Estudo para Identificação de Promotores Procarióticos Através de Classificação Bayesiana

Ilueny Constâncio C. dos Santos, Robinson Luis de S. Alves
Depto. de Informática, CEFET,
59015-000, Natal, RN
ilueny@yahoo.com.br, robinson@cefetrn.br

Carlos Alfredo G. Blaha
Depto. Microbiologia e Parasitologia/CB, UFRN,
59072-917, Natal, RN
carlos.blaha@pq.cnpq.br

Resumo: A Bioinformática ou Biologia Computacional diz respeito à utilização de técnicas e ferramentas de computação para a resolução de problemas da Biologia. Dentre as diversas áreas da Biologia, aquela em que a aplicação de técnicas computacionais tem se mostrado mais promissora é a Biologia Molecular e nos estudos de Biologia Genômica. Nesse amplo contexto, a computação pode ser aplicada na resolução de problemas na identificação e dos mecanismos de regulação gênica em organismos eucariotos e procariotos. Este trabalho apresenta um método de aprendizado para um classificador Bayesiano demonstrando sua aplicação na tarefa de reconhecimento de promotores procarióticos apresentando conceitos básicos de bioinformática necessários ao entendimento do problema e da solução apresentada. Nesta proposta foi implementado, em Matlab e Java, um modelo de classificador tendo por base o método de *Naive Bayes*, capaz de identificar promotores procarióticos reconhecidos pelo fator transcricional *Sigma70*. A base de treinamento foi composta por um conjunto de dados previamente selecionados do genoma de *Escherichia coli* K12 e durante os testes realizados o método demonstrou eficiência superior a 90% na identificação dos promotores desta bactéria.
Com este resultado, o trabalho proposto mostra sua utilidade contribuindo para estudos mais detalhados que envolvam o problema da identificação de promotores de procariotos.

Palavras-chave: Bioinformática 1. DNA 2. Classificação de Padrões 3. Reconhecimento de Promotores 4. Procariotos 5.

Introdução

Com a descoberta de que o DNA é estruturado com dupla hélice por Watson e Crick em 1953 [9], já foi possível prever o aparecimento da Bioinformática. Uma ciência que relaciona muito bem o estudo da vida (Biologia) com os conhecimentos da matemática, física, química e outros ramos das ciências exatas. Atualmente a Bioinformática ocupa lugar de destaque no mundo da ciência biológica aplicada (notadamente em todos os ramos da genética) quer seja automatizando tarefas, ou mostrando caminhos que seriam praticamente impossíveis de serem descobertos somente com a utilização da prática, na bancada do laboratório.

Com o surgimento dos seqüenciadores de DNA, por volta da segunda metade dos anos 90, a quantidade de dados biológicos disponíveis na internet aumentou exponencialmente e continua crescendo. Apesar do volume de dados, a quantidade de informação inteligível presente neles ainda é pequena, os cientistas ainda não conseguiram decifrar todos os mistérios que envolvem os genomas. Os benefícios que estes dados podem proporcionar, quando forem compreendidos em sua plenitude algum dia, são incalculáveis e com horizonte de aplicações sem limite, mas podemos citar a fabricação de remédios, vacinas revolucionárias, possibilidade de produção de alimentos mais resistentes à pragas, biotecnologias de remediação para o meio ambiente, entre outras.

O conhecimento da seqüência nucleotídica do DNA de um genoma por si só não tem significado sem a atribuição de uma função e isso nos leva a duas questões críticas: i) como identificar um gene e ii) como atribuir uma função ao produto protéico por ele codificado. No contexto da primeira questão a identificação de promotores é ponto fundamental no entendimento funcional do mecanismo primário na regulação gênica das células.

A aplicação, em conjunto, de modelos matemáticos e estatísticos aliados a métodos computacionais de resolução de problemas tem, ao longo dos últimos 20 anos, demonstrado pouca eficiência quando utilizada de forma convencional na abordagem de problemas de biologia molecular de alta complexidade [3]. As técnicas de Aprendizado de Máquinas (AM), entretanto, tem apresentado resultados bastante satisfatórios [11].

A utilização do modelo estatístico em que se baseia o método de *Naive Bayes* se justifica pela simplicidade de implementação associada à eficiência, comprovada quando aplicado sob um conjunto de dados apropriados. A técnica de classificação de *Naive* Bayes é baseada no também chamado, teorema Bayesiano que é particularmente eficiente quando a dimensão dos dados de entrada é alta. Leve e simples, o *Naive* Bayes pode, frequentemente, apresentar uma performance superior a muitos sofisticados métodos de classificação [13].

Metodologia

Este trabalho divide-se basicamente em 3 partes: **i)** a construção do referencial teórico metodológico; **ii)** a identificação de um problema de biologia molecular no qual se aplique os conhecimentos adquiridos no ponto anterior, fornecendo subsídios para o desenvolvimento de um software que auxilie na resolução do problema identificado; e **iii)** a aplicação de testes e análise dos algoritmos empregados no software.

Desenvolvimento

Construção do Referencial Teórico

A *E. coli* K12 é um organismo modelo que vem sendo amplamente estudado desde o surgimento da biologia molecular. Atualmente, na éra pós-genômica, pesquisas em Bioinformática, análise e modelagem de redes de regulação e outros sistemas biológicos, podem ser altamente beneficiados pelos estudos acumulados realizados ao longo desses anos com a *E. coli* K12 [10].

O Reconhecimento de Padrões (RP) é provavelmente uma das áreas da ciência de maior abrangência, estudada em diversos campos de pesquisa como biologia, psicologia, medicina, marketing, finanças, meteorologia, sensoriamento remoto, processamento de imagens, entre outros [12]. Em problemas complexos de RP, a abordagem comum utilizada é dividir o problema em dois módulos subseqüentes: a extração de características e o módulo de classificação. Tanto para o primeiro como para o segundo módulo já existem várias técnicas desenvolvidas que aplicam desde conhecimentos estatísticos e matemáticos até Inteligência Artificial (IA) [1]. O presente trabalho aborda a problemática proposta utilizando um classificador baseado em técnicas estatísticas.

Aplicando Reconhecimento de Padrões na resolução de um problema de biologia molecular

A compreensão do comportamento da enzima RNA Polimerase (RNAP), no desempenhar de suas funções, é de fundamental importância no estudo de uma forma eficaz de predizer regiões promotoras em determinados trechos de DNA. A caracterização dessas regiões nos promotores, conhecidas também como Sítios de Ligação, descrevendo quais as seqüências são mais comumente reconhecidas pelo fator, sigma ou então, até que ponto uma mutação no sítio de ligação impede o acoplamento do RNAP ao DNA, poderá servir como base na elaboração de um algoritmo capaz de simular o comportamento do RNAP reconhecendo os padrões dispostos na seqüência do sítio de ligação no promotor. Neste contexto, o trabalho hora apresentado, vem propor um software desenvolvido em linguagem de programação Java, que implementa um classificador de padrões baseado na regra de *Naive* Bayes objetivando a identificação de promotores reconhecidos pelo fator sigma.

O Classificador Estatístico e a Regra de *Naive* Bayes

Um método de aprendizagem é um algoritmo (usualmente implementado em software) que estima um mapeamento desconhecido (dependência) entre entrada do sistema e saídas baseado nos dados disponíveis, isto é, nas amostras (entradas/ saídas) conhecidas [4]. Um vez que a dependência tem sido estimada, ela pode ser usada para a predição das saídas futuras do sistema baseado nos valores de entrada conhecidos.

Em estatística, a tarefa de aprendizado preditivo (baseado nas amostras) é chamada de Estimação Estatística; ela resulta das propriedades de estimação de alguma distribuição estatística (desconhecida), tendo como base as amostras conhecidas ou dados de treinamento. Informações contidas nos dados de treinamento (experiências passadas) podem ser usadas para responder questões sobre amostras futuras [8].

A técnica de classificação de *Naive* Bayes é baseada no também chamado, teorema Bayesiano que é particularmente eficiente quando a dimensão dos dados de entrada é alta. Leve e simples, o *Naive* Bayes pode, freqüentemente, apresentar uma performance superior a muitos sofisticados métodos de classificação[13].

Os classificadores estatísticos baseados na Regra de *Naive* Bayes consideram que as entradas são independentes entre si, o que na maioria dos problemas

práticos não é verdade (daí o nome *naive* ou ingênuo). Esta suposição simplifica a abordagem do problema de classificação sem comprometer significativamente a precisão do resultado.

Determinação de características fundamentais dos promotores

A delimitação do espaço de características é de fundamental importância na montagem de um classificador eficiente.

Em *E. coli* K12, um grupo especial de promotores é identificado por uma enzima que sofre uma forte influência de um fator conhecido pelo nome de sigma. O fator *Sigma70* é a subunidade mais abundante que reconhece regiões promotoras de *E. coli*, mesmo quando ocorrem pequenas alterações na seqüência de nucleotídeos do DNA alvo. Adicionalmente, existem aproximadamente seis fatores sigma alternativos de acordo com as condições sobre as quais se observam as atividades da célula.

A RNAP reconhece os elementos hexaméricos[1] na seqüência promotora, posicionados a aproximadamente 7 pares de bases do início da transcrição, conhecidos como *box -10* e *-35* cujas seqüências consenso são (TATAAT) e (TTGACA) respectivamente [8]. Quando estas regiões são ocupadas pelo fator *Sigma 70* e a holoenzima, cobrem uma distância de 77 pares de bases (pb) [6]. As regiões -10 e -35 são separadas por uma distância variável que oscila entre 15 e 21 pb. Em alguns promotores, existe uma estrutura rica em AT (conhecida como elemento UP) que é situado a aproximadamente 20 nucleotídeos upstream da região -35 e que reconhece a subunidade α do RNAP [6]. A influência da interação entre o promotor e a RNAP depende diretamente de variantes no consenso2, bem como as modificações no espaço que as separam [6].

Dentre as características observadas as mais significativas e que portanto, preservam o poder discriminatório do padrão são as seguintes:

a) presença do hexamero -35;
b) presença do hexamero -10;
c) quantidade de pares de bases presentes entre os hexameros -35 e -10;

Obtenção de Dados Válidos

Basicamente, foram encontradas duas formas viáveis, para obtenção scores de seqüências reais de promotores conhecidos em *E. coli* K12 são elas: a) Pesquisa no banco de dados público RegulonDB, b) Através dos softwares de alinhamento WSONSENSUS e PATSER.

O RegulonDB contribui para o desenvolvimento de pesquisas como estas, armazenando em um banco de dados relacional dados detalhados de promotores que fazem parte do genoma da *E. coli* K12. Ele pode ser acessado através da URL: http://regulondb.ccg.unam.mx/.

Os softwares WSONSENSUS e PATSER foram desenvolvido por Gerald Z. Hertz do Departamento de Biologia Molecular, Celular e Desenvolvimento da Universidade do Colorado nos Estados Unidos.

O WCONSENSUS determina o padrão de consenso em seqüências não alinhadas. Seu algoritmo é baseado em uma matriz de representação para um padrão de consenso definido conforme as seqüências de entrada onde cada linha da matriz representa uma letra relevante do alfabeto.

O PATSER define os *scores* da seqüência indicada a partir de uma determinada matriz de pesos. Os elementos da matriz peso, podem muito bem ser usados para representar uma seqüência de DNA de determinado comprimento [14].

O uso associado dos softwares WCONSENSUS e PATSER na análise de seqüências de DNA permite, através dos *scores* gerados, estabelecer um nível de similaridade entre as seqüências analisadas.

Após os estudos realizados durante a fase de obtenção dos *scores* de regiões promotoras reais, das soluções encontradas, aquela que se mostrou mais adequada e segura para implementação do software foi a pesquisa a partir do banco de dados público RegulonDB pois, os *scores* lá disponíveis, já se encontram validados e inclusive servem de referência para diversas outras pesquisas já publicadas.

Determinação do tipo de classificador

Nesta fase o software de classificação de padrões foi inicialmente concebido. Baseado nos estudos realizados durante a fundamentação teórica, optou-se pelo classificador bayesiano e posteriormente foram realizados testes para confirmação desta decisão.

O desenvolvimento do classificador de padrões em GNU Octave

O GNU Octave é um software *Open Source* (de código aberto) distribuido sob licença GNU General Public License (GPL) escrito por John W. Eaton que tem funções bastante semelhantes ao MATLAB. O objetivo principal da implementação do classificador em GNU Octave foi a verificação da eficácia do método de classificação adotado na solução

1 Elemento formado por uma seqüência de 6 nucleotídeos.

2 Chama-se de região ou seqüência consenso um conjunto de nucleotídeos que ocorrem com maior freqüência em uma determinada região do DNA, RNA ou Proteína.

do problema de que trata este trabalho, sem, no entanto, ter que desempenhar um volume muito grande de trabalho nesta tarefa.

O desenvolvimento do classificador de padrões em Java

Processo de software utilizado

O ICONIX foi o processo de software seguido durante o desenvolvimento da aplicação. A seguir, são apresentadas algumas características do ICONIX que justificaram sua adoção no projeto. a) possuir uma metodologia prática, intermediária entre a complexidade do processo unificado (UP) e a simplicidade da Programação Extrema (XP); b)possuir uma abordagem flexível e aberta à UML; c)seu processo é interativo incremental e não exige marcos formais de projeto para o refinamento, isso é feito de forma natural durante o projeto.

Descrição do sistema

O sistema proposto possibilita a classificação eficiente de seqüências de DNA conhecidas como promotores. Os dados de entrada são fornecidos e o sistema, baseado em um conjunto de informações anteriormente catalogadas e caracterizadas para treinamento do software, determina qual tipo de promotor está presente na seqüência de DNA de entrada.

Diagramas de Casos de Usos

Esta tarefa tem como objetivo principal identificar as funcionalidades do sistema baseado nas interações dos atores envolvidos.

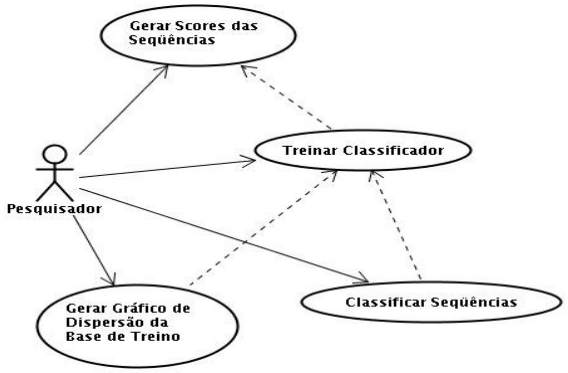

Figura1: Diagrama de Casos de Uso

Identificação dos Requisitos Funcionais

A lista de requisitos funcionais identificados foi a seguinte: a) o sistema deve receber uma seqüência de DNA promotora (ex. ATAGACCGA....), em um formato pré determinado, ou outro conjunto de dados que a represente (ex. matriz de scores), via formulário, em um web browser; b) o sistema deve ser capaz de classificar qual o tipo de promotor presente na seqüência de DNA informada. c) os resultados devem ser expressos graficamente facilitando a visualização dos dados pelo pesquisador.

Implementação e Testes

Durante a implementação do sistema alguns testes unitários foram desenvolvidos utilizando-se das facilidades oferecidas pelo *framework JUnit.* As interações se seguiram orientadas pelos testes até que modelo final atendesse os requisitos do usuário do software e conseqüentemente os teste fossem realizados sem erro algum.

Entrega do software para produção

O software resultante do processo de produção aqui descrito atendeu a de forma satisfatória aos requisitos propostos no início da análise, deixando pendente somente um maior aprimoramento do requisito de interface amigável com o usuário, no entanto, atendeu a todos os outros requisitos funcionais propostos e cumpriu com a função básica a que pretendia inicialmente.

Resultados

O Software, produto final deste trabalho de pesquisa, apresentou resultados bastante satisfatórios, do ponto de vista estatístico, se for considerar o volume de dados ao qual foi submetido (conjunto de 99 amostras de tipo de promotores, presentes em *E. coli* K12) [2];

Durante os testes observou-se o seguinte resultado:

a) Ao classificador, foi submetido um conjunto de 99 seqüências reais de promotores presentes no DNA de *E. coli* K12, divididos igualmente entre 3 classes: promotores reconhecidos pelo fator sigma 70, sigma 32 e sigma 24. A finalidade foi de treiná-lo para posterior classificação correta de outras seqüências, as quais ele não conhecia ainda, e;

b) Um conjunto de dados que não faziam parte do seu banco de dados de treino, lhe foi submetido para análise o, classificador apresentou o seguinte desempenho:

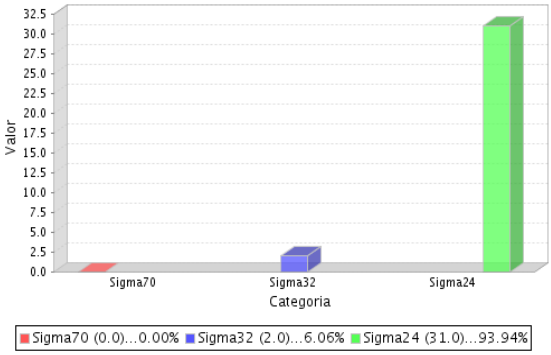

Figura 2: Testes de classificação para promotores reconhecidos pelo fator *Sigma24*.

Na Figura 2, pode ser observado o resultado de um dos testes do software com um conjunto de 33 promotores do tipo sigma24. Destes, todos[3] já haviam sido utilizados, durante a fase de treino do classificador. O software identificou:

- 31 amostras corretamente (93,9% de precisão);
- 2 amostras de forma errada (pouco mais de 6% de falso negativo).

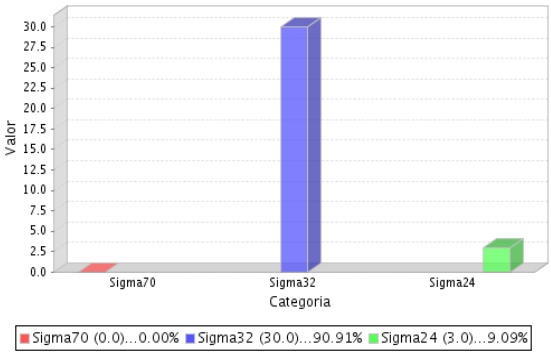

Figura 3: Teste de classificação para promotores *Sigma32*.

Na Figura 3 pode-se observar claramente que de um conjunto de 33 promotores do tipo *Sigma32*, onde somente pequena parte já haviam sido utilizadas na fase de treino do classificador, o software identificou:

- 30 amostras corretamente (90,9% de precisão);
- 3 falsos negativos (pouco mais de 9%)

3 Dadas as dificuldades de obtenção de dados reais disponíveis no banco de dados RegulonDB, no momento em que esta pesquisa foi realizada, os testes tiveram que ser executados com o mesmo grupo de dados com os quais foi realizado o treino.

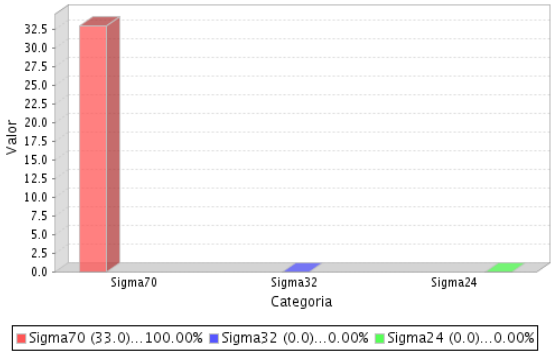

Figura 4: Testes de classificação para promotores *Sigma70*

De um conjunto de 33 promotores totalmente desconhecidos pelo classificador, mas que era sabido serem do tipo sigma 70, o software obteve 100% de classificação correta, não apresentando falsos positivos em nenhuma das 33 amostras analisadas. O resultado pode ser observado na Figura 4.

Pede-se observar ainda, através da Figura 5, o comportamento dos grupos de dados a que o classificador foi submetido durante os treinos. A análise deste gráfico ajuda a entender porque o software identificou erradamente parte dos promotores durante os testes.

A Figura 5 mostra que alguns elementos das classes *Sigma32* e *Sigma24* se misturam, em suas regiões limítrofes, "confundindo" o classificador no momento da tomada de suas decisões. Já a classe composta por amostras de promotores *Sigma70*, concentra-se em uma região bem definida, motivo pelo qual, o software classifica todos os promotores deste tipo sem cometer erro algum.

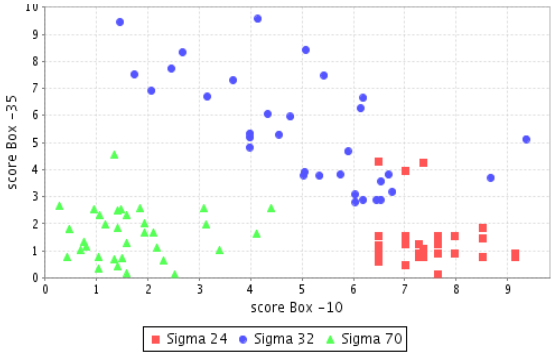

Figura 5: Gráfico de dispersão do conjuto de dados de treino do classificador.

Conclusões

Este trabalho abordou (principalmente) a forma como a informática, mais precisamente, as teorias envolvidas na área do reconhecimento de padrões, podem ser úteis na resolução de alguns problemas biológicos, para tanto, detalhou, por meio

de uma aplicação simples, desenvolvida em Java como isso é possível.

O software proposto apresentou resultados satisfatórios, conforme mostrado na seção anterior, se for levado em consideração o escopo inicialmente definido. Além disso, a aplicação desenvolvida pode ser utilizada para reconhecimento de padrões e classificação de características de outros organismos diferentes da bactéria *Eschericha Coli -K12,* com a qual foi testado.

O aprendizado adquirido foi de suma importância, principalmente porque permitiu adquirir conhecimentos também em outra área e possibilitou uma introdução, ainda de modesta, o vasto mundo de conhecimento proporcionado pela bioinformática.

Apesar do bom desempenho mostrado pelos testes realizados com o software proposto para solução do problema de reconhecimento de promotores, ficou claro que faz-se necessário um aprimoramento ainda maior do algoritmo de reconhecimento de padrões adotado na tentativa de reduzir, ainda mais, a taxa de erros de classificação. Outra melhoria que dever ser considerada diz respeito à adaptação do formato dos dados de entrada, de modo que, o pesquisador possa fornecer os dados não em uma matriz de *scores*, mas diretamente no formato de *string* (AGTAGCAGT).

Referências

[1] BISHOP, Christopher M. Neural Networks for Pattern Recognition. Oxford University Press, 1995.

[2] BOCKHORST, Joseph et al. A Bayesian Network Approach to Operon Prediction, Department of Biostatistics and Medical Informatics - University of Wisconsin, Junho de 2002.

[3] BORGES, Helyane Bronoski, Estudo comparativo entre algorítimos de seleção de atrubutos para dados de microarranjos em bioinformática; Pontifícia Universidade Católica do Paraná. Curitíba Abril de 2005.

[4] CHERKASSKG, Vladimir & MULIER, Filip. Learning from Data: Concepts, Theory, and Methods. New York: John Wiley & Sons, 1998, 441p.

[5] COLLADO-VIDES, Julio et al: Sigma70 promoters in *Escherichia Coli:* Specific Transcription in Dense Region of Overlapping Promoter-like Signals. J. Mol. Biol. 333, 261-278.

[6] DOMBROSKI, A.J., Walter, W.A., Record, M.T., Jr., Siegele, D.A. And Gross, C.A. 1992. Polypeptides containing highly conserved regions of transcription initiation factor sigma 70 exhibit specificity of binding to promoter DNA. Cel l 70, 501-12.

[7] DUDA, O., HART, P. E. Pattern classification and scene analysis. John Wiley & Sons, Inc., 1973.

[8] LACERDA, Willian Soares, BRAGA, Antônio de Pádua: Experimento de um Classificador de Padrões Baseado na Regra de *Naive* Bayes: Minas Gerais.

[9] LODISH, H. BERK, A.; MATSUDAIRA, P.; KAISER, C.A.; KRIEGER, M.; SCOTT, M.P.; ZIPURSKY, L. AND DARNELL. J. (2003). Molecular Cell Biology, Editora W. H. Freeman, New York, 5ta Edição.

[10] SALGADO H, Collado-Vides J. et al. RegulonDB (version 4.0): Transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res*. 2004 Jan 1; 32 Database issue: D303-6

[11] SOUTO, M. C. P. de, e Lorena, A. C., A. C. B. Delbem, A. C. P. L. F. de Carvalho: Técnicas de aprendizado de máquina para problemas de biologia molecular.

[12] SOUSA, de João Artur, Reconhecimento de padrões usando indexação recursiva – Tese submetida à Universidade Federal de Santa Catarina para a obtenção do título de Doutor em Engenharia da Produção. Florianópolis, setembro de 1999.

[13] STATSOFT, Inc Eletronic Statistics Textbook., Naive Bayes Classifier : http://www.statsoft.com/textbook/stathome.html: StatSoft, Inc. 1984-2004, acesso em: 12 dez. 2005.

[14] STORMO, G. D. (2000). DNA biding sites: representation and discovery. Bioinformatics, 16(1):16-23.

[15] SETÚBAL, J. C. e Meidanis. J. Introduction to Computational Molecular Biology. PWS Publishing Company, 1997.