Análise de Classificadores de Seqüências Projetados por Aprendizado Computacional Supervisionado e não Supervisionado

Caetano Jimenez Carezzato

Relatório Científico

Período: Julho/2001 – Janeiro/2003

Bolsa de Mestrado **FAPESP** nº 01/03975-5

Vinculado ao Projeto Temático **CAGE** — **FAPESP** nº 99/07390-0

Orientador: **Professor Doutor Junior Barrera**

Co-orientador: **Sandro José de Souza**

Departamento de Ciência da Computação

Instituto de Matemática e Estatística

Universidade de São Paulo

Sumário

1 Resumo do Projeto

Com o crescente número de genomas seqüenciados sendo disponibilizados atualmente [75], inclusive o humano[108, 54], um problema muito importante que surge imediatamente na área de Biologia Molecular é extrair informações desses enormes bancos de dados de seqüências.

A Biologia Molecular Computacional [67] consiste basicamente no desenvolvimento e uso de técnicas matemáticas e de Ciência da Computação para auxiliar a solução de problemas da Biologia Molecular.

Diversos problemas vêm sendo estudados nessa área: a comparação de seqüências de **DNA** [30, 118], montagem de fragmentos de **DNA** [2], mapeamento físico de **DNA** [3], árvores filogenéticas [119], reconhecimento de genes e partes de genes [95, 18], busca de homologia [48], clustering [29, 33], predição da estrutura de proteínas [91] etc.

O objetivo principal deste trabalho é estudar diversos métodos computacionais disponíveis atualmente para clustering e busca de homologia em seqüências de **DNA**, **RNA** e proteínas. Para tal, dentre os diversos modelos probabilísticos disponíveis para modelagem de seqüências [32], modelaremos os dados por Gramáticas Estocásticas [43] que serão estimadas a partir de dados reais utilizando-se diversas técnicas de reconhecimento de padrões [31, 109] e Aprendizado Computacional [112, 8].

Este trabalho está vinculado ao Projeto Temático **CAGE** [1] (do inglês, "*Cooperation for Analysis of Gene Expression*") (**FAPESP** nº 99/07390-0), que une esforços do Instituto de Química e do Instituto de Matemática e Estatística da Universidade de São Paulo com o objetivo de estudar os mecanismos de expressão gênica.

Uma versão eletrônica deste documento com links para alguns dos documentos citados pode ser encontrada em:

http://www.vision.ime.usp.br/∼caetano/mestrado/projeto/relatorio-fapesp-2003-01.pdf

Uma versão eletrônica do primeiro relatório do projeto pode ser encontrada em:

http://www.vision.ime.usp.br/∼caetano/mestrado/projeto/relatorio-fapesp-2002-07.pdf

Uma versão eletrônica da proposta original pode ser encontrada em:

http://www.vision.ime.usp.br/∼caetano/mestrado/projeto/proposta.pdf

[1] Para mais informações, sobre o grupo, http://www.vision.ime.usp.br/∼cage/

2 Plano de Trabalho e Cronograma (Inicial vs Realizado)

Ao longo dos últimos 3 semestres realizamos diversas atividades. Um resumo objetivo de tais tarefas, atualizado de acordo com o cronograma apresentado no relatório anterior, é apresentado a seguir.

2.1 Plano de Trabalho (Inicial vs Realizado)

Legenda

Executado	$\surd$
Parcialmente executado	$\circ$
A ser executado	$\bullet$

Atividades		
1	Disciplinas do programa de pós-graduação	$\surd$
2	Leitura supervisionada de [109]	$\surd$
3	Estudo de Biologia Computacional	$\surd$
4	Estudo de linguagens formais e gramáticas	$\surd$
5	Estudo do **Matlab** e **GCG** e suas ferramentas	$\surd$
6	Especificação do gerador estocástico de palavras	$\surd$
7	Especificação do Ambiente de Treinamento	$\circ$
8	Especificação dos Testes	$\circ$
9	Implementação do gerador estocástico de palavras	$\surd$
10	Implementação dos Ambientes Especificados	$\circ$
11	Implementação dos Testes	$\circ$
12	Primeira avaliação dos algoritmos	$\surd$
13	Segunda avaliação dos algoritmos	$\bullet$
14	Redação de relatórios semestrais para a FAPESP	$\surd$
15	Redação da proposta da dissertação	$\circ$
16	Seminários	$\circ$
17	Avaliação dos resultados	$\bullet$
18	Redação da dissertação	$\bullet$

2.2 Cronograma (Inicial vs Realizado)

proposto	Executado	$\surd$
	Parcialmente executado	$\circ$
	A ser executado	$\bullet$
	Não será executado	$\times$
não proposto	Executado	$\star$
	A ser executado	$\cdot$

2001/2002

	Ago	**Set**	**Out**	**Nov**	**Dez**	**Jan**	**Fev**	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**
1	$\surd$	$\surd$	$\surd$	$\surd$		$\surd$	$\surd$	$\surd$	$\surd$			
2					$\times$	$\times$	$\circ$	$\circ$	$\surd$	$\star$	$\star$	$\star$
3	$\surd$	$\surd$					$\surd$	$\surd$	$\star$	$\star$	$\star$	$\star$
4	$\surd$	$\surd$	$\surd$	$\surd$								
5		$\surd$	$\surd$	$\surd$	$\surd$		$\surd$		$\surd$		$\surd$	
6	$\surd$											
7					$\times$	$\times$	$\circ$				$\star$	$\star$
8								$\times$	$\times$	$\surd$	$\star$	$\star$
9		$\surd$										
10								$\circ$	$\circ$	$\circ$	$\surd$	$\surd$
14						$\times$						$\surd$

2002/2003

	Ago	**Set**	**Out**	**Nov**	**Dez**	**Jan**	**Fev**	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**
1	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$							
2	$\surd$	$\surd$										
7	$\circ$	$\circ$	$\star$	$\star$			$\bullet$	$\bullet$				
8	$\circ$		$\surd$		$\surd$		$\bullet$		$\bullet$			
10	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$		$\bullet$	$\bullet$			
11	$\circ$	$\circ$	$\circ$	$\star$	$\surd$	$\surd$		$\bullet$	$\bullet$			
12		$\circ$	$\circ$	$\surd$	$\surd$	$\star$						
13							$\bullet$	$\bullet$	$\bullet$			
14						$\surd$						
15	$\times$			$\times$	$\times$	$\cdot$	$\cdot$					
16			$\surd$	$\surd$	$\surd$	$\circ$	$\bullet$	$\cdot$	$\cdot$			
17									$\bullet$	$\bullet$	$\bullet$	
18										$\bullet$	$\bullet$	$\bullet$

3 Principais Realizações

O objetivo desta seção é apresentar um detalhamento das realizações efetuadas durante o projeto. Inicialmente apresentamos um resumo das tarefas concluídas no período do relatório anterior e logo em seguida iniciamos uma descrição das tarefas realizadas no período ao qual este relatório se refere.

3.1 Resumo do período anterior

Ao longo do período anterior a este relatório, o bolsista completou todos os crédito do programa de pós-graduação através da conclusão de 7 disciplinas (todas com o conceito **A**). Além disso, o bolsista foi aprovado no teste de inglês em 22/02/2002.

Dentre as principais tarefas relativas ao projeto realizadas no período, pode-se destacar o extenso estudo de técnicas de Biologia Computacional (principalmente os aspectos computacionais), bem como dos softwares **Matlab** [124], **GCG** [120] e suas ferramentas. Também especificamos e implementamos um gerador estocástico de palavras.

Foram também realizados a especificação e implementação dos ambientes de treinamento e testes.

Tendo em vista a extensa bagagem matemática e computacional do curso de graduação do bolsista, principalmente em relação aos aspectos teóricos da Computação, foi dado ênfase à leitura introdutória em Biologia Computacional e à leitura avançada na área de reconhecimento de padrões.

3.1.1 Resumos das leituras do período anterior

Apesar do bolsista ter realizado durante alguns anos um projeto de Iniciação Científica na área de reconhecimento de padrões, o mesmo não era muito familiar com Biologia (apesar de ser um pouco familiar com Biologia Computacional). Foi feito então um estudo introdutório de nivelamento e logo em seguida uma vasta revisão bibliográfica que é necessária para este trabalho. Foram lidos diversos livros, artigos, tutoriais e sítios na Internet visando o estudo aprofundado de técnicas de reconhecimento de padrões e aplicações na área de Biologia.

O bolsista participou de dois cursos em Biologia Computacional. Ambos os cursos foram ministrados na USP.

3.1.2 Resumos das especificações e implementações do período anterior

O objetivo do nosso projeto, é estudar ferramentas baseadas em Aprendizado Computacional, mas, muitas vezes, tais ferramentas são utilizadas em conjunto com ferramentas baseadas na construção analítica de algoritmos, ou seja, temos algorítmos híbridos. Resolvemos, então, especificar um ambiente de testes capaz de lidar com as diversas técnicas de Aprendizado Computacional, mas não só elas, de forma completamente transparente e simples de usar.

Para tal, acreditamos que o ponto mais forte da nossa especificação deve ser a forma com que os dados são armazenados e as relações entre os mesmos. Ou seja, estamos especificando um ambiente que seja flexível o suficiente para que se possa abordar qualquer problema na área sem a necessidade de se reescrever o ambiente, bastando extendê-lo de forma simples. Além disso, o modelo deve ser maduro o suficiente para acompanhar as mudanças na área que são muito rápidas, tanto em relação aos dados disponíveis quanto em relação aos tipos de dados disponíveis, ou seja, o modelo deve ser de fácil atualização ou até mesmo com mecanismos de atualização automática dos dados quando for o caso.

Ao longo do período em questão, realizamos três especificações principais. Especificamos o gerador estocásticos de palavras, especificamos o ambiente de trabalho e especificamos alguns testes. Realizamos também uma implementação dos ítens citados.

Implementamos o algoritmo para geração de imagens **CGR** a partir de seqüências de DNA descrito por Deschavanne [26]. Fomos capazes de reproduzir completamente o artigo e realmente os resultados obtidos foram muito interessantes. Criamos um programa que tanto gera as imagens descritas, como as tabelas brutas necessárias para gerar as imagens.

3.1.3 Alterações no Projeto propostas anteriormente

Após o término da fase inicial de estudos, nos familiarizamos melhor com as principais ferramentas e técnicas de Aprendizado Computacional utilizadas na área. Dentre tais ferramentas, podemos citar Redes Neurais, Modelos Escondidos de Markov, Gramáticas Estocásticas e algoritmos de Clustering. Também nos familiarizamos melhor com algumas técnicas analíticas, como o uso de estatísticas de variação na utilização de códons para predição de genes, por exemplo.

Tendo em vista os conhecimentos obtidos e os principais problemas atualmente enfrentados

na área, decidimos abordar o problema de predição de genes. Pretendemos seguir uma abordagem sugerida pelo co-orientador do projeto, cujo objetivo não é achar genes como se tenta fazer com o uso dos *softwares* de predição disponíveis atualmente na literatura, mas sim tentaremos encontrar genes "raros" (ou seja, genes que raramente se expressam) através da utilização de informações bem específicas, sendo possível, inclusive, a utilização de técnicas híbridas e altamente especializadas. Desta forma, optamos por abrir mão momentaneamente de dados simulados por gramáticas estocásticas como havíamos proposto originalmente. Passamos a utilizar dados reais devido à natureza do problema.

É importante ressaltar que fizemos uma ligeira correção nos rumos do projeto. O objetivo principal deixa de ser comparar diversos algoritmos disponíveis na área, passando a ser um estudo mais aprofundado das diversas ferramentas de reconhecimento de padrões e de algumas técnicas analíticas e estatísticas, bem como uma comparação entre as diversas formas de combinação de tais ferramentas na tentativa de resolver o problema em questão: predição de genes raros.

Outra pequena modificação que fizemos no trabalho foi em relação ao cronograma. Devido à dedicação praticamente exclusiva que foi dada às disciplinas do programa, houve um pequeno atraso em relação às outras tarefas, mas tal fato já foi compensado nos últimos meses. Além disso, apesar de todos os créditos já terem sido conluídos pelo bolsista, o mesmo optou por cursar uma disciplina extra chamada "Biologia Computacional".

3.2 Questão levantada pelos revisores

Os examinadores levantaram a seguinte questão a respeito do trabalho:

"Reserva Técnica: A memória extra no PC foi justificada por algoritmos de clustering, mas o bolsista vai agora procurar genes raros. O trabalho sobre clustering vai ficar de lado?"

O objetivo do nosso trabalho é estudar tanto técnicas supervisionadas quanto técnicas não-supervisionadas (*clustering*). Escolhemos a título de motivação o problema de procurar genes raros porque ambos os tipos de técnicas são necessárias para a abordagem de tal tarefa.

Num primeiro momento, experimentos com *clustering* são necessários para auxiliar na avaliação de um determinado caminho a ser seguido. Tais experimentos permitem que se tenha uma idéia da performance que algoritmos supervisionados obteriam. Um exemplo de experimentos que realizamos nesta linha é apresentado na seção 3.6.

De um ponto de vista mais teórico, podemos ver qualquer algoritmo de aprendizado computacional supervisionado como *clustering* seguido da definição de uma função de erro ou distância.

Por exemplo, no treinamento de uma rede neural, aparentemente não há *clustering*, mas, em sua essência, a definição da superfície de decisão que vai sendo construída conforme o algoritmo utilizado, nada mais é do que a definição de regiões (agrupamentos), ou seja, um *clustering*. Além disso, a superfície de decisão em si é a própria função de erro ou distância. Note que os pontos agrupados não são necessariamente os pontos dados como entrada, podem ser pontos produzidos a partir dos pontos de entrada como se faz numa redução de um problema a outro. Qualquer algoritmo de Aprendizado Computacional podem ser reduzido utilizando técnicas como essas.

Ao estudarmos *clustering*, estamos estudando também Aprendizado Supervisionado. Uma compreensão mais profunda de Aprendizado, principalmente supervisionado, só é alcançada estudando-se técnicas não supervisionadas. Tais estudos permitem ao aluno entender e compreender melhor as relações entre as diversas áreas de reconhecimento de padrões.

3.3 Resumo das realizações deste período

Dentre as principais tarefas realizadas no período, podemos destacar a conclusão de duas disciplinas extras, a realização de alguns experimentos e o início de um estudo mais aprofundado (do ponto de vista computacional) de técnicas de reconhecimento de padrões.

O bolsista também apresentou parte de seu trabalho no "1ºWorkshop do projeto Fenótipos".

Atualmente estamos realizando três tarefas principais:

1. Especificação e realização de alguns experimentos
2. Redação da proposta de dissertação
3. Leituras avançadas na área

3.4 Disciplinas Extras do Programa de Pós-Graduação

A seguir estão as disciplinas extras cursadas pelo bolsista ao longo do período deste relatório. Mais detalhes a respeito de cada disciplina podem ser encontrados no histórico escolar anexo ou no catálogo de disciplinas de Ciência da Computação [123].

MAC5700	Seminários em Ciência da Computação
MAC5726	Biologia Computacional

Ambas as disciplinas foram concluídas com o conceito **A**. Foram ao todo concluídas 9 disciplinas, todas com conceito **A**, totalizando 68 créditos.

Durante este semestre, tivemos a oportunidade de assistir à vários seminários, foram 10 no total. O bolsista apresentou também um seminário sobre algumas técnicas de aprendizado computacional, dando ênfase aos Modelos de Markov como os *Modelos Escondidos de Markov*[2] (**HMM**).

Durante a realização da disciplina *Biologia Computacional*, o bolsista teve a oportunidade de rever alguns conceitos básicos, como o de alinhamento, bem como estudar alguns tópicos em Biologia Computacional que não são abordados neste projeto, como montagem.

Além disso, o bolsista realizou uma implementação incrementada de árvores de sufixo que economiza memória proposta por Kurtz [61] . Acreditamos que tal implementação eventualmente seja útil para alguns experimentos que poderão ser realizados, como os que eventualmente usem o cálculo de tabelas de freqüência de palavras grandes.

3.5 Implementações Realizadas

Dentre as implementações realizadas, podemos destacar a finalização do ambiente de testes como descrito no relatório anterior.

Outra implementação importante foi a implementação de uma versão multi-escala e multi-resolução do **CGR**.

Tais implementações foram bem úteis para os nossos experimentos.

3.6 Experimentos Realizados

Ao longo deste semestre, realizamos dois conjuntos de experimentos principais que servirão para decidir os rumos dos próximos experimentos. Ambos os experimentos foram baseadas em técnicas **CGR**, só que com metodologias distintas. No primeiro conjunto de experimentos exploramos técnicas multi-resoluções. Já no segundo conjunto, foi utilizada uma técnica baseada em dimensão fractal que tem a propriedade de explorar a organização de imagens.

[2] Ou *Modelos Ocultosde Markov*, como alguns preferem

Experimentos baseados em CGR multi-resolução

Após implementarmos o **CGR** multi-resolução e multi-escala, realizamos diversos experimentos, variando inclusive diversos tipos de janela. A idéia de multi-resolução e multi-escala, é tentar variar o nível de detalhamento com que se observa um objeto. No caso de imagens, podemos definir informalmente a função de nível de detalhamento simplesmente como sendo um "zoom". Ou seja, se queremos pouco detalhe, dando ênfase para as estruturas principais, podemos olhar com um zoom pequeno. No caso de um mapa mundi, seria equivalente a olhar somente as águas e continentes. Caso se queira mais detalhe, ajusta-se a função zoom, ou seja, é possível aplicar um zoom maior num pequeno pedaço do continente para que se possa estudar melhor as propriedades de um determinado objeto que não podem ser determinadas avaliando-se somente o zoom menor. A idéia de multi-resolução é tentar reduzir custos computacionais, buscando primeiro no zoom menor e só depois no zoom maior. Funcionaria como um filtro, que podaria buscas desnecessárias que precisariam ser realizadas num mapa somente com zoom maior.

Chegamos à conclusão que tanto multi-escaola quanto multi-resolução como se faz em imagens (em geral com a função zoom) não se aplica muito bem em seqüências de DNA diretamente. Tal fato se deve principalmente porque não é claro para uma seqüência, o que é "mais detalhe" e o que é "menos detalhe". Provavelmente, no caso de seqüências, esse conceito só faz sentido num ambiente tridimensional, ou seja, no zoom menor, teríamos um cromossomo em 3D "visto de longe", conforme vamos aumentando o zoom, ou seja, o nível de detalhe, surgem as primeiras dobras e estruturas tridimensionais, aumentando-se ainda mais o nível de detalhe, surgem novas estruturas tridimensionais que vão sendo responsáveis por guardar informações, por exemplo. Aumentando ainda mais o zoom, chega-se ao código de DNA em si.

O problema é que para calcular essa função de detalhamento, teria-se que conhecer as estruturas tridimensionais das seqüências. Só que essa informação está longe de estar disponível, o que inviabiliza uma abordagem multi-resolução do ponto de vista mais purista como descrito aqui.

O que tentamos fazer, foi uma abordagem simplista, simplesmente cortando pedaços de DNA, como um zoom cortaria uma imagem (ou algo também parecido com os **HMM** que utilizam histórico de tamanho variável), mas não foram obtidos resultados animadores. Um fato no entanto curioso, é que apesar dos cortes serem suficiente para deformar a estrutura da imagem, tal fato não acontece, o que talvez esteja sugerindo que a estrutura baseada em freqüência de bases é uma

assinatura razoavelmente confiável da espécie.

Tendo em vista um resultado não satisfatório neste experimentos, optamos por buscar técnicas que utilizam elementos mais concretos.

Uma visão multi-resolução interessante baseada em **HMM** é a proposta por David Kulp [60]. Ele procura dividir os cromossomo em diversas regiões e tenta localizar essas regiões usando três níveis de detalhamento. Talvez essa abordagem seja a mais razoável dadas as limitações computacionais existentes atualmente.

Experimentos baseados em CGR e dimensão fractal

Após participarmos do "1ºWorkshop do projeto Fenótipos" surgiu uma idéia muito interessante de avaliar imagens **CGR** utilizando técnicas de dimensão fractal propostas por Luciano da Fontoura Costa [23].

Realizamos então uma série de 25 experimentos de *clustering* na tentativa de avaliar se tal medida é interessante para se procurar genes. Aparentemente os primeiros resultados são animadores.

Para efeito comparativo, comparamos 5 tipos de características:

- Imagem CGR da seqüência
- Imagem CGR da seqüência normalizada usando logaritmo
- Curva de dimensão fractal da imagem CGR da seqüência
- Curva de dimensão fractal da imagem CGR da seqüência normalizada usando logaritmo
- Conteúdo GC puro (equivalente à calcular a imagem CGR com janela de tamanho 1)

Foram extraídas as características acima de 526 genes do cromossomo 22 (seqüência completa) e de 655 regiões contínuas do mesmo cromossomo em que se acredita com forte confiança que não há genes ou pedaços de genes nas mesmas.

A idéia do experimento foi tentar separar tais regiões utilizando-se somente uma única característica por vez, dentre as descrita acima.

Para cada uma das características, foi calculada uma matriz de distâncias e foram utilizados os 5 algoritmo hierárquicos disponíveis no Matlab (totalizando 25 experimentos) para se tentar

realizar a separação. Importante observar que para calcular a matriz de distância, foram utilizados os seguinte métodos:

Para o conteúdo GC e as imagens CGR normalizadas linearmente e por log simplesmente realiza-se-se a somatória do módulo da subtração ponto a ponto. Esse é o valor da distância. Para as curvas baseadas em dimensão fractal, calcula-se a soma da integral das curvas e subtraí-se duas vezes o valor da intersecção entre as duas integrais. Não foram usadas nenhuma informação adcional que é disponibilizada pela técnica de dimensão fractal.

Cada um dentre os 25 experimentos consistiu no cálculo de 1180 *clusters* usando o mesmo método com número de agrupamentos variando de 2 a n.

Foi calculado então o "erro" de cada cluster da seguinte forma: "Se para cada agrupamento do cluster for necessário associar um único rótulo, qual o erro que será obtido no final?". Atribui-se então o rótulo "gene" à um agrupamento se no mesmo há mais fragmentos que são genes ou rotula-se como "não-gene" caso contrário. O erro é simplesmente o número de elementos de cada agrupamento cujo rótulo é diferente do rótulo do agrupamento, dividindo-se o total pelo número de pontos que no caso é 1181. Obtém-se assim a taxa de erro.

Foram calculados também o número de falsos positivos (FP - Atribuição de rótulo de gene para seqüências que não são gene) e o número de falsos negativos (FN - Atribuição de rótulo de não gene para seqüência que são genes).

Na Figura 1 temos um gráfico somente com as taxas de acerto calculadas para as cinco medidas apresentadas anteriormente utilizando-se o método de *clustering* Complete. O eixo das abcissas representa a porcentagem do número de agrupamento de cada *cluster* em relação ao número total de fragmentos. O eixo das ordenadas representa a taxa de acerto.

Interessante observar que independentemente do método de *cluster* utilizado, as curvas obtidas são bem semelhantes às apresentadas na Figura 1. Em todos os experimentos, a curva azul claro (correspondente ao método baseado em dimensão fractal com normalização pela função log) dá resultados bem interessantes e ligeiramente superiores às outras curvas, principalmente em relação à azul escuro, que serve de controle.

Importante observar que só faz sentido comparar as curvas nos primeiros $10 - 20\%$, visto que para clusters com número de agrupamentos maior do que 20% do número total de pontos ocorre o problema de overffiting.

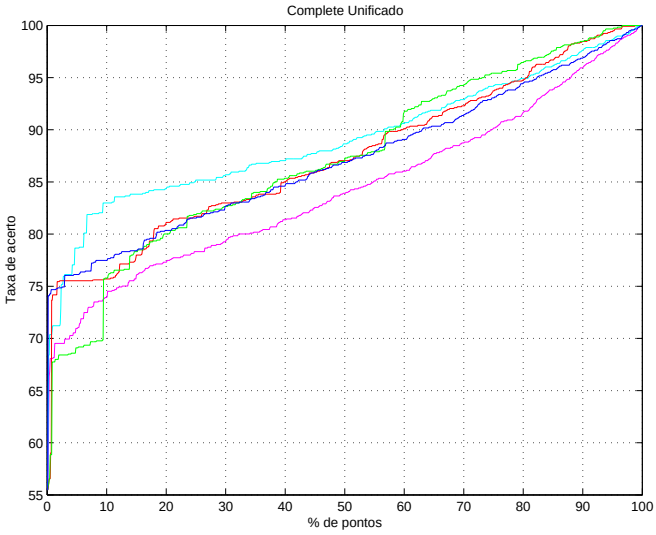

Figura 1: Resumo dos resultados obtidos com o método Complete. Imagem CGR em vermelho, imagem CGR normalizada por log em verde, dimensão fractal em roxo, dimensão fractal normalizada por log em azul claro, conteúdo GC em azul escuro.

A medida baseada em dimensão fractal normalizada por log parece ter dado resultados bem interessantes, enquanto que a mesma medida com normalização linear foi a pior de todas.

Os resultados parecem ser interessantes, porque não utilizam nenhuma informação biológica adcional e, apesar da separação obtida não ser muito boa, ela é significativamente superior à técnica baseada em conteúdo GC que são muito utilizadas por algoritmos baseadas em **HMM**, o que sugere que talvez tal medida possa ser utilizada em substituição ou complemento ao conteúdo GC nesses algoritmos.

De qualquer forma, esses ainda são resultados preliminares. É necessário ainda uma série de experimentos para validarem essa medida. É importante observar que o experimento é realizado de forma simples, e equivale a treinar e aplicar o classificador nos próprios dados de treinamento, o que em geral não tem muito significado.

3.7 Estudos Realizados

Após a conclusão dos experimentos baseados em **CGR** multi-resolução logo no início do semestre, ficamos ligeiramente desanimados com os resultados, o que nos motivou a iniciar mais uma rodada de estudos teóricos, só que desta vez mais aprofundados e voltados para as principais técnicas utilizadas na área, principalmente abordagens mais concretas, como as baseadas em

HMM.

Como não tínhamos muito conhecimento sobre o assunto, optamos por reiniciar a nossa revisão bibliográfica, lendo uma quantidade razoavelmente grande de artigos e livros.

Artigos

Fizemos uma larga busca e encontramos mais de 100 artigos interessante sobre aspectos teóricos que estamos interessados. Como o número de artigos é muito grande, agrupamos os mesmos em classes e lemos alguns representantes de cada classe. Atualmente estamos concluindo a leitura desses representantes.

Inicialmente, começamos nossos estudos dando ênfase em técnicas estatísticas. Começamos por alguns artigos até encontrarmos um excelente tutorial escrito por Rabiner [88]. Apesar deste artigo não ser voltado para aplicações em biologia, nos foi muito útil. Após concluirmos a leitura de uma série de artigos mais teóricos, iniciamos a leitura de artigos mais práticos, como é o caso do modelo de gene utilizando-se **GHMM** proposto por David Kulp [60].

Foram lidos muitos artigos. A lista completa desses artigos está na bibliografia.

Também localizamos uma série de artigos mais práticos que poderão servir de inspiração. Como esses artigos também são muitos, pretendemos agrupá-los e estudar somente alguns após a conclusão da fase teórica. Nesta linha, já lemos uma série de artigos sobre programas para predição de genes. Uma lista de artigos mais práticos podem ser encontrados na página web mantida por Wentian Li [63].

Livros Após lermos alguns artigos, percebemos que precisríamos de alguns livros mais teóricos e especializados. Foi realizada uma série de leituras, entre elas podemos citar os excelentes livros de Durbin [32] e Baxevanis [11], o livro de Pierre Baldi [8] que fornece uma bela e leve visão sobre técnicas estatísticas da área, livros mais técnicos como os de Fu [43], Devroye [27], Prum [87], Bremaud [17], Grant [46], Koski [58], Cristianini [24], Fishman [41]. A lista completa está na bibliografia.

Lemos também algumas dissertações de mestrado, como a do Rogério Feris [40], Teófilo Campos [20] e Ariane Oliveira [76].

Alguns dos livros citados ainda não foram completamente lidos, outros estão sendo lido em

grupo e serão apresentados em seminários. De qualquer forma, acreditamos que com a conclusão das leituras prevista para o início deste semestre tenhamos uma bela base teórica sobre o assunto.

4 Próximos experimentos

Nesta seção pretendemos apresentar uma breve proposta para os próximos experimentos baseados nos experimentos anteriores. Pretendemos dividir os experimentos em três categorias:

1. Como visto na Seção 3.6, foram realizados alguns experimentos preliminares com técnicas baseadas em dimensão fractal. Pretendemos fazer uma série de experimentos com o objetivo de tentar validar a técnica. Para tal, pretendemos realizar mais alguns experimentos com outros dados e também utilizar alguns métodos como *bootstrap*. Dependendo dos resultados obtidos, poderemos optar por realizar o mesmo tipo de experimento com outros cromossomos humanos.

   No caso da técnica se mostrar positiva, poderemos abordar também alguns problemas técnicos, como a normalização das imagens **CGR**.

   Caso a técnica seja validada, ela poderá ser utilizada como substituição ou complemento à técnicas baseadas em conteúdo GC que são largamente utilizadas como parte de softwares para predição de genes.

2. A abordagem acima necessita da definição de uma forma de construir um classificador. Atualmente estamos desenvolvendo algumas abordagens na tentativa de resolver tal questão. A nossa idéia mais básica é simplesmente utilizar como classificador o *cluster* obtido. Para a classificação de um novo ponto, poderia-se atribuir o rótulo do agrupamento menos distante ao ponto em questão, por exeplo.

3. Poderemos realizar também experimentos com adaptações da abordagem acima para outros subproblemas, como discriminação das regiões intergênicas, ou mesmo determinação de fronteiras intron/exon, exon/intron.

5 Alterações no Projeto

Tendo em vista um resultado não muito animador nos nossos primeiros experimentos, optamos por dar mais ênfase nos estudos teóricos. Mas, como os experimentos seguintes deram resultados interessantes, pretendemos continuar realizando uma série de experimentos. Note que ainda não realizamos experimentos supervisionados, que deverão acontecer neste próximo semestre.

Acreditamos que um estudo mais teórico é importante, porque sem ele, fica muito difícil entender os modelos que são usados na área.

A maior dificuldade que encontramos neste período é o excesso de artigos sobre um mesmo tema na área. Na bibliografia é apresentada somente os artigos e livros lidos efetivamente, os artigos que foram simplesmente folheados ficaram de fora. Mesmo assim, o número apresentado é bem grande.

Retardamos também em alguns meses a redação da proposta de dissertação, porque acreditamos no meio do semestre que precisávamos estudar mais alguns tópicos, bem como realizar mais alguns experimentos antes de apresentar uma proposta. Acreditamos que agora é o momento ideal para iniciar a redação desta proposta visto que com a conclusão da segunda safra de experimentos os rumos experimentais ficaram bem claros.

Pretendemos também realizar uma série extra de seminários informais justamente para que se possa compartilhar com os colegas de laboratório o conteúdo dos estudos realizados e que serão realizados no próximo semestre.

6 Plano de Trabalho e Cronograma (Etapas Seguintes)

Atualmente o projeto está numa fase de aplicação dos estudos teóricos já realizados, bem como aprofundamento dos estudos. Durante o próximo semestre pretendemos concluir o mestrado.

6.1 Plano de Trabalho (Etapas Seguintes)

Atividades	
1	Finalização da Especificação do Ambiente de Treinamento
2	Especificação dos Testes
3	Finalização da Implementação dos Ambientes Especificados
4	Implementação dos Testes
5	Segunda avaliação dos algoritmos
6	Redação da proposta da dissertação
7	Seminários
8	Avaliação dos resultados
9	Redação da dissertação

6.2 Cronograma (Etapas Seguintes)

2003

	Jan	**Fev**	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**
1		•	•				
2		•		•			
3	•		•	•			
4	•		•	•			
5		•	•	•			
6	•	•					
7	•	•	•	•			
8				•	•	•	
9					•	•	•

São Paulo, 10 de janeiro de 2003.

Caetano Jimenez Carezzato

Bolsista

Junior Barrera

Orientador

Referências

[1] M. D. Adams, C. Fields, and J. C. Venter, editors. *Automated DNA Sequencing and Analysis*. Academic Press, 1994.

[2] S. S. Adi. Ferramentas de Auxilio ao Sequenciamento de DNA por Montagem de Fragmentos: um estudo comparativo. Master's thesis, Instituto de Matemática e Estatística da Universidade de São Paulo, 2000. http://www.ime.usp.br/∼said/prjdiss.ps [8/Jan/2001].

[3] F. Alizadeh, K. Karp, L. Newberg, and D. Weisser. Physical mapping of chromosome: A combinatorial problem in molecular biology. In *the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, ACM Press, pages 371–381, 1993. http://citeseer.nj.nec.com/alizadeh93physical.html [21/Feb/2001].

[4] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. In *Nucleic Acids Research*, volume 25, pages 3389–3402, 1997. http://nar.oupjournals.org/cgi/content/full/25/17/3389 [31/Jan/2001].

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[6] D. Angluin. Computational Learning Theory: Survey and Selected Bibliography. In *Proceedings of the twenty-fourth annual ACM Symposium on Theory of Computing*, pages 351–369, 1992.

[7] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1):45–48, 2000.

[8] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach (Adaptive Computation and Machine Learning)*. MIT press, 1998.

[9] S. Batzoglou, L. Pachter, J. Mesirov, B. Berger, and E. Lander. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Research*, 10:950–958, 2000.

[10] A. D. Baxevanis. The molecular biology database collection: 2002 update. *Nucleic Acids Research*, 30(2):1–12, 2002.

[11] A. D. Baxevanis and B. F. Ouellette, editors. *Bioinformatics: A Pratical Guide to the Analysis of Genes and Proteins*. Wiley-Interscience, 1998.

[12] G. Bejerano and G. Yona. Modeling protein families using probabilistic suffix trees. In S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceedings of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 15–24, Lyon, France, 1999. ACM Press.

[13] M. J. Bishop and C. J. Rawlings, editors. *DNA and protein sequence analysis: a pratical approach*. The Pratical Approach Series. Oxford University Press, 1st edition, 1997.

[14] E. Bolten, A. Schliep, S. Schneckener, D. Schomburg, and R. Schrader. Clustering protein sequences - structure prediction by transitive homology, 2000. ftp://ftp.zpr.uni-koeln.de/pub/paper/pc/zpr2000-383.zip [16/Fev/2001].

[15] J. Bondy and U. Murty. *Graph Theory with Applications*. MacMillan, London, 1976.

[16] M. Bot and W. Langdon. Application of genetic programming to induction of linear classication trees. In *European Conference on Genetic Programming EuroGP2000*, pages 247–258, Apri 2000. http://www.cs.vu.nl/∼mbot/mijnpapers/euroGP2000/paper.ps [6/Fev/2001].

[17] P. Bremaud, editor. *Markov Chains*. Springer Verlag, 1999.

[18] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78–94, 1997.

[19] A. Campbell, J. Mrazek, and S. Karlin. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. In *Proceedings of the National Academy of Sciences*, volume 96, pages 184–9189, 1999.

[20] T. E. Campos. Técnicas de seleção de características com aplicações em reconhecimento de faces. Master's thesis, Instituto de Matemática e Estatística – Universidade de São Paulo, SP - Brasil, maio 2001.

[21] N. Chomsky. On certain formal properties of grammars. *Information and Control*, 2:137–167, 1959.

[22] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, 2nd edition, 1998.

[23] L. F. Costa and A. G. C. Blanchi. A outra da dimensão fractal. *Ciência Hoje*, 31(183):40–47, 2002.

[24] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[25] R. E. Davis. Introduction to bioinformatics and genomics, 2002. http://www.library.csi.cuny.edu/∼davis/Bioinfo_326/ [29/Jun/2002].

[26] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.*, 16(10):1391–1399, 1999.

[27] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic Theory of Pattern Recognition*. Springer, 1996.

[28] R. Diestel. *Graph Theory*. Springer, 2nd edition, 2000.

[29] E. R. Dougherty, J. Barrera, M. Brun, S. Kim, R. M. Cesar, Y. Chen, M. Bittner, and J. M. Trent. Inference from clustering with application to gene-expression microarrays. *Journal of Computational Biology*, 9(1):105–126, 2002.

[30] D.Sankoff and J. Kruskal. *An overview of sequence comparison, Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Massachusetts: Addison-Wesley, 1983.

[31] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2000.

[32] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of protein and nucleic acids*. Cambridge University Press, 2000.

[33] A. J. Enright and C. A. Ouzounis. Generage: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, 16(5):451–457, 2000. http://bioinformatics.oupjournals.org/cgi/reprint/16/5/451 [6/Fev/2001].

[34] A. B. et al. Plasmodb: the plasmodium genome resource. an integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished). *Nucleic Acids Research*, 30(1):87–90, 2002.

[35] D. A. B. et al. Genbank. *Nucleic Acids Research*, 30(2):17–20, 2002.

[36] D. L. W. et al. Database resources of the national center for biothecnology information: 2002 update. *Nucleic Acids Research*, 30(2):13–16, 2002.

[37] M. Farach-Colton, F. S. Roberts, M. Vingron, and M. Waterman, editors. *Mathematical Support for Molecular Biology*, volume 47 of *DIMACS series in discrete mathematics and theorical computer science*. America Mathematical Society, 1999.

[38] D. Fasulo. An analysis of recent work on clustering algorithms, April 1999. http://www.cs.washington-.edu/homes/dfasulo/clustering.ps [26/Jan/2001].

[39] I. Felger, V. M. Marshal, J. C. Reeder, J. A. Hunt, C. S. Mgone, and H.-P. Beck. Sequence diversity and molecular evolution of the merosoite surface antigen 2 of plasmodium falciparum. *Journal of Molecular Evolution*, 45:154–160, 1997.

[40] R. S. Feris. Rastreamento eficiente de faces em um espaço wavelet. Master's thesis, Intituto de Matemática e Estatística – Universidade de São Paulo, SP - Brasil, maio 2001.

[41] G. S. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer Series in Operations Research. Springer Verlag, 1996.

[42] C. Frontali. Genome plasticity in plasmodium. *Genetica*, pages 91–100, 1994.

[43] K. S. Fu. *Syntactic Pattern Recognition and Applications*. Prentice-Hall, 1982.

[44] C. Gibas and P. Jambeck. *Desenvolvendo bioinformática: ferramentas de software para aplicações em biologia*. Campus – O'Reilly, 2001. Tradução de: Developing bioinformatics computer skills.

[45] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley, 1992.

[46] G. R. Grant and W. J. Ewens. *Statistical Methods in Bioinformatics: An Introduction*. Springer Verlag, 2001.

[47] D. Gusfield. *Algorithms on strings, trees, and sequences. Computer Science and Computational Biology*. Cambridge University Press, 1999.

[48] B. K. Hall, editor. *Homology: The Hierarchical Basis of Comparative Biology*. Academic Press, 1994.

[49] M. H. Hassoun. *Fundamentals of ARTIFICIAL NEURAL NETWORKS*. MIT Press, 1995.

[50] J. Henderson, S. Salzberg, and K. Fasman. Finding genes in DNA with a Hidden Markov Model. *Journal of Computational Biology*, 4(2):121–141, 1997. http://citeseer.nj.nec.com/179996.html [21/Feb/2001].

[51] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA*, 89:10915–10919, 1992.

[52] D. Higgins and P. Sharpe. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. In *Gene*, volume 73, pages 237–244, 1988.

[53] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3), September 1999.

[54] B. R. Jasny and D. Kennedy. The human genome. *Science*, 291(5507), February 2001. http://www.sciencemag.org/genome2001/1153.html [16/Fev/2001].

[55] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1997.

[56] J. Kim, L. Ohno-Machado, and I. Kohane. Unsupervised learning from complex data: The matrix incision tree algorithm. In *Pacific Symposium on Biocomputing*, volume 6, pages 30–41, 2001. http://www.smi.stanford.edu/projects/helix/psb01/kim.pdf [20/Jan/2001].

[57] S. Knudsen. Promoter2.0: for the recognition of polii promoter sequences. *Bioinformatics*, 15(5):356–361, 1999.

[58] T. Koski and T. Koskinen. *Hidden Markov Models for Bioinformatics*. Kluwer Academic Publishers, 2001.

[59] A. Krogh, I. Mian, and D. Haussler. A Hidden Markov Model that finds genes in E. coli DNA. *Nucleic Acids Research*, 22:4768–4778, 1994.

[60] D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman. A Generalized Hidden Markov Model for the recognition of human genes in DNA. In *Fourth International Conference on Intelligent Systems for Molecular Biology*, 1996.

[61] S. Kurtz. Reducing the space requirement of suffix trees. *Software – Practice and Experience*, 29(13):1149–1171, 1999.

[62] H. R. Lewis and C. H. Papadimitriou. *Elements of the Theory of Computation*. Prentice-Hall, 2nd edition, 1998.

[63] W. Li. A bibliography on computational gene recognition. http://linkage.rockefeller.edu/wli/gene/[29/Ago/2002].

[64] D. Lipman and W. Pearson. Rapid and sensitive protein similarity searches. *Science*, pages 1435–1441, 1985.

[65] A. V. Lukashin and M. Borodovsky. Genemark.hmm: new solutions for gene finding. *Nucleic Acids Research*, 26(4):1107–1115, 1998.

[66] Q. Ma and J. T. L. Wang. Application of bayesian neural networks to biological data mining: A case study in dna sequence classification. In *Twelfth International Conference on Software Engineering and Knowledge Engineering*, pages 23–30, 2000.

[67] J. Meidanis and J. C. Setubal. *Introduction to Computational Molecular Biology*. PWS Publishing Co., 1997.

[68] E. Mendelson. *Álgebra Booleana e Circuitos de Chaveamento*. Coleção Schaum. McGraw-Hill, 1977.

[69] Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, 3rd rev. and extended edition, 1999.

[70] T. Minka. A statistical learning/pattern recognition glossary. http://www-white.media.mit.edu/∼tpminka/statlearn/glossary/ [20/May/2002].

[71] A. A. Mironov, J. W. Fickett, and M. S. Gelfand. Frequent alternative splicing of human genes. *Genome Research*, 9:1288–1293, 1999.

[72] B. Morgenstern. Dialign 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218, 1999.

[73] B. Morgenstern, A. Dress, and T. Werner. Multiple dna and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA*, 93:12098–12103, 1996.

[74] D. W. Mount. *Bioinformatics: Sequence and Genome Analysis*. CSHL Press, 1st edition, 2001.

[75] M. Moura. O novo produto brasileiro. *Pesquisa FAPESP*, 55:8–15, julho 2000. http://www.fapesp.br/-capa551.htm [4/Set/2000].

[76] A. M. L. Oliveira. Laboratório de geração de classificadores de seqüências. Master's thesis, Instituto de Matemática e Estatística – Universidade de São Paulo, SP - Brasil, julho 2002.

[77] T. Oliveira and A. Brunstein. HIV sequence analysis and bioinformatics course, November 2001. http://www.vision.ime.usp.br/∼tulio/ [29/Jun/2002].

[78] J. L. Oliver, P. Bernaola-Galván, J. Guerrero-García, and R. Román-Roldán. Entropic profile of DNA sequences trought chaos-game-derived images. *Journal of Theorical Biology*, 160:457–470, 1993.

[79] C. H. Papadimitriou. *Computational Complexity*. Addison-Wesley, New York, 1994. Reprinted August, 1995.

[80] C. H. Papadimitrou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, 1982.

[81] W. Pearson and D. Lipman. Improved tools for biological sequence comparison. In *Proceedings of National Academy of Sciences of the USA*, volume 85, pages 2444–2448, 1988.

[82] W. R. Pearson. Rapid and sensitive sequence comparisions with FASTP and FASTA. *Methods in Enzymology*, 183:63–98, 1990.

[83] M. Pertea, X. Lin, and S. Salzberg. Genesplicer: a new computational method for splice site prediction. *Nucleic Acids Research*, 29(5):1185–1190, 2001.

[84] E. Pizza, S. Liuni, and C. Frontali. Detection of latent sequence periodicities. *Nucleic Acids Research*, 18(13):3745–3752, 1990.

[85] E. Pizzi and C. Frontali. Low-complexity regions in plasmodium falciparum proteins. *Genome Research*, 11:218–229, 2001.

[86] W. K. Pratt. *Digital Image Processing*. Wiley-Interscience, 1991.

[87] B. Prum. Markov models and hidden markov models in genome analysis. 6th Brazilian School of Probability, Praia das Tininhas - Ubatuba, São Paulo, August 5–10 2002.

[88] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77(2), pages 257–284, 1989.

[89] S. Rich, M. Ferreira, and F. Ayala. The origin of antigenic diversity in plasmodium falciparum. *Parasitology Today*, 16(9):390–396, 2000.

[90] S. M. Rich, R. R. Hudson, and F. J. Ayala. Plasmodium falciparum antigenic diversity: Evidence of clonal population structure. *Proc. Natl. Acad. Sci. USA*, 94:13040–13045, 1997.

[91] F. Richards. The protein folding problem. *Scientific American*, 264(1):54–63, January 1991.

[92] Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjolander, R. C. Underwood, and D. Haussler. Recents Methods for RNA Modeling Using Stochastic Context-Free Grammars. In *Combinatorial Pattern Matching, 5th Annual Symposium*, pages 289–306, 1994. ftp://ftp.cse.ucsc.edu/pub/rna/cpm94.ps.Z [21/Feb/2001].

[93] Y. Sakakibara, M. Brown, R. Underwood, I. S. Mian, and D. Haussler. Stochastic context-free grammars for modeling RNA. In *Proceedings of the 27th Hawaii International Conference on System Sciences*, pages 284–283, Honolulu, 1994. IEEE Computer Society Press.

[94] M. K. Sakharkar, T. W. Tan, and S. J. de Souza. Generation of a database containing discordant intron positions in eukaryotic genes (midb). *Bioinformatics*, 17(8):671–675, 2001.

[95] S. Salzberg, A. Delcher, K. Fasman, and J. Henderson. A decision tree system for finding genes in DNA. *Journal of Computational Biology*, 5(4), 1998. http://www.tigr.org/∼salzberg/morgan.ps.gz [21/Feb/2001].

[96] R. Sandberg, G. Winberg, C.-I. Bränden, A. Kaske, I. Ernberg, and JoakimCöster. Capturing whole-genome characteristics in short sequences using a naïve bayesian classifier. *Genome Research*, 11:1404–1409, 2001.

[97] J. R. Searle. *Minds, Brains and Science*. Harvard University, March 1986.

[98] D. Searls. String variable grammar: a logic grammar formalism for the biological language of DNA. *The Journal of Logic Programming*, 12:1–30, 1993. http://citeseer.nj.nec.com/searls93string.html [21/Feb/2001].

[99] D. Searls. Linguistic approaches to biological sequences. *CABIOS*, 13(4):333–344, 1997.

[100] D. Searls. Formal language theory and biological macromolecules. *Series in Discrete Mathematics and Theoretical Computer Science*, 47:117–140, 1999. http://citeseer.nj.nec.com/searls99formal.html [21/Feb/2001].

[101] D. Searls and S. Dong. A syntactic pattern recognition system for DNA sequences. In *Second International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis*, pages 89–101, 1993. http://citeseer.nj.nec.com/searls93syntactic.html [21/Feb/2001].

[102] I. Simon. *Linguagem Formais e Autômatos*. Segunda Escola de Computação, 1981. Instituto de Matemática, Estatística e Ciência da Computação da UNICAMP.

[103] P. Smyth. Clustering sequences with hidden markov models. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 648. The MIT Press, 1997.

[104] D. P. Snustad, M. J. Simmons, and J. B. Jenkins. *Principles of Genetics*. John Wiley and Sons, 1997.

[105] G. Stoesser. The EMBL nucleotide sequence database. *Nucleic Acids Research*, 30(2):21–26, 2002.

[106] J. E. Tabaska, R. V. Davuluri, and M. Q. Zhang. Identifying the 3'-terminal exon in human dna. *Bioinformatics*, 17(7):602–607, 2001.

[107] T. A. Thanaraj. Positional characterisation of false positives from computational prediction of human splice sites. *Nucleic Acids Research*, 28(3):744–754, 2000.

[108] The Genome International Sequencing Consortium. Initial sequencing and analysis of the human genome, February 2001. http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v409/-n6822/full/409860a0_fs.html [16/Fev/2001].

[109] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 1999.

[110] J. Thompson, D. Higgins, and T. Gibson. improving the sensitivity of progressive multiple sequence alignment through sequence weighting. *Nucleic Acids Research*, 22:4673–4680, 1994.

[111] N. S. Tomita. Programação Automática de Máquinas Morfológicas Binárias baseada em Aprendizado PAC. Master's thesis, Instituto de Matemática e Estatística - Universidade de São Paulo, SP - Brasil, março 1996.

[112] L. G. Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[113] D. Voet and J. G. Voet. *Biochemistry*. John Wiley and Sons, 2nd edition, 1995.

[114] E. O. Voit. *Computational Analysis of Biochemical Systems: a pratical guide for biochemists and molecular biologists*. Cambridge University Press, 2000.

[115] J. Wang, X. Wang, K. Lin, D. Shasha, B. Shapiro, and K. Zhang. Evaluating a class of distance-mapping algorithms for data mining and clustering. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 307–311, San Diego CA, August 1999. ACM. http://www.msci.memphis.edu/∼linki/_mypaper/kdd99.ps.gz [26/Jan/2001].

[116] J. T. L. Wang, Q. Ma, D. Shasha, and C. H. Wu. Application of neural networks to biological data mining: A case study in protein sequence classification. In *The Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 305–309, 2000.

[117] J. T. L. Wang, Q. Ma, D. Shasha, and C. H. Wu. New techniques for extracting features from protein sequences. *IBM Systems Journal, Special Issue on Deep Computing for Life Sciences*, 40(2), 2001.

[118] M. S. Waterman. *Mathematical Methods for DNA sequences*. Boca Raton, FL: CRC Press, 1989.

[119] J. J. Wiens, editor. *Phylogenic Analysis of Morphological Data*. Smithsonian Series in Comparative Evolutionary Biology. Smithsonian Institution Press, 2000.

[120] Wisconsin package version 10.0. Genetics Computer Group (GCG). Madison, Wisc.

[121] C. H. Wu. The protein information resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Research*, 30(2):35–37, 2002.

[122] A. Zaha. *Biologia Molecular Básica*. Mercado Aberto, 1996.

[123] Catálogo de disciplinas de pós-graduação em ciência da computação. http://sistemas.usp.br/fenixweb/fexDisLista?codare=45134&codcpg=45 [28/Jun/2002].

[124] Matlab 6. http://www.mathworks.com/products/matlab/ [28/Jun/2002].

[125] Bielefeld university bioinformatics server. http://bibiserv.techfak.uni-bielefeld.de/ [29/Jun/2002].

[126] Bioinformatics – oxford journals online. http://bioinformatics.oupjournals.org/ [29/Jun/2002].

[127] Ensembl genome browser. http://www.ensembl.org/ [29/Jun/2002].

[128] geneid web server. http://www1.imim.es/geneid.html [29/Jun/2002].

[129] Human genome project working draft. http://www.genome.ucsc.edu/ [29/Jun/2002].

[130] I latin american course on bioinformatics for tropical disease research. http://icb.ime.usp.br/tdr/ [29/Jun/2002].

[131] National center for biotechnology information. http://www.ncbi.nlm.nih.gov/ [29/Jun/2002].

[132] Researchindex. http://citeseer.nj.nec.com/cs [29/Jun/2002].

[133] VSNS biocomputing division multiple alignment resource page. http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/welcome.html [29/Jun/2002].