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

Caetano Jimenez Carezzato

Relatório Científico

Período: Agosto/2001 – Julho/2002

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 [56], inclusive o humano[79, 44], 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 [50] 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** [24, 85], montagem de fragmentos de **DNA** [1], mapeamento físico de **DNA** [2], árvores filogenéticas [86], reconhecimento de genes e partes de genes [69, 16], busca de homologia [39], clustering [23, 28], predição da estrutura de proteínas [65] etc.

O objetivo principal deste trabalho é comparar 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 [27], modelaremos os dados por Gramáticas Estocásticas [35] que serão estimadas a partir de dados reais utilizando-se diversas técnicas de reconhecimento de padrões [25, 80] e Aprendizado Computacional [83, 9].

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-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)

No nosso cronograma inicial, a previsão para o início da bolsa era abril de 2001. Como a bolsa só começou em agosto de 2001, apresentamos abaixo uma versão com as datas atualizada do cronograma original.

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 [80]	$\circ$
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	$\bullet$
13	Segunda avaliação dos algoritmos	$\bullet$
14	Redação de relatórios semestrais para a FAPESP	$\circ$
15	Redação da proposta da dissertação	$\bullet$
16	Seminários	$\bullet$
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**
3	$\bullet$	$\bullet$					$\bullet$	$\bullet$				
5	$\bullet$		$\bullet$		$\bullet$		$\bullet$		$\bullet$			
10	$\bullet$	$\bullet$	$\cdot$	$\cdot$	$\bullet$	$\bullet$		$\bullet$	$\bullet$			
11	$\bullet$	$\bullet$	$\bullet$		$\bullet$	$\bullet$		$\bullet$	$\bullet$			
12		$\cdot$	$\bullet$	$\bullet$	$\bullet$							
13							$\bullet$	$\bullet$	$\bullet$			
14						$\bullet$						
15	$\times$			$\cdot$	$\cdot$							
16	$\times$	$\times$	$\bullet$	$\bullet$	$\bullet$	$\cdot$	$\cdot$					
17									$\bullet$	$\bullet$	$\bullet$	
18										$\bullet$	$\bullet$	$\bullet$

3 Principais Realizações

O objetivo desta seção é apresentar um detalhamento das realizações efetuadas no período em que o relatório se refere.

3.1 Resumo

Ao longo de seu mestrado, 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** [91], **GCG** [87] e suas ferramentas. Também especificamos e implementamos um gerador estocástico de palavras.

Atualmente estamos realizando duas tarefas principais:

1. Especificação e implementação dos ambientes de treinamento e testes
2. Finalização de algumas leituras

3.2 Disciplinas do Programa de Pós-Graduação

A seguir estão as disciplinas cursadas pelo bolsista ao longo do programa de pós-graduação. Mais detalhes a respeito de cada disciplina podem ser encontrados no histórico escolar anexo ou no catálogo de disciplinas de Ciência da Computação [90].

Ao todo foram cursadas 7 disciplinas, o que completa todos os créditos exigidos pelo programa de mestrado. As seguintes disciplinas foram cursadas:

MAC5722	Introdução à Teoria da Complexidade de Algoritmos
MAC5720	Teoria dos Autômatos Finitos
MAC5727	Algoritmos de Aproximação
MAC5828	Tópicos em Complexidade Computacional
MAC5711	Análise de Algoritmos
MAC5749	Análise e Reconhecimento de Formas: Teoria e Prática
MAC5832	Aprendizagem Computacional: modelos, algoritmos e aplicações

Todas as disciplinas foram concluídas com o conceito **A**. A dedicação às disciplinas foi bastante grande para se chegar ao desempenho obtido. Em alguns períodos, o trabalho em cima do projeto foi interrompido para poder haver um empenho maior nas atividades exigidas pelas disciplinas. Tais períodos de interrupção foram mais freqüentes ao longo do ano de 2001, o que comprometeu ligeiramente o cronograma como pode ser observado na seção 2.2, mas nada que não possa ser recuperado até o final do próximo semestre.

3.3 Estudos Realizados

O objetivo desta seção é apresentar um resumo das principais leituras e cursos realizados. Ao longo do período de estudos, foram realizadas diversas leituras de livros, artigos e páginas na Internet. O bolsista também assisstiu a dois cursos ministrados no próprio Insituto de Matemática e Estatística.

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.

3.3.1 Leituras

Livros

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 Bomputacional). Tendo em visto tal deficiência, resolvemos começar as leituras por [77] na área de Biologia e [38] na área de Computação. Foram dois livros muito interessantes porque o bolsista teve a oportunidade de entender melhor Biologia e também algumas técnicas computacionais mais a fundo, como as árvores de sufixo, algoritmos de "matching" exatos e com erros e algumas técnicas de programação dinâmica e heurísticas usadas em Biologia Computacional.

Na seqüência, fizemos uma releitura de [50] para fixar melhor alguns conceitos e finalizamos a leitura de [55], um livro muito interessante na área de Biologia Computacional que dá uma visão da área um pouco diferente dos outros livros que lemos.

Um livro que não era bem o que esperávamos foi [39], que aborda aspectos biológicos da homologia e não computacionais como esperávamos.

Tendo concluído a fase introdutória de leituras, passamos para uma fase mais técnica e computacional. Iniciamos tal fase com o livro [9], que é introdutório, mas resume em um só lugar diversas técnicas de Aprendizado Computacional aplicadas à Biologia Computacional. Para entender melhor a essência de tais técnicas, realizamos a leitura de [45] e [7], dois livros que abordam aspectos mais teóricos da área de Aprendizado Computacional. Nesta mesma linha de estudos teóricos (mas não tão teórico), realizamos a leitura de [26], um excelente livro na área de reconhecimento de padrões e também lemos [40], um livro sobre Redes Neurais.

Em paralelo com todas essas leituras acima, realizamos a leitura dos manuais do **GCG** e **Matlab** bem como o uso de suas ferramentas. Chegamos a escrever alguns programas de testes e realizar alguns experimentos para nos familiarmos com esses softwares que serão muito úteis ao longo do projeto.

Após todas estas leituras, resolvemos ler [80], também um excelente livro na área de reconhecimento de padrões, dando ênfase nas técnicas de "clustering". Realizamos também as leituras parciais de [11, 36, 12], tais leituras, entre outras, serão finalizadas ao longo do próximo semestre.

Artigos

Ao longo deste primeiro ano de bolsa, foram lidos algumas dezenas de artigos. Uma lista completa dos artigos lidos está na bibliografia. Vamos somente citar aqui os que mais nos interessaram, foram eles [22], [70] e [17]. Todos abordam o mesmo assunto que é a assinatura genômica das espécies. Tais artigos propõem classificadores baseados na freqüência de utilização de bases no genoma de cada espécie. Além disso é proposto uma forma de representação dessas tabelas de freqüência através do uso de imagens. A representação resultante é muito interessante, principalmente porque, entre outras coisas, permite a utilização de técnicas de reconhecimento de padrões em imagens na abordagem do problema de reconhecimento de padrões em seqüências (como, por exemplo, a detecção de genes).

Sítios na Internet

Durante o período de leituras, navegamos na Internet em busca de informações atualizadas a

respeito da área que estamos estudando. Acessamos desde algumas páginas clássicas como [98, 94] até páginas com informações detalhadas sobre algoritmos, principalmente algoritmos para predição de genes, como por exemplo o **geneid** [95] e o **GENSCAN** [16].

Um curso muito interessante de Biologia Computacional que acompanhamos é [20]. A lista completa de links acessados será omitida, e um pequeno resumo da mesma está apresentada na bibliografia.

3.3.2 Cursos

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

O primeiro curso [21], assistido entre os dias 14 e 17 de novembro de 2001, abordava alguns aspectos básicos de análise de seqüências e filogenia.

O segundo curso [97], ministrado entre os dias 18 de fevereiro e primeiro de março de 2002, abordava tanto aspectos de informática como aspectos de Biologia. Preferimos assistir somente a parte de Biologia do curso.

3.4 Especificações e Implementações

Após a conclusão da fase de estudos mais teóricos, conceitos muito utilizados em Biologia Computacional ficaram bem mais claros. Podemos dividir as principais ferramentas utilizados na área em dois tipos principais: ferramentas baseadas em Aprendizado Computacional (como por exemplo, a utilização de Modelos Escondidos de Markov para modelar genes) e ferramentas baseadas na construção analítica de algoritmos (como é o caso da programação dinâmica aplicada ao alinhamento de seqüências ou algoritmos para predição de genes baseados na estatística de utilização da terceira componente de cada códon, por exemplo).

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.

Como motivação e forma de orientação, escolhemos o problema clássico de localização de genes devido à sofisticação que tal problema alcança na área. Na tentativa de resolver este problema, diversas técnicas de Aprendizado Computacional são utilizadas bem como técnicas híbridas, o que nos faz acreditar que ele seja um excelente problema a ser analisado. Aqui é importante observar que o objetivo principal do projeto não é achar genes ou entender como funciona um determinado algoritmo para tal tarefa, mas sim entender as técnicas e ferramentas principais que sustentam os algoritmos da área.

3.4.1 Especificação do Ambiente

Ao longo do período em questão, realizamos três especificações principais.

- Primeiramente especificamos o gerador estocástico de palavras. Durante sua especificação, levamos em conta os padrões de gramáticas estocásticas disponíveis, bem como as necessidades do nosso grupo, visto que tal gerador será usado em testes dos algoritmos que estão sendo desenvolvidos pelo nosso grupo e por este projeto. O pograma lê um conjunto de gramáticas estocásticas, alguns parâmetros de entrada, como o número de seqüências de saída, tamanho máximo, formato da saída, entre outros e a saída do programa é simplesmente o conjunto de seqüências desejado.

- Partimos então para a especificação do nosso ambiente de trabalho. Tendo em vista as necessidades já citadas, optamos por construí-lo baseado no sistema de arquivos do Linux e a utilização de *Makefiles* e *scripts* para sua atualização. Tal forma de armazenamento facilita o desenvolvimento rápido de filtros que permitem a utilização de tudo quanto é tipo

de programa disponível atualmente, mesmo que rodem em sistemas operacionais distintos. Além disso os dados estão rapidamente disponíveis, visto que estão no HD do micro de testes.

Dividimos este ambiente em quatro módulos principais:

Dados de entrada

Este módulo é o que consideramos o mais importante. Como base deste módulo temos os dados brutos. Tais dados são filtrados de forma a extrair somente as informações que nos interessam. Não só isso, muitas vezes filtramos dados de diversas fontes de forma que eles sejam unificados. Com esses dados unificados e de formato simples em mãos, fica facilitada a confecção de scripts que convertam tais dados para o formato de entrada de algoritmos específicos. Não só isso, fica facilitada a seleção dos dados que se quer utilizar.

Este módulo é desenvolvido todo baseado em *Makefiles* e filtros de forma que qualquer alteração nos dados brutos (como por exemplo, a alteração dos arquivos do genoma humano no **NCBI**) possa facilmente ser incorparada, bastando fazer download dos novos dados e rodar o utilitário *make*, ações que podem até mesmo serem colocadas num script automático. Resumindo este módulo foi projetado para ser de fácil atualização, modificação e extensão.

Dados de teste

Neste módulo guardamos uma série de filtros necessários para a utilização dos diversos programas. A função de tais filtros é gerar a partir dos dados de entrada filtrados, conjuntos de dados de testes prontos para serem rodados pelos programas a serem utilizados.

Outro objetivo deste módulo é também manter um histórico dos experimentos realizados.

Programas

Neste módulo mantemos todos os programas utilizados no projeto, tanto os desenvolvidos por nós como os disponibilizados por terceiros. Serve simplesmente como um índice das opções disponíveis.

Relatórios

Neste último módulo mantemos relatórios sobre tudo o que estamos fazendo, desde os experimentos realizados até os principais links acessados por nós.

- Por último, realizamos uma especificação de como nossos testes devem ser, ou seja, como vamos testar os algoritmos e técnicas que estamos avaliando. Esta especificação ainda não foi concluída.

3.4.2 Implementações realizadas

Realizamos uma implementação parcial dos ítens citados na seção anterior. Entre elas:

- Implementação completa e testes do gerador estocástico de palavras. Ele foi implementado a partir da especificação e tem se mostrado muito útil em alguns testes, principalmente nos testes envolvendo gramáticas estocásticas do grupo.

- Implementamos parcialmente o ambiente de trabalho. Finalizamos o módulo principal que lida com os dados de entrada.

  Como estamos abordando o problema de predição de genes, estamos utilizando dados do **NCBI** [98], **Ensembl** [94], Genome Project Working Draft [96] e dados fornecidos pelo nosso co-orientador para gerar dois tipos de dados principais:

  1. Um mapa completo do genoma humano (respeitando as limitações atuais quanto à completude do genoma humano) em que a posição de genes reais e preditos bem como outros elementos são assinalados.
  2. Um diretório para cada gene conhecido contendo diversas informações sobre o gene, como a posição dos seus exons, tabelas de freqüências etc.
  3. Mapas extras dos cromossomos e contigs contendo informações como áreas mascaradas, de freqüência alterada, entre outros ítens.

  Tais dados serão muito úteis para os nossos experimentos, principalmente porque estão num formato muito simples apesar de ocuparem bastante espaço em relação às outras representações utilizadas.

  Também implementamos parcialmente os filtros para a geração dos dados de testes.

Uma informação curiosa a respeito desta implemantação é que foi uma das partes mais difíceis do projeto até agora, não pela dificuldade para se escrever os filtros e *Makefiles*, mas sim pela dificuldade de encontrar os dados que precisamos. Por exemplo, a localização dos genes num cromossomo, é uma informação muito importante para nós, mas tal informação é difícil de se obter através dos *sites* do **NCBI**, entre outros. Só conseguimos extrair tal informação ao pegarmos os arquivos disponíveis no ftp dos *sites* e filtrarmos os mesmos. Outra informação difícil de se extrair é se um gene foi predito por algum programa ou se o gene está lá porque ele alinha razoavelmente bem com alguma proteína ou mRNA conhecidos. Outro problema é em relação à nomenclatura dos genes, em cada site que você vai muitas vezes o mesmo gene aparece com nomes distintos. São questões simples mas que nos tomaram diversos meses, muito mais do que o esperado.

Infelizmente a área de Biologia Computacional ainda tem muito a evoluir no sentido de armazenamento de informações, mas acreditamos que conseguimos contornar esses problemas através dos nossos filtros após passar um bom tempo estudando o problema. Além disso, o **Ensembl** acabou de lançar no dia primeiro de julho uma nova versão de seu banco de dados que, através de uma excelente interface escrita na linguagem **perl**, facilita a obtenção de diversas informações que precisamos para o nosso projeto.

- Ainda não realizamos a implementação dos testes, mas tal tarefa será a próxima a ser executada.

- Implementamos o algoritmo descrito em [22]. Fomos capazes de reproduzir completamente o artigo e realmente os resultados obtidos foram muito interessantes. Criamos um programa que gera tanto as imagens descritas, como as tabelas brutas necessárias para a realização de alguns experimentos.

- O bolsista também ajudou na implementação de um Cluster de 16 nós baseado em Linux e na implementação da atual rede de computadores utilizada pelo grupo.

4 Alterações no Projeto

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 estamos fazendo 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.

A opção que fizemos por estudar o problema de predição de genes se justifica porque é um problema em que se utilizará diversas técnicas de reconhecimento de padrão, o que também acaba servindo como motivação. Nosso objetivo é, num primeiro momento, abordar tal problema utilizando as técnicas básicas e próprias que desenvolvemos justamente para "sentirmos" o limite de tais técnicas. Num segundo momento, pretendemos utilizar algumas novas informações extras sobre o modelamento do problema fornecidas pelo co-orientador na tentativa de melhorar os resultados obtidos pelas nossas técnicas. Num terceiro momento, pretendemos estudar a organização utilizada pelos principais algoritmos da área. Acreditamos que assim vamos ter uma melhor visão das ferramentas disponíveis. Não só isso, tal abordagem permitirá uma melhor compreensão das

ferramentas fundamentais utilizadas na área, o que nos permitirá atacar mais facilmente qualquer outro problema que exija técnicas de reconhecimento de padrões.

Uma das maiores dificuldades que não havíamos previsto originalmente e que enfrentaremos na próxima etapa do projeto é a ausência de dados negativos de treinamento. Dada uma região, é difícil (e custoso) decidir se existe ou não um gene na mesma. Nos bancos de dados mundiais, depositam-se as regiões em que há fortes motivos para se acreditar que há genes nas mesmas, mas não se deposita (ou aponta-se) regiões em que há fortes motivos para se acreditar que não existe genes nas mesmas, como, por exemplo, uma tentativa mal sucedida na bancada de um laboratório de se encontrar algum gene numa determinada região (aqui é importante observar que as regiões mascaradas não são exemplos negativos muito bons). Tendo em vista que muitas técnicas de reconhecimento de padrão dependem de exemplos negativos para a obtenção de bons resultados, pretendemos desenvolver técnicas específicas para contornar este empecilho nos próximos meses.

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 será compensado nos próximos 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”. Acreditamos que tal disciplina será útil na formação do bolsista.

5 Plano de Trabalho e Cronograma (Etapas Seguintes)

Atualmente o projeto está numa fase de transição dos estudos teóricos para a parte prática. Durante o próximo semestre pretendemos concluir esta transição, bem como escrever a nossa proposta de dissertação para a Comissão de Pós-Graduação do Instituto.

5.1 Plano de Trabalho (Etapas Seguintes)

Atividades	
1	Disciplina Extra do Programa de Pós-Graduação
2	Finalização da Leitura Supervisionada de [80]
3	Finalização da Especificação do Ambiente de Treinamento
4	Especificação dos Testes
5	Finalização da Implementação dos Ambientes Especificados
6	Implementação dos Testes
7	Primeira avaliação dos algoritmos
8	Segunda avaliação dos algoritmos
9	Redação de relatórios semestrais para a FAPESP
10	Redação da proposta da dissertação
11	Seminários
12	Avaliação dos resultados
13	Redação da dissertação

5.2 Cronograma (Etapas Seguintes)

2002/2003

	Ago	Set	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul
1	•	•	•	•								
2	•	•										
3	•	•					•	•				
4	•		•		•		•		•			
5	•	•	•	•	•	•		•	•			
6	•	•	•		•	•		•	•			
7		•	•	•	•							
8							•	•	•			
9						•						
10				•	•							
11			•	•	•	•	•					
12									•	•	•	
13										•	•	•

São Paulo, 10 de julho de 2002.

Caetano Jimenez Carezzato

Bolsista

Junior Barrera

Orientador

Referências

[1] 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].

[2] 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].

[3] 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].

[4] 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.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 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] M. H. G. Anthony and N. Biggs. *Computational Learning Theory: An Introduction*, volume 30 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1997.

[8] 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.

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

[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] 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.

[13] 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].

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

[15] 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].

[16] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic dna. *J. Mol. Biol.*, 268:78–94, 1997.

[17] 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.

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

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

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

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

[22] 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.

[23] 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. *Submetido*, 2001.

[24] 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.

[25] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, 1973.

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

[27] S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1999.

[28] 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].

[29] 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.

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

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

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

[33] 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.

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

[35] K. S. Fu. *Syntactic Pattern Recognition and Applications*. Prentice-Hall, Inc., Englewood Cliffs,NJ, 1982.

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

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

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

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

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

[41] 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].

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

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

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

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

[46] 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].

[47] 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.

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

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

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

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

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

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

[54] 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.

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

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

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

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

[59] 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.

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

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

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

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

[64] 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.

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

[66] 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].

[67] Y. Sakakibara, M. Brown, I. Mian, R. Underwood, and D. Haussler. Stochastic context-free grammars for modeling RNA. In *Hawaii Intl. Conf. on System Sciences*. IEEE Computer Society Press, 1993. http://citeseer.nj.nec.com/sakakibara93stochastic.html [30/Jan/2001].

[68] 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.

[69] 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].

[70] 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.

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

[72] 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].

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

[74] 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].

[75] 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].

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

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

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

[79] 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].

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

[81] 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.

[82] 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.

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

[84] 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].

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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