as moléculas de proteína, revelou-se incorreta (LINDEN, 2008).

Em 1928, os bacteriologistas Fredrick Griffith e Oswald Avery, deram início ao trabalho que levou à correção dessa suposição equivocada. Após longas pesquisas, demonstraram que o DNA possuía as informações da hereditariedade e não o RNA ou as proteínas. Nesta época, entretanto, não se sabia exatamente como o processo funcionava. Somente após o trabalho de Francis Crick e James Watson, que desvendaram a dupla hélice do DNA e a maneira como os ácidos nucleicos se ligam dentro desta molécula, foi que tudo se esclareceu (CESAR, 2005).

Segundo o trabalho de Crick e Watson, as duas cadeias helicoidais antiparalelas, com a “coluna vertebral” de açúcar e fosfato na parte externa e as bases (adenina, timina, guanina e citosina) no interior como demonstra a Figura 1. Devido aos ângulos em que as substâncias químicas do DNA se ligam umas às outras, todas as moléculas de DNA consistem em duas faixas paralelas espiraladas, como corrimão de uma escada em espiral – daí o nome que imediatamente se celebrou com a descoberta de Crick-Watson: a hélice dupla.

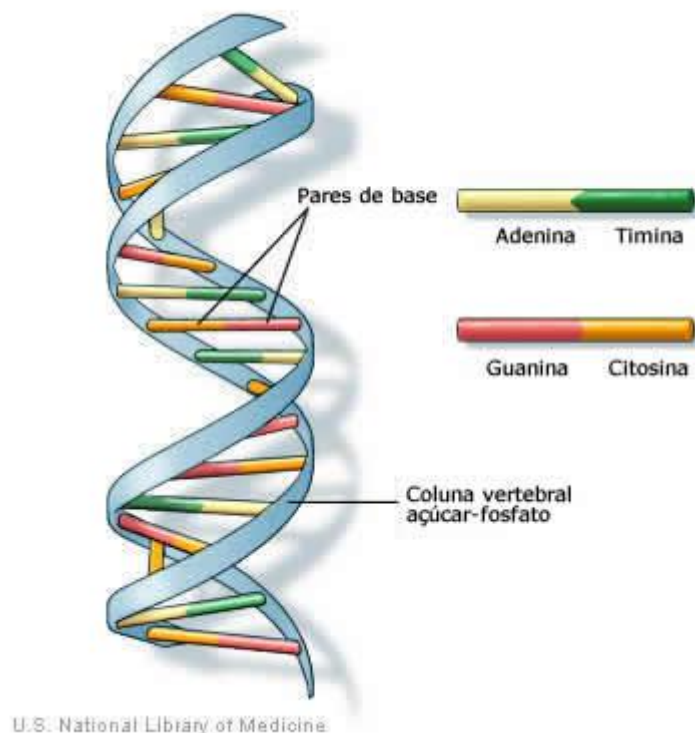


Figura 1 - Estrutura do DNA [NLM, 2009]

Sinteticamente, cada indivíduo, animal, vegetal ou até mesmo fungos e bactérias, é formado por uma ou mais células, e dentro de cada uma delas, o organismo possui uma cópia completa do conjunto de um ou mais cromossomos que definem o organismo, e a este conjunto deu-se o nome de genoma (LINDEN, 2008).

2.4. DNA E RNA

Segundo Rocha et al. (2003), o DNA foi descoberto como principal substância química do núcleo, aproximadamente ao mesmo tempo em que Mendel e Darwin publicaram seus trabalhos. Entretanto, no início do século 20, as proteínas eram mais consideradas como moléculas capazes de transmitir grandes quantidades de informação hereditária ao longo das gerações. Embora o DNA fosse conhecido por ser uma molécula muito grande, pensava-se que seus quatro componentes químicos estivessem unidos num padrão monótono como um polímero sintético. Além disso, até então não se achava nenhuma função celular específica para o DNA. Por outro lado, as proteínas

eram importantes como enzimas e como componentes estruturais de células vivas. As proteínas também eram conhecidas como polímeros de vários aminoácidos. Estes polímeros são chamados polipeptídios. Mas mais importante é que o alfabeto de proteínas de 20 aminoácidos possivelmente podia formar estruturas transportadoras de informação mais singulares do que o alfabeto de apenas quatro letras do DNA.

A estrutura do DNA é composta de duas fitas que são longos polímeros formados por milhões de nucleotídeos ligados uns aos outros. Cada nucleotídeo consiste de um açúcar (desoxirribose) ligado a um lado para um grupo de fosfato e ligado ao outro lado para uma base de nitrogênio. Existem duas classes de bases de nitrogênio chamadas purinas (estruturas aneladas duplas) e pirimidinas (estruturas aneladas simples). As quatro bases no alfabeto do DNA são: Adenina (A) - purina Guanina (G) - purina Citosina (C) - pirimidina Timina (T) – pirimidina (ROCHA et al., 2003).

O RNA é uma molécula também formada por um açúcar (ribose), um grupo fosfato e uma base nitrogenada (U) uracila, (A) adenina, (C) citosina ou (G) guanina. Um grupo reunindo um açúcar, um fosfato e uma base é um "nucleotídeo".

Os aminoácidos são os componentes básicos das proteínas. Eles se organizam ao redor das quatro ligações do átomo de carbono, considerando, é claro, que o carbono seja tetra valente. Isto significa que o carbono possui quatro elétrons sem par na casca externa, e isso lhe permite fazer essas ligações e o torna o átomo e o elemento químico mais importante da biologia. Embora existam apenas vinte variedades de aminoácidos, longas repetições de sequências múltiplas permitem dezenas de milhares de combinações de aminoácidos para formar uma grande variedade de proteínas. Realmente, existem cerca de 50 mil tipos diferentes de proteínas no corpo humano. Os mesmos vinte aminoácidos em 50 mil combinações diferentes estão ligados aos outros em longas cadeias dobradas sobre si mesmas (CESAR, 2005).

Para Vogt (2003), o DNA e o RNA são considerados a última grande e revolucionária descoberta científica da humanidade dos últimos 50 anos, permitindo a abertura de novos caminhos para o desenvolvimento das ciências da vida e para o

nascimento de áreas multidisciplinares de estudo e pesquisa, antes desconhecidas. A própria biologia, com o assombroso desenvolvimento da genômica e, mais recentemente, da proteômica, vem observando transformações que têm mudado o seu paradigma teórico e metodológico. Aproximando-a assim, sob esses aspectos, das chamadas ciências duras, para as quais a materialidade de seu objeto e a quantificação de seu conhecimento, são condições constitutivas do rigor dos procedimentos e da verdade dos resultados produzidos pela investigação (VOGT, 2003).

A informação contida no DNA, o código genético, está registrada na sequência de suas bases na cadeia (Timina sempre ligada à Adenina, e Citosina sempre com Guanina). A sequência indica outra sequência, a de aminoácidos, substâncias que constituem as proteínas. O DNA não é o fabricante direto das proteínas; para isso ele forma um tipo específico de RNA, o RNA mensageiro, no processo chamado transcrição. O código genético, na forma de unidades conhecidas como genes, está no DNA, no núcleo das células. Já a "fábrica" de proteínas fica no citoplasma celular em estruturas específicas, os ribossomos, para onde se dirige o RNA mensageiro. Na transcrição, apenas os genes relacionados à proteína que se quer produzir são copiados na forma de RNA mensageiro.

Cada grupo de três bases (ACC, GAG, CGU etc.) é chamado códon e é específico para um tipo de aminoácido. Um pedaço de ácido nucléico com cerca de mil nucleotídeos de comprimento pode, portanto, ser responsável pela síntese de uma proteína composta por centenas de aminoácidos. Nos ribossomos, o RNA mensageiro é por sua vez lido por moléculas de RNA de transferência, responsável pelo transporte dos aminoácidos até o local onde será montada a cadeia proteica. Essa produção de proteínas com base em um código é a base da Engenharia genética.

2.5. ADENOMA

Os adenomas colorretais são tumores benignos com um potencial para desenvolvimento de câncer de colo de intestino em um período de 5 a 15 anos (SALCES et. al.,2004). Desta forma acredita-se que o câncer de colo de intestino, na

maioria dos casos, seja precedido por adenomas. Segundo Giancarlo Marra (2007), embora o adenoma seja uma lesão pré-cancerosa, eles foram submetidos a uma pesquisa clínica extensa, em uma análise patológica e molecular. Pouco se sabe sobre as alterações de expressão gênica global acompanham a sua formação.

Para caracterizar os processos moleculares subjacentes à transformação do epitélio do cólon normal, foram comparados os transcriptomas coletados prospectivamente de 32 adenomas com os de mucosa normal dos mesmos indivíduos. Surgiram diferenças importantes, não só entre a expressão dos perfis dos tecidos normais e adenomatosos, mas também entre os adenomas pequenos e grandes (MARRA, 2007).

Os perfis de transcriptoma normal da mucosa do cólon e adenomas do colo de intestino lançam uma nova luz sobre estágios iniciais da tumorigênese colorretal, que é o nascimento dos tumores de colo de intestino (ARBER, 2006).

2.6. DESCOBERTA DO CONHECIMENTO

Nas últimas décadas, a capacidade de gerar e armazenar dados vem aumentando com extrema rapidez. Este crescimento descontrolado na quantidade de dados armazenados induziu a necessidade de evoluir na busca de novas técnicas e ferramentas automatizadas, que pudessem auxiliar na transformação desses dados em informação útil e conhecimento. A quantidade de dados disponíveis em repositórios, em conjunto com a necessidade por ferramentas de análise, é conhecida como uma situação rica em dados, mas pobre em informações (HAN e KAMBER, 2001).

Percebeu-se, no decorrer dos últimos anos, uma desproporção, na quantidade de dados produzidos e a quantidade de dados compreendidos, o que gerou uma crescente expectativa de que os dados, após analisados e apresentados de forma inteligente, tornaram-se um recurso valioso, podendo ser usado como uma vantagem competitiva (FRAWLEY et al., 1992). Assim nasceu a Descoberta de Conhecimento em Bases de Dados ou KDD, do termo em inglês *Knowledge Discovery in Databases*. KDD foi

inicialmente definida como a extração de informação implícita, previamente desconhecida e potencialmente útil a partir de dados (FRAWLEY et al., 1992).

Mesmo que muitas e diferentes informações possam ser descobertas nos dados, o foco sempre foi referente aos padrões expressos nas linguagens de alto nível. Linguagem de processamento natural é muitas vezes desejável, pela perspectiva humana, mas nem sempre é conveniente para ser utilizada na manipulação pelos algoritmos de descoberta. Representações lógicas são mais naturais para a computação e, quando necessário, podem ser traduzidas para um formato em linguagem de processamento natural (FRAWLEY et al., 1992).

A definição inicial de KDD foi posteriormente revisada:

“Descoberta de conhecimento em bases de dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em repositórios de dados.” (FAYYAD et al., 1996a,).

Assim definido, o termo processo indica que KDD compreende duas fases:

- Informação: Qualquer fato ou conhecimento do mundo real que pode ou não ser registrado e/ou armazenado (ALVES, 2004);
- Dado: é um conjunto de fatos, é a representação da informação, que pode estar registrada (ALVES, 2004), (e.g. registros de uma base de dados (a idade de uma pessoa));

Padrões podem ser expressões de alguma linguagem específica, que possa descrever um subconjunto dos dados, ou ainda pode ser um modelo que possa ser aplicado a subconjuntos dos dados; os padrões que porventura sejam descobertos devem ter validade nos novos dados, com certo grau de certeza; é sempre desejável que estes padrões descobertos sejam ‘novos’ para o usuário, e sejam também ‘potencialmente úteis’, que possam levar a algum benefício; finalmente, estes padrões precisam ser ‘compreensíveis’, se não de imediato, então, após ser feito algum pós-processamento (FAYYAD et al., 1996a, 1996b, 1996c).

A área de KDD evoluiu, e continua a evoluir, da intersecção de áreas de pesquisa tais como aprendizado de máquina, reconhecimento de padrões, banco de dados, estatística e visualização de dados. Pode ser imaginada como a confluência dessas disciplinas (FAYYAD et al., 1996b).

KDD se baseia fortemente em técnicas conhecidas de aprendizado de máquina, de reconhecimento de padrões e de estatística para encontrar os padrões nos dados. A estatística também oferece métodos de quantificação da incerteza inerente, quando se procura inferir padrões gerais a partir de amostras de uma população. As técnicas de visualização de dados estimulam naturalmente a percepção e a inteligência humana, aumentando a capacidade de entendimento e de associação de novos padrões (MONARD et al., 2003). O termo “mineração de dados” é muitas vezes usado como um sinônimo de KDD. Alternativamente, a mineração de dados é considerada uma etapa essencial do processo de KDD (HAN e KAMBER, 2001). Segundo Fayyad et al. (1996b), KDD refere-se ao processo global de descoberta de conhecimento a partir de dados, enquanto a mineração de dados é uma fase desse processo. Por essa visão, a mineração de dados refere-se à aplicação de algoritmos específicos para extrair os padrões dos dados. As outras fases do processo são também importantes para se garantir que conhecimento útil seja derivado dos dados.

A mineração de dados está intimamente associada à noção de extração de conhecimento a partir de um grande volume de dados. Entretanto, o processo de KDD pode ser realizado independentemente da quantidade de dados disponível, em todas as suas fases.

A Figura 2 apresenta uma visão geral das fases do processo de KDD (FAYYAD et al., 1996a). Primeiro, antes de se começar a tratar os dados, é preciso compreender o domínio de aplicação e identificar a meta da descoberta de conhecimento pelo ponto de vista do usuário; em seguida, os dados de interesse são selecionados, é feito um pré-processamento nos dados (e.g. eliminação de ruídos e tratamento de dados ausentes), os dados sofrem transformações (e.g. conversão de dados e derivação de novos atributos) e é realizada então, a mineração de dados (extração de padrões); ao final, é feita a interpretação e a avaliação dos resultados obtidos, e o conhecimento descoberto é

distribuído conforme se tenha definido no planejamento. Esse processo pode envolver várias iterações e quase sempre é necessário o retorno para fases anteriores.

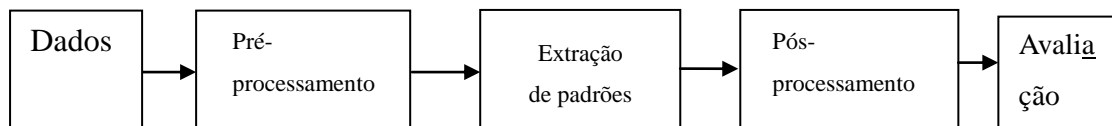


Figura 2 - Visão geral das fases do processo de KDD adaptado de (FAYYAD et al., 1996a).

Para Monard et al., (2003) o processo de mineração de dados é dividido em três etapas: pré-processamento, extração de padrões e pós-processamento. Inclui-se, ainda, uma fase anterior ao processo, referente ao conhecimento do domínio e à identificação do problema, e outra fase posterior, relacionada com a utilização do conhecimento obtido. O processo de KDD é centrado na cooperação entre os seus diversos atores, e o seu sucesso depende, em parte, dessa cooperação. Os atores do processo podem ser divididos em três classes (MONARD et al., 2003):

- **Especialista do domínio:** pessoa que deve possuir amplo conhecimento do domínio de aplicação e deve fornecer apoio para a execução do processo.
- **Analista de dados:** pessoa responsável pela execução do processo de KDD. Este usuário deve conhecer a fundo as etapas que compõem o processo.
- **Usuário final:** representa a classe de usuários que vai utilizar o conhecimento extraído como auxílio em um processo de tomada de decisão.

2.7. APRENDIZADO DE MÁQUINA

Neste capítulo serão apresentadas definições sobre o aprendizado de máquina. Este pode ser do tipo indutivo ou do tipo não indutivo, conhecido por dedutivo. O aprendizado de máquina do tipo indutivo emprega o princípio de inferência chamado

indução. Desta forma podemos obter conclusões genéricas partindo de um conjunto de exemplos. Neste aprendizado, chamado indutivo, podemos dividir em dois grupos principais, o aprendizado supervisionado e o aprendizado não supervisionado.

No aprendizado supervisionado, possuímos a figura de um professor externo, ou também chamado de indutor, que apresenta o conhecimento do ambiente através de conjuntos de exemplos da seguinte forma: entrada e saída desejada (HAYKIN, 1999). O algoritmo de aprendizado de máquina deve extrair a representação do conhecimento a partir destes conjuntos de exemplos, cujo objetivo é que a representação gerada possa ser capaz de produzir saídas corretas para quaisquer entradas que possam ser apresentadas (HAYKIN, 1999).

O aprendizado não supervisionado não possui esta presença do professor, logo não existem exemplos rotulados. Neste caso o algoritmo de aprendizado de máquina deve aprender a representar ou agrupar as entradas, que foram submetidas através de uma medida de qualidade. Utilizam-se estas técnicas principalmente quando se tem por objetivo, encontrar padrões ou tendências que auxiliem no entendimento dos dados (SOUTO, 2003).

Neste trabalho utilizou-se o aprendizado supervisionado. Neste caso, dado um conjunto de exemplos rotulados na forma $(x_i; y_i)$, em que x_i representa um exemplo e y_i denota o seu rótulo, deve-se produzir um classificador, também denominado modelo, preditor ou hipótese, capaz de prever precisamente o rótulo de novos dados. Esse processo de indução de um classificador a partir de uma amostra de dados é denominado treinamento. O classificador obtido também pode ser visto como uma função f , a qual recebe um dado x e fornece uma predição y (LORENA et al., 2007), como pode-se observar na Figura 3.

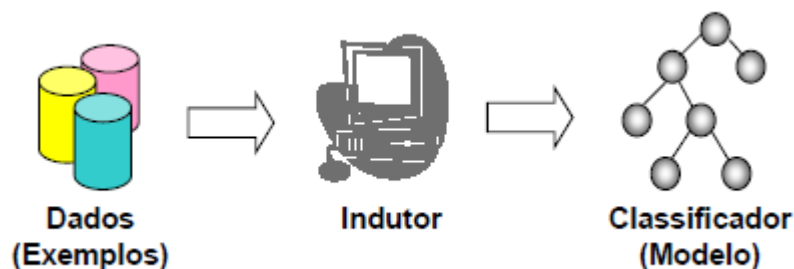


Figura 3 - Aprendizado de máquina Supervisionado, Adaptado de Lorena et al., 2007

2.8. DATA MINING

“A mineração de dados é o processo de descoberta de novas correlações significativas, padrões e tendências de peneirar grandes quantidades de dados armazenados em repositórios, usando tecnologias de reconhecimento de padrões, bem como técnicas de estatística e matemática” (LAROSE, 2005, Pg.2).

Segundo Goebel e Gruenwald (1999), o termo KDD é usado para representar o processo de tornar dados de baixo nível em conhecimento de alto nível, enquanto mineração de dados pode ser definida como a extração de padrões ou modelos de dados observados.

Na prática, como se pode observar na Figura 4, os dois objetivos principais da mineração de dados são a predição e a descrição. A predição envolve o uso de variáveis com valores conhecidos para prever um valor desconhecido ou futuro de outra variável (atributo meta). A descrição caracteriza propriedades gerais encontradas nos dados, com foco em padrões interpretáveis pelo ser humano. Esses objetivos podem ser alcançados por meio de vários tipos de tarefa. A escolha de uma ou mais tarefas depende do problema em questão. As tarefas tradicionais de mineração de dados estão representadas na Figura 4 e são brevemente descritas a seguir (HAN e KAMBER, 2001; FAYYAD et al., 1996b).

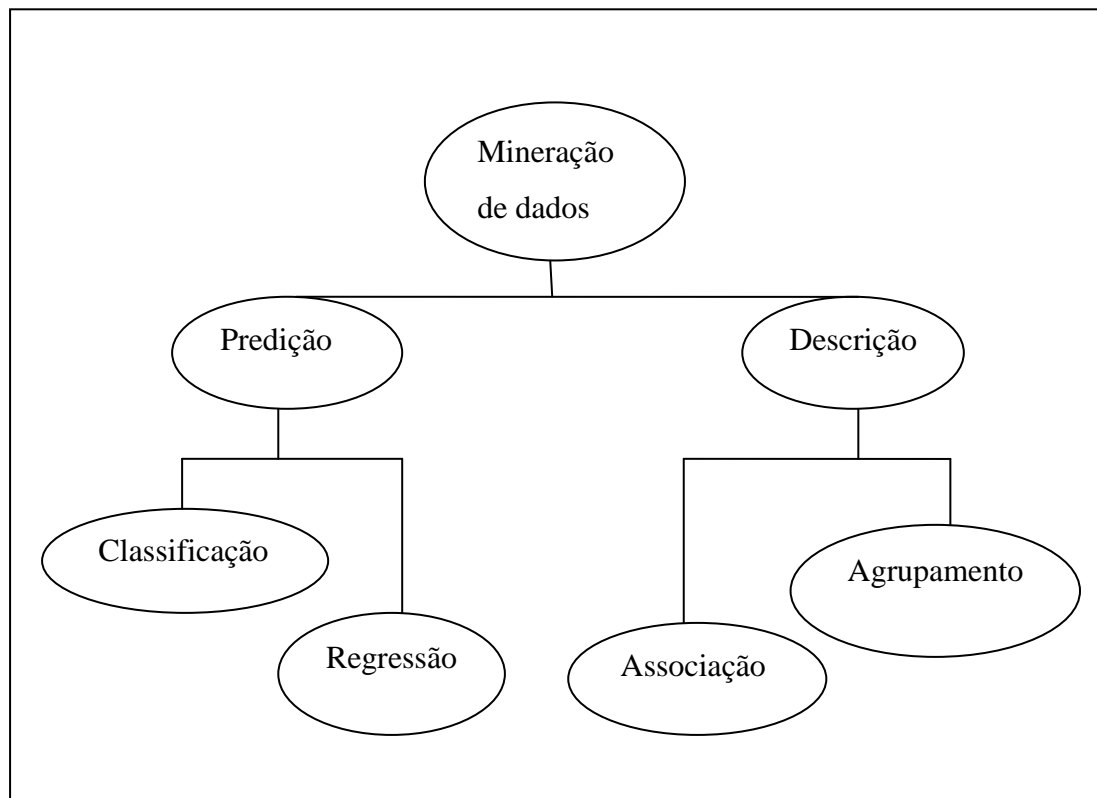


Figura 4 - Objetivos e tarefas de mineração de dados (adaptada de REZENDE et al., 2003).

- **Classificação:** Para Goldschmidt (2005), a classificação tem como ideia principal, examinar as características de um objeto ou situação de forma a atribuir a ela uma classe pré-definida. Assim, a classificação consiste na construção de modelos que permitam o agrupamento de dados em classes. Esta é uma tarefa considerada preditiva, pois uma vez que as classes são definidas, ela pode prever automaticamente a classe de um novo dado.
- **Regressão:** Para Dunham (2003), a regressão pode ser definida como uma técnica que utilizando uma saída de dados tenta encontrar a fórmula utilizada para sua classificação. É quase que como uma equação matemática, porém não se resolve a equação para determinar o resultado, mas sim, através do resultado, busca-se a equação. Muitas técnicas são

utilizadas para encontrar a origem da saída, a divisão dos dados em regiões é uma forma de classificar os dados baseados em características do conjunto de dados.

- **Associação:** é uma técnica cujo objetivo é a descoberta de regras de associação de acordo com atributos que venham a ocorrer com frequência e em conjunto. Por exemplo: uma análise das transações de compra de um supermercado pode encontrar itens que tendem a ocorrer juntos em uma mesma compra, como fraldas e cervejas. Os resultados desta análise podem ser úteis na hora de distribuir estes produtos nas prateleiras. Desta forma os clientes que adquiram um produto, visualizem o outro com facilidade. Esta é uma tarefa considerada descritiva, pois objetiva encontrar padrões em dados históricos.

- **Agrupamento (*clustering*):** Agrupamento ou análise de grupos consiste em agrupar os dados em classes ou *clusters*, tal que os elementos de uma classe tenham alta similaridade entre si e sejam diferentes dos elementos das outras classes. Devem também ter um significado dentro do grupo, e serem úteis (TAN, 2009). Ao contrário da classificação, o rótulo de classe de cada elemento não é conhecido de antemão. O próprio algoritmo descobre as classes a partir das alternativas encontradas na base de dados, agrupando então um conjunto de objetos em classes de objetos semelhantes.

A mineração de dados combina métodos e ferramentas das seguintes áreas: aprendizagem de máquina, estatística, banco de dados, sistemas especialistas e visualização de dados (CRATOCHVIL, 1999).

Mineração de dados é a exploração e a análise, por meio automático ou semiautomático, de grandes quantidades de dados, a

fim de descobrir padrões e regras significativos (BERRY e LINOFF, 1997, p.5).

Os principais objetivos da mineração de dados são descobrir relacionamentos entre dados e fornecer subsídios para que possa ser feita uma previsão de tendências futuras baseada no passado.

Os resultados obtidos com a mineração de dados podem ser usados no gerenciamento de informação, processamento de pedidos de informação, tomada de decisão, controle de processo e muitas outras aplicações.

A mineração de dados pode ser aplicada de duas formas: como um processo de verificação e como um processo de descoberta (GROTH, 1998). No processo de verificação, o usuário sugere uma hipótese acerca da relação entre os dados e tenta prová-la aplicando técnicas como análises estatísticas e multidimensionais sobre um banco de dados contendo informações passadas. No processo de descoberta não é feita nenhuma suposição antecipada.

Esse processo usa técnicas, tais como descoberta de regras de associação, árvores de decisão, algoritmos genéticos, redes neurais e máquinas de vetores de suporte.

2.9. AMOSTRAGEM

A amostragem é um conjunto de atividades que realizadas de forma adequada, permitem inferências a respeito das propriedades de um universo maior, a partir do estudo de uma pequena porção deste, a amostra. Utilizar a totalidade dos dados disponíveis provocaria uma carga muito penosa aos recursos computacionais disponíveis. Assim, nesta dissertação, foi adotada a utilização de uma amostra do universo de dados disponíveis.

Segundo Tan (2009), quando se pretende analisar um conjunto de dados, é bastante comum à utilização da amostragem. Em estatística, por exemplo, é extensamente utilizada não somente para uma investigação preliminar dos dados quanto na parte de análise final dos mesmos.

Ainda conforme Tan (2009) a amostragem também pode ser utilizada na mineração de dados. Entretanto “*as motivações para amostragem em estatística e mineração de dados são muitas vezes diferentes*” (TAN, 2009, pg. 56).

O fato é que os Estatísticos utilizam a amostragem em função de que se obter um volume grande de dados é extremamente difícil e com custo muito elevado. Já na mineração de dados, este custo pode ficar na capacidade de processamento dos computadores, pois quanto mais se aumenta a quantidade de dados disponíveis, mais próximo da verdade se estará, conclui Tan (2009).

Em Guyon (2002), observa-se que em seu recente trabalho, por não haver uma determinação de quantidade de número do conjunto de testes e do conjunto de treino, ela optou por cinquenta por cento para treinar o algoritmo e aplicou os testes nos cinquenta por cento restantes.

O princípio chave para amostragens eficazes é o seguinte: Usar uma amostra funcionará quase tão bem quanto usar o conjunto inteiro de dados se a amostra for representativa. Por sua vez, uma amostra é representativa se tiver aproximadamente a mesma propriedade (de interesse) do conjunto original de dados. Se a média dos objetos de dados for a propriedade de interesse, então uma amostra é representativa se tiver uma média que seja próxima àquela dos dados originais (TAN, 2009, p56).

2.10. BANCOS DE DADOS

Segundo Silberschatz (2006), um Banco de dados é uma coleção de informações, que sejam relevantes a um determinado assunto, como uma empresa, por exemplo. Ainda segundo Silberschatz (2006), um sistema gerenciador de banco de dados (DBMS) é uma base de informações inter-relacionadas e um conjunto de programas para acessar estas.

Para Waymire (1999), um banco de dados pode ser imaginado como um conjunto de dados relacionados, mas não apenas em um arquivo, mas em vários. É na verdade um conceito lógico, baseado em um conjunto de objetos relacionados.

Dentre os Sistemas Gerenciadores de Bancos de dados (DBMS), existe uma variedade de softwares disponíveis no mercado, como o MYSQL, o Microsoft SQL Server, o Oracle, etc.. Neste projeto foi escolhido o MYSQL, por ser um DBMS relacional de código aberto, estando disponível nos sistemas operacionais Windows e Unix. Também influenciou esta escolha o fato de o MYSQL possuir um conjunto de recursos rico e completo (GIBAS e JAMBECK, 2002, Pg. 371).

2.11. BANCOS DE DADOS BIOLÓGICOS

De acordo com Gibas e Jambeck (2002), a internet mudou de forma substancial a maneira como os cientistas buscam e trocam informações. Dados que antigamente precisavam ser escritos no papel, hoje são digitalizados e distribuídos de forma rápida, a partir de bancos de dados centralizados. Revistas acadêmicas são publicadas diretamente na internet e praticamente, todos os grupos de pesquisas têm uma página na internet.

Segundo Arthur Lesk (2008), o nosso conhecimento sobre sequências e estruturas biológicas está muito longe de ser chamado de completo, mesmo assim, ele já apresenta um tamanho respeitável e continua em franca expansão. Há um numero muito grande de cientistas, trabalhando na geração de dados ou na execução de projetos de pesquisa, na parte de análise de resultados. A armazenagem e redistribuição destes dados ficam a cargo de companhias que mantêm estes bancos de dados específicos.

Ainda de acordo com Arthur Lesk (2008), a área de bioinformática originalmente armazenava seus dados através de grupos de pesquisas individuais, que tinham sua motivação no interesse da ciência associada àqueles dados. Com o aumento do número de pesquisadores e também da capacidade dos equipamentos em gerar informações, este armazenamento teve sua responsabilidade atribuída a projetos específicos nacionais e até mesmo internacionais.

A quantidade de bancos de dados públicos, em Biologia Molecular disponibilizados na internet, vem crescendo de forma exponencial nos últimos anos, e o aspecto funcional da bioinformática é a representação, o armazenamento e a distribuição de dados. Os objetivos destes bancos variam e podem ser utilizados para armazenar e

disponibilizar bioseqüências, funções moleculares, estruturas de proteínas, modelos metabólicos, entre outros, oferecendo em alguns casos informações mais amplas, ou seja, cobrindo uma área mais significativa da Biologia, e em outros, informações menos detalhadas. Em muitos casos, as informações biológicas são obtidas através da análise computacional de outros bancos de dados; em outros casos, através da literatura ou até mesmo, por informações definidas por pesquisadores (LESK, 2008).

Há muito tempo, os cientistas tem confiado na qualidade dos artigos publicados em revistas científicas impressas, em função de estas revistas contratarem revisores. Esses revisores comentam os manuscritos e muitas vezes, sugerem revisões ou adendos antes do artigo ser aceito para a publicação. Estas revistas, por sua vez, cada vez mais estão publicando suas edições em formato eletrônico, disponibilizando estas informações através da internet.

Outra tendência muito utilizada é a de as revistas não estarem mais mantendo suas versões impressas, e publicando diretamente seus artigos na internet. Nesta mesma linha de publicação, existe um servidor público para pesquisa de literatura científica sobre publicações biológicas, patrocinado pelo NCBI (*National Center for Biotechnology Information* – Centro Nacional de Informações sobre Biotecnologia) da Biblioteca Nacional de Medicina dos Estados Unidos. Este servidor permite buscas aos seus bancos de dados, gratuitamente, através de um navegador web. Existem outros bancos de dados biológicos de boa qualidade na internet, mas na grande maioria, não são gratuitos.

Segundo Gibas e Jambeck (2002), um dos grandes problemas enfrentados pela biologia molecular é a nomenclatura dos genes, que comumente são conhecidos por nomes não sistemáticos. Alguns genes podem ter nomes como *flightless*, *shaker* e *antennapedia*, isto devido aos efeitos no desenvolvimento que eles podem causar em algum animal específico. Outros nomes são escolhidos por biólogos celulares e representa a própria função do gene em nível celular, como o *homeobox*, esse tipo de nomenclatura confusa significa que geralmente, após um cientista que trabalhe com um gene específico, o produto genético ou o processo bioquímico do qual ele é parte pode reconhecer imediatamente o que o nome comum do gene se refere.

Alguns dos bancos de dados biológicos públicos mais conhecidos são o PUBMED, o GENBANK e o GEO, como ferramenta de busca entre os bancos de dados biológicos mantidos pelo NCBI, o ENTREZ é muito eficiente.

O PUBMED é um dos bancos de dados disponíveis na internet, mantido pelo NCBI com um dos maiores acervos de artigos científicos para biólogos. Possui a indexação de mais de 4.000 revistas científicas, incluindo as mais respeitadas revistas sobre biologia celular e molecular, bioquímica, genética e áreas afins (GIBAS e JAMBECK, 2002).

Segundo Lesk (2008), uma das características do PUBMED é a possibilidade de efetuar buscas por artigos relacionados. Sendo esta uma maneira muito rápida de se atualizar da literatura a respeito do tópico. Nos dias atuais, a maioria dos periódicos científicos disponibiliza suas listas de conteúdos e, na grande maioria das vezes, até suas edições completas, em sites da Internet (LESK, 2005).

O sistema ENTREZ oferece acesso por meio das seguintes divisões de bancos de dados: Proteínas, Peptídeos, Nucleotídeo, Gene, Estrutura, Genoma, dentre outros. “Um dos pontos fortes deste sistema do NCBI são a sua conexão entre os vários bancos de dados” (LESK, 2005). ENTREZ é o ponto inicial para que se possam recuperar as sequências de estruturas. Pode-se acessar a ferramenta ENTREZ através do endereço eletrônico [HTTP://www.ncbi.nlm.nih.gov/Entrez/](http://www.ncbi.nlm.nih.gov/Entrez/).

Segundo o próprio endereço eletrônico (GEO, 2010), O GEO (Gene Expression Omnibus) é um repositório público funcional de dados sobre genoma, onde são distribuídos gratuitamente dados de sequências genômicas e de micro matrizes. Além do armazenamento de dados e de aplicativos, uma coleção de interfaces baseadas na web está disponível para ajudar os usuários a fazer consultas e facilitar os *downloads* das experiências e dos padrões de expressão gênica armazenada no GEO.

O banco de dados GEO recebe dados das mais diferentes fontes. Tenta disponibilizar estes dados de forma simples para que outros pesquisadores possam tratá-los de forma acessível e prática (GEO, 2010).

O banco de dados público GEO é mantido pelo NCBI (*National Center for Biotechnology Information*), que é um órgão americano que maior número de bases de dados públicas mantém.

2.12. MICROARRANJOS (*MICROARRAY*)

Neste capítulo será apresentada a definição de microarranjos (*microarrays*) e seu relacionamento com a bioinformática.

Os Microarrays de DNA (ou chips de genes) são laboratórios miniaturizados para o estudo da expressão gênica. Em cada chip contém um arranjo projetado deliberadamente com sondas moleculares que podem se vincular a segmentos específicos de DNA ou em mRNA. A marcação do DNA ou do RNA com moléculas fluorescentes permite que o nível de expressão de qualquer gene em uma preparação celular seja mensurado quantitativamente. Os *microarrays* (microarranjos) também tem outras aplicações na biologia molecular, mas sua utilização no estudo da expressão gênica inaugurou uma nova forma de mensurar as expressões do genoma (GIBAS E JAMBECK, 2002, Pg. 151 – 152).

Microarranjos de DNA são instrumentos usados para testar a presença de várias sequências de DNA, ao mesmo tempo. Eles podem ser usados para determinar padrões de expressão em diferentes proteínas através da detecção de mRNAs. Podem ainda genotipar, através de diferentes sequências variantes de genes. Incluindo aí, os poliformismos de um único nucleotídeo, embora sem se limitar a ele (LESK, 2005).

Para a Bioinformática, Microarranjos de DNA são, ainda segundo Lesk (2008), outra fonte prolífera de geração de dados. Demandando um planejamento eficaz dos arquivos e dos sistemas de processamento das informações.

2.13. AFFYMETRICS

A empresa Affymetrix, localizada na cidade de Santa Clara, na Califórnia/EUA, foi uma das primeiras a utilizar microarranjos, com um acervo de publicação superior a

vinte mil artigos. Em constante evolução, provê novas ferramentas de análise gênica e também reagentes para exploração, descoberta e validação de testes genéticos. A contribuição da Affymetrix, com uma completa solução para a análise dos estudos do genoma e outros tantos produtos para a análise celular e proteínas, permite e facilita a evolução das pesquisas biomédicas, de acordo com o site <http://www.affymetrix.com>.

No mundo todo, as ferramentas da Affymetrix são utilizadas na luta para compreender a real ligação dos genes com a formação e evolução das doenças e suas formas de tratamento. Descobertas têm acontecido através das tecnologias de microarranjos, auxiliando a ciência na investigação das causas do câncer, do diabetes, do HIV e outras muitas doenças que se disseminam e assustam a população dos mais diversos países. Neste trabalho, foi utilizada a plataforma GPL570 da Affymetrix para identificação dos genes correspondentes as sondas encontradas nos repositórios obtidos no GEO. A plataforma GPL570 é a mais completa dentre as plataformas mantidas pela Affymetrix, por esta razão, foi escolhida.

2.14. SVM

O SVM (do inglês *Support Vector Machine*) é um algoritmo de classificação de alto desempenho que nos últimos anos vem recebendo bastante atenção da comunidade científica na classificação de dados de grande dimensão, inclusive evitando o problema da dimensionalidade, pois supera os demais algoritmos no tempo de resposta e na acuracidade da classificação. O SVM é uma técnica que usa a Teoria do Aprendizado Estatístico (VAPNIK, 1995) e apresenta resultados empíricos promissores em diversas aplicações práticas, como reconhecimento de padrões de imagens, classificação de textos e também na classificação gênica, foco desta dissertação. O SVM, apresenta uma característica que aponta os limites através de subconjuntos de decisão, chamados de vetores de suporte, é um método de classificação baseado em um algoritmo de otimização que define hiperplanos de separação ótimos entre as amostras.

Recentemente, o SVM foi aplicado de forma muito ampla no campo da biologia computacional, em problemas de reconhecimento de padrões, incluindo a análise de expressão de micro matrizes de genes, o reconhecimento dos locais de início de tradução, classificação funcional de regiões promotoras, na previsão de interações

proteína-proteína, e identificação de peptídeos a partir de dados de espectrometria de massa.

SVM é uma recente técnica de Aprendizado de Máquina cujo objetivo inicial esteve ligado à resolução de problemas de Reconhecimento de Padrões. Esta técnica, introduzida por Vapnik (VAPNIK, 1998) mostra-se muito poderosa, e em pouco tempo tem superado muitos sistemas em uma ampla variedade de aplicações (CRISTIANINI; SHAW-TAYLOR, 2000). Sua ideia principal é mapear não linearmente a informação (vetores do espaço de entrada) para um espaço de característica de alta dimensionalidade através de um mapeamento escolhido a priori. Nesse espaço uma superfície de decisão linear é construída, constituindo assim um hiperplano de separação ótima entre exemplos (SOUZA, 2006).

2.14.1. FRONTEIRAS LINEARES

O SVM obtém fronteiras lineares para separação dos dados que pertençam a duas classes distintas, isto é obtido de tal forma que todos os integrantes de uma classe sejam separados da outra, através de um hiperplano de separação. Porém, há a possibilidade de se encontrar muitos hiperplanos distintos que possam separar as classes em duas, precisa-se então identificar qual dos hiperplanos permite a melhor separação, qual deles consegue aumentar a sua largura, e ficar então com a margem maior.

Inicialmente, observa-se na Figura 5, onde existem vários círculos e vários quadrados, a necessidade de serem separados e pode-se imaginar: qual a melhor forma de separar os círculos dos quadrados.

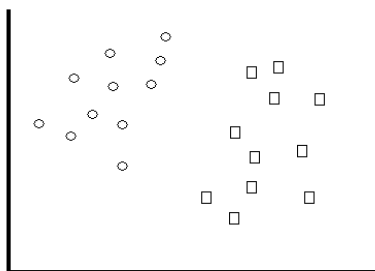


Figura 5 - Conjunto de dados com círculos e quadrados, adaptado de Tan et al., 2009.

Através desta imagem podem-se traçar inúmeras linhas retas que separam perfeitamente as duas classes, como é possível observar na Figura 6.

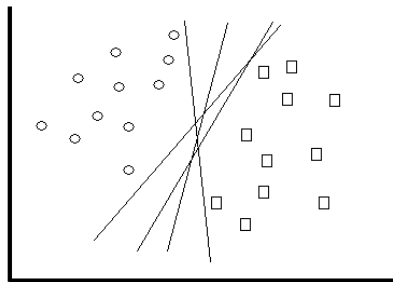


Figura 6 - Círculos e quadrados separados por várias linhas, adaptado de Tan et al. 2009.

Na Figura 6, pode-se observar que todos os hiperplanos têm a capacidade de separar as duas classes. Por esta razão deve-se minimizar a possibilidade do erro de separação. Encontrando-se e escolhendo-se o hiperplano de maior margem, pois não se possui garantias que os hiperplanos sejam executados de forma homogênea. Considerando as diferenças das larguras das margens, o classificador é o responsável pela escolha do melhor hiperplano, por escolher a melhor fronteira de decisão, e que traga os melhores e mais confiáveis resultados. O SVM não cria linhas para separação de classes, mas sim fronteiras lineares para separá-las, e procura encontrar a posição de separação onde a maior margem pode ser localizada (BURGES, 1998), como é possível ver na Figura 7.

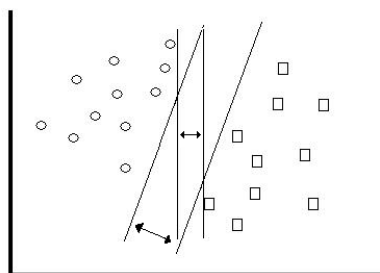


Figura 7- Separação dos Grupos por margens, adaptado de Burges, 1998.

Ainda segundo Burges (1998), estas margens são expandidas ao máximo, até que se encostem a alguns elementos das classes que estão tentando separar, e estes elementos então são chamados de vetores de suporte e a margem de maior distancia é escolhida (Figura 8).

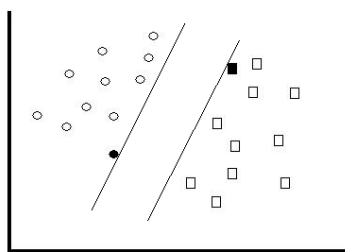


Figura 8- Vetores de Suporte, adaptado de Tan et al., 2009.

Os elementos em pretos (quadrado e círculo) na Figura 8 são então os vetores de suporte e as linhas a margem linear máxima.

Um SVM linear é um classificador que procura um hiperplano com a maior margem, motivo pelo qual muitas vezes é conhecido como classificador de margem máxima (TAN et al., 2009).

Define-se agora como o SVM descobre os limites e decide por qual deles é o melhor.

Analisando-se um problema de classificação linear consistindo de N exemplos de treinamento, cada exemplo é caracterizado por (x_i, y_i) ($i=1,2,...,N$), onde $x_i = (x_{i1}, x_{i2}, ..., x_{id})^T$ corresponde ao conjunto de atributos para o exemplo de índice i , supõe-se então que por convenção, $Y_i \in \{-1, 1\}$ represente o seu rótulo de classe, tendo w e b como parâmetros do modelo, pode-se então definir que o limite de decisão do classificador linear seja:

$$w \cdot x + b = 0 \quad (\text{equação 1})$$

e que para:

$$w \cdot x + b \geq 0 \text{ pertence a classe } y_i = 1,$$

$$w \cdot x + b \leq 0 \text{ pertence a classe } y_i = -1.$$

Seja x_1 um ponto no hiperplano $H_1: w \cdot x + b = +1$ e x_2 um ponto no hiperplano $H_2: w \cdot x + b = -1$, conforme ilustrado na Figura 9. Projetando-se $x_1 - x_2$ na direção de w , perpendicular ao hiperplano separador $w \cdot x + b = 0$, é possível obter-se a distância entre os hiperplanos H_1 e H_2 . Tem-se então um limite de decisão que separa em duas partes iguais os exemplos de treinamento nas suas respectivas classes através da reta sólida $w \cdot x + b = 0$.

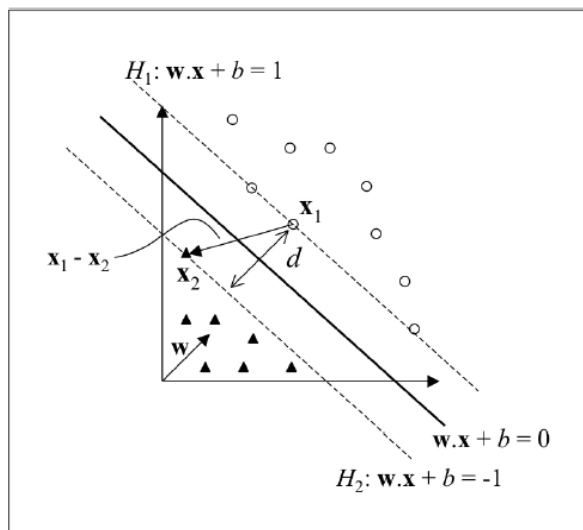


Figura 9- Limites de Decisão, adaptado de Lorena et al. 2007.

Um exemplo localizado pelo limite de decisão deve satisfazer a equação 1, por exemplo, se x_1 e x_2 forem dois pontos localizados no limite de decisão, então:

$$w \cdot x_1 + b = 0,$$

$$w \cdot x_2 + b = 0$$

Subtraem-se as duas equações e tem-se o seguinte resultado:

$$w \cdot (x_2 - x_1) + (b - b) = 0, \text{ logo,}$$

$$w \cdot (x_2 - x_1) = 0$$

onde $x_2 - x_1$ é um vetor paralelo ao limite de decisão e é direcionado de x_1 para x_2 . Como o produto de ponto é zero, a direção de w deve ser perpendicular ao limite de decisão, como se observa na figura 9.

Para quaisquer x_s localizados no limite de decisão, podemos mostrar que:

$$w \cdot x_s + b = k,$$

onde $k' > 0$. De forma semelhante, para qualquer círculo x_c localizado abaixo do limite de decisão, pode-se demonstrar que:

$$w \cdot x_c + b = k'$$

onde $k' < 0$. Se rotularem-se todos os círculos como classe + 1 e todos os triângulos como classe -1, então se pode prever o rótulo de classe y para qualquer exemplo de teste z da seguinte forma:

$$y = \begin{cases} +1, & \text{se } w \cdot z + b > 0; \\ -1 & \text{se } w \cdot z + b < 0. \end{cases}$$

Porque então utilizar o SVM? O SVM se apresenta como um excelente classificador devido a sua base matemática ser definida na TAE (Teoria do Aprendizado Estatístico), que proporciona um classificador, que delimita suas margens evitando o *overfitting* e o *underfitting* como mostra a Figura 10.

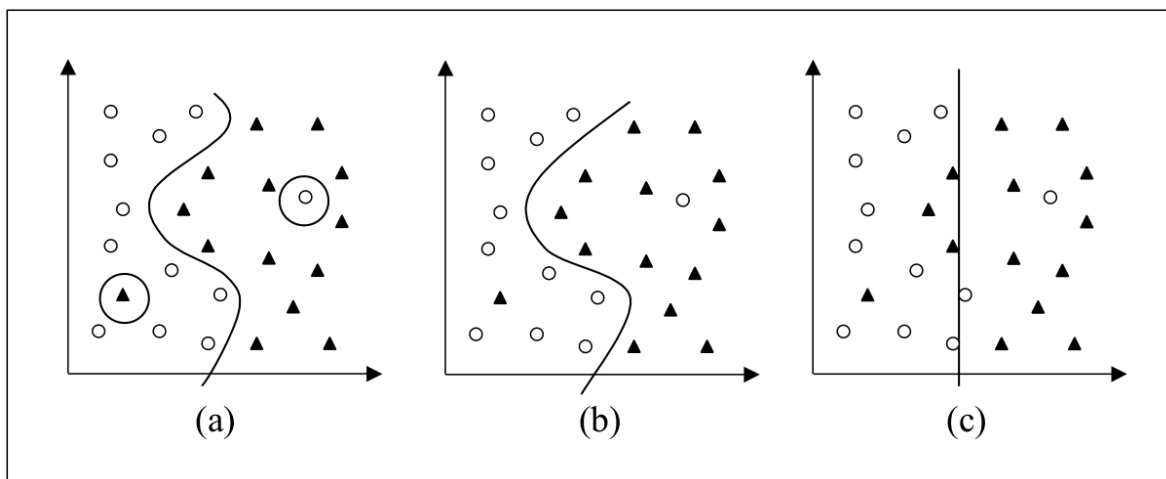


Figura 10 - SVM - adaptado de Lorena et al. 2007 (a) *Overfitting* (b) *SVM* (c) *Underfitting*
O Overfitting exagera na limitação de margens necessitando de grande capacidade computacional, já o Underfitting, economiza nas limitação de margens diminuindo a assertividade, o SVM provê um equilíbrio entre os dois casos.

2.15. TIPOS DE KERNELS SVM

O svm foi desenvolvido baseado na teoria do aprendizado estatístico (TAE) e em alguns casos, sofreu algumas alterações ou adaptações do SVM original. a seguir algumas destas variações:

2.14.1. SVM E-1071 (LIBSVM)

O SVM E-1071 é o pacote SVM desenvolvido para a ferramenta R, atuando de acordo com a TAE, que atualmente causa grande impacto, gerando muito entusiasmo, pois em pouco tempo tem superado muitos sistemas em uma ampla variedade de aplicações (CRISTIANINI; SHAW-TAYLOR, 2000), assim como foi quando se criou as Redes Neurais Artificiais (MEYER, 2001). O SVM representa uma poderosa técnica de classificação não linear, de regressão e detecção de ruídos (*outliers*) com a representação do modelo intuitivo.

O libsvm foi codificado para trabalhar com a ferramenta R e denominado e1071, cuja interface da ferramenta, foi projetada para ser tão intuitiva quanto possível. Os modelos estão enquadrados e novos dados são previstos como de costume, e tanto vetor/matriz quanto interface são implementados. Como o R é um software estatístico,

ele usa um tipo de variável dependente (y): se (y) é uma classe, ele trabalha como um classificador, caso contrario, ele trabalha como uma máquina de regressão, se não houver (y), ele tenta prever as novidades.

2.14.2. SVM – FISCHER

È uma variação do SVM original, utilizado para detectar similaridades entre sequências de proteína. Todos os resultados estatísticos, no entanto, são gerados a partir de exemplos positivos, isto é, a partir de sequências que são conhecidas ou postas a ser evolutivamente relacionadas entre si. Jaakkola et al. (1999) deu início a quarta fase do desenvolvimento de algoritmos de detecção de homologia com um artigo que ganhou o “*Best Paper Award*” na conferência anual “*Intelligent Systems for Molecular Biology*”. Sua visão preliminar é que precisões adicionais podem ser obtidas por modelagem das diferenças entre exemplos positivos e negativos. Porque a tarefa exige discriminar a homologia entre as sequências ligadas e não ligadas, explicitamente, modelando a diferença entre estes dois conjuntos de sequências, torna-se um método extremamente poderoso.

2.14.3. MOTIF KERNELS

Também é uma variação do SVM original, utilizado para detectar similaridade entre sequencias de proteínas. Um inconveniente para este modelo é a simplicidade na representação da proteína (LOGAN et al. 2001), este *kernel* tende a corresponder as regiões funcionalmente importantes das proteínas.

2.15. A FERRAMENTA R

A ferramenta R é na realidade um sistema e uma linguagem de programação desenvolvida a partir da linguagem S (que é usada numa versão comercial – o S-Plus), oriunda dos laboratórios da AT&T no final dos anos 80. Já em meados de 1995, dois professores da Universidade de Auckland, na Nova Zelândia, iniciaram o “Projeto R”, com o objetivo de desenvolver um programa poderoso baseado na linguagem S e de domínio público. O nome R foi baseado nos nomes dos professores que deram início a este projeto, o professor Robert Gentleman e o professor Ross Ihaka do Departamento

de Estatística da Universidade de Auckland em Nova Zelândia, mais conhecidos por “R & R”.

Um fato importante para a difusão desta poderosa ferramenta é a sua compatibilidade com várias plataformas, atualmente o R está disponível para a família UNIX (incluindo LINUX), Windows 95, 98, NT, 2000, Me, XP e também para a maior parte do MacOS. Por esta versatilidade e pela qualidade desta ferramenta, ela foi escolhida para a realização deste trabalho.

O R está disponível na internet no website do CRAN – que é o *Comprehensive R Archive Network* ou “Rede Completa de Arquivos do R”, no seguinte endereço: <http://www.r-project.org>.

2.16. HARDWARE E SOFTWARE

O computador utilizado nesta pesquisa foi um Note Book marca Sony/Vaio, com processador Dual Core 2.16 Ghz, com 3 Gb de memória Ram e HD de 320 Gb.

O sistema Operacional utilizado foi o Microsoft Windows 7 com o pacote Microsoft Office 2010.

O Banco de Dados utilizado foi o MYSQL e o software de classificação foi a ferramenta R com o algoritmo SVM.

3. METODOLOGIA

Nesta dissertação, a metodologia científica utilizada foi do tipo pesquisa-ação combinada com revisão bibliográfica e estudo de caso quantitativo. De acordo com Stringer (1996), a pesquisa-ação é envolvida em uma rotina que possui três ações principais: a observação, para permitir reunir informações e construir um cenário; o pensamento, para explorar, analisar e interpretar os fatos; e a ação, implementando e avaliando as atividades. Dentro desta mesma ideia, pode-se dividir o processo de pesquisa-ação em quatro principais etapas (THIOLLENT, 1997): fase exploratória, onde se pesquisou a teoria da mineração de dados, o impacto das ferramentas e dos algoritmos, fase principal, onde se montou as tabelas, aplicou os filtros e a escolha dos genes utilizados, a fase de ação, onde se aplicou as ferramentas e o algoritmo na base depurada e a fase de avaliação, onde se estudou os resultados encontrados. A pesquisa bibliográfica, para Arnal et al. (1992), é o tipo de pesquisa desenvolvida a partir de referências teóricas que apareçam em livros, artigos, documentos, etc.. O estudo de caso quantitativo, que segundo Caruso (2002), é definido como aquele que, utilizando instrumentos de coleta de informações numéricas, medidas ou contadas, aplicados a uma amostra representativa de um universo a ser pesquisado, fornece resultados numéricos, probabilísticos e estatísticos.

Neste contexto, este trabalho busca classificar os dados dos microarranjos de RNA oriundos de tecido epitelial humano. Primeiramente, dividindo os pacientes em dois grupos, os que serão usados para treino da ferramenta e os que serão utilizados para teste. Na sequência, o classificador, com base nos padrões identificados no treino, poderá indicar um a um dos pacientes incluídos no grupo de testes, se o padrão da atividade gênica condiz com o padrão de portador ou não portador de adenoma de colo de intestino.

3.1 ESTUDO DE CASO

Neste trabalho, aplicaram-se técnicas de mineração em uma base de dados biológica pública, que contém informações de pacientes com adenoma. Estas informações são a leitura da expressão do gene extraídas do RNA do tecido epitelial com adenoma, através de sondas. Há também a mesma leitura extraída nos mesmos

pacientes em células sadias. Objetivando que, através dos padrões levantados pelo aprendizado de máquina, pudessem fazer com que o classificador faça a separação em duas classes, para auxiliar na apresentação dos genes que possam estar envolvidos na presença do adenoma.

3.2 DESCRIÇÃO DAS TÉCNICAS APLICADAS

O desenvolvimento de pesquisas nas áreas biológicas vem enriquecendo as bases de dados públicas com dados expressivos de toda a ordem. Este trabalho partiu de dados existentes no GEO (Gene Expression Omnibus), que são informações de 32 (trinta e dois) pacientes com adenoma de colo de intestino, estas informações são efetivamente pequenos arquivos em formato texto, mostrados parcialmente na Tabela 1, onde estão as informações de toda a atividade gênica dos 54675 (cinquenta e quatro mil e seiscentos e setenta e cinco) sondas ativas nas células epiteliais, cada arquivo possui as informações de uma célula, cada paciente possui dois arquivos, um de uma célula saudável e outro de uma célula com o adenoma.

A Tabela 1 é composta pelas informações do paciente número um com adenoma.

A base de dados foi criada no Sistema Gerenciador de Banco de Dados MYSQL e nomeada “PACIENTES”, onde foram importados os dois arquivos de todos os 32 pacientes na qualidade normal e na qualidade adenoma, totalizando 64 arquivos, no seguinte formato:

- ID, que representa a identificação da sonda;
- Gene_ID, que representa a atividade medida na sonda;
- Abs, que pode ter três representação: (A)usente, (P)resente e (M)arginal, este indica que a leitura está em um limiar muito tênue entre os dois anteriores;
- P_Value, que é o coeficiente de erro que esta leitura da sonda pode ter.

	<i>ID</i>	<i>Gene_ID</i>	<i>ABS</i>	<i>P-Value</i>
1	AFFX-BioB-5_at	1476.68	P	0.000224668
2	AFFX-BioB-M_at	1929.9	P	7,01E+00
3	AFFX-BioB-3_at	1378.71	P	5,17E+00
4	AFFX-BioC-5_at	2636.85	P	5,17E+00
5	AFFX-BioC-3_at	2976.87	P	4,43E+00
6	AFFX-BioDn-5_at	3844.12	P	4,43E+00
7	AFFX-BioDn-3_at	13572.3	P	4,43E+00
8	AFFX-CreX-5_at	31840.9	P	4,43E+00
9	AFFX-CreX-3_at	42101.5	P	4,43E+00
10	AFFX-DapX-5_at	1526.97	P	4,43E+00
11	AFFX-DapX-M_at	2064.24	P	0.000753643
12	AFFX-DapX-3_at	2112.55	P	8,14E+00
13	AFFX-LysX-5_at	120.174	P	0.00359458
14	AFFX-LysX-M_at	260.268	P	0.026111
15	AFFX-LysX-3_at	383.183	P	0.000146581
16	AFFX-PheX-5_at	261.533	P	0.000126798
17	AFFX-PheX-M_at	395.377	P	0.000195116
18	AFFX-PheX-3_at	382.701	P	0.000753643
19	AFFX-ThrX-5_at	331.746	P	0.000169227
20	AFFX-ThrX-M_at	536.651	P	4,43E+00
21	AFFX-ThrX-3_at	770.254	P	5,17E+00
22	AFFX-TrpnX-5_at	482.797	A	0.544587
23	AFFX-TrpnX-M_at	608.628	A	0.834139

Tabela 1 - Modelo parcial do arquivo disponível no GEO

Todos os 64 arquivos de texto foram importados para esta única tabela. Como originalmente não possuíam identificação de paciente, nem tampouco de serem oriundas de células saudáveis ou de células com o adenoma, (pois estas informações se obtinham através da nomenclatura do arquivo), como pode ser visto na Tabela 1, foram inseridos, então, dois campos na base. O primeiro foi chamado de “status”, que permite somente um caractere de identificação, “N” para normal ou “A” para adenoma. O segundo é o campo “idp” que contempla a identificação do paciente, utilizado para controle da pesquisa. Estas alterações podem ser visualizadas na Tabela 2.

	<i>ID</i>	<i>Gene_ID</i>	<i>ABS</i>	<i>P_value</i>	<i>Status</i>	<i>Idp</i>
1	AFFX-LysX-3_at	383.183	P	0.000146581	A	1
2	AFFX-PheX-5_at	261.533	P	0.000126798	A	1
3	AFFX-PheX-M_at	395.377	P	0.000195116	A	1
4	AFFX-PheX-3_at	382.701	P	0.000753643	A	1

Tabela 2 - Modelo parcial do arquivo, com os campos Status e IDP, neste caso, entende-se que todos os dados são do paciente numero 1 com adenoma

Com todos os 64 arquivos importados para a base, esta ficou extremamente grande e de difícil manuseio. Deu-se então o início do tratamento dos dados. Criou-se uma nova base, com a mesma estrutura, onde foram transferidas apenas as sondas que possuíam *P_value* inferior a 0,04, pois os coeficientes maiores que este foram considerados muito altos, podendo classificar erroneamente o conjunto de dados.

Na próxima etapa, fez-se uma comparação com os dados existentes na nova base com os dados fornecidos na tabela GPL570. A tabela GPL570 é fornecida pela Affymetrix e possui informações que associam as sondas a nomes de genes, com a seguinte estrutura:

- SEQ., É apenas um identificador da linha para referencia, não faz parte da tabela original;
- ID, Identificador da sonda;
- Gene Symbol, Nome oficial do gene para a Affymetrix;
- ENTREZ_GENE_ID, código do gene para o Entrez.

O modelo da tabela GPL570 pode ser vista na Tabela 3. Nesta tabela fornecida pela Affymetrix (sessão 3.5.1), os nomes das sondas têm sua referência ao gene que ela representa e ao código de identificação do gene utilizado pelo ENTREZ. Pode acontecer de uma sonda ser associada a mais de um gene, como pode ser observado na Tabela 4. A

tabela GPL570 foi importada em uma base de dados de mesmo nome, para que pudesse ser utilizada em comparações pelo MYSQL.

<i>SEQ.</i>	<i>ID</i>	<i>Gene Symbol</i>	<i>ENTREZ_GENE_ID</i>
1	1007_s_at	DDR1	780
2	1053_at	RFC2	5982
3	117_at	HSPA6	3310
4	121_at	PAX8	7849
5	1255_g_at	GUCA1A	2978
6	1294_at	UBA7	7318
7	1316_at	THRA	7067
8	1320_at	PTPN21	11099
9	1405_i_at	CCL5	6352
10	1431_at	CYP2E1	1571
11	1438_at	EPHB3	2049
12	1487_at	ESRRA	2101
13	1494_f_at	CYP2A6	1548
14	1552256_a_at	SCARB1	949
15	1552257_a_at	TTLL12	23170
16	1552258_at	NCRNA00152	112597
17	1552261_at	WFDC2	10406
18	1552263_at	MAPK1	5594

Tabela 3 - Tabela GPL570 – fornecida pela Affymetrix

<i>SEQ.</i>	<i>ID</i>	<i>Gene Symbol</i>	<i>ENTREZ_GENE_ID</i>
655	1553181_at	DDX31	64794
656	1553183_at	UMODL1	89766
657	1553185_at	RASEF	158158
658	1553186_x_at	RASEF	158158
659	1553188_s_at	PARD3B	117583
660	1553190_s_at	PARD3B	117583
661	1553191_at	DST	667
662	1553192_at	ZNF441	126068
663	1553193_at	ZNF441	126068
664	1553194_at	NEGR1	257194
665	1553196_a_at	FCRL3	115352
666	1553197_at	WDR21C	138009
667	1553199_at	WDR21C	138009
668	1553202_at	STOX1	219736
669	1553204_at	C20ORF200	253868
670	1553205_at	C20ORF200	253868
671	1553207_at	ARL10	285598
672	1553209_at	RNFT2	84900
673	1553211_at	ANKFN1	162282

Tabela 4 - Tabela GPL570 – fornecida pela Affymetrix com nomes duplos

Diante da situação de existir mais de uma sonda associada a um mesmo gene, foi então realizada a comparação entre a base de dados e a tabela da Affymetrix, e, somente as sondas que possuíam correspondência na tabela GPL570 foram consideradas e copiadas para uma terceira base, de estrutura igual à primeira e à segunda. Em casos que mais de uma sonda fizesse referência a um mesmo gene, como é o caso das linhas 657 e 658, ou 659 e 660 na Tabela 4. Desta forma foi então realizada uma média dos valores de atividade das sondas e associada a média ao gene. Considerando como exemplo as linhas 657 e 658 que fazem referência ao gene RASEF, com valores das sondas 1553185_at e 1553186_x_at, possuindo valores respectivamente de 3976.05 e 1860.55

como pode ser visto na Tabela 5. Neste caso, fazendo-se a média dos dois valores, obtemos o valor 2918.30, que é então atribuído como valor da atividade ao gene RASEF.

<i>ID</i>	<i>Gene_ID</i>	<i>ABS</i>	<i>P_Value</i>
1553180_at	2.61349	A	0.74707
1553181_at	85.6442	A	0.0805664
1553183_at	9.55577	A	0.696289
1553185_at	3976.05	P	0.000732422
1553186_x_at	1860.55	P	0.000732422
1553188_s_at	5.94567	A	0.932373
1553190_s_at	107.152	A	0.432373
1553191_at	17.4371	A	0.0952148

Tabela 5 - Adenoma1 - parcial

Desta forma então foi criada uma nova base de dados, com os seguintes campos: Nome do gene, atividade, identificação do paciente e status, que identifica se é paciente normal ou com adenoma, como pode ser observado na tabela 6.

<i>ZWILCH</i>	<i>ZSCAN29</i>	<i>ZSCAN16</i>	<i>Status</i>	<i>idp</i>
10896000	4940850	4704880	A	1
3778950	6626560	5447820	N	1
10527500	5546950	5488130	A	2
5533430	4046670	4534900	N	2

Tabela 6 - Pacientes identificados por Status e idp

De posse desta nova base de dados, foram selecionados os 50 genes que possuíam maior atividade gênica demonstrados na Tabela 7, e divididos em dois conjuntos, o de treino e o de testes. A ferramenta R foi então utilizada para treinar o

algoritmo SVM utilizando o conjunto de treino e logo depois, aplicados ao conjunto remanescente.

ZZZ3	ZZEF1	ZYX	ZYG11B	ZXDC	ZXDB	ZWINT	ZWILCH
ZW10	ZUFSP	ZSWIM7	ZSWIM6	ZSWIM5	ZSWIM1	ZSCAN5A	ZSCAN29
ZSCAN22	ZSCAN21	ZSCAN20	ZSCAN2	ZSCAN18	ZSCAN16	ZSCAN12	ZRSR2
ZRANB2	ZRANB1	ZNRF3	ZNRF2	ZNRF1	ZNRD1	ZNHIT6	ZNHIT3
ZNHIT1	ZNFX1	ZNF93	ZNF92	ZNF91	ZNF862	ZNF853	ZNF850P
ZNF85	ZNF846	ZNF844	ZNF843	ZNF84	ZNF839	ZNF830	ZNF83
ZNF829	ZNF828						

Tabela 7 - Relação dos 50 genes de maior expressão

Alguns problemas foram encontrados nesta fase, pois quando um gene estava presente em um único paciente, ou em alguns poucos, o SVM não conseguia classificar, pois nenhum padrão conseguia ser criado.

A próxima etapa foi, então, partindo dos 50 genes mais ativos de cada paciente, fazer uma avaliação da frequência com que cada gene aparecia, promovendo que, os genes selecionados tivessem inclusive, representação em um grande número de pacientes. Desta forma, o classificador poderia efetivamente realizar a definição de padrões e nos dar resultados mais assertivos.

Os resultados foram muito satisfatórios e reduziu-se mais a quantidade de genes por paciente, e arrojadamente, realizou-se a classificação com os 5 genes mais ativos. O índice de acertos neste caso foi desastroso e passou-se então a ampliar gradativamente o numero de genes a ser classificados, passando para 10 e posteriormente para 15. Onde tivemos então os resultados que foram considerados satisfatórios.

3.3 RESULTADOS

No Software “R”, primeiramente se carrega todo o conjunto de dados sem distinção entre treino e testes. Em um segundo momento, através de comandos, é informado ao software, “quem”, dentro do conjunto de dados será usado como conjunto de treino. De posse do algoritmo treinado, inicia-se a fase de testes, que são aplicados individualmente a cada um dos indivíduos que não participaram do treino.

O resultado da ferramenta para a avaliação do indivíduo aplicado ao algoritmo treinado, se dá na forma de uma matriz 2 x 2, como mostrado na Tabela 8, onde na diagonal principal aparece o resultado considerado correto, e na diagonal secundária, o erro.

<i>Pred</i>	<i>A</i>	<i>N</i>
<i>A</i>	1	0
<i>N</i>	0	0

Tabela 8 - Exemplo de Predição correta

Observando a Tabela 8, pode-se interpretá-la da seguinte forma: Na linha há a informação constante na base de dados e na coluna o resultado da predição. A leitura no caso acima significa que o algoritmo acha que o caso é de Adenoma e a linha informa que na base de dados, este caso era de Adenoma. A predição está correta.

Um possível caso de erro seria apresentado como podemos ver na Tabela 9.

<i>Pred</i>	<i>A</i>	<i>N</i>
<i>A</i>	0	1
<i>N</i>	0	0

Tabela 9 - Exemplo de Predição incorreta

No caso da Tabela 9 o Algoritmo classificou o paciente como Adenoma, mas na base de dados ele consta como Normal. A classificação estará correta quando na matriz, tanto a linha quanto a coluna coincidirem.

Quando se trabalhou com os 50 genes mais ativos, vistos na Tabela 7, os treinos foram realizados com 32 amostras sendo 16 amostras com adenoma e 16 amostras de tecido normal. Os testes foram realizados nas 32 amostras restantes. Os resultados foram assertivos em 100 % das situações, como se pode observar na Tabela 10 e no Gráfico 1.

<i>Pacientes</i>	<i>Adenoma</i>	<i>Normal</i>	<i>%</i>
Total	32	32	100,00
Treino	16	16	50,00
Teste	16	16	50,00
Acerto	16	16	100

Tabela 10 - 50 Genes, Treino com 32, Testes com 32



Gráfico 1 - 50 Genes, Treino com 32, Teste em 32

Diminuindo-se o conjunto de treino, passou-se a utilizar 15 amostras para treino e testes nas 49 amostras restantes, mantendo-se o numero de 50 Genes. Os resultados foram de 100 % de acerto nas predições. Como pode ser observado na Tabela 11 e no Gráfico 2.

<i>Pacientes</i>	<i>Adenoma</i>	<i>Normal</i>	<i>%</i>
------------------	----------------	---------------	----------

Total	32	32	100,00
Treino	8	7	23,44
Teste	24	25	76,56
Acerto	24	25	100

Tabela 11 - 50 Genes, Treino com 15, testes com 49

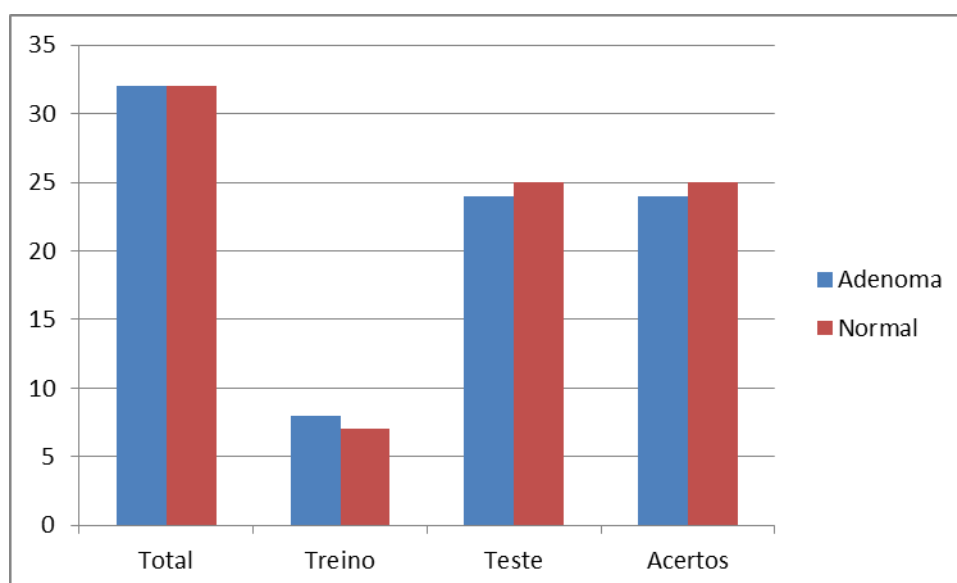


Gráfico 2- 50 Genes, Treino com 15, testes com 49

Diminuindo-se ainda mais o conjunto de treino. Realizou-se a classificação com os mesmos 50 genes mais ativos, porém os treinos foram realizados com 5 amostras, sendo 3 com adenoma e 2 amostras de tecido normal, obteve-se 100 % dos casos com adenoma que o SVM conseguiu classificar corretamente, porém somente 2 amostras de tecido normal foram assertivas, as 28 amostras restantes foram classificadas também como se fossem de tecido com adenoma. Como descreve a Tabela 12 e Gráfico 3.

<i>Pacientes</i>	<i>Adenoma</i>	<i>Normal</i>	<i>%</i>
Total	32	32	100
Treino	3	2	7,81
Teste	29	30	92,19
Acerto	29	2	52,54

Tabela 12 – 50 Genes, Treino com 5, Teste com 59

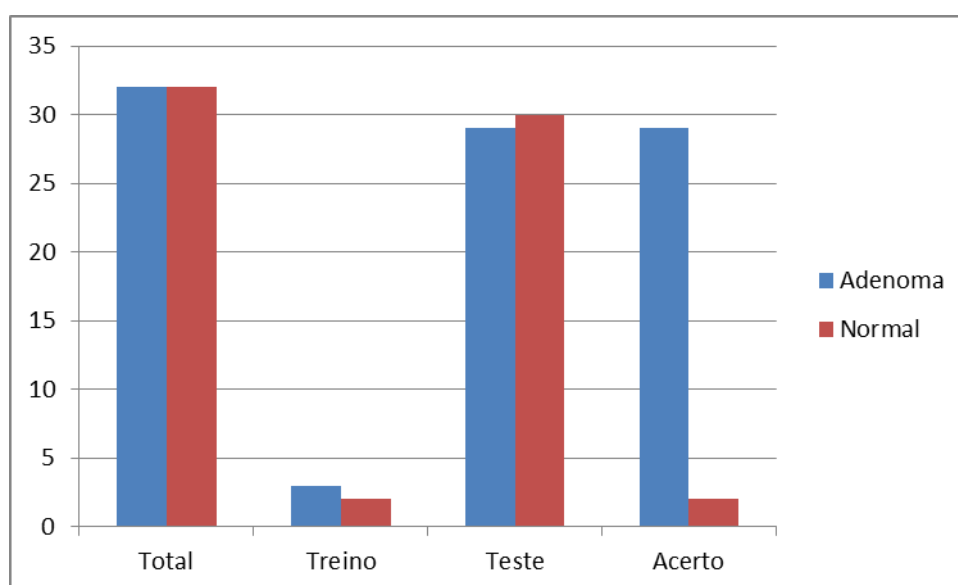


Gráfico 3 - 50 Genes, Treino com 5, Teste com 59

Passando-se o conjunto de dados para os 20 Genes mais ativos, foi iniciado o processo de treino com 15 amostras, sendo 8 de tecido com adenoma e 7 de tecido normal, e testado em 49 amostras restantes. Como resultado obteve-se 2 erros apenas, com a amostra 22 e a com a amostra 64, ambas amostras de tecido normal, que foram classificadas como adenoma. Estes resultados são demonstrados na Tabela 13 e no Gráfico 4.

<i>Pacientes</i>	<i>Adenoma</i>	<i>Normal</i>	<i>%</i>
Total	32	32	100
Treino	8	7	23,44
Teste	24	25	76,56
Acerto	24	23	92,59

Tabela 13 - 20 Genes, Treino com 15, Teste com 49

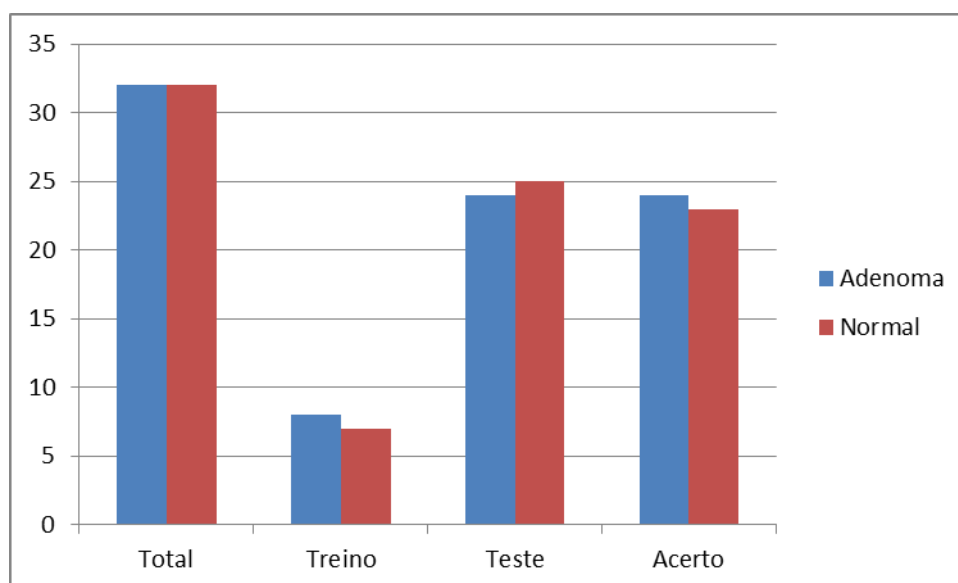


Gráfico 4- 20 Genes, Treino com 15, Teste com 49

Mantendo-se o número de Genes, porém aumentando-se o conjunto de treino para 32 amostras, e testados nas 32 amostras restantes, obteve-se um resultado um pouco melhor que com o conjunto de testes menor. A assertividade nestas condições foi de 31 casos, mantendo a amostra da linha 64 com erro, que era um caso normal e foi classificado como adenoma. Pode-se observar estes resultados na Tabela 14 e no Gráfico 5.

<i>Pacientes</i>	<i>Adenoma</i>	<i>Normal</i>	<i>%</i>
Total	32	32	100
Treino	16	16	50,00
Teste	16	16	50,00
Acerto	16	15	96,88

Tabela 14 - 20 Genes, Treino com 32, Testes com 32

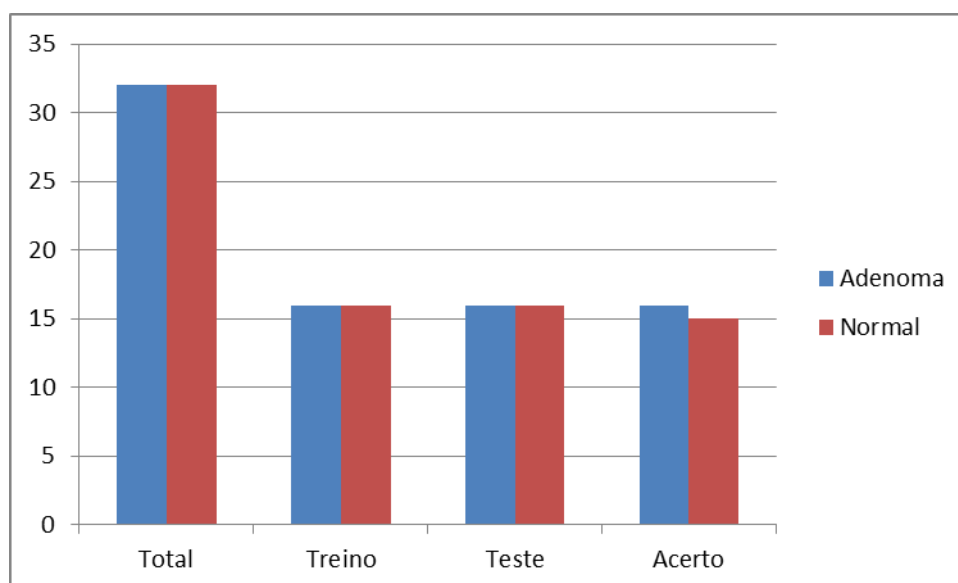


Gráfico 5 - 20 Genes, Treino com 32, Testes com 32

Passando-se o número de Genes para 15, obteve-se um excelente resultado ao utilizar um conjunto de Treino com 32 amostras, sendo 16 de tecido normal e 16 de tecido com adenoma. Os testes realizados nas 32 amostras restantes apresentaram assertividade de 100 % dos casos. A Tabela 15 e o Gráfico 6 demonstram estes resultados.

<i>Pacientes</i>	<i>Adenoma</i>	<i>Normal</i>	<i>%</i>
Total	32	32	100
Treino	16	16	50,00
Teste	16	16	50,00
Acerto	16	16	100,00

Tabela 15 - 15 Genes, Treino com 32, Teste com 32

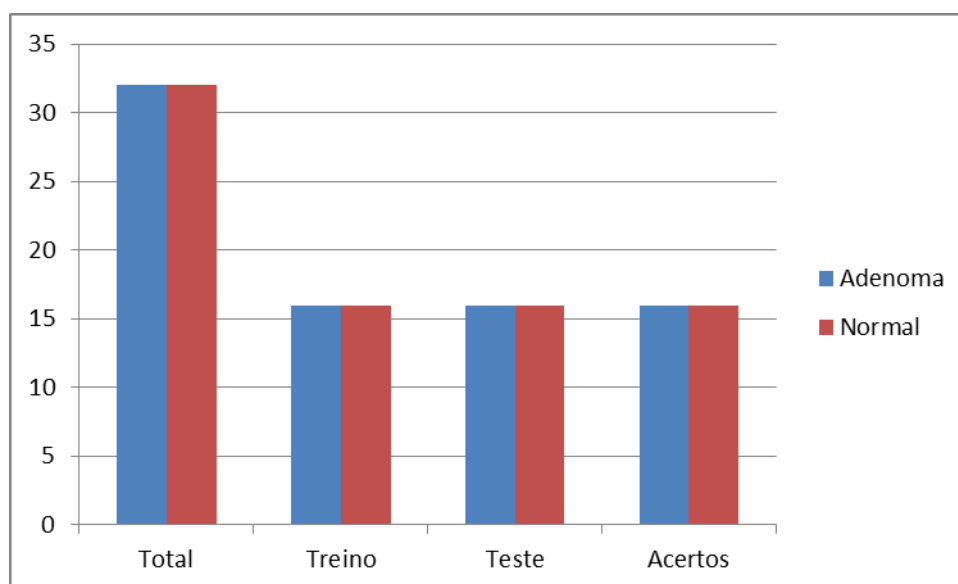


Gráfico 6 - 15 Genes, Treino com 32, Teste com 32

Mantendo-se o número de Genes em 15, alcançou-se igualmente um excelente resultado com um conjunto de treino de 15 amostras, das quais 8 eram de tecido com adenoma e 7 de tecido normal. Os testes foram feitos nas 49 amostras restantes e nenhum erro foi encontrado na classificação. Estes resultados estão apresentados na Tabela 16 e no Gráfico 7.

<i>Pacientes</i>	<i>Adenoma</i>	<i>Normal</i>	<i>%</i>
Total	32	32	100
Treino	8	7	23,44
Teste	24	25	76,56
Acerto	24	25	100,00

Tabela 16 - 15 Genes, Treino com 15, teste com 49

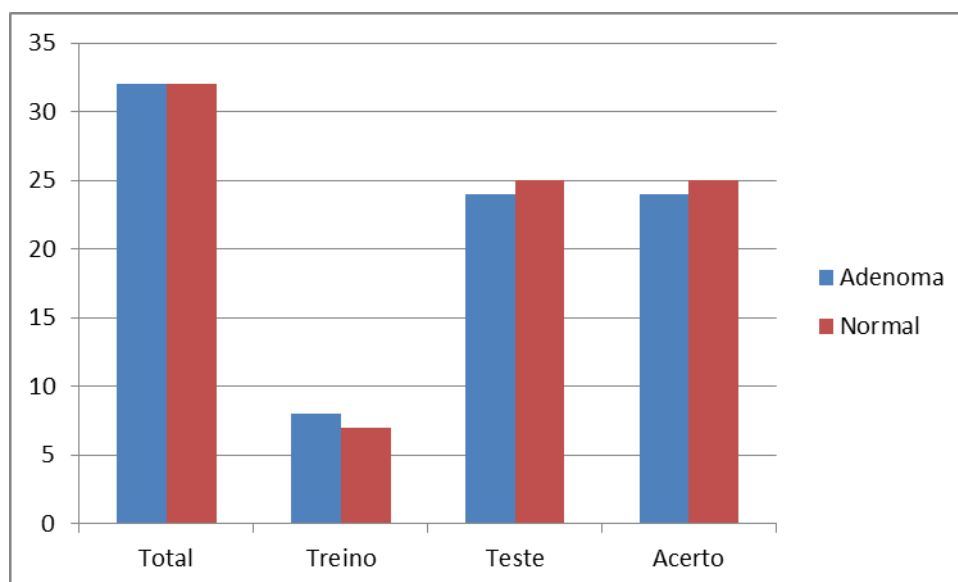


Gráfico 7- 15 Genes, Treino com 15, teste com 49

Reduzindo-se ainda mais o conjunto de amostras no treino, trazendo-o para apenas 5, sendo 3 com adenoma e 2 com tecido normal e testadas as 59 amostras restantes obteve-se um resultado interessante. Apenas 44 amostras foram classificadas corretamente e 15 foram classificadas com erro. Os erros apareceram nas linhas 62, 60, 56, 54, 52, 50, 48, 42, 40, 38, 34, 28, 26, 24 e 22. Ao se perceber que todas as linhas eram pares, pode-se entender que todas as amostras classificadas com erro são de tecidos normais e foram classificados como adenoma. Talvez em função do reduzido número de amostras no treino, as regras criadas para classificação tenham sido tão amplas que permitiram identificar como adenoma algumas amostras que estejam mais

próximas dos valores de tecidos com adenoma. Estes resultados podem ser observados na Tabela 17 e no Gráfico 8.

<i>Pacientes</i>	<i>Adenoma</i>	<i>Normal</i>	<i>%</i>
Total	32	32	100
Treino	3	2	7,81
Teste	29	30	92,19
Acerto	29	15	74,58

Tabela 17- 15 Genes, Treino com 5, Teste com 59

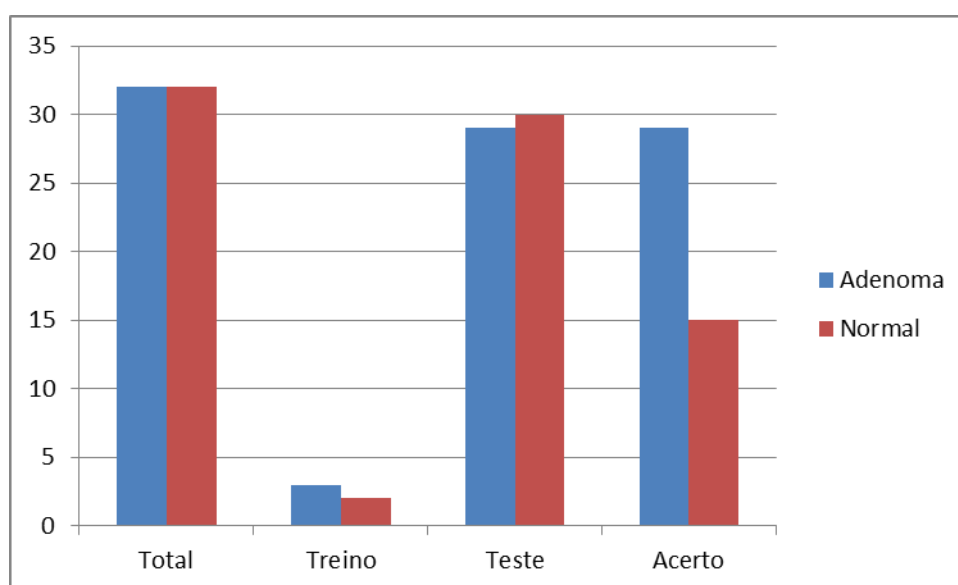


Gráfico 8 - 15 Genes, Treino com 5, Teste com 59

O conjunto de treino foi reduzido ainda mais, para 10 Genes com maior atividade. Inicialmente o treino foi realizado com a metade das amostras e testado na metade restante. Nenhuma amostra foi classificada incorretamente como se pode observar na Tabela 18 e no Gráfico 9.

<i>Pacientes</i>	<i>Adenoma</i>	<i>Normal</i>	<i>%</i>
Total	32	32	100,00
Treino	16	16	50,00
Teste	16	16	50,00
Acerto	16	16	100,00

Tabela 18 - 10 Genes, Treino com 32, Testes com 32

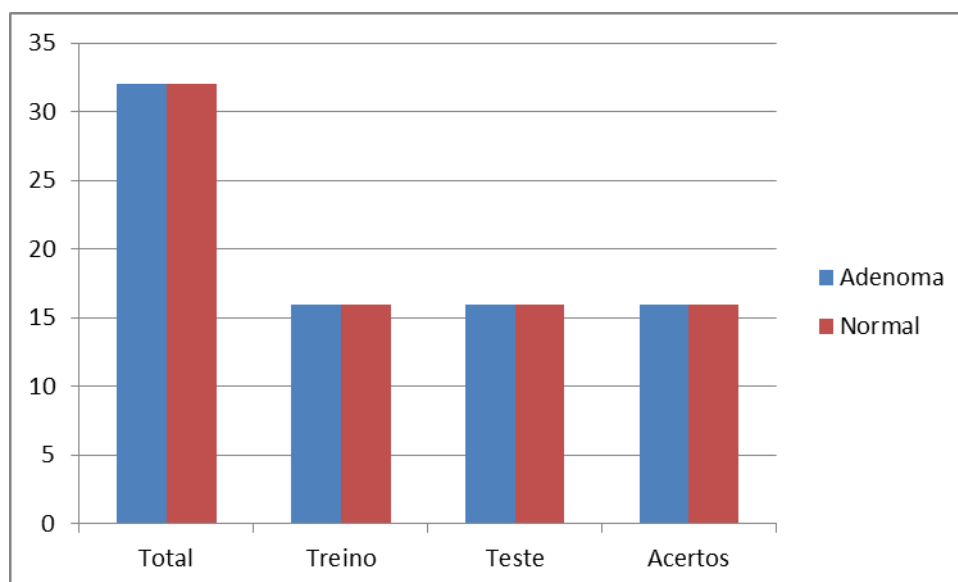


Gráfico 9 - 10 Genes, Treino com 32, Testes com 32

Neste conjunto de dados com os 10 Genes mais ativos, Foi realizado o Treino em 15 amostras, sendo 8 de tecido com adenoma e 7 com tecido normal. Os testes foram realizados nas 49 amostras restantes e apenas uma amostra não foi classificada corretamente. A amostra incorretamente classificada ficava na linha 22, portanto era uma amostra de tecido normal que foi classificada como adenoma. Os resultados do teste podem ser observados na Tabela 19 e no Gráfico 10.

<i>Pacientes</i>	<i>Adenoma</i>	<i>Normal</i>	<i>%</i>
Total	32	32	100,00
Treino	8	7	23,44
Teste	24	25	76,56
Acerto	24	24	97,96

Tabela 19 - 10 Genes, Treino com 15, Teste com 49

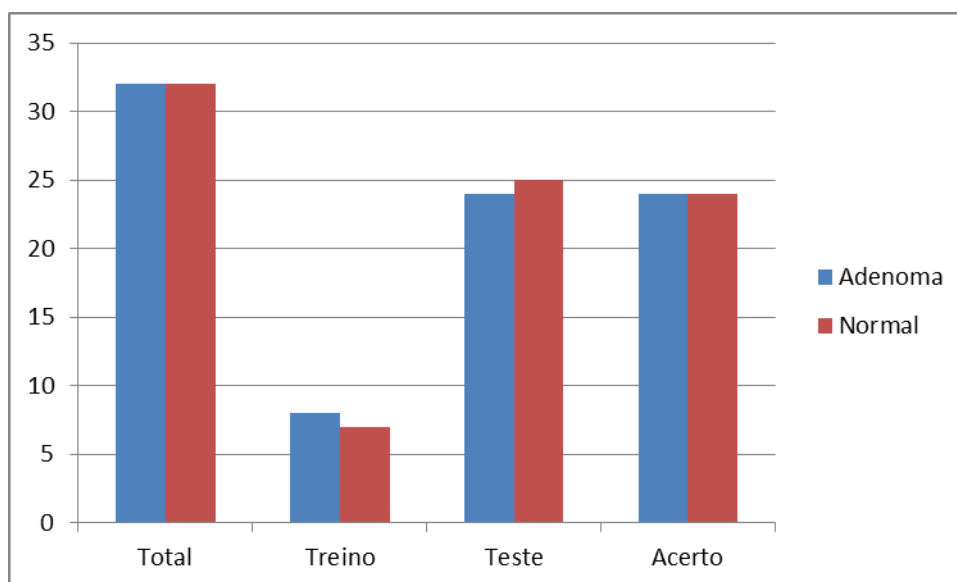


Gráfico 10 - 10 Genes, Treino com 15, Testes com 49

A Tabela 20 apresenta um comparativo entre os resultados encontrados.

<i>Genes (a)</i>	<i>Treino (b)</i>	<i>Teste (c)</i>	<i>Erros (d)</i>	<i>Acertos (e)</i>
50	32	32	0	100
50	15	49	0	100
20	32	32	1	96,88
20	15	49	2	95,92

15	32	32	0	100
15	15	49	0	100
15	5	59	15	74,58
10	32	32	0	100
10	15	49	1	97,96

Tabela 20 - Comparativo dos resultados

A Tabela 20 possui na coluna (a) quantidade de Genes utilizados na base de dados, na coluna (b) o conjunto de amostras utilizadas no treino do algoritmo, na coluna (c) a quantidade de amostras utilizadas nos testes, na coluna (d) o numero de amostras classificadas com erro e na coluna (e) o percentual de acertos do classificador.

A observação destes resultados permitiu que se chegasse a algumas conclusões que são descritas no capítulo 4.

3.4 DIFICULDADES ENCONTRADAS

Alguns problemas foram encontrados no desenvolver deste trabalho dos quais se podem citar:

- a) Inicialmente utilizou-se a ferramenta Weka. Mas a quantidade de leituras das sondas tornaram as bases de dados bastante grandes, fazendo com que a ferramenta passasse mais de 10 horas processando e esgotasse sua capacidade de memória. Alterando as configurações da ferramenta acabou-se por tornar o computador inoperante. Os testes com a ferramenta R trouxeram resultados mais rapidamente e por isto optou-se por utiliza-la.
- b) Quando forma filtrados os genes de maior atividade no MYSQL, muitos foram selecionados em função de uma sonda ou algumas poucas, o que deixava muitas lacunas em branco, promovendo corrupções de processamento. Então se selecionou as com maior atividade conjuntamente com as que possuíam informações mais completas.

- c) Escassez de documentação da ferramenta R, de forma clara e objetiva, o que promoveu uma demanda de tempo alta para poder utiliza-la.

4. CONCLUSÕES

Uma pergunta frequente com a mineração de dados é: "qual é o melhor método para...? A resposta pode ser decepcionante. "*There is no free lunch for anyone*". Qualquer abordagem estatística é tão boa diante das características que ela possui (DUBITZKY et al. ,2007. Pg. 32).

A área de mineração de dados é muito ampla, possuindo uma gama enorme de algoritmos. Estes algoritmos são projetados e escritos para casos específicos e muitos deles se adequam a outros casos, permitindo sua reutilização com outras áreas ou enfoques.

Neste trabalho foi possível analisar e preparar os dados que foram obtidos junto ao GEO, utilizando-se várias tabelas no MYSQL. Importando-se os arquivos texto e manipulando-os. Atingiu-se o objetivo de isolar os genes com atividade mais alta e com as informações completas dentro do que se considerou aceitável.

Dentre os genes selecionados, foi possível isolar em arquivos separados, os genes 50 genes com maior atividade nas células com adenoma quando comparados a células normais. Estes arquivos foram submetidos à ferramenta R.

Após os testes com os 50 Genes mais ativos, foram criadas as bases menores com 20, 15 e finalmente com os 10 Genes mais ativos, cada base foi submetida à ferramenta R, e os resultados foram apresentados na seção anterior.

A análise realizada nos dados disponibilizados pelo GEO, contendo as informações dos pacientes portadores de adenoma e normais, permitiu observar que há eficácia no classificador escolhido. Pode-se verificar isto nos resultados apresentados na Tabela 17 que demonstra esta eficácia.

A partir dos resultados obtidos através da classificação dos dados com o SVM, percebeu-se que as características dos genes são bem distintas e que a atividade varia bastante de gene para gene. Porém, isto ocorre de forma padronizada, o que permitiu que o algoritmo pudesse identificar estes padrões e realizar as predições.

4.1 TRABALHOS FUTUROS

A pesquisa realizada foi muito importante, pois permitiu a obtenção de resultados utilizando-se os grupos de 50, 20, 15 e 10 genes com atividade mais elevada. Entretanto, outras abordagens podem ser consideradas e pesquisas mais profundas podem ser realizadas, como por exemplo:

- a) Realizar a pesquisa com os genes de menor atividade, para comparação com os resultados aqui apresentados.
- b) Identificar os genes que são biologicamente responsáveis pelo adenoma. E a partir deles, ou dele, identificar os padrões da sua evolução no decorrer do tempo. Permitindo assim predizer os prazos de surgimento do Adenoma.
- c) Realizar uma pesquisa comparativa com outros algoritmos para verificar sua eficácia com bases de dados biológicas.
- d) Fazer um comparativo entre as ferramentas para identificar se o mesmo algoritmo apresenta os mesmos resultados com ferramentas diferentes.
- e) Realizar os testes com todos os conjuntos possíveis de Genes (como por exemplo, 2, 3, 4, 5...) até que se obtenha o melhor resultado em um menor conjunto de dados utilizados.

As questões aqui apresentadas sugerem algumas alternativas de pesquisas para serem realizadas futuramente. Entretanto, outras pesquisas podem ser desenvolvidas inspiradas nessa, em função de outros interesses.

5. REFERENCIAS BIBLIOGRAFICAS

AFFYMETRICS, Site oficial, Disponível em: www.affymetrics.com; Acessado em 27 de Novembro de 2010.

ALVES, W. P. **Fundamentos de Bancos de Dados**. São Paulo, Editora Érica. 2004.

ARBER, N. EAGLE, C.J. SPICAK, J. RÁCZ, I. DITE, P. HAJER, J. for the PreSAP Trial Investigators. **Celecoxib for the Prevention of Colorectal Adenomatous Polyps**. New Engl J Med 2006;355:885-95.

ARNAL, J et al.. *Investigación Educativa*. Barcelona: Editorial Labor. 1992.

BERRY, M.J.A.; LINOFF, G. **Data mining techniques**. New York: John Wiley & Sons, Inc., 1997.

BOLSTAD, B. M.; Pré-processin DNA Microarray Data. In DUBITZKY, W. (Org.), **Fundamental of Data Mining in Genomics and Proteomics**. Springer Science, New York, NY. 2007. pg. 51 a 78.

BORIN, A.; **Aplicações de Máquinas de vetores de suporte por mínimos quadrados (LS-SVM) na quantificação de parâmetros de qualidade de matrizes lácteas**. 2007. 128f. Tese (Doutorado em Ciências - Área de concentração: Química Analítica) – UNICAMP, São Paulo.

BURGES, C. J. C. **Data Mining and Knowledge Discovery (1998) Volume: 2, Issue: 2, Publisher: Springer, Pages: 121-167. ISSN: 13845810. ISBN: 0818672404. DOI: 10.1023/A:1009715923555**

CARVALHO, J. V. SAMPAIO, M. C. MONGIOVI, G.; **Utilização de Técnicas de “Data Mining” para o Reconhecimento de Caracteres Manuscritos, Simpósio Brasileiro de Banco de Dados, Florianópolis, SC – 1999 disponível em: <http://www.inf.ufsc.br/sbbd99/>**

CARUSO, P.. **Metodologia da Investigação Científica**. 30/04/2002. [Disponível na Internet:<http://atlas.ucpel.tche.br/~pdmc/link1.html>]. 2002.

CESAR, R. M. Jr., **A estrutura da molécula de DNA**, São Paulo. 2005, Disponível em: <http://www.ime.usp.br/~cesar/projects/lowtech/setemaiores/dna.htm>. Acessado em: 18 de novembro de 2010.

CRATOCHVIL, A. **Data mining techniques in supporting decision making**. Master thesis, Universiteit Leiden, 1999.

CRISTIANINI, N. and SHAW-TAYLOR, J.; **An Introduction to Support Vector Machines and other kernel-based learning methods**. Cambridge University Press, 2000.

DARWIN, C., 2004, “A origem das espécies”, Ed. Martin Claret, São Paulo, Brasil.

DUBITZKY, W. GRANZOW, M. BERRAR, D., **Fundamental of Data Mining in Genomics and Proteomics**. Springer Science, New York, NY. 2007).

DUNHAM, M. H., - **Data Mining – Introductory and Advanced Topics** – Prentice Hall, Upper Saddle River, New Jersey 2003.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in knowledge discovery and data mining**. Menlo Park: AAAI Press, 1996a.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. **From data mining to knowledge discovery in databases**. AI Magazine, v. 17, n. 3, p. 37-54, 1996b.

FAYYAD, U., PIATETSKY-SHAPIO, G.; SMYTH, P. **Knowledge discovery and data mining: towards a unifying framework**. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2., 1996. Menlo Park: AAAI Press, p. 82-88, 1996c.

FELDENS, M.A. et al. **Towards a methodology for the discovery of useful knowledge combining data mining, data warehousing and visualization**. In: CLEI (Conferência Latino-Americana de Informática), 24., 1998, Equador. Proceedings... Equador, 1998.

FILHO, F. C.; PROSDOCIMI, F.; CERQUEIRA, G. C.; BINNECK, E.; SILVA, A. F.; REIS, A. N.; JUNQUEIRA, A. C. M., SANTOS, A. C. F.; JÚNIOR, A. N.; WUST, C. I.;

KESSEDJIAN, J. L.; PETRETSKI, J. H., CAMARGO, L. P., FERREIRA, R. G. M.; LIMA, R. P.; PEREIRA, R. M.; JARDIM, S.; SAMPAIO, V. S., FOLGUERASFLATSCHART, A. V. **Bioinformática: Manual do Usuário**, 2002. Disponível em: <http://www.bioteecnologia.com.br/revista/bio29/bioinf.pdf>. Acesso em: 11 de outubro de 2009.

FONTANA, D. R. MARIM, L. R.; **Sistema de Autenticação/Identificação Pessoal Biométrica através da Palma da Mão Com o Auxílio de Redes Neurais Artificiais**; Anais do 14º Encontro de Iniciação Científica e Pós Graduação do ITA – XV ENCITA/2009; Instituto Tecnológico da Aeronáutica, São José dos Campos, SP, Brasil, Outubro, 19 a 22, 2009.

FRAWLEY, W. J.; PIATETSKY-SHAPIO, G.; MATHEUS, C. J. **Knowledge discovery in databases: an overview**. AI Magazine, v. 14, n. 3, p. 57-70, 1992.

FUREY, T. S. CRISTIANINI, N. DUFFY, N. BEDNARSKI, D. W. SCHUMMER, M. and HAUSSLER, D. ; **Support vector machine classification and validation of cancer tissue samples using microarray expression data**; *Bioinformatics*, Vol. 16 nº 10 2000, Pg 906-914, 2000.

GEO, Site oficial, Disponível em: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM215086>, Acessado em 20 de dezembro de 2010.

GIBAS, C. e JAMBECK, P. **Desenvolvendo Bioinformática: ferramentas de software para aplicações em biologia**. Rio de Janeiro : Campus, 2001.

GOEBEL, M.; GRUENWALD L. **A survey of data mining and knowledge discovery software tools**. In: ACM SIGKDD Explorations Newsletter. 1. ed. vol. 1 1999.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: um guia prático**. Rio de Janeiro: Elsevier, 2005.

GREENMAN, C. STEPHENS, P. SMITH, R. DALGLIESH, G. L. HUNTER, C. BIGNELL, G. DAVIES, H. TEAGUE, J. BUTLER, A. STEVENS, C. EDKINS, S. O'MEARA, S. VASTRIK, I. SCHMIDT, E. E. AVIS, T. BARTHORPE, S. BHAMRA, G. BUCK, G. CHOUDHURY, B. CLEMENTS, J. COLE, J. DICKS, E. FORBES, S.

GRAY, K. HALLIDAY, K. HARRISON, R. HILLS, K. HINTON, J. JENKINSON, A. JONES, D. MENZIES, A. MIRONENKO, PERRY, T. J. RAINE, K. RICHARDSON, D. SHEPHERD, R. SMALL, A. TOFTS, C. VARIAN, J. WEBB, T. WEST, S. WIDAA, S. YATES, A. CAHILL, D. P. LOUIS, D. N. GOLDSTRAW, NICHOLSON, P. A. G. BRASSEUR, F. LOOIJENGA, L. WEBER, B. L. CHIEW, Y. DEFAZIO, A. GREAVES, M. F. GREEN, A. R. CAMPBELL, P. BIRNEY, E. EASTON, D. F. CHENEVIX-TRENCH, G. TAN, M. H. KHOO, S. K. TEH, B. T. YUEN, S. T. LEUNG, S. Y. WOOSTER, R. P. FUTREAL, A. STRATTON, M. R.; **Patterns of somatic mutation in human cancer genomes**; *Nature* **446**, 153-158 (8 March 2007) | doi:10.1038/nature05610; Received 7 September 2006; Accepted 18 January 2007

GROTH, R. **Data mining**. New Jersey: Prentice Hall, Inc., 1998.

GUYON, I. WESTON, J. BARNHILL, S. VAPNIK, W.; **Gene Selection for Cancer Classification using Support Vector Machines**; *Machine Learning*, 46, 389–422, 2002; Kluwer Academic Publishers. Manufactured in The Netherlands. 2002.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques**. San Francisco: Morgan Kaufmann Publishers, 2001.

HAYKIN, S. **Neural Networks - A Comprehensive Foundation**., 2nd edition; Prentice-Hall; New Jersey; 1999.

JAAKKOLA, T. Diekhans, M.; and Haussler, D.; **Using the Fisher kernel method to detect remote protein homologies**. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 149/158, Menlo Park, CA, AAAI Press. 1999.

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining**. New Jersey: John Wiley & Sons, p. 4. 2005.

LINDEN, R. **Algoritmos genéticos, Uma importante ferramenta da Inteligência Computacional**; 2.ed. Rio de Janeiro; Brasport; 2008.

LESK, A. M. **Introdução a Bioinformática**, 2, Ed. – Porto Alegre, Artmed, 2005.

LESK, A. M. **Introduction to Bioinformatics**, 3, Ed. – Oxford, Oxford, 2008.

LOGAN, B. MORENO, P. SUZEK, B. WENG, Z. KASIF, S. **A study of remote homology detection**. Relatório técnico, Cambridge Research Laboratory, Junho 2001. Disponível em: <http://www.hpl.hp.com/techreports/Compaq-DEC/CRL-2001-5.html>.

LORENA, A. C. CARVALHO, A. C. P. L. F. **Uma Introdução às Support Vector Machines**. RITA Volume XIV Número 2 Pg 43 – 67, 2007

MAGATÃO, M. G. S. JUNIOR, E.F.S. **Educação para a ciência: Uma proposta de intervenção diferenciada no ensino de biologia**. 2008.

MARRA, G. **Biopsy of a colorectal adenoma from patient #4**. Molecular Cancer Research, University of Zurich. Agosto de 2007. Disponível em (GEO, 2010).

MEYER, D. **Support Vector Machines - The Interface to libsvm in package e1071**.

R-News, Vol.1 pg 3 - 9. Wien, Austria. 2001.

MONARD, M. C. BARANAUSKAS, J. A. **Conceitos sobre aprendizado de máquina**. In: REZENDE, S. O. (Org.). **Sistemas inteligentes: fundamentos e aplicações**. Barueri: Editora Manole. 2003. Cap. 4. pg. 89-114.

NLM. The DNA Structure, 2009. Disponível em <http://ghr.nlm.nih.gov/handbook/illustrations/dnastructure>. acesso em 10 de dezembro de 2010.

NYTIMES, Jornal “ The new York Times”, versão para web, publicado na seção de ciencias em 27 de junho de 2000, disponível em:

<http://partners.nytimes.com/library/national/science/062700sci-genome.html?scp=1&sq=%22Genetic%20Code%20of%20Human%20Life%20is%20cracked%20by%20Scientists%22&st=cse> acessado em 22 de dezembro de 2010.

REZENDE, S. O. PUGLIESI, J. B. MELANDA, E. A. PAULA, M. F. **Mineração de dados**. In: REZENDE, S. O. (Org.). **Sistemas inteligentes: fundamentos e aplicações**. Barueri: Editora Manole. 2003. Cap. 12. pg. 307- 335.

ROCHA, A. de R. CARVALHO, A. A. REZENDE, A. G. ALVES, J. C.; **Computação baseada em DNA**; Universidade Federal de Lavras, Lavras/MG. 2003.

SALCES, I. VEGH, I. RODRÍGUEZ-MUÑOZ, S. COLINA, F. PÉREZ, A. SOTO, S. SÁNCHEZ, F. DE LA CRUZ, J. and SOLÍS-HERRUZO, J. A. **tissue ca-19.9 content in colorectal adenomas and its value in the assesment of dysplasia**, Revista Española De Enfermedades Digestivas, versión impresa issn 1130-0108 rev. esp. enferm. dig. v.96 n.4 Madrid abr. 2004.

SCHUCH, Regis. DILL, Sergio. PADOIN, Edson. SAUSEN, Paulo. Mineração de Dados para o Perfil dos Pacientes Ambulatoriais com Câncer, *VII CONGED, ISSN 1981-8882, Ponta Grossa, 07 a 09/04/2010*.

SETUBAL, J. C. **A origem e o sentido da bioinformática**, Revista eletrônica ComCiência.br, 2003. Disponível em <http://www.comciencia.br/reportagens/bioinformatica/bio10.shtml>. Acesso em 19 de julho de 2010.

SHAWE-TAYLOR, J. BARLETT, P.L. WILLIAMSON, R.C. and ANTHONY, M. *Structural risk minimization over data-dependent hierarquies*. IEEE Transactions on Information Theory, 44(5):1926–1940, 2000.

SILBERDCHATZ, A. KORTH, H. F. SUDARSHAN, S. **Sistema de Banco de Dados**, Tradução da 5. ed., São Paulo; Campus; 2006.

SILVA, C. F., VIEIRA, R., **Grupos gramaticais e sintáticos em categorização automática com Support Vector Machines** ; XXV Congresso da Sociedade Brasileira de Computação, 22 a 29 de julho de 2005; UNISINOS – São Leopoldo/RS; 2005.

SOUTO M. C. P., LORENA, A. C. DELBEM, A. C. B. and CARVALHO. A. C. P. L. F.; **Técnicas de Aprendizado de Máquina para problemas de Biologia Molecular**, paginas 103–152. Minicursos de Inteligência Artificial, Jornada de Atualização Científica em Inteligência Artificial, XXIII Congresso da Sociedade Brasileira de Computação, 2003.

SOUZA, M. A. **SQL, PL/SQL, SQL *Plus**. Rio de Janeiro, Editora Ciência Moderna, 2004.

SOUZA, B. F. CARVALHO, A. C. P. L. F. CALVO, R. ISHII, R. P. **Multiclass svm model selection using particle swarm optimization**. In Proceedings of 6th HIS, Pg. 31–36, 2006.

STRINGER, E. T. **Action Research: a Handbook for Practitioners**. Sage, 1996.

TAN, P-N. STEINBACK, M. KUMAR, V.; **Introdução ao Data Mining**, Rio de Janeiro, Ciência Moderna, 2009.

THIOLLENT, M. **Pesquisa-Ação nas Organizações**. São Paulo: Atlas, 1997.

VAPNIK, V. N.; **The nature of Statistical learning theory**. Springer-Verlag, New York, 1995.

VAPNIK, V. N.. **Statistical Learning Theory**. John Wiley and Sons, 1998.

VOGEL, C. **Bioinformática, genes e inovação**, Revista eletrônica ComCiência.br, 2003. Disponível em <http://www.comciencia.br/reportagens/bioinformatica/bio01.shtml>. Acesso em 19 de Julho de 2010.

VOGT, C. **A espiral da cultura científica**. *ComCiência* . 2003. Disponível em: <http://www.comciencia.br>. Acesso em: jul. 2010.

WYMIRE, R. **Microsoft SQL Server 7.0, Aprenda em 21 dias**; Rio de Janeiro; Campus; 1999.