SISTEMA DE GERENCIAMENTO E ANÁLISE DE DADOS POR BIOINFORMÁTICA

Pablo Rodrigo Sanches

OK

Dissertação apresentada ao Instituto de Física de São Carlos, da Universidade de São Paulo, para a obtenção do Título de Mestre em Ciências: Física Aplicada – Opção: Física Computacional.

Orientador: Prof. Dr. Luciano da Fontoura Costa

São Carlos
2006

Sanches, Pablo Rodrigo

"Sistema de Gerenciamento e Análise de Dados por Bioinformática"
Pablo Rodrigo Sanches – São Carlos, 2006

Dissertação (Mestrado) – Área de Física Aplicada do Instituto de Física de São Carlos da Universidade de São Paulo
2006 – Páginas: 82

Orientador: Prof. Dr. Luciano da Fontoura Costa

1. Pipeline; 2. Bioinformática; 3. Expressão gênica

I. Título

IFSC UNIVERSIDADE DE SÃO PAULO
Instituto de Física de São Carlos

Caixa Postal 369
13560-970 São Carlos, SP
Av. Trabalhador São-carlense, 400,
13560-970 São Carlos, SP

Fone Fax 16 3373 9777

MEMBROS DA COMISSÃO JULGADORA DA DISSERTAÇÃO DE MESTRADO DE ***PABLO RODRIGO SANCHES*** *APRESENTADA AO INSTITUTO DE FÍSICA DE SÃO CARLOS, UNIVERSIDADE DE SÃO PAULO, EM 10/10/2006.*

COMISSÃO JULGADORA:

Prof. Dr. Luciano da Fontoura Costa (Orientador e Presidente) – IFSC/USP

Profa. Dra. Helaine Carrer – ESALQ/USP

Prof. Dr. João Eduardo Ferreira – IFSC/USP

DEDICATÓRIA

Dedico esta dissertação à toda minha família, em especial meus pais Claudinei e Thelma, minha mulher Yanê e meu filho Pedro pelo amor e companheirismo.

AGRADECIMENTOS

Ao Prof. Dr. Luciano da Fontoura Costa pela orientação dedicada na realização deste trabalho, pela amizade e confiança em mim depositada.

À Profa. Dra. Nilce Maria Martinez Rossi, pelo apoio, incentivo, suporte e oferecimento de seu laboratório para que eu pudesse me envolver em pesquisas da área de Bioinformática.

Aos meus pais, Claudinei e Thelma, pelo apoio, suporte e incentivo que sempre recebi, tendo sempre como conselho, a calma necessária e a obstinação por aquilo que desejava.

À minha mulher Yanê e meu filho Pedro, pelo apoio, incentivo, alegria e amor que recebi mesmo quando tive que dividir meu tempo entre eles e os estudos.

Aos Pós-Graduandos Nalu, Henrique, Jeny, Fernando, Fernanda Paião, Fernanda Maranhão, Diana, Juliana e Luciene pelos ensinamentos e incentivo em biologia molecular.

Ao amigo e funcionário do Departamento de Genética Mendelson pelos favores e prestações de serviços de seqüenciamento de DNA.

Ao amigo Pedro do Departamento de Genética que me ajudou na pós sempre que precisei.

À secretária de pós-graduação do IFSC, Wladerez, por sempre me auxiliar e solucionar os problemas burocráticos do mestrado.

Às secretárias Cleusa, Susie e Maria Aparecida do Departamento de Genética pelo auxílio enquanto estou em Ribeirão Preto.

Ao Conselho Técnico Administrativo da Faculdade de Medicina de Ribeirão Preto por tornar possível meu afastamento para realização do mestrado, já que sou funcionário desta faculdade.

À todos os meus amigos, pelos momentos de "espairecimento da mente".

À todos aqueles que colaboraram direta ou indiretamente para a realização deste trabalho.

À Deus por tudo...

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Os projetos para estudo de genomas ou genes expressos partem de uma etapa de seqüenciamento no qual são gerados em laboratório dados brutos, ou seja, seqüências de DNA sem significado biológico. Estas seqüências de DNA possuem códigos responsáveis pela produção de RNAs e proteínas.

O grande desafio dos pesquisadores consiste em analisar essas seqüências e obter informações biologicamente relevantes. Durante esta análise diversos programas de computador, além de um grande volume de dados armazenados em fontes de dados biológicas, são utilizados. Assim sendo, o presente trabalho propôs a elaboração de um sistema computacional que permite a análise de dados sobre biologia molecular e facilite a instanciação do software dependendo do ambiente de trabalho e tipo de projeto de análise. Para este sistema foi dado o nome de *Sistema de Gerenciamento de Análise de Dados por Bioinformática* - SGADBio.

O trabalho apresenta o desenvolvimento do sistema baseado em metodologias de Engenharia de Software, além dos módulos e funções disponíveis.

Seqüências oriundas de um projeto de ESTs do fungo dermatófito *Trichophyton rubrum*, geradas em um laboratório de biologia molecular, foram submetidas ao sistema para análise. Os resultados são expressivos, demonstrando que o sistema é adequado e capaz de adaptar-se a projetos envolvendo seqüenciamento.

ABSTRACT

Projects involving the study of genomes and expressed genes typically initiate with the raw data generated by laboratory sequencing of DNA, devoid of any biological meaning. However, such sequences contain the codes for the production of RNAs and proteins. One of the great challenges faced by researchers is the analysis of such sequences in order to obtain biologically meaningful information. Several computer programs and auxiliary databases are used for that purpose.

The present work reports on the development of a computational system capable of supporting biology data analysis and it can be instantiated in order to suit specific working environments and analyses projects. This system has been called *Management and Data Analysis System for Applications in Bioinformatics* - SGADBio.

This work presents the development of the system based on Software Engineering methodologies, as well as the involved modules and functionalities.

Sequences from an EST project involving the dermatophyte fungus *Trichophyton rubrum*, generated in a molecular biology laboratory, were submitted to the system for analysis. The results are expressive, corroborating the versatility of the system for adaptation to sequencing projects.

I – INTRODUÇÃO E OBJETIVOS

Os projetos para estudo de genomas ou genes expressos partem de uma fase de seqüenciamento no qual são gerados em laboratório dados brutos, ou seja, seqüências de DNA sem significado biológico. As seqüências de DNA possuem códigos responsáveis pela produção de proteínas e RNAs, enquanto que as proteínas participam de todos os fenômenos biológicos, como a replicação celular, produção de energia, defesa imunológica, contração muscular, atividade neurológica e reprodução. Como estas seqüências possuem um papel fundamental em todos os organismos, espera-se que o seu entendimento leve a uma revolução em inúmeras áreas, como a medicina, biologia, agricultura, pecuária, entre outras.

O grande desafio dos pesquisadores consiste em analisar essas seqüências e obter informações biologicamente relevantes. Durante esta análise, os pesquisadores utilizam diversas ferramentas, programas de computador, e um grande volume de informações armazenadas em fontes de dados de Biologia Molecular. De fato, o crescente volume, a distribuição das fontes de dados e a implementação de novos processos em Bioinformática facilitaram enormemente a fase de análise. Porém, criaram uma demanda por ferramentas e sistemas semi-automáticos para lidar com tal volume e complexidade.

Esta dissertação aborda a construção de um sistema que facilite a fase de análise de seqüências e o gerenciamento dos processos utilizados na análise por ferramentas de Bioinformática.

A dissertação apresenta, inicialmente, um levantamento de requisitos para estes tipos de sistemas, com informações básicas sobre ferramentas e métodos de análise, em seguida propõe o desenvolvimento de um sistema que possibilite a análise de dados de projetos de EST's (*Expression Sequence Tags*), o gerenciamento dos dados e instanciações quanto ao ambiente de trabalho e a otimizações de processos.

Por fim, a dissertação descreve um estudo de caso de uma implantação do sistema em um laboratório de genética e biologia molecular aplicado ao estudo de fungos e alguns resultados experimentais obtidos através do sistema.

1.1 Organização do Trabalho

A dissertação está organizada em 6 capítulos.

O Capítulo 2 contém uma discussão sucinta dos contextos biológicos e computacional, necessária para o entendimento e motivação deste trabalho. Esta discussão dá uma visão geral dos assuntos tratados em Bioinformática relevantes a esta dissertação, como os programas de análise, os bancos de dados, os sistemas de anotação, a integração dos dados e aplicativos, os *pipelines* e os sistemas de gerenciamento e análise de dados por Bioinformática, especificando os requisitos que eles devem atender.

O Capítulo 3 apresenta os passos para o desenvolvimento do software, com um levantamento de requisitos, as etapas de desenvolvimento, as ferramentas e métodos utilizados.

O Capítulo 4 exibe o software desenvolvido, seus módulos com suas respectivas funções.

O Capítulo 5 apresenta um estudo de caso como teste de funcionamento do software implantado no Laboratório de Genética e Biologia Molecular de Fungos do Departamento de Genética da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo.

Por fim, o Capítulo 6 apresenta as contribuições e sugere trabalhos futuros.

II - REVISÃO E CONCEITOS BÁSICOS

Este capítulo apresenta os contextos biológico e computacional necessários para o entendimento deste trabalho.

Os principais conceitos relacionados ao projeto genoma, transcriptoma e proteoma serão apresentados na discussão da parte biológica. Uma visão mais detalhada sobre o assunto pode ser encontrada em (Wasinger, 2006).

Quanto à discussão do contexto computacional serão dados elementos mais abrangentes da Bioinformática, que é vista como uma área que auxilia aos pesquisadores em Biologia, criando, melhorando, ampliando, desenvolvendo e manipulando ferramentas e bancos de dados de Biologia Molecular, que são utilizados por estes pesquisadores na interpretação e organização dos dados obtidos experimentalmente nos laboratórios (Kim, 2002).

Todos os assuntos relevantes a esta dissertação terão o foco da Bioinformática voltada aos programas de análise, os bancos de dados, os sistemas de anotação, a integração dos dados e aplicativos, os *pipelines* e o gerenciamento de dados em Bioinformática. O capítulo conclui com uma discussão envolvendo aspectos de gerenciamento e análise de dados por Bioinformática.

2.1 Contexto Biológico

Cada célula de um organismo vivo contém cromossomos, que são compostos de uma seqüência escrita em um alfabeto formado por A,C,G,T denominados nucleotídeos. O conjunto dessas seqüências ou DNA inteiro de um organismo vivo forma o genoma, código que traz instruções para controle da replicação celular e do funcionamento do organismo. O genoma costuma variar em tamanho de acordo com a espécie, desde milhões de pares de bases, no caso de bactérias, até

bilhões de pares de bases como é o caso do genoma humano (Waterman, 1996).

De um modo geral, quando dizemos pares de base, estamos enfatizando a estrutura do DNA que é formado por uma fita dupla, onde as fitas têm orientações opostas e complementares, base A (adenina) sempre irá parear com T (timina) e C (citosina) com G (guanina). As orientações são ditas serem de extremidade a extremidade 5′ → 3′ e 3′ → 5′ (Figura 1). O tamanho de um trecho de DNA de fita dupla é medido pelo seu número de pares de base, denotados por bp (*base-pairs*) (Zaha, 2000).

Figura 1. Exemplo de duas fitas pareadas de DNA.

O Projeto Genoma Humano é um programa coordenado pelo U.S. Department of Energy (DOE, 2006) e National Institutes of Health (NIH, 2006) que foi oficialmente iniciado na década de 90 e que tem dentre várias diretrizes, mapear e seqüenciar completamente o genoma humano, identificar aproximadamente trinta mil genes no DNA humano, armazenar esta informação em bancos de dados, contribuir com a pesquisa além do setor privado e tratar de conseqüências éticas e legais que surgirem com o projeto.

Além do genoma humano, outros organismos vêm sendo seqüenciados, criando assim novas tecnologias de análise de dados de Biologia Molecular com o objetivo de disponibilizar ainda mais informações biológicas que trarão avanços em diversos campos, como Biologia, Medicina e agricultura. As seqüências destes outros organismos contribuem no Projeto Genoma Humano, pois facilitam a descoberta de funções de genes, já que há um princípio biológico que diz que no caso de duas seqüências, sejam elas nucleotídicas ou protéicas, similares, é

razoável supor que suas funções também sejam similares (Venter et al., 2001).

Alguns objetivos do projeto genoma são (Sousa, 2001):

- Compreender de forma mais ampla a organização do genoma e a função dos genes humanos, através da comparação com outros genomas já seqüenciados;
- Diagnosticar antecipadamente doenças;
- Desenvolver drogas específicas;
- Curar doenças genéticas;
- Determinar a predisposição genética de indivíduos a doenças como o câncer.

No Brasil o Ministério da Ciência e Tecnologia e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) lançaram, em dezembro de 2000, o Projeto Genoma Brasileiro com a participação de 25 laboratórios de biologia molecular, distribuídos em todas as regiões geográficas do país.

Com o objetivo de ampliar a competência em nível nacional das atividades de pesquisa sobre genômica a Rede Genoma Brasileiro (Brazilian-Genome, 2006) vem oferecendo a formação de recursos humanos especializados e desenvolvendo trabalhos multi-institucionais.

Até o momento a Rede Genoma Brasileiro atuou em três grandes projetos. O primeiro se trata do seqüenciamento da bactéria *Chromobacterium violaceum*, freqüentemente encontrada no solo e na água em regiões tropicais e sub-tropicais. O segundo organismo seqüenciado foi a bactéria *Mycoplasma synoviae*, causadora de doenças endêmicas que são transmitidas verticalmente através de ovos contaminados de aves infectadas. O terceiro, em fase de andamento, é o seqüenciamento do Genoma de *Anopheles darlingi*, mosquito responsável pela transmissão da malária.

Em projetos genômicos, várias são as etapas a serem realizadas em seu desenvolvimento. Na etapa de seqüenciamento, apesar dos genomas variarem de tamanho, independentemente do tipo de organismo, as reações químicas que os pesquisadores utilizam para

decodificar os pares de bases do DNA são precisas para se obter em média 600 nucleotídeos por vez. Sendo assim, um dos processos mais utilizados nesta etapa é conhecido como *shotgun*, que se inicia com a quebra do DNA em milhões de fragmentos aleatórios, que são então alimentados ao seqüenciador de DNA.

A partir deste momento inicia-se a fase computacional de análise dos dados, onde, logo no primeiro passo, as bases de DNA dão lugar a quatro curvas representando um sinal para cada leitura de cada um dos tipos de nucleotídeos. Para esta representação é dado o nome de cromatograma ou eletroferograma. Em seguida inicia-se o processo de *base calling* no qual o cromatograma é convertido em seqüência de bases (também chamada a partir desta etapa de *read*) por um programa específico. Alguns exemplos de programas que realizam essa tarefa incluem o Phred (Ewing and Green, 1998; Ewing et al., 1998) e o Chromas (Technelysium-Pty, 2006). Logo após esta etapa, é feita a reconstrução da seqüência original através da combinação (montagem) dos *reads*, esta fase, normalmente chamada de *assembling* é realizada por programas como CAP3 (Huang and Madan, 1999) e Phrap (Green, 2006). Este processo muitas vezes não é simples e pode conter erros, seja por causa de limitações na tecnologia de seqüenciamento ou até mesmo por falha humana.

Cada conjunto de *reads* agrupados é chamado de *contig*. O objetivo final em projetos genoma é unir os *reads* gerados em apenas um *contig* representando o cromossomo completo. A esta etapa chamamos de *finishing* que visa fechar os buracos entre os vários *contigs* obtidos durante o processo de seqüenciamento com o objetivo de formar um único *contig*.

Neste momento teremos os dados brutos do genoma, ou seja, ainda sem significado biológico ou funções conhecidas. É neste momento que os pesquisadores entram em outra etapa de análise, chamada de fase de anotação, cujo objetivo consiste em atribuir a esses dados informações biologicamente relevantes.

Vários dos programas e bancos de dados em Biologia Molecular utilizados nesta fase serão abordados no decorrer desta dissertação. Porém cabe ressaltar que o processo de anotação é uma combinação entre programas de computador e interpretações humanas, tornando-se cada vez maior o desafio computacional para que estas análises se tornem eficientes e cuidadosas.

Nem todos os projetos têm o objetivo de seqüenciar completamente o genoma do organismo. Muitas vezes prefere-se realizar o seqüenciamento apenas das regiões gênicas, utilizando informações oriundas de RNA mensageiro (mRNA), produzindo-se assim pequenas seqüências que irão representar pedaços dos genes expressos no momento da extração do mRNA, a este tipo de estudo denominamos genômica funcional ou projeto transcriptoma (Prosdocimi et al., 2002). Essas pequenas seqüências são denominadas etiquetas de genes expressos, ou ESTs (*Expressed Sequence Tags*), esta preferência muitas vezes é tomada levando-se em consideração simplicidade e redução de tempo e custo dos projetos.

EST é uma seqüência obtida de forma aleatória, geralmente incompleta, de DNA, que representa um gene expresso que é aquele cujo produto, seja uma proteína ou um RNA, está sendo produzido em um dado momento em uma célula (Griffiths et al., 2000).

Como foi dito, estes projetos beneficiam-se do fato de ser relativamente simples fazer cópias de DNA a partir de mRNA durante o processo de tradução, processo onde o DNA dará origem ao RNA. Os projetos de ESTs são mais comuns para eucariotos, organismos pluricelulares, principalmente devido ao baixo custo além é claro de fornecer informações importantes quanto a uma estimativa do número de genes que codificam proteínas que é uma primeira medida útil para descobrirmos a complexidade molecular do organismo em estudo (Ewing and Green, 2000).

Além das etapas de anotação nos projetos de genomas funcionais, procura-se estudar a expressão gênica, ou seja, se um gene é expresso

ou não, assim com as diferenças no seu nível de expressão em uma célula ou tecido, em determinado momento (quantificação do mRNA).

Avanços tecnológicos têm contribuído para o desenvolvimento de técnicas de quantificação de mRNA em larga escala, possibilitando o estudo paralelo de centenas ou milhares de genes. Entre estas técnicas poderíamos destacar *Differential Display* (DD), *Serial Analysis of Gene Expression* (SAGE), *Suppression Subtractive Hybridization* (SSH) e DNA microarray.

Para entender a função de todos os genes em um organismo, é necessário conhecer não só quais genes são expressos, quando e onde, mas também quais são os produtos da expressão e em que condições esses produtos (proteínas) são sintetizados em certos tecidos. O projeto proteoma tenta descrever o conjunto completo de proteínas produto da expressão do genoma (James, 1997).

Apesar da seqüência de aminoácidos de uma proteína ser definida pelas informações contidas no DNA, não é possível deduzir o proteoma conhecendo-se o genoma. Cada tipo de célula possui apenas parte do total de genes do genoma apta para formar proteínas e após a tradução, as proteínas podem sofrer modificações químicas que não estavam codificadas no genoma, portanto, cada célula possui seu proteoma específico, fazendo com que a variedade de proteínas aumente, o que ajuda a compreender porquê células com genomas iguais desempenham funções diferentes.

2.2 Programas e Métodos de Análise

O primeiro passo na descoberta de informações sobre seqüências moleculares dentro de um laboratório é verificar se outros pesquisadores já estudaram ou obtiveram seqüências similares à obtida experimentalmente. Neste caso a ferramenta computacional mais usada em Biologia é o BLAST – Basic Local Alignment Tool (Altschul et al., 1990; Altschul, 1998) – que através de uma heurística de comparação,

procura em bancos de dados, como o Genbank (Benson et al., 2003; NCBI, 2006a) todas as seqüências similares a uma determinada seqüência.

O alinhamento de seqüências é uma operação básica em Bioinformática que dá suporte a várias aplicações na Biologia Molecular, tais como, descoberta de homologia entre seqüências de proteína e DNA, predição de genes, comparação de genomas, predição de estruturas, análise filogenética, reconhecimento de padrões, entre outros.

O alinhamento de seqüências pode ser definido em 2 tipos distintos: alinhamento local e alinhamento global (Figura 2).

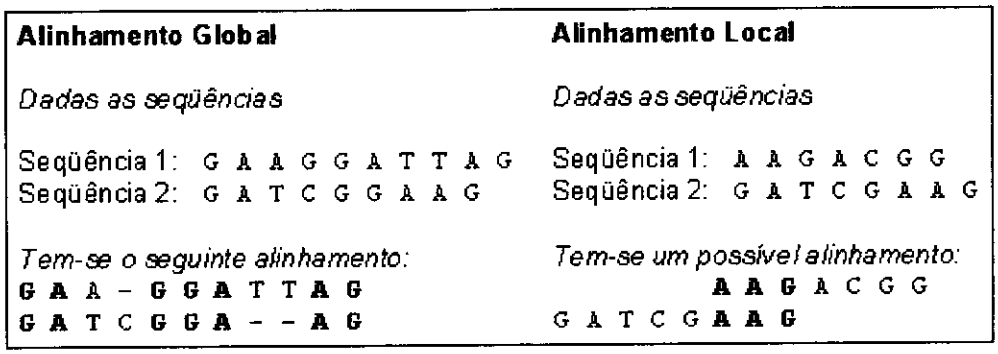


Figura 2. Exemplo de alinhamento global e local de seqüências.

No alinhamento local procura-se alinhar apenas as regiões mais conservadas entre seqüências, independente da localização relativa de cada região. Dessa forma, o alinhamento resulta em uma ou mais regiões similares entre as seqüências. Já no alinhamento global, as seqüências devem ser alinhadas de um extremo ao outro, resultando em apenas uma região de similaridade (Baxevanis and Ouellete, 2001).

O primeiro algoritmo projetado para realizar alinhamento global foi o Needleman-Wunsch (Needleman and Wunsch, 1970). Enquanto que para o alinhamento local foi construída uma variante chamada Smith-Waterman (Smith and Waterman, 1981). Estes dois algoritmos trabalham com tempo de execução proporcional ao produto dos comprimentos das seqüências a serem alinhadas.

Como na maioria dos projetos desejamos encontrar similaridade entre seqüências de proteínas e DNA e estas similaridades ocorrem apenas em segmentos das seqüências, utilizamos o alinhamento local normalmente baseado em Smith Waterman. No entanto, muitas vezes existe a necessidade de analisar um número muito grande de seqüências com tamanhos diversos, neste caso os programas FASTA (Pearson and Lipman, 1988) e BLAST são mais adequados, pois usam heurísticas para concentrar seus esforços computacionais nas regiões das seqüências com maiores probabilidades de estarem relacionadas, diminuindo o tempo de execução do alinhamento.

Dentre outras ferramentas que utilizam métodos de alinhamento de seqüência local e global, podemos citar:

- CLUSTALW (Thompson et al., 1994) e MultAlign (CBRG, 2006) com a função de alinhamento múltiplo de seqüências (Corpet, 1988);
- ORFFinder (NCBI, 2006b) para predição de genes;
- Modeller (Sali, 2006) para predição de estruturas de proteínas;
- PHYLIP (Felsenstein, 2006) para análise filogenética;
- Cross_Match (Green, 2006) para alinhamento local e muitas vezes utilizado na retirada de regiões de contaminantes em seqüências de insertos clonados;
- CAP3 e Phrap para montagem (agrupamento) de seqüências;
- GCG Wisconsin Package (Accelrys-Software, 2006) e EMBOSS (Rice et al., 2000) como pacotes com várias funções de alinhamento e análise de seqüências.

2.3 Bancos de Dados em Biologia Molecular

A quantidade de bancos de dados em Biologia Molecular vem crescendo exponencialmente nos últimos anos. Os objetivos destes

bancos variam e podem ser utilizados para armazenar e disponibilizar bioseqüências, funções moleculares, estruturas de proteínas, modelos metabólicos, entre outros, oferecendo em alguns casos informações mais amplas, ou seja, cobrindo uma área mais significativa da Biologia, e em outros, informações menos detalhadas.

Em alguns casos, as informações biológicas são obtidas por análise computacional de outros bancos de dados; já em outros, através da literatura ou informações definidas por pesquisadores.

Dentre os bancos de dados mais comuns, é possível destacar:

- Genbank (Benson et al., 2005), EMBL (Kulikova et al., 2004) e DDBJ (Tateno et al., 1998) exemplos de bancos de dados de seqüências de nucleotídeos;
- Swiss-Prot (Boeckmann et al., 2003), PIR (Wu et al., 2002) e Uniprot (Apweiler et al., 2004; Bairoch et al., 2005) bancos de seqüências de aminoácidos;
- PDB (Protein Data Bank) (Westbrook et al., 2002) banco de estruturas terciárias de proteínas.

Existem também outros bancos de dados, muitas vezes derivados destes primeiros e com funções bem específicas, alguns dos quais podemos citar:

- PROSITE (Falquet et al., 2002) armazena padrões de seqüências protéicas conservadas, associados a funções específicas;
- KEGG (Kanehisa, 1997; Kanehisa and Goto, 2000; Kanehisa et al., 2006) banco funcional que disponibiliza mapas metabólicos de organismos com genoma completamente ou parcialmente seqüenciados.

Cabe ressaltar que com o crescente número de dados biológicos que vem sendo gerado, vários bancos de dados têm surgido e anualmente a revista *Nucleic Acids Research* publica uma lista atualizada com a classificação de todos os bancos de dados biológicos disponíveis (Galperin, 2006).

2.4 Sistemas de Anotação

Uma das tarefas mais importantes realizadas após a obtenção das seqüências gênicas de projetos de seqüenciamento de genes expressos ou até mesmo de projetos genoma é a interpretação dos dados experimentais com o objetivo de se obter conhecimento biológico a respeito dos mesmos.

Nesta fase os pesquisadores são auxiliados por programas de análise de seqüências, que buscam em fontes de dados externas ou internas, informações para esclarecer e interpretar os dados obtidos na bancada.

Existem atualmente várias fontes de dados, assim como vários programas de análise destes dados. As características das anotações, as fontes de dados e os programas de análise vão variar dependendo do tipo de projeto ou tipo de organismo.

Dentre as características de um sistema de anotação, ele deve oferecer (Andrade et al., 1999):

- Sistema Gerenciador de Banco de dados para armazenar informações obtidas por programas de análise, facilitando assim o acesso a estes resultados. Esta estratégia é útil quando a comunidade de pesquisadores planeja acessar, visualizar e analisar estes resultados repetidas vezes;
- Ferramentas para captura de dados internos e externos, armazenados em seu sistema gerenciador de banco de dados ou obtidos através das diversas fontes de dados disponíveis. Da mesma forma que as fontes de dados externas estão sempre sendo atualizadas, o sistema de anotação deve importar periodicamente tais dados para seu gerenciador;
- Mecanismos de controle de execução de programas que estejam em sítios externos ou locais;

- Ferramentas de busca e análise com interfaces amigáveis, tabelas e gráficos, além de desenhos que representam os elementos genômicos e seus atributos;
- Controle de acesso aos seus dados, devido à utilização por vários pesquisadores, dos quais podem obter informações públicas ou restritas ao seu grupo de trabalho.

Atualmente existem vários sistemas de anotações, destacando-se GeneQuiz (Hoersch et al., 2000), Artemis (Rutherford et al., 2000), EDITtoTrEMBL (Moller et al., 1999), GBrowser (Stein et al., 2002), Cancer Annotation Project (LBI, 2006), NCBI's Genome Annotation Project (Agarwala, 2006). Grande parte destes sistemas não atende a todos os requisitos levantados anteriormente.

2.5 Integração

Devido à grande quantidade de programas de análise de dados e fontes de dados, e devido a estes dados e programas muitas vezes serem heterogêneos, criou-se a necessidade de integração dos mesmos, a fim de se obter um conhecimento completo sobre o objeto de estudo.

Modelos de dados armazenados em sistemas gerenciadores de banco de dados ou até mesmo em simples arquivos textos necessitam de integração.

Geralmente as fontes não oferecem uma documentação detalhada do esquema do banco de dados, ou seja, o simples fato de conhecer os dados disponíveis e seus domínios pode se tornar uma tarefa não muito trivial a quem necessita de uma determinada informação em alguns destes bancos.

Um dos grandes desafios da Bioinformática é a construção de ferramentas de integração das informações que residem nestas diversas fontes de dados.

2.6 Pipeline

Muitas tarefas descritas pelos pesquisadores envolvem a composição de vários programas. Uma coleção de dados produzida por um programa pode, dada uma semântica apropriada, ser a coleção de entrada de outro programa. Este conceito é denominado *Pipeline*.

Estes sistemas dizem respeito à automatização de procedimentos, onde informações ou tarefas são passadas entre programas participantes de acordo com um conjunto pré-definido de regras para se alcançar um objetivo global.

A composição destes programas não é uma tarefa simples de ser realizada por pesquisadores da área biológica, tornando-se uma grande barreira para análises mais complexas. Daí a importância de profissionais envolvidos na área de informática no desenvolvimento e gerenciamento destes sistemas.

A utilização em estudos comparativos e a necessidade de acomodar diferentes fontes de dados com diferentes formatos e modos de acesso, além do "tsunami" de dados que requer sistemas cada vez mais seguros e robustos, e com o estado evolutivo de novos algoritmos e paradigmas na análise de dados por Bioinformática, fica cada vez mais evidente a aplicação e o estudo de sistemas que atendam esta demanda (Hoon et al., 2003).

Existem diversos pacotes de programas com esta finalidade, disponíveis para os pesquisadores em biologia. Estes pacotes muitas vezes possuem *scripts* que definem o *pipeline* de análise com parâmetros *default*, dificultando que o pesquisador realize modificações nestes *scripts* e até mesmo interprete os resultados obtidos. Por exemplo, o pacote Phred/Phrap possui um *script* que, dado um ou vários cromatogramas, chama todos os programas que geram *reads* e *contigs*, incluindo conceitos que alguns usuários podem não tomar conhecimento, como no caso do mascaramento de vetores, ou até mesmo nos parâmetros utilizados para definir regiões de sobreposição utilizadas na formação dos *contigs*.

Um exemplo de *pipeline* mais elaborado para projetos de ESTs, incluindo análise dos dados e armazenamento dos resultados em um sistema gerenciador de banco de dados pode ser visto a seguir (Figura 3).

O primeiro passo inicia-se com os cromatogramas, obtidos no laboratório, que constituem entradas para o programa Phred. Os *reads* (no formato phd) são então encaminhados ao programa phd2fasta para que os mesmos sejam convertidos em formato FASTA. Alguns *pipelines* possuem esta próxima etapa automatizada, como é o caso do *pipeline* desenvolvido para o Laboratório de Genética e Biologia Molecular de Fungos do Departamento de Genética da Faculdade de Medicina de Ribeirão Preto, estudo de caso deste trabalho. Nesta etapa são gerados filtros para identificar as seqüências de boa qualidade baseadas nos dados extraídos do programas Phred, juntamente com o filtro de contaminantes (vetores e *primers*) dos *reads* (realizado pelo programa Cross_Match).

A partir do momento em que os *reads* tiverem suas regiões de contaminantes e de baixa qualidade cortadas (*trimming*), três novas etapas serão realizadas na análise dos dados. A primeira etapa inicia-se com o BLASTX, que faz a tradução da seqüência de nucleotídeos do *read* nos seis *frames* possíveis, gerando seis seqüências de aminoácidos, e realiza a comparação destas seqüências com o banco de dados Não-Redundante do Genbank (NR). Os seis possíveis *frames* são as fases de leitura em que o DNA é traduzido em proteína, através da "leitura" de trincas de nucleotídeos de forma seqüencial. Isso faz com que seja possível a leitura de seis diferentes fases para a mesma molécula de DNA, sendo três fases para cada fita.

O BLASTX, além de mostrar seqüências similares do NR, sugere qual é o frame correto para que seja feita a tradução do read para uma seqüência de aminoácidos, neste caso utiliza-se o programa Transeq (utilitário do pacote EMBOSS) para esta função.

A seqüência de aminoácidos é então alimentada ao programa RPS-BLAST (Reverse Position Specific - BLAST) (NCBI, 2006c) para fazer a comparação com o banco de dados CDD (*Conserved Domain*

Database) (Marchler-Bauer et al., 2003) encontrando domínios conservados em regiões da seqüência estudada.

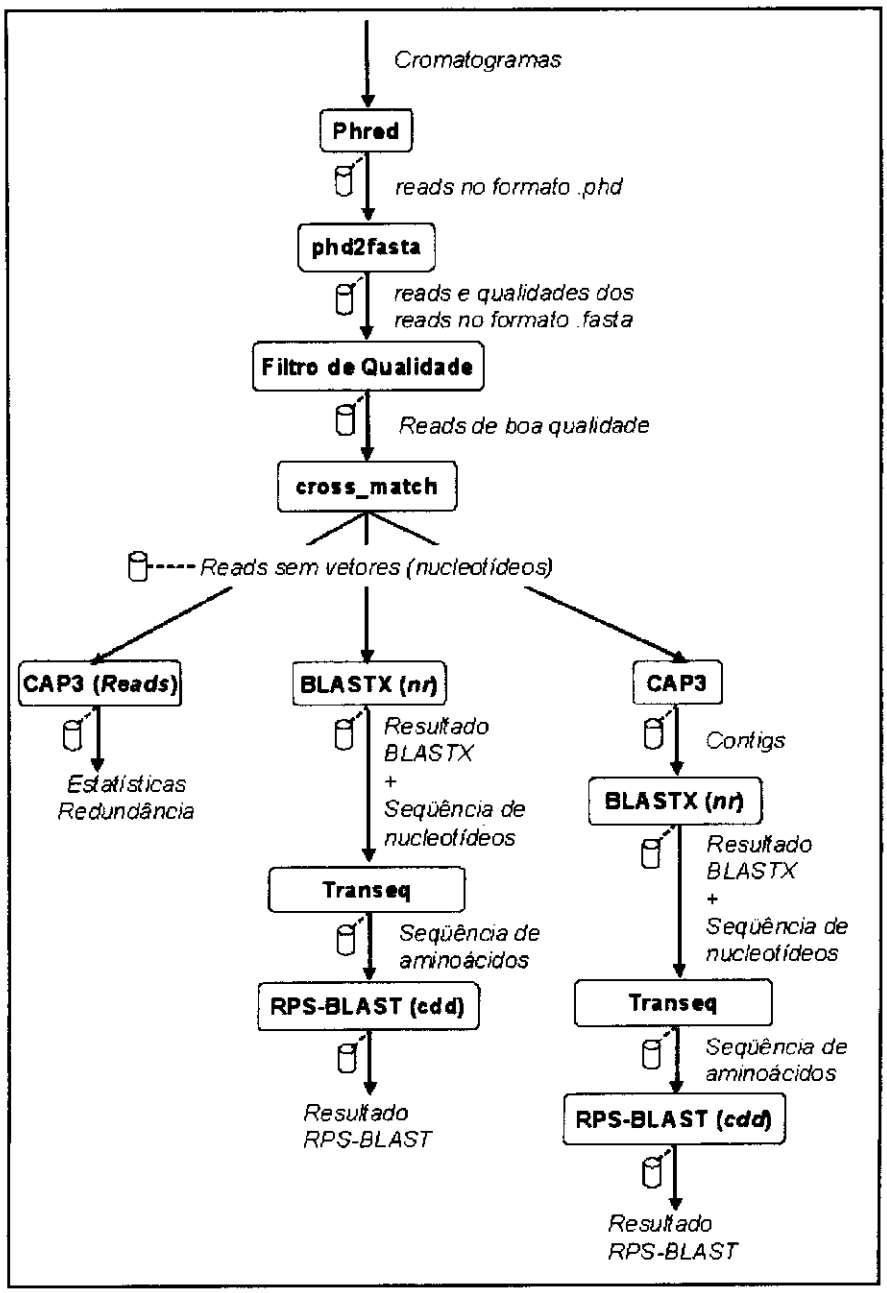


Figura 3. Esquema de *pipeline* para projetos de ESTs.

A segunda etapa trata da comparação dos novos *reads* (que acabaram de ser seqüenciados) com todos os outros *reads* já seqüenciados. Esta comparação é feita com o CAP3 e o objetivo é descobrir se o *read* novo já foi seqüenciado antes. Esta etapa é

importante para mostrar quando o processo de seqüenciamento de um projeto de EST não está mais produzindo seqüências novas (identificação da redundância da biblioteca).

A terceira etapa é a montagem de fragmentos feita também com o programa CAP3 os chamados *contigs*, que serão analisados com o objetivo de identificar seqüências mais precisas e completas do cDNA ou parte de um cDNA, já que se trata de um *pipeline* para projetos de ESTs. Depois da montagem, faz-se com os *contigs* (cDNA) as mesmas comparações feitas com os *reads* individuais, ou seja, compara-se aos bancos de dados NR e CDD.

2.7 Gerenciamento de Pipelines

Não podemos esquecer que o acréscimo ou remoção de qualquer um dos passos ou processos em projetos de *pipeline* pode ser uma tarefa não muito trivial para a maioria dos usuários.

Sendo assim, torna-se importante a criação de sistemas que facilitem este processo não apenas de análise dos dados, mas também o gerenciamento dos passos, facilitando a inclusão ou remoção de programas e fontes de dados no decorrer de cada projeto de análise.

O gerenciamento de *pipelines* oferece automatização dos procedimentos de um projeto de análise através do gerenciamento de sua seqüência de trabalho.

Alguns requisitos importantes neste tipo de aplicação:

- O sistema deve incluir os processos, fontes e recursos normalmente usados e oferecer mecanismos de extensibilidade de novos processos, fontes e recursos. No caso dos processos podemos citar a grande variedade de programas de análise por Bioinformática. Já no caso das fontes de dados, a grande quantidade de bancos de dados em Biologia Molecular disponíveis atualmente;

- O sistema deve ajudar os usuários na definição e redefinição dos processos mais importantes, ou seja, que devem ser considerados úteis pelos pesquisadores;
- O sistema deve oferecer ferramentas para validação de dados. Verificando se as entradas e saídas geradas possuem coerência;
- O sistema deve ser otimizado de acordo com a arquitetura que está sendo utilizada;
- O sistema deve oferecer agendamento da execução das análises. Permitindo que o *pipeline* de análise dos dados seja executado em horários específicos e de menos fluxo de trabalho no servidor;
- O sistema deve conter controle de usuários e permissões para definição e redefinição de seus processos.

2.8 Sistemas de Gerenciamento e Análise de Dados por Bioinformática

Nesta dissertação, denominamos Sistema de Gerenciamento e Análise de Dados por Bioinformática (SGADBio), um sistema capaz de analisar seqüências moleculares, armazenar resultados analisados, extrair diversos tipos de informações e permitir o instanciamento dos processos envolvidos nos diversos tipos de análise por Bioinformática.

Esta dissertação apresenta o desenvolvimento de um software com estas características, que poderá ser configurado para diversos ambientes de trabalho em diversos tipos de projeto de seqüenciamento.

2.9 Comentários Finais

Este capítulo apresentou uma discussão sucinta dos contextos biológico e computacional, apresentando uma visão geral dos assuntos em Bioinformática relevantes a esta dissertação, como os programas de

análise, os bancos de dados, os sistemas de anotação, a integração dos dados e aplicativos, os *pipelines*, o gerenciamento de *pipelines* e os sistemas de gerenciamento e análise de dados por Bioinformática (SGADBio).

De acordo com o que foi argumentado, mostrou-se a necessidade do desenvolvimento de um SGADBio que atendesse aos diversos requisitos expostos neste capítulo.

Os passos para o desenvolvimento deste sistema serão descritos no próximo capítulo.

III – O DESENVOLVIMENTO DO SISTEMA DE GERENCIAMENTO E ANÁLISE DE DADOS POR BIOINFORMÁTICA - SGADBIO

Este capítulo apresenta as fases de desenvolvimento do projeto proposto, incluindo todo o processo de desenvolvimento do software, ou seja, o conjunto de métodos, técnicas e ferramentas utilizadas para analisar, projetar e gerenciar o desenvolvimento e a manutenção do software.

3.1 Visão geral

O SGADBio objetiva realizar, de forma geral, a análise de dados de Biologia Molecular (também chamadas de bioseqüências).

O sistema disponibiliza uma ferramenta para gerenciar os processos envolvidos em projetos de seqüenciamento, permitindo assim a criação de diversas instâncias e métodos de análise para os diferentes tipos de projetos de pesquisa existentes.

Informações como usuários, *primers*, vetores, mRNA e tipos de biblioteca são cadastradas no sistema.

Seqüências são submetidas para análise, esta previamente configurada com uma seqüência de processos, fontes de dados e atividades, dos quais os resultados são então armazenados em sistemas gerenciadores de banco de dados e disponibilizados em um ambiente *Web*, facilitando assim o interfaceamento com o usuário final.

Informações de grande importância para pesquisas em biologia molecular podem ser geradas pelo sistema, tais como:

- Identificação de regiões contaminantes em seqüências;
- Identificação do padrão de qualidade de bibliotecas e seqüências;

- Identificação do grau de redundância do seqüenciamento de bibliotecas;
- Identificação de similaridade da seqüência com diversos bancos de dados de Biologia Molecular;
- Estatísticas de análise como tamanho médio de clones, número de seqüências de baixa e alta qualidade, porcentagem de similaridade entre seqüências analisadas, quantidade de domínios conservados, além de diversos gráficos representativos dos resultados.

O sistema permite ainda:

- Controlar o fluxo de processos a serem executados na etapa de análise de dados, ou seja, através de uma interface com o usuário, pode-se definir a seqüência de passos que a seqüência molecular será submetida;
- Controle de usuários e permissões, tornando o gerenciamento do software mais confiável;
- Registro de *Logs* de operações, permitindo que seja identificado o usuário responsável pelas diversas operações disponíveis no sistema.

Outro importante recurso do sistema é o de agendar a análise dos dados através de um controle para identificar os horários de menos fluxo de trabalho no servidor.

3.2 Arquitetura

A arquitetura de software é uma estrutura que serve para o entendimento de componentes de um sistema e seus inter-relacionamentos (Silva Filho, 2002).

A arquitetura proposta para o SGADBio é mostrada graficamente (Figura 4). O acesso para pesquisadores e administradores do sistema ocorre através de requisições *http* para o servidor, que interage com

scripts desenvolvidos na linguagem de programação Perl, que por sua vez faz o acesso aos dados armazenados no banco de dados Mysql.

O modelo proposto está baseado na arquitetura do tipo cliente-servidor, no qual o servidor *Web* representa a figura do analisador e armazenador de dados e os pesquisadores são representados através de clientes utilizando seus navegadores.

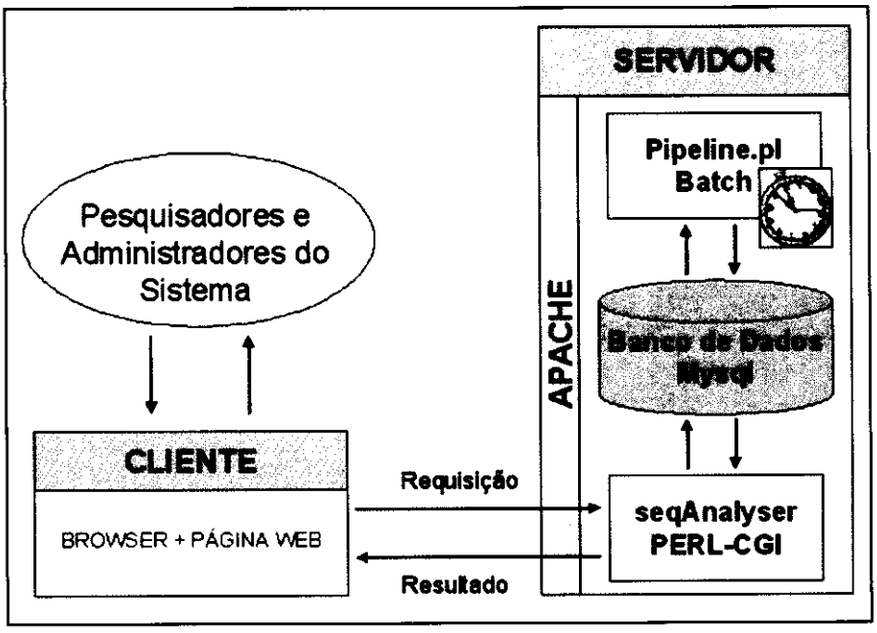


Figura 4. Arquitetura do SGADBio.

O acesso à ferramenta é realizado via *Web* e o acesso ao *script* de análise de dados (*Pipeline*) é feito em modo *batch*[1] através de um "cron", serviço de agenda de tarefas disponível nos sistemas Linux que é carregado durante o processo de *boot* do sistema.

3.3 Elicitação de Requisitos

Na etapa de elicitação de requisitos, procura-se capturar os requisitos do sistema, buscando obter um conhecimento do domínio do problema (Vasconcelos, 2004).

[1] Modo de processamento de dados no qual os dados de entrada são coletados em grupos, ou lotes, e periodicamente processados em seqüência por um ou mais *jobs*.

Foram realizadas várias entrevistas com pesquisadores envolvidos em projetos de seqüenciamento, principalmente os que atuam em análise de expressão gênica em eucariontes[2], mais especificamente em fungos. Foram relatadas várias informações relevantes ao processo de análise de dados de biologia molecular, os objetivos e as limitações funcionais e organizacionais, com o objetivo de identificar e criar soluções que se adequem à realidade mais comumente encontrada nos laboratórios de pesquisa.

3.4 Modelo conceitual

Para melhor entendimento do SGADBio e para fornecer documentação para futuras implementações ou atualizações do software, foi criado um modelo conceitual, de forma que sejam satisfeitos os requisitos propostos, proporcionando portabilidade para qualquer ambiente, uma vez que o modelo desenvolvido é independente de linguagem de programação.

Uma das técnicas que foi utilizada para apresentar os processos do sistema é o diagrama de fluxo de dados (DFD), que é utilizado como o primeiro passo em um projeto estruturado e apresenta o fluxo de dados global em um sistema.

Esquemas representativos, diagramas entidade-relacionamento e textos explicativos, também foram utilizados na etapa de análise do sistema.

3.4.1 Diagrama de Fluxo de Dados

O Diagrama de Fluxo de Dados (DFD) é uma ferramenta de modelagem que permite imaginar um sistema como uma rede de

[2] De forma simplificada, um organismo procariótico é aquele cujo material genético não está organizado em um "envelope" nuclear e não possui organela ligada a membrana; ao contrário do organismo eucariótico, que possui uma membrana ligada ao núcleo, cromossomos múltiplos e organelas internas.

processos funcionais, sendo assim, ele é uma ótima ferramenta para o entendimento e manipulação de um sistema, ao nível lógico, de qualquer complexidade.

Procurou-se logo após o levantamento dos requisitos, documentar o que foi abordado através de uma representação esquemática de processos. Alguns diagramas foram explorados, dos quais um breve resumo está sendo mostrado nesta dissertação através do diagrama de contexto, usado para ilustrar as fronteiras de um sistema (Figura 5), e do diagrama de fluxo de dados nível 0. Este último dividido em partes para melhor entendimento do trabalho.

Cada parte do DFD nível 0 representa a interação do sistema com uma de suas entidades externas. Pesquisador (Figura 6), Gerente de Projetos (Figura 7), Administrador do Sistema (Figura 8) e Comunidade Externa (Figura 9).

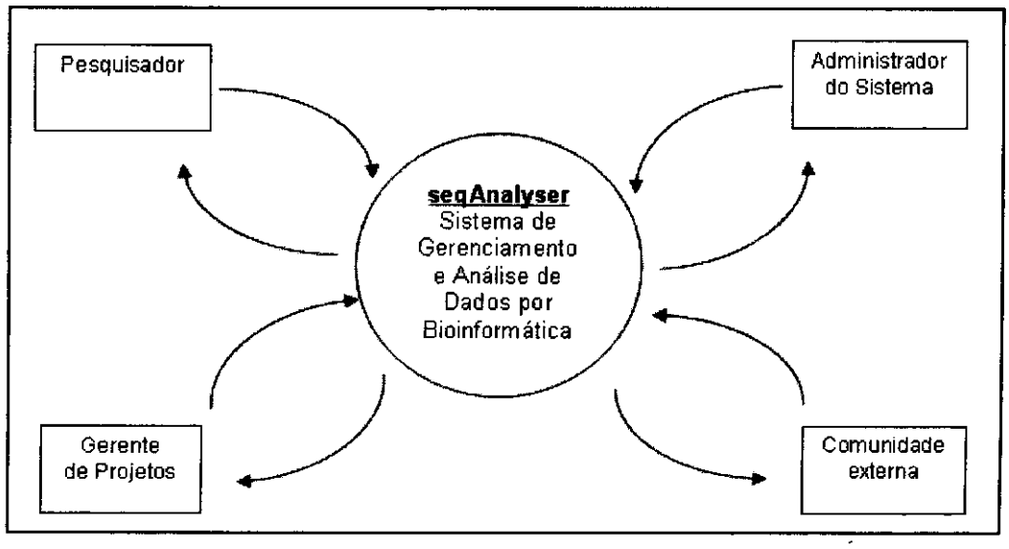


Figura 5. Diagrama de contexto do SGADBio.

Como se pode observar, o sistema interage com quatro entidades externas. A entidade denominada Pesquisador é responsável pela submissão de cromatogramas e análise dos dados dos projetos de seqüenciamento. No caso da entidade Gerente de Projetos, esta é responsável pela configuração do *pipeline*, ou seja, da definição dos processos que serão utilizados na fase de análise dos dados. É ela quem

tem o papel de configurar os parâmetros necessários para a execução do *pipeline*.

As outras duas entidades também possuem um importante papel de interação com o sistema, é o caso da entidade Administrador do Sistema que possui a função principal de administrar e controlar os usuários e tipos de acesso ao software. Não esquecendo da entidade Comunidade Externa que através de solicitações de consulta através do *website* do projeto pode ter acesso a diversas informações disponíveis à comunidade em geral.

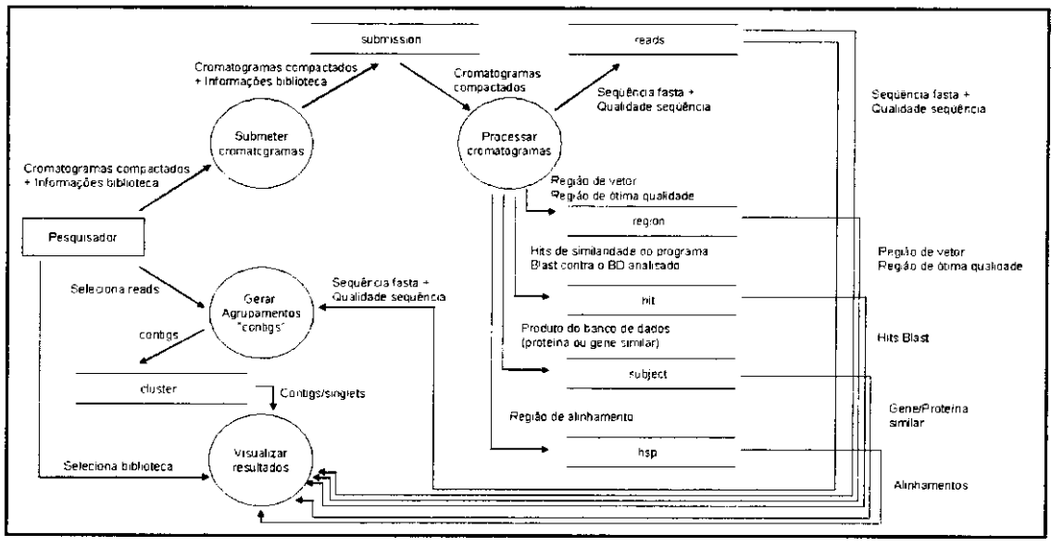


Figura 6. Diagrama de fluxo de dados nível 0 do SGADBio, interação com entidade Pesquisador.

Na Figura 6 a entidade externa Pesquisador possui fluxos de dados apontando para três diferentes processos. Cada um apresenta sua interação com os diversos depósitos de dados contidos no sistema.

No primeiro processo (Submeter cromatogramas), o pesquisador deve encaminhar os cromatogramas compactados em conjunto com as informações pertencentes à biblioteca de seqüênciamento. Estes dados são então armazenados em um depósito de dados chamado submission. Os cromatogramas são então dirigidos à fase de processamento, onde, programas de Bioinformática são responsáveis por extrair informações relevantes e encaminhá-las para armazenamento.

O processo Gerar Agrupamentos consiste em definir os *contigs* da biblioteca. Para isso, as seqüências obtidas dos cromatogramas em formato FASTA são processadas por um programa de montagem de seqüências, que gera os *contigs* para em seguida serem armazenados no depósito de dados cluster.

Outra tarefa do pesquisador é visualizar os resultados. Neste caso ele deve selecionar a biblioteca desejada e o processo responsável por esta função se encarrega de buscar nos diversos depósitos de dados as informações necessárias.

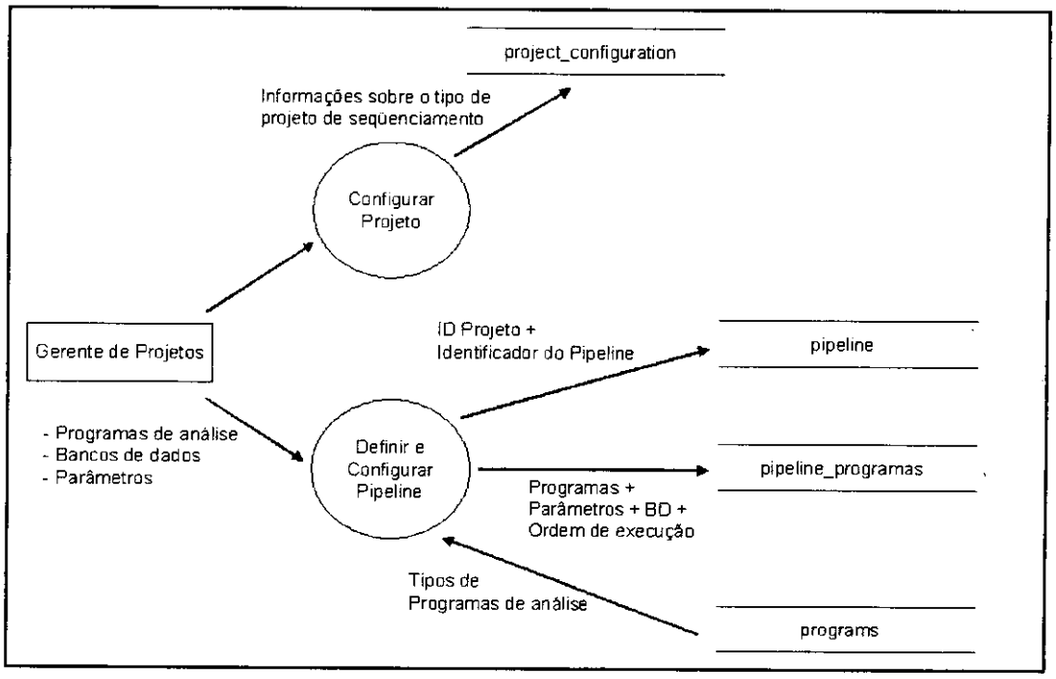


Figura 7. Diagrama de fluxo de dados nível 0 do SGADBio, interação com entidade Gerente de Projetos.

A Figura 7 apresenta o comportamento do sistema em resposta a entidade Gerente de Projetos. Neste modelo, processos referentes à configuração do *pipeline* quanto aos programas de Bioinformática, aos bancos de dados biológicos, aos parâmetros e à ordem de execução são contemplados.

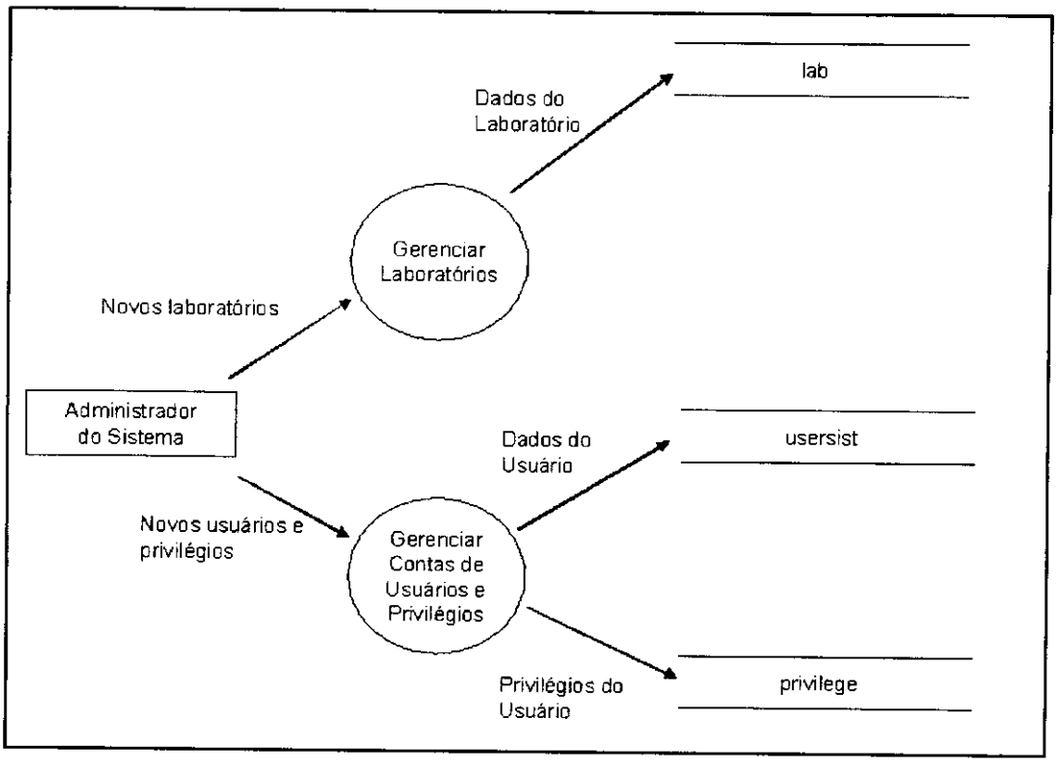


Figura 8. Diagrama de fluxo de dados nível 0 do SGADBio, interação com entidade Administrador do Sistema.

Dados associados com processos que visam o gerenciamento de contas de usuário para acesso ao sistema são demonstrados no diagrama da Figura 8.

O administrador do sistema além de interagir com o processo de gerenciamento destas contas é responsável na manutenção dos laboratórios pertencentes ao ambiente do sistema.

Na Figura 9 tem-se o papel da Comunidade Externa que possui um único processo (nível 0).

Este processo é responsável por atender a solicitação de consulta, realizada através da *homepage* do projeto, e disponibilizar os dados provenientes de análises que possuam acesso livre e aberto a comunidade.

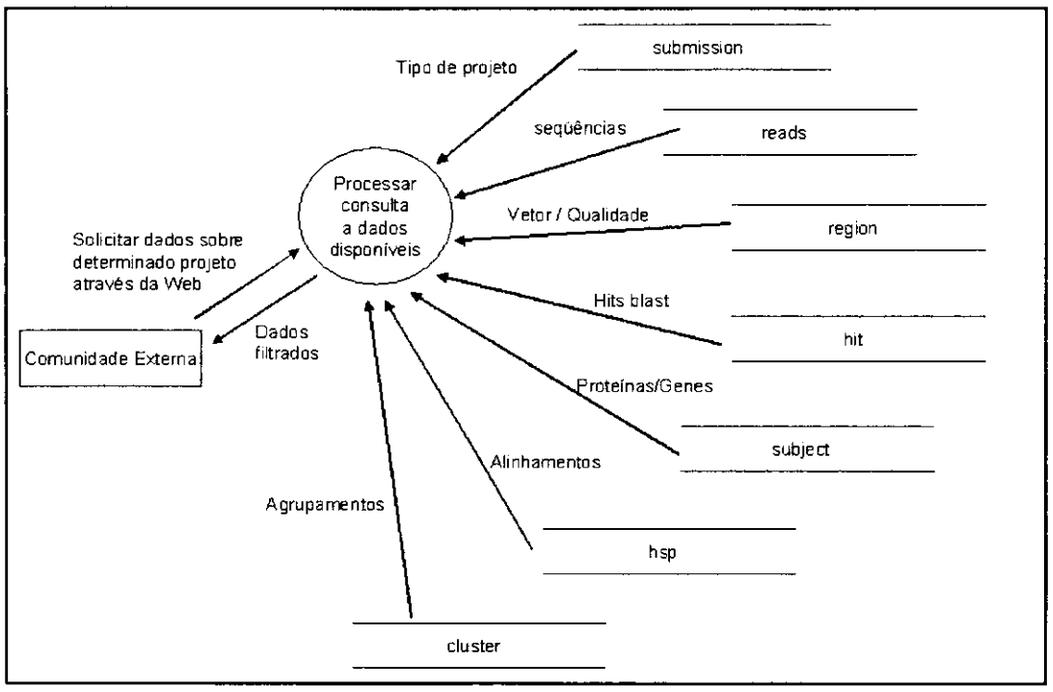


Figura 9. Diagrama de fluxo de dados nível 0 do SGADBio, interação com entidade Comunidade Externa.

Nem todos os fluxos de dados estão representados no DFD nível 0 devido a complexidade do modelo. O objetivo principal nesta fase foi demonstrar os principais fluxos e suas interações com as entidades externas do sistema.

O refinamento do DFD ou explosão dos processos, mais comumente chamado, não foi realizado nesta etapa. Procurou-se a partir deste ponto realizar o armazenamento do conhecimento organizacional, ligando a análise, o projeto e a implementação através do Diagrama Entidade-Relacionamento (DER).

3.4.2 Diagrama Entidade-Relacionamento

O Diagrama Entidade-Relacionamento (DER) é um modelo em rede que descreve a diagramação dos dados armazenados de um sistema em alto nível de abstração (Teorey et al., 2006).

Através do diagrama entidade-relacionamento do SGADBio (Figura 10; Figura 11), pode-se observar que são contemplados inúmeros aspectos que fazem com que os dados armazenados possam gerar informações variadas, devido ao grau de normalização dos dados e a arquitetura proposta baseada em uma estrutura cliente-servidor *Web* e um sistema gerenciador de banco de dados que dê suporte a linguagem de consulta SQL.

Dada a grande quantidade de entidades necessárias no desenvolvimento do software, o diagrama foi dividido em duas partes. O primeiro diagrama representa as entidades e relacionamentos necessários para o armazenamento dos dados analisados após o processamento das seqüências de biologia molecular (Figura 10), já o segundo contém os objetos necessários para manter as configurações sobre o *Pipeline*, ou seja, dos fluxos e processos pertencentes a cada projeto de análise dos dados (Figura 11). Nem todas entidades estão representadas nestas figuras, sendo assim, foram escolhidas as que são envolvidas nos principais aspectos e funções do software.

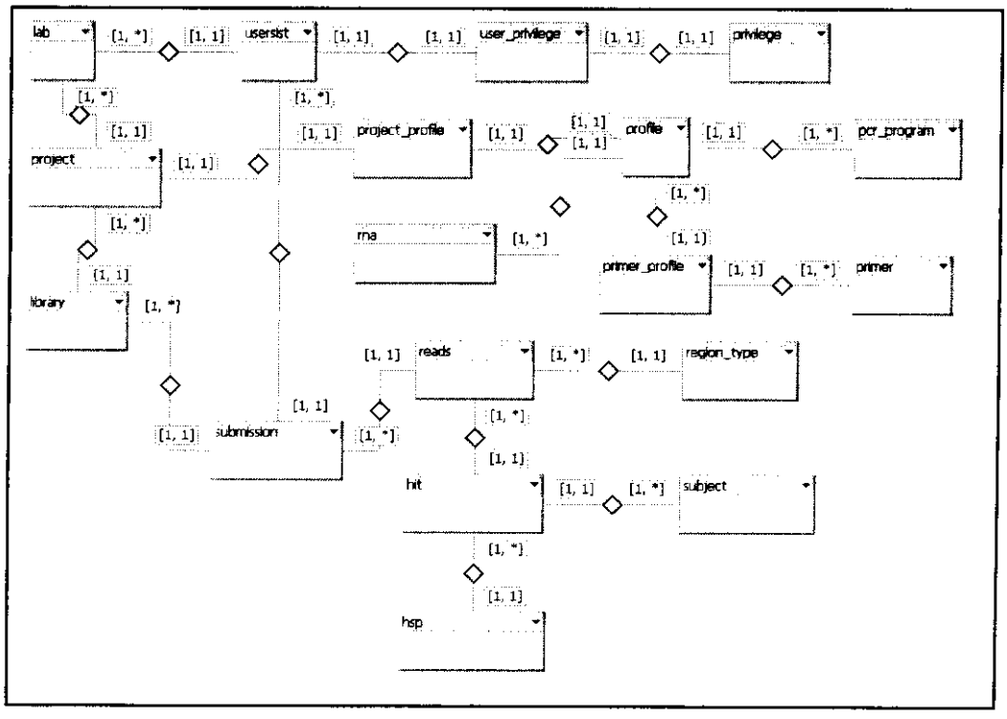


Figura 10. Diagrama entidade-relacionamento para armazenamento de dados analisados pelo SGADBio.

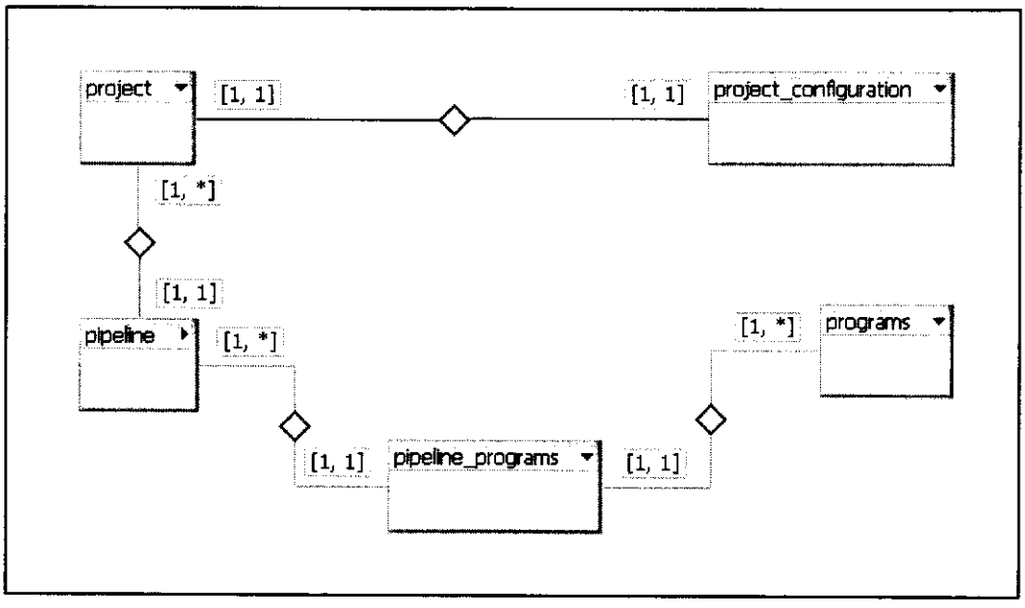


Figura 11. Diagrama entidade-relacionamento para gerenciamento de processos e configurações por tipo de projeto do SGADBio.

Diversas extensões e notações sobre o modelo entidade-relacionamento foram definidas nos últimos anos. O software DB Designer (ferramenta *open source* utilizada para criação de modelos de bancos de dados) foi utilizado para confeccionar os diagramas deste trabalho baseados na notação original de Peter Chen, porém com a ausência dos atributos pertencentes as entidades. O objetivo principal foi facilitar a compreensão dos diagramas.

3.5 Implementação

Durante esta etapa, o projeto do software é implementado como um conjunto de unidades de uma linguagem de programação (Pressman, 1991).

Esta etapa baseia-se totalmente no uso de ferramentas e/ou ambientes de apoio à programação (ex: compiladores, depuradores de código e editores sintáticos) (Pressman, 1991).

3.5.1 Plataforma

A plataforma utilizada no desenvolvimento do SGADBio foi o Linux distribuição Fedora Core 3 sobre um microcomputador com processador Pentium IV 3.0GHz, com 512MB de memória RAM e HD de 80GB.

O Linux é uma implementação livre do UNIX (Welsh, 1999), desenvolvido em meados de 1991 não só por Linus Torvalds (que na época era um estudante de Ciência da Computação da Universidade de Helsinque, na Finlândia), mas por vários programadores ao redor do mundo, projetado para fornecer a usuários de computador pessoal um sistema operacional de alta performance, eficiência e sem custos.

Para o desenvolvimento do SGADBio foi escolhido o Linux devido a vários aspectos positivos, tais como: multitarefa, multiusuário, multiplataforma, segurança, confiabilidade e é claro por se tratar de um sistema livre.

Compõe ainda a plataforma o Servidor *Web* Apache, instalado no sistema Linux para disponibilizar os serviços do sistema através de um ambiente de interface *Web*.

A escolha do Apache foi feita principalmente por se tratar de um software altamente configurável, poder ser executado em diferentes plataformas, ser flexível e também ser livre.

3.5.2 Linguagem de Programação

A programação e a Bioinformática estão relacionadas, tanto na obtenção de dados, quanto no desenvolvimento de ferramentas que resolvam os problemas e obstáculos encontrados na pesquisa em desenvolvimento. A elaboração de ferramentas para acessar bancos de dados e realizar tarefas importantes para a otimização das análises das seqüências, tornou-se possível com a utilização de programação em linguagem Perl (*Practical Extraction and Reporting Language*), a qual é

uma linguagem muito utilizada em Bioinformática por sua facilidade na manipulação de *strings*, conexão a bancos de dados e acesso via *Web*.

A Perl é gratuita e possui uma diversidade de módulos (bibliotecas) de funções específicas para tratar de diversos assuntos como gráficos, estatísticas e até mesmo um módulo específico de Bioinformática o BioPerl (utilizado neste projeto).

Para o desenvolvimento de aplicações *Web* utiliza-se o conceito de programação CGI (*Common Gateway Interface*). Um programa ou *script* CGI é um aplicativo que reside em um servidor *Web*, quando o programa CGI é chamado por um usuário remoto, o aplicativo é executado no servidor que, em seguida, encaminha ao usuário (cliente) a resposta em formato HTML (Hypertext Transport Protocol) (Figura 12).

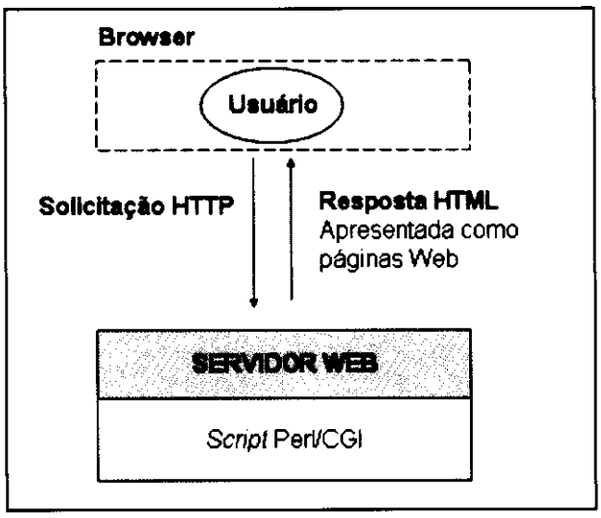


Figura 12. Diagrama esquemático de interação CGI - Usuário.

A Perl possui um módulo para auxílio no desenvolvimento de aplicativos CGI o Perl-CGI.

Além do módulo CGI, três importantes módulos utilizados neste trabalho foram o DB (Módulo para interfaceamento com bancos de dados), o GD (Módulo de funções gráficas) e o BioPerl (Módulo de funções aplicadas a problemas de Bioinformática). Cabe ressaltar que este último possui grande importância no crescimento da utilização da linguagem Perl em projetos de Bioinformática, por ser um grande

repositório de módulos computacionais rotineiramente usados em Bioinformática.

Funções como tradução de seqüências de nucleotídeos em seqüências de aminoácidos, "*parser*" para resultados de programas como Blast, Phred, Phrap, CAP3, ClustalIW entre outros são facilmente implementados com a utilização deste conjunto de ferramentas.

3.5.3 Sistema Gerenciador de Banco de Dados

O banco de dados é uma ferramenta de fundamental importância na Bioinformática, tanto na busca como no armazenamento de informações biológicas.

O Sistema Gerenciador de Banco de Dados (SGBD) utilizado para dar suporte ao SGADBio é o MySQL.

O MySQL é um SGBD relacional que oferece integração com diversas aplicações através de um subconjunto da popular linguagem de consulta SQL.

A escolha do MySQL como SGBD para o SGADBio deve-se ao fato de ele oferecer tempos de acesso pequenos (por não utilizar o conceito de transações), por tratar-se de um sistema gratuito e ter grande disponibilidade de suporte.

IV – SGADBio UM SISTEMA DE GERENCIAMENTO E ANÁLISE DE DADOS POR BIOINFORMÁTICA

Neste capítulo são apresentados os módulos e funções do sistema de gerenciamento e análise de dados por Bioinformática – SGADBio.

Produzido neste trabalho como resultado de um projeto de análise e implementação definida com a utilização de técnicas de Engenharia de Software, o SGADBio é um sistema desenvolvido para análise de dados de projeto de seqüenciamento de genes expressos.

Interfaces do sistema para manutenção e visualização dos dados, serão apresentadas neste capítulo.

4.1 Módulos e funções

O sistema SGADBio consiste de cinco módulos, sendo que o acesso a eles dependerá do tipo de usuário e seus privilégios.

Os módulos estão organizados da seguinte forma:

- ***Administrator*** – módulo responsável pelo gerenciamento dos laboratórios, usuários e seus privilégios de acesso ao sistema;
- ***Project Manager*** – configuração do projeto, definição dos processos e fluxo de dados envolvidos no *pipeline*;
- ***Profile Tools*** – permite a associação de um perfil a cada amostra a ser analisada, informando ao sistema dados como mRNA utilizado, programa de PCR, *primers* e outras informações referente à construção da biblioteca;
- ***Analyser*** – utilizado para submissão de cromatogramas, "clusterização" de dados e visualização de resultados;
- ***Query*** – módulo capaz de gerar relatórios e gráficos que possibilitam total acompanhamento dos projetos analisados.

Informações detalhadas sobre cada um dos módulos e suas principais funções serão descritas no decorrer desta seção.

Inicialmente, o sistema possui uma interface que consiste de uma tela de "*Logon*" onde o usuário deverá fornecer seu nome de usuário e senha, *Username* e *Password* respectivamente (Figura 13).

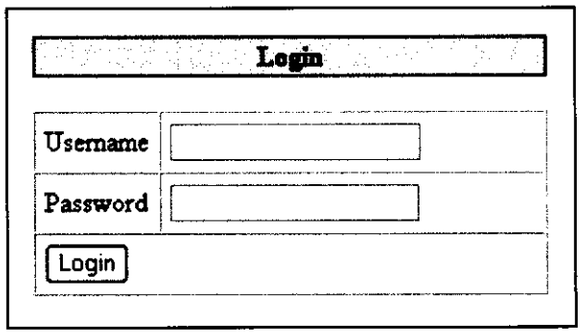


Figura 13. Tela *Logon* do SGADBio.

Logo após o *Login* do sistema, é apresentado um menu principal com diversas opções (Figura 14).

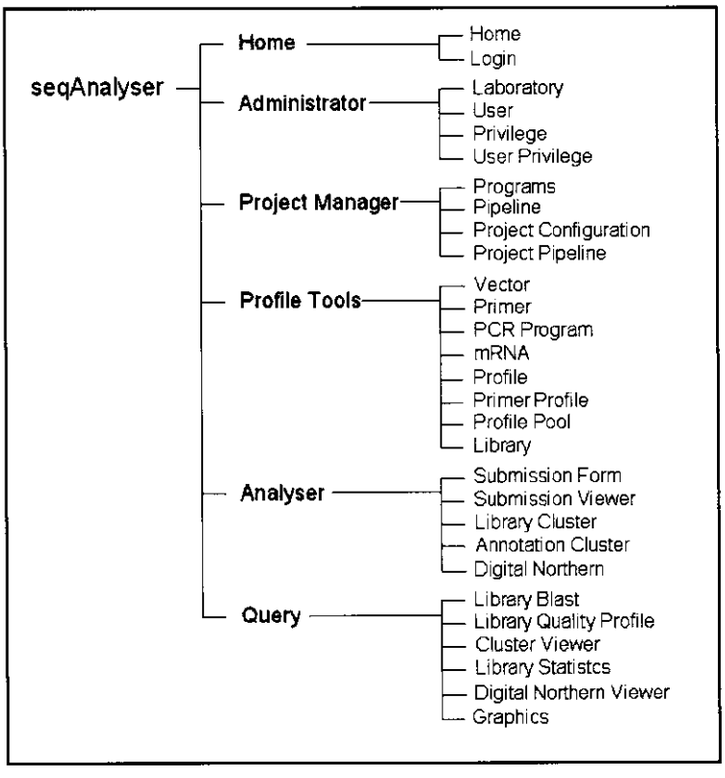


Figura 14. Opções do menu principal do SGADBio.

A seguir, os módulos são apresentados para melhor entendimento de suas funções.

4.1.1 Módulo Administrator

O módulo *Administrator* (Administrador do Sistema) permite o gerenciamento dos laboratórios, usuários e seus privilégios.

A tela de gerenciamento dos dados é comum na maioria das opções do sistema, permitindo a inclusão, alteração, exclusão e consulta dos dados armazenados em seus respectivos depósitos do banco de dados.

O cadastro de laboratórios permite o gerenciamento dos laboratórios que terão acesso aos projetos inclusos no sistema. Cada projeto será vinculado a um determinado laboratório, onde permissões a ele serão dadas de acordo com o tipo de acesso de seus usuários.

Neste módulo, disponibiliza-se ainda um cadastro de usuários e privilégios, opção responsável pelo gerenciamento dos usuários e os tipos de acesso ao sistema. Com isso pode-se permitir, por exemplo, que um pesquisador tenha acesso à submissão de cromatogramas pertencentes a um determinado projeto, porém não possa realizar a mesma tarefa para um outro qualquer.

A Figura 15 ilustra o cadastro de usuários do sistema com suas interfaces de consulta (A), inclusão (B) e alteração (C) de dados.

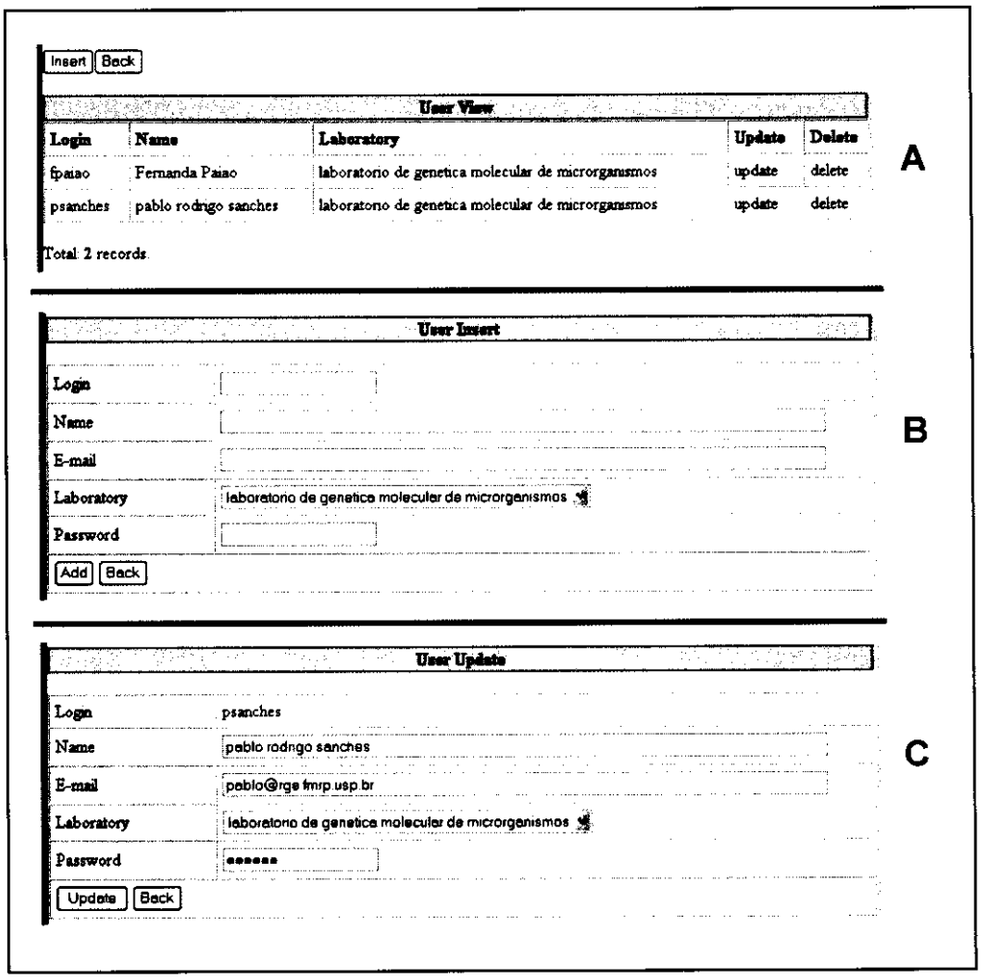


Figura 15. Cadastro de Usuários do SGADBio.

4.1.2 Módulo Project Manager

O módulo *Project Manager* (Gerenciador de Projetos) é responsável pelo gerenciamento dos projetos de análise de seqüências, manipulados através do sistema.

Dentre as várias funções, pode-se destacar o gerenciamento de *pipelines* configuráveis por tipo de projeto. Esta função é de fundamental importância nos projetos de análise de ESTs, porém a grande maioria dos softwares desta categoria possuem seus processos e parâmetros de análise fixos. Neste caso, para qualquer tipo de particularidade em seu

projeto de seqüenciamento, o usuário deverá contar com um especialista da área de software e programação para efetuar as mudanças.

O software desenvolvido neste trabalho possui esta função configurável e de fácil gerenciamento, ou seja, qualquer usuário que tenha permissão de gerenciar um determinado projeto de análise poderá através deste módulo configurar seu tipo de *pipeline* e os programas que serão utilizados na fase de análise das seqüências sem envolver-se com qualquer tipo de linguagem de programação.

O primeiro passo para tornar um *pipeline* configurável é incluir, através da função *programs*, os programas disponíveis para análise de dados por Bioinformática, assim como seus tipos de parâmetros de entrada. Em seguida, na função *pipeline*, vincula-se uma ordem de execução dos processos aos programas disponíveis configurados no sistema, sem esquecer de seus parâmetros de execução. Por último associa-se cada projeto ao tipo de *pipeline* que será utilizado para análise dos dados.

Por tratar-se de um módulo de grande importância no desenvolvimento deste trabalho, serão descritos e exemplificados alguns de seus procedimentos e funções mais detalhadamente.

4.1.2.1 Programs

A opção *Programs, pertencente ao* módulo *Project Manager* é responsável pelo cadastramento dos programas de análise em Bioinformática no sistema, permitindo que os mesmos tornem-se disponíveis na definição do *pipeline.*

Para esta opção, o Gerente de Projetos deverá definir três informações básicas: o nome do programa (*name*), uma descrição sobre o programa (*description*) e por último a sintaxe de execução deste programa (*syntax*) que será utilizada no momento em que o *pipeline* for executado.

A Figura 16 apresenta o formulário de cadastro de programas e a Tabela 1 alguns exemplos de dados que podem ser inseridos através desta opção.

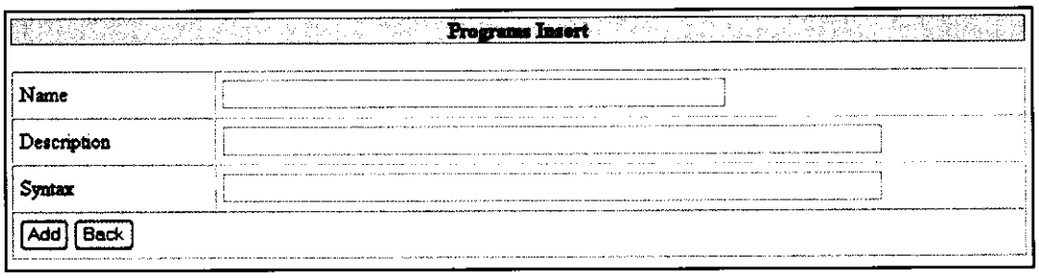


Figura 16. Cadastro de Programas.

Tabela 1. Exemplos de programas que podem ser configurados no SGADBio.

Name	***Description***	***Syntax***
Phred	Phrep	phrep -id /$pasta/chromat_dir/ -pd /$pasta/phd_dir/
Phd2Fasta	Phred to Fasta	phd2fasta -id /$pasta/phd_dir/ -os /$pasta/edit_dir/$seq.fasta -oq /$pasta/edit_dir/$seq.fasta.qual
Cross_Match	Cross_Match	cross_match /$pasta/edit_dir/$seq.fasta /$pasta/vector/$vector.fasta -minmatch <p_minmatch> -minscore <p_minscore> -screen
Cap3	CAP3	cap3 /$pasta/edit_dir/$seq.fasta
BlastN	Blast (DNAxDNA)	blastall -p blastn -d <p_bd> -i /$pasta/edit_dir/$seq.fasta.screen -e <p_evalue>
BlastX	Blast (DNAxProteína)	blastall -p blastx -d <p_bd> -i /$pasta/edit_dir/$seq.fasta.screen -e <p_evalue>

Pode-se observar na Tabela 1 que na coluna *Syntax*, alguns valores estão entre os símbolos de menor (<) e maior (>). O objetivo foi

preparar a sintaxe de execução dos programas, permitindo que o usuário defina, de acordo com o projeto, os valores para estes parâmetros.

Alguns exemplos de parâmetro, listados na tabela, para os programas BlastN e BlastX são o <p_evalue> e <p_bd>. O primeiro define o corte de aceitação do e-value (valor utilizado como estimativa de similaridade entre duas seqüências) dos *Hits* encontrados pelo programa Blast. O segundo permite que o usuário selecione o banco de dados biológico que será utilizado na análise por similaridade.

Os valores para estes parâmetros serão configurados na próxima etapa deste módulo de gerenciamento. Esta etapa se dá na opção *Pipeline*.

4.1.2.2 Pipeline

A opção *Pipeline* permite a configuração do *pipeline* quanto aos passos de execução de cada processo de análise, assim como a definição dos valores para os parâmetros definidos na opção *Programs*.

Nesta opção o Gerente de Projetos tem o papel de configurar tipos de *pipeline*, definindo a ordem nas quais serão executados os diversos tipos de programa de Bioinformática.

Para configurar o *pipeline*, uma interface é apresentada ao usuário pedindo para que sejam informados o programa (cadastrado na opção *Programs*), a ordem de execução e os valores para cada parâmetro, quando for o caso. A Figura 17 ilustra um tipo de configuração de *pipeline* cadastrada no sistema. Esta configuração pode ser utilizada em diversos projetos de análise de seqüenciamento de genes expressos. Um esquema representando como o software converte os dados digitados em um *script* de execução (arquivo em formato texto) está sendo mostrado nesta figura.

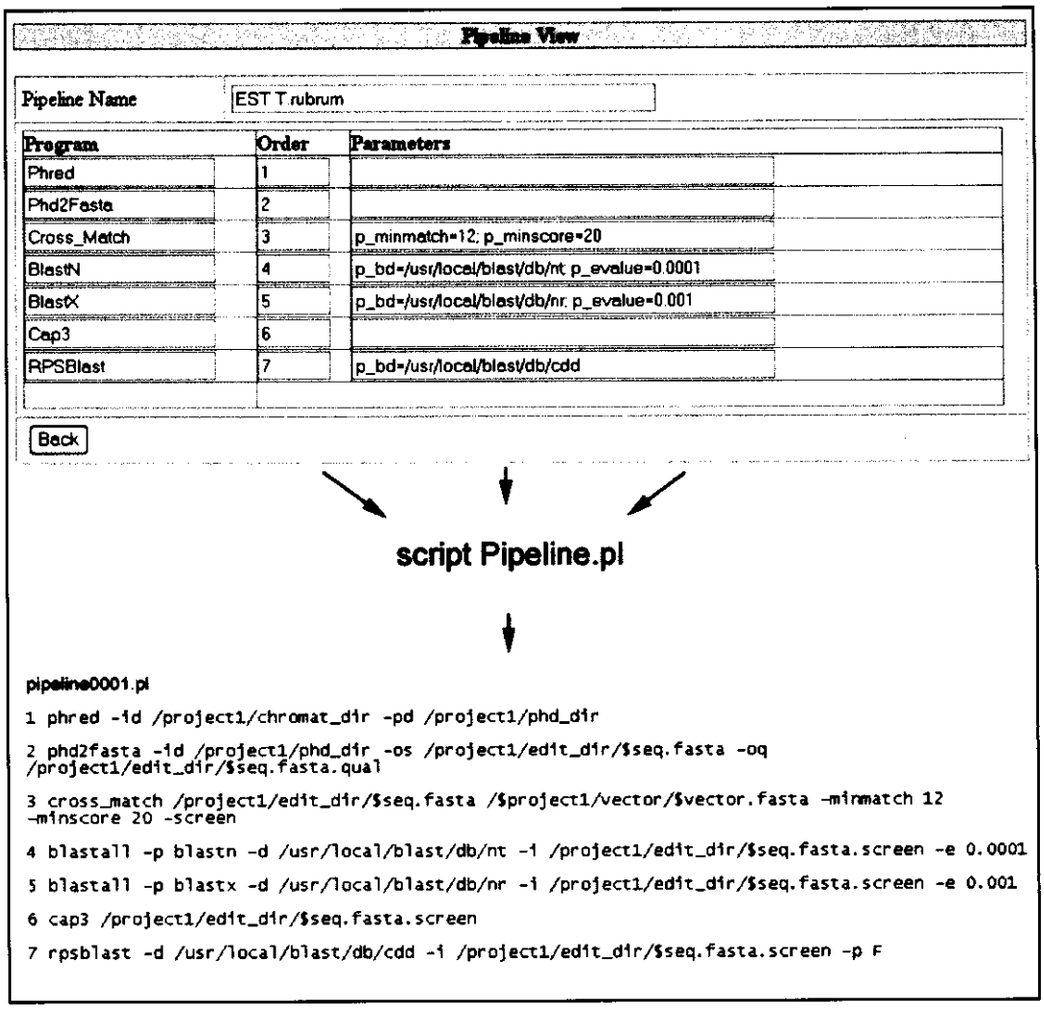


Figura 17. Configuração do *pipeline* e *script* gerado após a interpretação dos dados pelo software.

4.1.2.3 Project Configuration

A opção *Project Configuration* permite que o usuário realize algumas configurações sobre cada projeto de análise quanto aos diretórios de execução e armazenamento de resultados, as especificações do equipamento e o sistema operacional utilizado.

Nesta opção o usuário deverá informar as especificações do microcomputador responsável pelo processamento dos dados, principalmente quanto ao número de processadores deste equipamento, já que o programa Blast, normalmente utilizado como um dos passos na

análise dos projetos de seqüenciamento, envolve um maior custo computacional e pode ser executado de forma paralela, tornando o tempo de análise do projeto mais satisfatório (Bealer, 2004). Deve-se ainda informar nesta opção os diretórios onde alguns programas de análise estão instalados e onde alguns resultados serão armazenados.

4.1.2.4 Project Pipeline

Nesta última opção do módulo, o Gerente de Projetos deverá vincular cada projeto de análise ao tipo de *pipeline* que será utilizado.

Esta estrutura foi criada para permitir que uma mesma configuração ou tipo de *pipeline* possa ser relacionado a um ou vários projetos de análise.

O usuário poderá, por exemplo, configurar as etapas de processamento de um *pipeline* básico com a execução, na seguinte ordem, dos programas, Phred, Phd2fasta, Cross_Match, Cap3 e BlastX (Genbank-NR). Esta mesma configuração poderá ser então utilizada em vários projetos de análise de ESTs, independente do organismo ou condição estudada.

4.1.3 Módulo Profile Tools

O próximo módulo a ser apresentado é o *Profile Tools.* Este é responsável pela manutenção dos dados sobre o perfil de cada projeto de análise, ou seja, através deste módulo o pesquisador irá informar ao sistema dados sobre o projeto, destacando-se informações referente à fase de preparo do seqüenciamento, tipo e condições da biblioteca, localização física do material biológico, etc.

O conjunto completo de opções deste módulo permite o cadastramento da biblioteca, os *primers* utilizados na reação de seqüenciamento, o programa de PCR e o mRNA utilizado. O pesquisador

deverá sempre que for criado um novo projeto, incluir estes dados no sistema. Estas informações serão então vinculadas a cada uma das seqüências biológicas analisadas, facilitando assim uma melhor documentação do projeto de seqüenciamento.

A Figura 18 apresenta uma das telas de cadastro do módulo *Profile Tools* (Cadastro de mRNA).

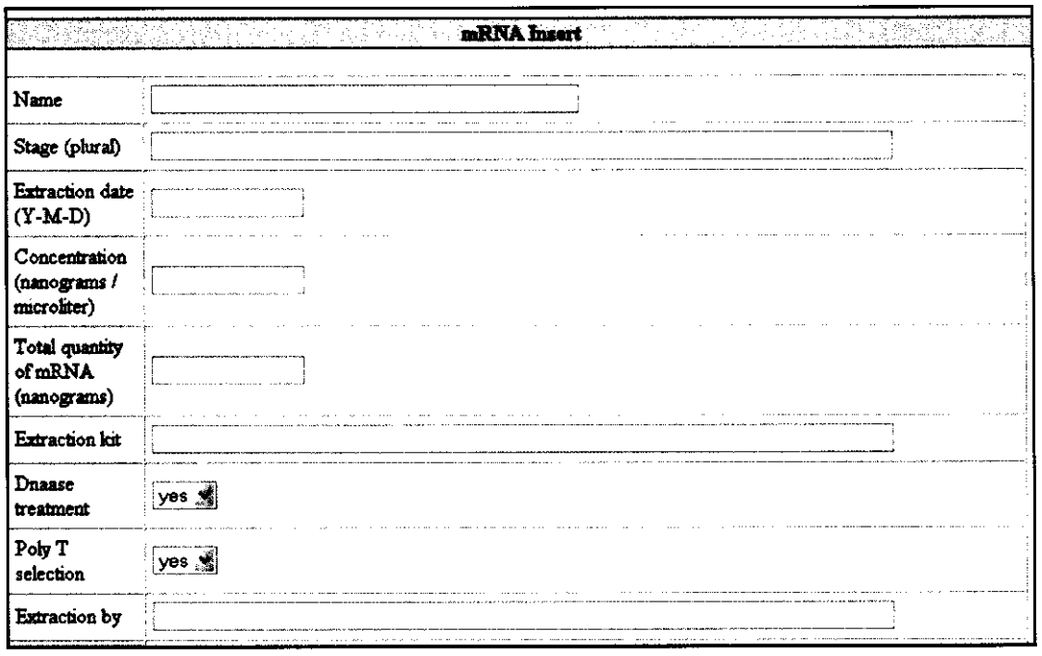


Figura 18. Tela de cadastro de mRNA (módulo *Profile Tools*).

A documentação é parte integrante de qualquer projeto de seqüenciamento. A preservação de alguns dados é de fundamental importância para que os diversos pesquisadores tenham conhecimento da problemática como um todo.

Este tipo de documentação pode auxiliar os pesquisadores, por exemplo, a saber, qual tipo de kit de seqüenciamento foi utilizado, quem extraiu o RNA mensageiro e qual a localização física do material biológico. Estas informações detalhadas podem facilitar o trabalho de outros pesquisadores.

4.1.4 Módulo Analyser

O módulo *Analyser* permite que o pesquisador submeta as seqüências biológicas para análise, agrupe seqüências por similaridade e realize comparações entre duas bibliotecas indicando genes diferencialmente expressos (*Digital Northern*).

Faz parte também deste módulo o *script que agenda a* execução do *pipeline* no serviço de agendamento do sistema linux (cron). As principais funções deste módulo serão detalhadas a seguir.

4.1.4.1 Submission Form

Opção responsável pela submissão dos arquivos (cromatogramas) a serem analisados.

Para esta opção deve-se compactar o conjunto de cromatogramas, normalmente 96 amostras em um arquivo do formato zip (formato originalmente criado por Phil Katz, fundador do PKWARE), em seguida, incluir em um formulário as informações referentes à biblioteca ao qual pertence o arquivo para análise.

O programa responsável pela submissão dos cromatogramas realiza uma série de validações sobre o arquivo enviado. Em outras palavras, esta opção possui formas para verificar se o que foi submetido contém realmente cromatogramas e se o arquivo foi corretamente transmitido.

Assim que submetido, o arquivo é então agendado em uma fila de execução. A Figura 19 apresenta a interface de submissão de cromatogramas para análise.

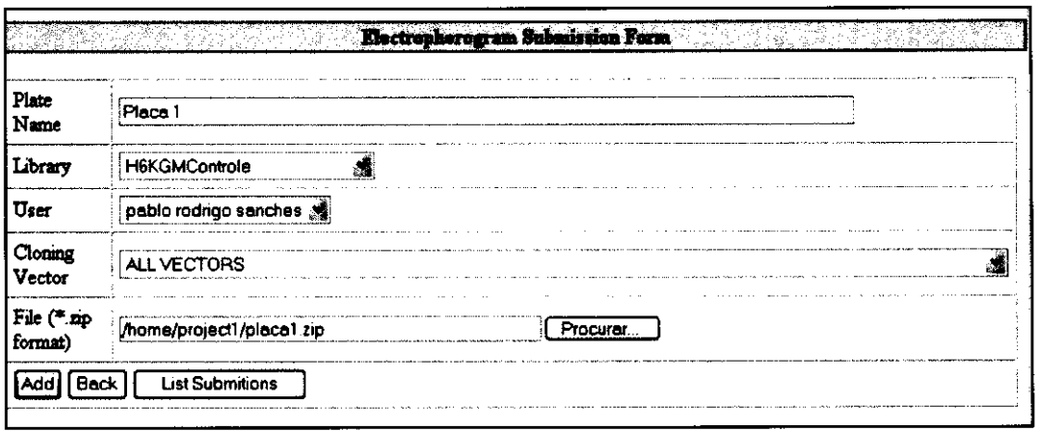


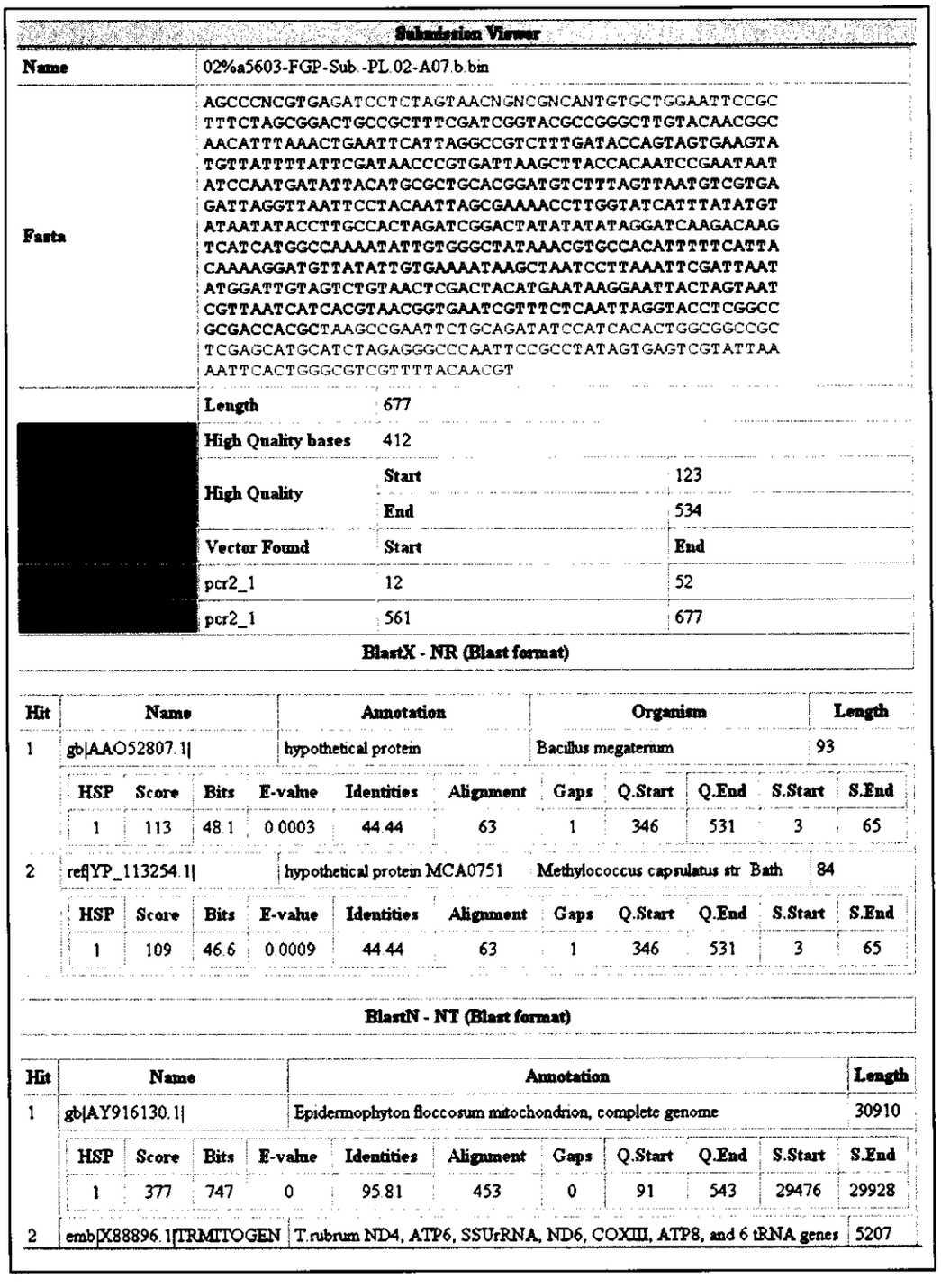
Figura 19. Tela de submissão de cromatogramas para análise.

A fila de execução é analisada por um *script* agendado em um serviço de agendamento de tarefas disponível nos sistemas Linux (cron), cuja responsabilidade é a de verificar o momento em que se pode iniciar o processamento de um novo arquivo. Caso exista um arquivo em processamento, o *script* aguarda um próximo momento para realização da nova análise. Isso evita processamentos excessivos, o que pode tornar o Sistema Operacional do microcomputador menos vulnerável a falhas.

4.1.4.2 Submission Viewer

Opção utilizada para visualização dos dados analisados para um determinado cromatograma. Criada para facilitar a verificação de informações como a seqüência do DNA obtida, a região de alta qualidade da seqüência, a região de contaminantes (vetor e *primer*) e o grau de similaridade entre a seqüência e outros bancos de dados biológicos analisados.

Com esta opção o usuário deve selecionar a biblioteca desejada, e dependendo de suas permissões, ele poderá selecionar a placa e um dos cromatogramas analisados para obter uma visualização geral dos dados (Figura 20).



Submission Viewer

Name	02%a5603-FGP-Sub.-PL.02-A07.b.bin
Fasta	AGCCCNCGTGAGATCCTCTAGTAACNGNCGNCANTGTGCTGGAATTCCGC TTTCTAGCGGACTGCCGCTTTCGATCGGTACGCCGGGCTTGTACAACGGC AACATTTAAACTGAATTCATTAGGCCGTCTTTGATACCAGTAGTGAAGTA TGTTATTTTATTCGATAACCCGTGATTAAGCTTACCACAATCCGAATAAT ATCCAATGATATTACATGCGCTGCACGGATGTCTTTAGTTAATGTCGTGA GATTAGGTTAATTCCTACAATTAGCGAAAACCTTGGTATCATTTATATGT ATAATATACCTTGCCACTAGATCGGACTATATATATAGGATCAAGACAAG TCATCATGGCCAAAATATTGTGGGCTATAAACGTGCCACATTTTTCATTA CAAAAGGATGTTATATTGTGAAAATAAGCTAATCCTTAAATTCGATTAAT ATGGATTGTAGTCTGTAACTCGACTACATGAATAAGGAATTACTAGTAAT CGTTAATCATCACGTAACGGTGAATCGTTTCTCAATTAGGTACCTCGGCC GCGACCACGCTAAGCCGAATTCTGCAGATATCCATCACACTGGCGGCCGC TCGAGCATGCATCTAGAGGGCCCAATTCCGCCTATAGTGAGTCGTATTAA AATTCACTGGGCGTCGTTTTACAACGT

	Length	677
	High Quality bases	412
	High Quality	**Start** 123
		End 534
	Vector Found	**Start** **End**
	pcr2_1	12 52
	pcr2_1	561 677

BlastX - NR (Blast format)

Hit	Name	Annotation	Organism	Length
1	gb|AAO52807.1|	hypothetical protein	Bacillus megaterium	93

HSP	Score	Bits	E-value	Identities	Alignment	Gaps	Q.Start	Q.End	S.Start	S.End
1	113	48.1	0.0003	44.44	63	1	346	531	3	65

Hit	Name	Annotation	Organism	Length
2	ref|YP_113254.1|	hypothetical protein MCA0751	Methylococcus capsulatus str. Bath	84

HSP	Score	Bits	E-value	Identities	Alignment	Gaps	Q.Start	Q.End	S.Start	S.End
1	109	46.6	0.0009	44.44	63	1	346	531	3	65

BlastN - NT (Blast format)

Hit	Name	Annotation	Length
1	gb|AY916130.1|	Epidermophyton floccosum mitochondrion, complete genome	30910

HSP	Score	Bits	E-value	Identities	Alignment	Gaps	Q.Start	Q.End	S.Start	S.End
1	377	747	0	95.81	453	0	91	543	29476	29928

2	emb|X88896.1|TRMITOGEN	T.rubrum ND4, ATP6, SSUrRNA, ND6, COXIII, ATP8, and 6 tRNA genes	5207

Figura 20. Tela de visualização de cromatogramas analisados.

Os *Hits* encontrados através do programa Blast utilizado na comparação da seqüência com os diversos bancos de dados biológicos, são apresentados ao usuário por ordem de similaridade. O alinhamento entre as seqüências também pode ser visto nesta opção (Figura 21).

```
gb|AAO52807.1| hypothetical protein [Bacillus megaterium] >gi|28...    48   3e-04
ref|YP_113254.1| hypothetical protein MCA0751 [Methylococcus cap...    47   9e-04

>gb|AAO52807.1| hypothetical protein [Bacillus megaterium]
 ref|NP_799510.1| hypothetical protein [Bacillus megaterium]
          Length = 93

 Score = 48.1 bits (113), Expect = 3e-04
 Identities = 28/63 (44%), Positives = 35/63 (55%), Gaps = 1/63 (1%)
 Frame = +1

Query: 346 TSHHGQNIVGYKRATFFITKGCYIVKIS*SLNSI-NMDCSL*LDYMNKELLVIVNHHVTV 522
           ++HH    +GY RAT   TKGC   + S S  +I + DC L L YM  E LVI + H  V
Sbjct: 3   SNHHAPYDLGYTRATMDGTKGCKTARSSQSHKTILSSDCRLQLAYMKLESLVIADQHAAV 62

Query: 523 NRF 531
           N F
Sbjct: 63  NTF 65
```

Figura 21. Alinhamento entre duas seqüências de aminoácidos.

No alinhamento demonstrado, o *Query* representa a seqüência de aminoácidos, gerada a partir da tradução da seqüência de nucleotídeos, submetida para análise. O *Subject* a seqüência da proteína similar. Entre as duas seqüências, encontram-se as bases similares e o sinal (+) que indica que os aminoácidos daquela posição não são idênticos, porém fazem parte do mesmo grupo. Como exemplo tem-se os aminoácidos Treonina (T) e Serina (S).

4.1.4.3 Library Cluster

Na primeira etapa de agrupamento dos dados utiliza-se o programa CAP3 para "montagem" das seqüências pertencentes a uma determinada biblioteca, formando *contigs* (seqüências produzidas da sobreposição de diversos fragmentos de seqüências menores) e *singlets* (seqüências que não tiveram sobreposição).

Esta opção se faz necessária para identificação do grau de redundância do seqüenciamento da biblioteca. A redundância é uma medida utilizada para definir o momento onde deve-se interromper a etapa de seqüenciamento do projeto, ou seja, caso os resultados parciais do seqüenciamento obtiverem um grau muito elevado de redundância,

sugere-se que para se obter novas seqüências gênicas um número muito grande de clones devem ser seqüenciados, o que torna o custo do projeto cada vez mais elevado.

A redundância é equivalente à redução informativa com respeito à quantidade de informação que poderia ser transmitida por meio da mesma quantidade de sinais se todas elas fossem escolhidas como igualmente prováveis (informação máxima de uma fonte). A redundância se expressa como:

$$R = \frac{H_0 - H}{H_0}$$

Sendo:

H – informação efetiva de uma fonte;

H_0 – informação máxima;

A interface com o usuário é realizada através de um formulário (Figura 22) onde se pode selecionar a biblioteca a ser "clusterizada", em seguida as seqüências que deverão participar desta etapa. O software identifica ainda as seqüências de má qualidade que podem ou não, a critério do usuário, serem selecionadas para "clusterização".

Cada agrupamento realizado por esta opção é registrado no banco de dados do software com um nome para facilitar sua identificação.

Os agrupamentos são realizados pelo programa CAP3 onde, diferentemente do Phrap, produz-se seqüências consenso menores, porém com menor taxa de erro (Telles and Silva, 2001). Atualmente o programa CAP3 tem grande aceitação em projetos de análise de ESTs. Testes realizados durante a confecção deste trabalho indicam que o programa CAP3 é mais preciso que o Phrap quando trata de seqüências pequenas.

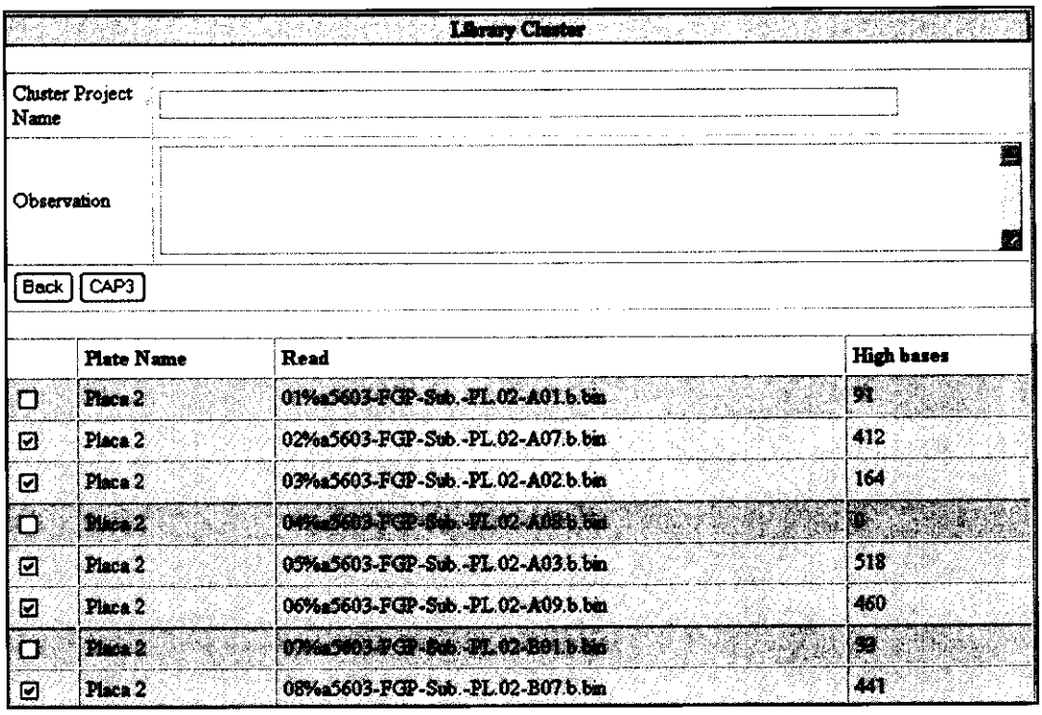


Figura 22. Tela de agrupamento de seqüências – CAP3.

4.1.4.4 Annotation Cluster

Opção também utilizada para agrupamento de seqüências selecionadas de uma ou várias bibliotecas através de palavras-chaves disponíveis em suas anotações.

O usuário pode, através desta opção, comparar várias bibliotecas identificando um mesmo produto gênico com o objetivo de comparar a expressão ou mesmo de procurar uma cobertura mais abrangente do mesmo.

Para esta opção o usuário deverá informar em um campo palavras relacionadas a nomes de proteínas. O software então localiza as seqüências anotadas onde suas proteínas similares satisfaçam os critérios de busca. Tendo-se em mãos as seqüências selecionadas, as mesmas são agrupadas através do programa CAP3.

Este agrupamento pode ser realizado para um único projeto de seqüenciamento ou para vários ao mesmo tempo.

A Figura 23 ilustra um agrupamento realizado por esta opção, onde as palavras-chaves pesquisadas em um determinado projeto de ESTs foram *Hypothetical protein*.

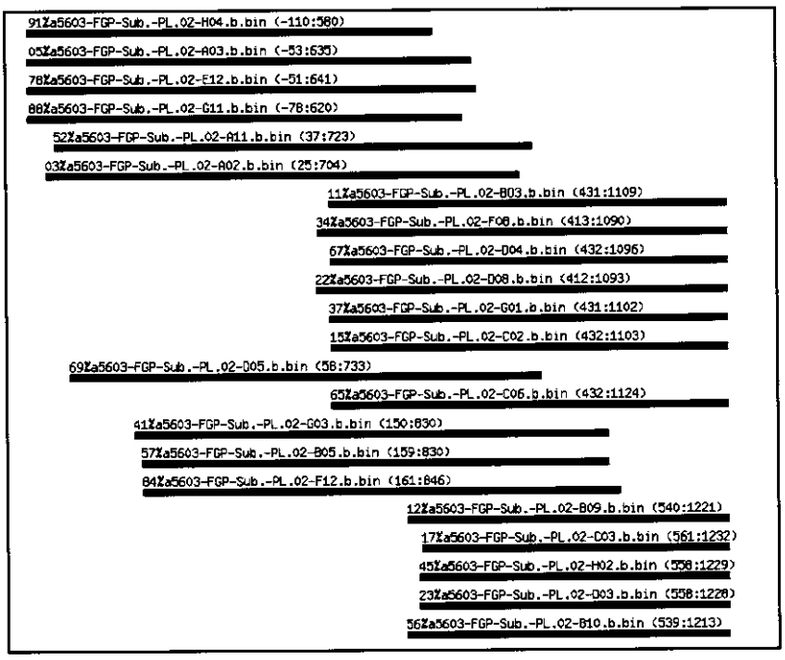


Figura 23. Agrupamento de seqüências anotadas como Proteínas Hipotéticas na análise de uma biblioteca subtrativa de *T. rubrum*.

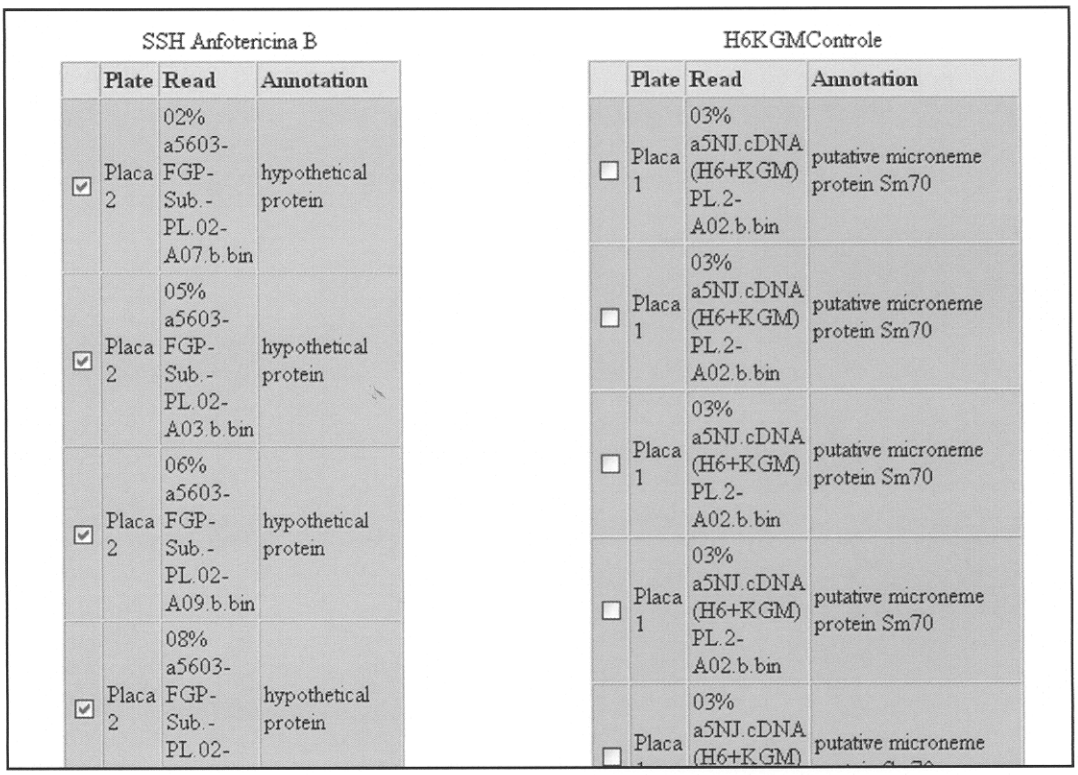
Nesta figura é possível observar que várias seqüências foram agrupadas. Isto auxilia a pesquisa, já que pode indicar através da quantidade de seqüências agrupadas, o nível de expressão deste gene. Outra informação importante é que, através desta montagem dos fragmentos menores, foi possível chegar a uma seqüência de consenso maior.

4.1.4.5 Digital Northern

A opção *Digital Northern* permite selecionar genes com padrão diferente de expressão entre duas bibliotecas com base nos parâmetros de razão e diferença. Estes parâmetros são suficientes para a seleção de genes candidatos a serem validados por métodos laboratoriais.

O método para obtenção deste dado é baseado em um agrupamento por similaridade, realizado também pelo programa CAP3, de um *pool* de seqüências das duas bibliotecas comparadas. Um *script* compara os agrupamentos realizados e identifica o diferencial de expressão gênica entre as bibliotecas.

A Figura 24 ilustra a interface de seleção dos dados para realização do processo de agrupamento para a obtenção do *Northern* Digital.



SSH Anfotericina B

	Plate	Read	Annotation
☑	Placa 2	02% a5603-FGP-Sub.-PL.02-A07.b.bin	hypothetical protein
☑	Placa 2	05% a5603-FGP-Sub.-PL.02-A03.b.bin	hypothetical protein
☑	Placa 2	06% a5603-FGP-Sub.-PL.02-A09.b.bin	hypothetical protein
☑	Placa 2	08% a5603-FGP-Sub.-PL.02-	hypothetical protein

H6KGMControle

	Plate	Read	Annotation
☐	Placa 1	03% a5NJ.cDNA (H6+KGM) PL.2-A02.b.bin	putative microneme protein Sm70
☐	Placa 1	03% a5NJ.cDNA (H6+KGM) PL.2-A02.b.bin	putative microneme protein Sm70
☐	Placa 1	03% a5NJ.cDNA (H6+KGM) PL.2-A02.b.bin	putative microneme protein Sm70
☐	Placa 1	03% a5NJ.cDNA (H6+KGM) PL.2-A02.b.bin	putative microneme protein Sm70
☐	Placa 1	03% a5NJ.cDNA (H6+KGM)	putative microneme

Figura 24. Interface para seleção das bibliotecas a serem comparadas através do Northern Digital.

Uma melhor visualização dos resultados sobre Northern Digital será mostrada no próximo módulo do sistema (Query).

4.1.5 Módulo Query

O módulo *Query* permite a visualização de diversos relatórios e gráficos sobre os resultados das análises realizadas pelo sistema.

Dentre as opções podemos citar:

- ***Library Blast*** – Relatório contendo o primeiro Hit de cada seqüência comparada através do programa Blast com os diversos bancos de dados analisados;
- ***Library Quality Profile*** – Apresenta um perfil sobre a qualidade do seqüenciamento da biblioteca, indicando seqüências de boa e má qualidade, e representando graficamente a qualidade da seqüência e seu cromatograma;
- ***Cluster Viewer*** – Visualizador dos agrupamentos gerados em uma determinada biblioteca. Informações sobre a montagem das seqüências, com a utilização do programa CAP3, podem ser mostradas nesta opção;
- ***Library Statistics*** – Apresentação de alguns dados estatísticos extraídos dos resultados da análise de cada biblioteca;
- ***Digital Northern Viewer*** – Visualizador dos *Northerns* Digitais gerados e armazenados no banco de dados do sistema;
- ***Graphics*** – Disponibiliza diversos gráficos que demonstram o comportamento dos dados analisados.

Nesta seção serão apresentadas figuras e tabelas com exemplos de resultados extraídos através de algumas destas opções.

A Tabela 2 contém um exemplo de relatório sintético sobre resultados de similaridades encontrados entre as seqüências analisadas e os diversos bancos biológicos. Esta tabela foi gerada pela opção *Library Blast* do software.

Tabela 2. Alguns dados obtidos através da opção *Library Blast*.

Read	*Annotation*	*Organism*	*Leng.*	*E-value*	*Ident.%*
01%a5603 -FGP- Sub.- PL.02- A01.b.bin	Epidermophyton floccosum mitochondrion, complete genome		30910	6e-41	92.42
02%a5603 -FGP- Sub.- PL.02- A07.b.bin	hypothetical protein	Bacillus megaterium	93	0.0003	44.44
02%a5603 -FGP- Sub.- PL.02- A07.b.bin	Epidermophyton floccosum mitochondrion, complete genome		30910	0	95.81
03%a5603 -FGP- Sub.- PL.02- A02.b.bin	Epidermophyton floccosum mitochondrion, complete genome		30910	2e-47	94.66
05%a5603 -FGP- Sub.- PL.02- A03.b.bin	hypothetical protein	Bacillus megaterium	93	0.0002	44.44
05%a5603 -FGP- Sub.- PL.02- A03.b.bin	Epidermophyton floccosum mitochondrion, complete genome		30910	0	97.18
06%a5603 -FGP- Sub.- PL.02- A09.b.bin	hypothetical protein	Bacillus megaterium	93	0.0003	42.86
06%a5603 -FGP- Sub.- PL.02- A09.b.bin	Epidermophyton floccosum mitochondrion, complete genome		30910	0	98.68

* As cores da coluna *Read* representam a opção de análise por mais de um banco de dados biológico.

Na Tabela 3, o objetivo principal é informar ao usuário dados referentes à qualidade do seqüênciamento. Alguns dados como tamanho da seqüência, região e tamanho do inserto, qualidade das bases, região

de alta qualidade, vetor e identificação se a seqüência está completa ou não, são mostradas pela opção *Library Quality Profile.*

Tabela 3. Dados sobre qualidade do seqüenciamento.

Read	*Bases*	*Insert*	*Size*	*Full*	*Qual >= 20*	*Qual >= 30*	*HQ*	*HQ total*	*Vector*
01%a5603-FGP-Sub.-PL.02-A01.b.bin	681	48-209	162	Y	68	20	104-194	91	12-47; 210-588;
02%a5603-FGP-Sub.-PL.02-A07.b.bin	677	53-560	508	Y	246	91	123-534	412	12-52; 561-677;
03%a5603-FGP-Sub.-PL.02-A02.b.bin	669	52-215	164	Y	223	75	33-399	164	11-51; 216-669;
04%a5603-FGP-Sub.-PL.02-A08.b.bin	802	1-459	459	N	34	2	549-568	0	460-801;
05%a5603-FGP-Sub.-PL.02-A03.b.bin	667	1-595	595	N	394	199	53-570	518	596-667;
06%a5603-FGP-Sub.-PL.02-A09.b.bin	676	71-560	490	Y	398	200	53-530	460	45-70; 561-676;

* HQ = High Quality → Região de alta qualidade da seqüência.

Nesta opção ainda, para cada seqüência existe um *link*, cujo objetivo é apresentar ao usuário uma tela com a visualização detalhada da seqüência, seus valores de qualidade atribuídos em um gráfico de linhas e seu cromatograma (Figura 25).

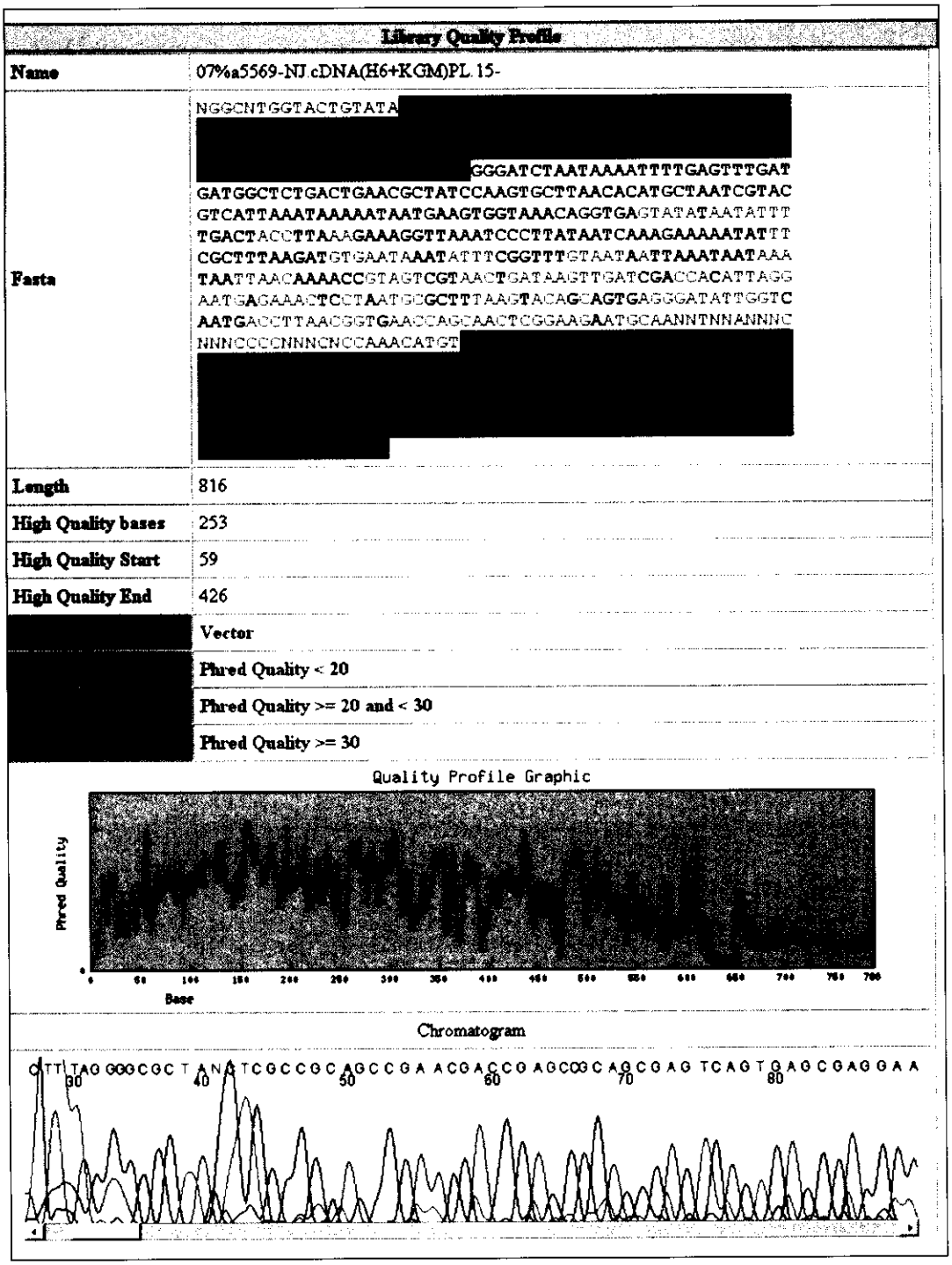
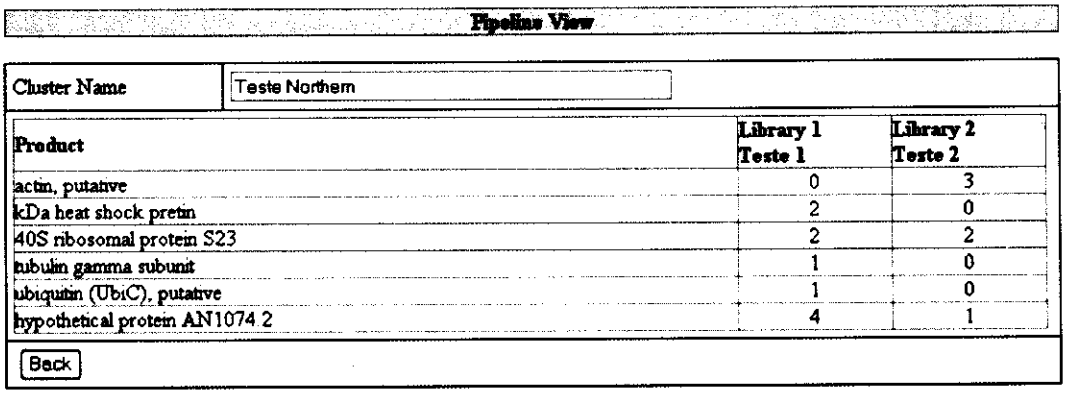


Figura 25. Visualização do perfil de qualidade de uma seqüência analisada pelo software.

Uma importante opção do módulo é a *Digital Northern Viewer*. Ela permite visualizar genes com padrão diferente de expressão entre duas bibliotecas com base nos parâmetros de razão e diferença. Uma interface com o produto gênico e suas quantidades expressas é apresentada ao

usuário para que possa realizar a comparação de duas bibliotecas (Figura 26).



Pipeline View

Cluster Name	Teste Northern	
Product	**Library 1 Teste 1**	**Library 2 Teste 2**
actin, putative	0	3
kDa heat shock protein	2	0
40S ribosomal protein S23	2	2
tubulin gamma subunit	1	0
ubiquitin (UbiC), putative	1	0
hypothetical protein AN1074.2	4	1

Back

Figura 26. Exemplo de Northern Digital gerado com dados teste.

Outra opção pertencente ao módulo *Query* é a *Library Statistics*. Esta informa através dos dados da Figura 27 algumas estatísticas sobre cada biblioteca analisada pelo software.

Library Statistics

Total Reads	95
Acceptable Reads	72 (75.79%)
Average Read Size	412.0972
Full Insert Reads	58 (80.56%)
Average Full Insert Reads Size	381.9828
Number of genes with similarity BlastX-NR	43 (59.72%)
Insignificant similarity (10E-3 < E) BlastX-NR	29 (40.28%)
Number of genes with similarity BlastN-NT	69 (95.83%)
Insignificant similarity (10E-3 < E) BlastN-NT	3 (4.17%)

Library Cluster

Cluster Name	Agrupa_Anfo_High_bases_maior_100bases
Observation	Agrupamento das sequencias com regiao de alta qualidade (menos vetor) maior de 100 bases.

Analized Reads	Contigs	Singlets	Unigenes	Redundancy
72	3	6	9	87.50%

Figura 27. Exemplo de resultados obtidos através da opção *Library Statistics* (módulo *Query*).

Cabe ressaltar que, por se tratar de um software integrado a um banco de dados, todas as opções de relatórios e gráficos são dinâmicas, ou seja, basta algum tipo de alteração em um parâmetro ou inclusão de qualquer seqüência na análise da biblioteca, que estas várias opções terão refletidas as novas informações.

Por último, mas não menos importante, está a opção *Graphics*. Esta opção permite a apresentação de diversos gráficos sobre os dados analisados para uma ou várias bibliotecas. Os gráficos são importantes ferramentas que facilitam a familiarização com as informações.

Dentre as informações geradas através desta opção, destacam-se na Figura 28, a quantidade de similaridades atingidas por tipo de programa e banco de dados (A), o número de seqüências com insertos completos (vetor em ambos os lados da seqüência) por biblioteca (B), o número de seqüências por organismo similar (C), o número de seqüências de boa e má qualidade por biblioteca (D), o tamanho médio das seqüências por biblioteca (E), além de outras.

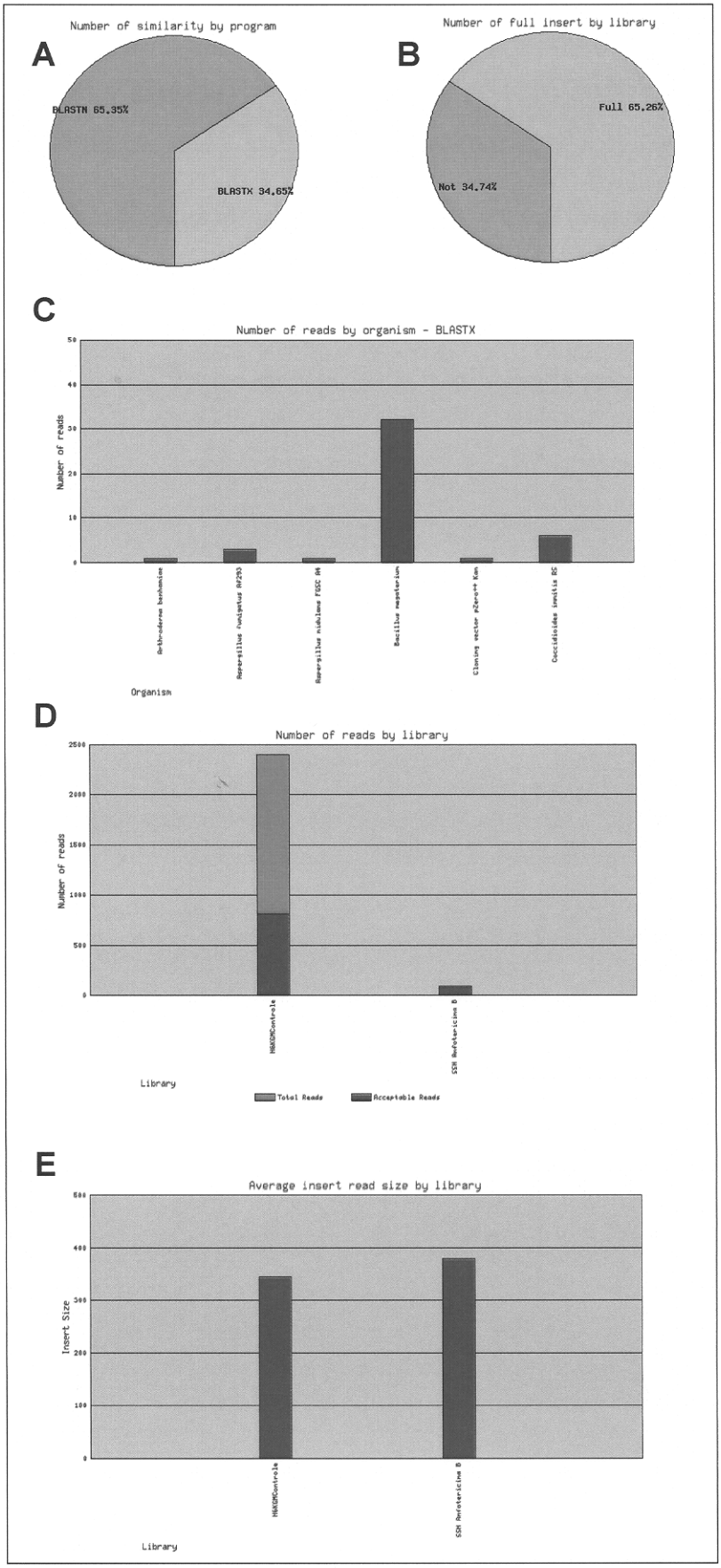


Figura 28. Alguns gráficos gerados pela opção *Graphics*.

Uma outra opção para gerar gráficos e outros dados estatísticos é exportar, em formato texto, os dados analisados pelo software para que outras ferramentas estatísticas possam utilizá-los. Ferramentas como o Microsoft Excel ou a linguagem de programação R podem importar os dados facilmente.

V – ESTUDO DE CASO: ANÁLISE DO TRANSCRIPTOMA DO FUNGO *TRICHOPHYTON RUBRUM*

Este capítulo apresenta um estudo de caso ao qual foi submetido o sistema SGADBio relativamente a um projeto de análise de bibliotecas de ESTs representativas da expressão gênica de *Trichophyton rubrum* durante sua interação *in vitro* com queratinócitos. Neste estudo de caso serão apresentados os dados referentes a biblioteca de ESTs obtidas na condição controle (Linhagem de *T. rubrum* crescida em meio para cultura de queratinócitos mas na ausência dessas células).

Este projeto, objeto do estudo de caso desta dissertação, é parte integrante do projeto de Doutorado da aluna Nalu Teixeira de Aguiar Peres do Laboratório de Genética e Biologia Molecular de Fungos do Departamento de Genética da Faculdade de Medicina de Ribeirão Preto - USP coordenado pela Profa. Dra. Nilce Maria Martinez Rossi.

Serão abordados neste capítulo aspectos biológicos que justificam este projeto, a aplicação do software SGADBio e seu desempenho, além de ilustrar alguns dos resultados obtidos.

5.1 Contexto Biológico

Existem aproximadamente 50.000 espécies de fungos descritas, das quais 300 correspondem a fungos patogênicos (Odom, 1993). Dentre os patogênicos há cerca de 30 espécies de dermatófitos identificados em relação ao seu habitat como antropofílicos, zoofílicos e geofílicos e classificados em três gêneros anamórficos (assexuado ou imperfeito) que são os gêneros *Epidermophyton, Microsporum* e *Trichophyton* (Emmons, 1934).

Os dermatófitos causam os principais tipos de micoses superficiais que são denominadas dermatofitoses. Esses fungos são especializados

em infectar substratos queratinizados como pele, unha e cabelo sendo capazes de utilizar a queratina, uma proteína insolúvel presente na camada córnea da epiderme, como fonte de carbono e energia. Usualmente não invadem outros tecidos, entretanto em indivíduos imunodeprimidos pode haver a evolução para uma micose profunda (Richardson and Warnock, 1993; Martinez-Rossi et al., 2004).

Dentre as espécies de dermatófitos, *Trichophyton rubrum* é a mais comumente encontrada e associada como causa de lesões superficiais de pele e unha, causando tínea corporis, tínea cruris, tínea pedis e tínea unguium também conhecida por onicomicose, raramente infecta cabelo causando tínea capitis (Costa et al., 2002; Vella Zahra et al., 2003; Foster et al., 2004).

Um dos importantes mecanismos de defesa da epiderme contra microrganismos é o processo de queratinização que envolve a renovação de parte da camada córnea da epiderme. Essa renovação é realizada pelos queratinócitos e tem como consequência a descamação da camada mais superficial da pele, o que pode levar a remoção dos microrganismos que acometem esses sítios. Os queratinócitos são as células mais numerosas na epiderme, participam do processo de queratinização, têm um importante papel estrutural formando uma barreira física contra microrganismos e também medeiam a resposta imune.

Atualmente, o aumento significativo do número de doenças fúngicas tem causado grande preocupação aos órgãos de saúde. Estima-se que aproximadamente 80-93% das dermatofitoses crônicas ou recorrentes são causadas por *T. rubrum* (Weitzman and Summerbell, 1995) e este pode ter comportamento invasivo causando infecções oportunistas atípicas em pacientes com o sistema imune deprimido, como aqueles com AIDS, transplantados e outros que estejam sendo submetidos a tratamentos quimioterápicos, por exemplo (Sturtevant, 2000; Liang and Pardee, 2003). O uso intenso e inapropriado de antimicrobianos também tem contribuído para esse quadro de aumento das infecções fúngicas (Masia Canuto and Gutierrez Rodero, 2002).

Estudos de expressão gênica têm levado a um maior entendimento da biologia e patogênese de diversos microorganismos. A construção de bibliotecas de ESTs têm sido uma importante ferramenta utilizada em estudos de expressão gênica (Felipe et al., 2003; Felipe et al., 2005; Ribichich et al., 2005). Além dessa metodologia, outras têm sido utilizadas, como por exemplo, as técnicas de microarranjos (microarrays), DDRT-PCR *(Differential Display Reverse Transcription-Polymerase Chain Reaction)*, SAGE (*Serial Analysis of Gene Expression*), hibridação subtrativa e biblioteca subtrativa seguida de PCR supressivo (Martin and Pardee, 2000; Sturtevant, 2000; Rocha et al., 2002; Hwang et al., 2003).

O estudo do cariótipo e da expressão gênica abriu o campo à novas perspectivas para um melhor entendimento de diversos processos biológicos de *Trichophyton rubrum* (Pereira et al., 1998; Fachin et al., 1999; Fachin et al., 2001; Cervelatti et al., 2004). Projetos como este visam responder quais os genes de *T. rubrum* que são expressos na interação fungo-queratinócitos e quais são aqueles diferencialmente expressos nessa situação.

A investigação das bases moleculares envolvidas neste tipo de interação pode contribuir diretamente para um maior entendimento da interação *T. rubrum* e hospedeiro, bem como, fornecer informações que indiretamente venham a ajudar no desenvolvimento de novas estratégias terapêuticas ou mesmo revelar possíveis alvos terapêuticos. Ademais, bibliotecas de ESTs produzidas neste trabalho podem servir como uma importante base de dados a ser consultada por outros pesquisadores envolvidos na investigação da biologia e patogênese de *T. rubrum*.

5.2 Construção e Seqüenciamento das Bibliotecas de ESTs

Por se tratar de parte de um estudo de caso da aplicação de um software na etapa de análise de dados, as etapas de construção e seqüenciamento da biblioteca de ESTs terão uma abordagem básica,

ficando a fase de análise dos dados como objetivo principal desta dissertação.

A construção da biblioteca de ESTs do fungo em condição controle foi realizada seguindo parâmetros já estabelecidos em kits de desenvolvimento deste tipo de biblioteca. Conhecimentos anteriormente adquiridos por pesquisadores da área de biologia molecular foram empregados nesta fase.

Com relação ao seqüenciamento, o mesmo foi realizado em um seqüenciador automático ABI Prism 377 (Perkin Elmer) seguindo os protocolos recomendados pela fabricante do kit de seqüenciamento.

Um total de 2.400 clones (25 placas com 96 amostras cada uma) dessa biblioteca de cDNA controle foram seqüenciados, fornecendo um perfil da expressão gênica de *T. rubrum* cultivado na ausência de células humanas.

5.3 Análise das ESTs de T. rubrum

A análise das seqüências de ESTs obtidas na construção da biblioteca foi realizada através do software SGADBio (desenvolvido neste trabalho).

O software foi instalado em um microcomputador com um processador Pentium IV 2.8GHz, 512MB de memória RAM e HD de 160GB. O sistema operacional utilizado foi o Linux Fedora Core 3. Foram instalados e configurados o servidor *Web* Apache 2.0, a linguagem de programação Perl 5.8.5, os módulos CGI, DB e Bioperl, a biblioteca gráfica GD e o sistema gerenciador de banco de dados Mysql 3.23.

Através de outros microcomputadores instalados no laboratório o software foi utilizado. Sistemas Operacionais Windows, Macintosh e Linux com navegadores (*browsers*) Internet Explorer e Mozilla serviram de interface para o acesso ao sistema.

O primeiro passo para a análise dos dados foi criar um tipo de *pipeline*, definindo os programas, os bancos de dados biológicos e os

parâmetros que usaríamos para a análise dos dados. Nesta etapa vários trabalhos de análise deste tipo de biblioteca foram pesquisados. Vários testes foram realizados com diversas opções para os parâmetros de entrada dos programas de Bioinformática.

Através do módulo *Project Manager* e suas opções (*Programs*, *Pipeline*, *Project Configuration* e *Project Pipeline*) o gerente responsável pelo projeto de análise das ESTs de *T. rubrum* definiu um *pipeline* que foi configurado como seguem os dados da Tabela 4.

Tabela 4. Configuração do *pipeline* para análise das ESTs de *T. rubrum* no gerenciador de projetos do SGADBio.

Programa	***Parâmetros de entrada***	***Ordem execução***
Phred		1
Phd2Fasta		2
Cross_Match	minmatch=12; minscore=18	3
QualityFilter		4
BlastN	bd=/usr/local/blast/db/nt; evalue=0.001	5
BlastN	bd=/usr/local/blast/db/dbest; evalue=0.001	6
BlastX	bd=/usr/local/blast/db/nr; evalue=0.001	7
CorrectFrame		8
Transeq		9
RPS-Blast	bd=/usr/local/rpsblast/db/cdd	10
CAP3		11
Blast2GO		12

Na Tabela 4 observa-se que o *pipeline* possui 12 processos. Alguns são processos cujo objetivo é a execução de programas públicos de Bioinformática. Outros são *scripts* em linguagem perl criados no decorrer do desenvolvimento deste trabalho.

Segue um breve detalhamento sobre os processos e alguns de seus parâmetros de entrada:

1. **Phred** – Lê o arquivo gerado pelo seqüenciador e realiza o *base-calling*[3] e a atribuição de qualidades das bases, gerando como resultado um arquivo no formato phd[4];
2. **Phd2Fasta** – Processa o arquivo phd e converte as seqüências e os valores de qualidade para o formato FASTA;
3. **Cross_Match** – Realiza o alinhamento da seqüência em formato FASTA com o banco de vetores, mascarando (substituindo por X) eventuais bases que possam ser parte de um vetor;
4. **QualityFilter** – *Script* desenvolvido neste trabalho como filtro para identificação da seqüência de boa qualidade baseada nos dados extraídos dos programas Phred e Cross_Match. O filtro definido para as seqüências de *T. rubrum* foi selecionar apenas seqüências com mais de 50 bases não vetor, com qualidade Phred acima de 20;
5. **BlastN (*nt*)** – Partindo da seqüência de boa qualidade, é realizado um BlastN contra o banco de dados de nucleotídeos do Genbank-NCBI com o objetivo de encontrar similaridades com seqüências de genes depositados. O valor escolhido do e-value para definir se duas seqüências podem ser similares foi 10e-3;
6. **BlastN (*dbest*)** – Com a mesma seqüência faz-se então a comparação com o banco de ESTs do NCBI a procura de ESTs similares. O valor para o e-value também foi 10e-3;
7. **BlastX (*nr*)** – A seqüência é comparada através do programa BlastX com o banco não redundante de proteínas do Genbank-NCBI. Esta comparação é feita com a tradução da seqüência de nucleotídeos nas seis *frames* de leitura. O e-value utilizado foi 10e-3;

[3] Arquivo de cromatogramas é processado e transformado em uma seqüência de bases.
[4] Formato padrão de resultados do programa Phred

8. **CorrectFrame** – *Script* desenvolvido neste trabalho, tem o objetivo de detectar com os dados extraídos do relatório do resultado do BlastX, a frame correta de leitura para tradução da seqüência de nucleotídeos em aminoácidos;
9. **Transeq** – Utilitário do pacote EMBOSS que informada a seqüência de nucleotídeo e a frame de leitura, o programa gera então a seqüência de aminoácidos traduzida;
10. **RPS-Blast (*cdd*)** – É feita a procura de domínios conservados através da comparação da seqüência de aminoácidos com o banco cdd através do programa RPS-Blast;
11. **CAP3** – A seqüência de nucleotídeos é comparada com todas as outras seqüências pertencentes à mesma biblioteca com o objetivo de identificar a redundância;
12. **Blast2GO** – Através do relatório gerado pelo BlastX, o programa relaciona vários bancos de dados cujo resultado é a classificação via Gene Ontology (GO) da seqüência analisada.

Assim que definida a configuração do *pipeline*, os dados com o perfil da biblioteca sobre sua condição de preparo, kit de seqüenciamento utilizado, programa de PCR e outras informações, foram inseridos no sistema (módulo *Profile Tools*). À biblioteca foi dada o nome de H6KGMControle, que descreve uma biblioteca controle de uma linhagem de *T. rubrum* (H6) cultivada em meio KGM (meio ideal para cultura de queratinócitos).

Os cromatogramas foram então, separados por placa de seqüenciamento, submetidos ao sistema.

Após a realização do processamento das seqüências, 1.131 foram identificadas como de má qualidade ou muito pequenas (menor que 50 bases). Das 1.269 seqüências aceitas, o que corresponde a 52,9% do número inicial de seqüências analisadas, o tamanho médio obtido foi de 353 nucleotídeos por seqüência, eliminando sua região contaminante.

Um resumo da análise extraída pela opção *Library Statistics* do módulo *Query* é apresentado na Tabela 5.

Tabela 5. Características gerais das seqüências expressas de *T. rubrum* obtidas na condição controle analisadas pelo software.

Número total de seqüências	2400
Seqüências aceitas	1269 (52,87%)
Tamanho médio das seqüências (nucleotídeos)	353,33
Número de seqüências inteiras	702 (55,32%)
Tamanho médio das seqüências inteiras (nucleotídeos)	187,08
Numero de seqüências com similaridade BlastN-NT	545 (42,95%)
Numero de seqüências com similaridade insignificante (10e-3 < E) BlastN-NT	724 (57,05%)
Numero de seqüências com similaridade BlastN-DBEST	733 (57,76%)
Numero de seqüências com similaridade insignificante (10e-3 < E) BlastN-DBEST	536 (42,24%)
Numero de seqüências com similaridade BlastX-NR	413 (32,55%)
Numero de seqüências com similaridade insignificante (10e-3 < E) BlastX-NR	856 (67,45%)
Número de *contigs*	94
Número de *singlets*	677
Número de unigenes	771
Redundância	39,20%

Como o objetivo desta dissertação não é apresentar o projeto de análise das ESTs de *T. rubrum* e sim demonstrar alguns benefícios que o software desenvolvido pôde trazer para pesquisas desta área, não serão discutidos estes dados em particular, e sim, apenas apresentados como meio de visualização dos recursos disponíveis na ferramenta e benefícios trazidos aos pesquisadores.

Os dados mostrados na Tabela 5 são de extrema importância em projetos de seqüenciamento, pois demonstram uma visão geral sobre a biblioteca em estudo.

Outros recursos do software utilizados na análise das ESTs do fungo são apresentados a seguir.

Tabela 6. Algumas seqüências dos clones obtidos a partir da biblioteca de cDNA das seqüências expressas na condição controle e seus resultados obtidos pela etapa BlastX (NR) do *pipeline*.

Identificação putativa pelo BLASTX	Código do Genbank	Organismo	*E-value*	Identidade
12 kDa heat shock protein	gb|EAL89085.1|	Aspergillus fumigatus Af293	5,00E-22	65,06%
14-3-3-like protein 2	gb|AAR24348.1|	Paracoccidioides brasiliensis	4,00E-81	86,03%
40S ribosomal protein S23	ref|XP_658949.1|	Aspergillus nidulans FGSC A4	7,00E-20	94,00%
60s ribosomal protein L24, putative	gb|EAL85795.1|	Aspergillus fumigatus Af293	2,00E-52	69,18%
60S ribosomal protein L3	gb|AAF15600.1|	Emericella nidulans	2,00E-24	85,48%
60S ribosomal protein L44 (60S ribosomal protein L41)	sp|P52809|RL44	Pichia jadinii	8,00E-52	90,57%
60S ribosomal protein l5, putative	emb|CAF32004.1|	Aspergillus fumigatus	3,00E-46	79,65%

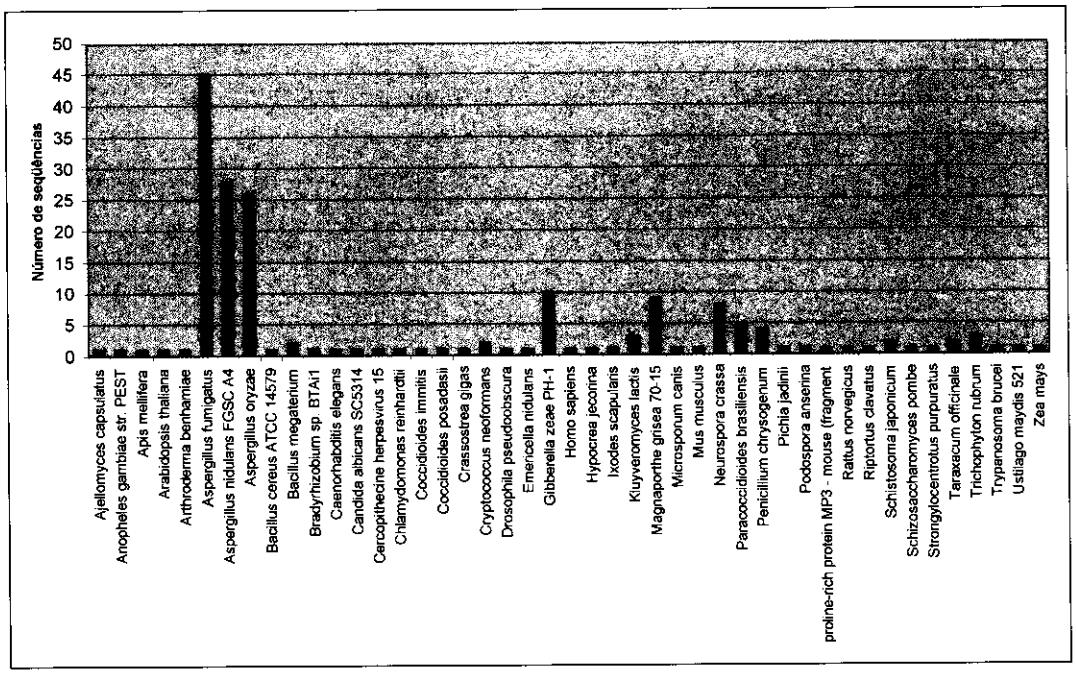


Figura 29. Representação esquemática que relaciona o número de seqüências a seus organismos correspondentes obtidos após a análise dos dados contra o banco GenBank.

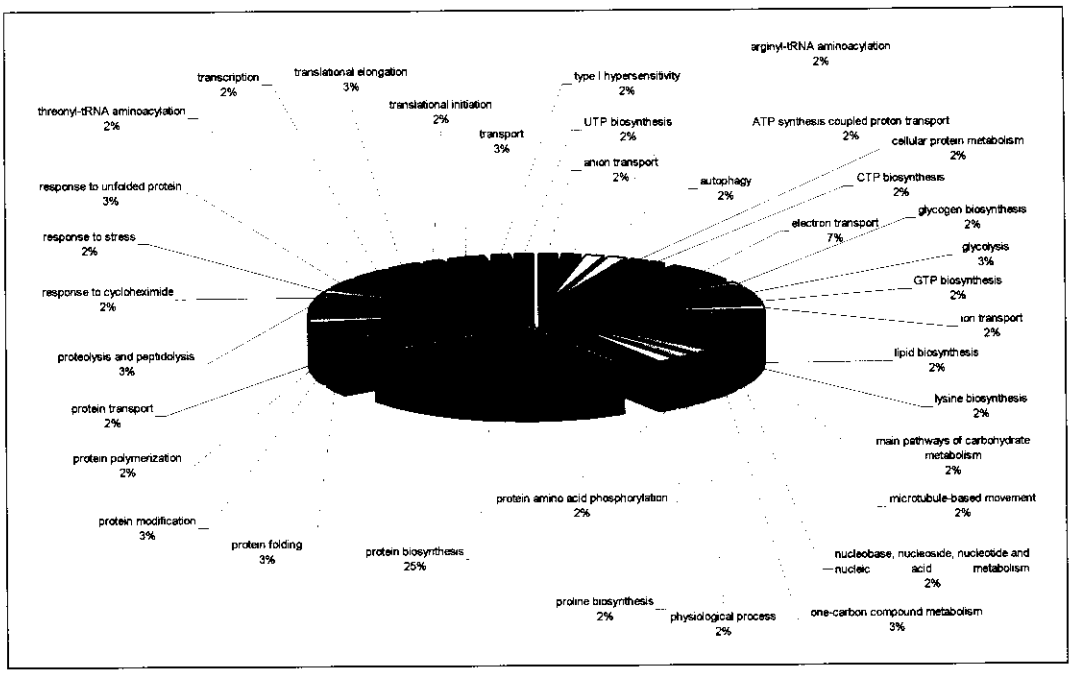


Figura 30. Representação esquemática da análise das sequências quanto aos processos biológicos catalogados pelo consorcio Gene Ontology.

Todos estes recursos foram disponíveis devido à configuração realizada pare este *pipeline* em específico. Programas como BlastX e Blast2GO fizeram parte dos processos envolvidos no *pipeline*. Isso mostra mais uma vez a importância dos Sistemas de Gerenciamento e Análise de Dados por Bioinformática, pois através deste tipo de sistema, foi possível configurar etapas específicas para este projeto de análise de *T. rubrum* e que facilmente poderiam ser reconfiguradas para análise de outros organismos ou projetos.

5.4 Desempenho do Software SGADBio

O desempenho do software, medido pela quantidade de atividades executadas em um determinado intervalo de tempo, atividades que neste caso se trataram da análise de uma determinada seqüência, depende do poder de processamento do equipamento responsável pela análise dos dados.

Neste estudo de caso, o equipamento utilizado foi um microcomputador de arquitetura simples com apenas um processador.

A análise de uma seqüência levou em média 11 minutos para realização dos processos descritos no *pipeline* deste estudo. Sabendo-se que cada placa contem 96 amostras, um total de 17,6 horas foi consumido na análise de cada placa. Como o projeto contou com 25 placas, um total de 440 horas de processamento ininterrupto foi realizado.

Por estes números, dá para se ter uma idéia que projetos desta grandeza dependem de um grande poder de processamento.

Nos testes realizados, detectou-se que o programa Blast foi o grande responsável por este tempo de processamento, haja visto que o mesmo foi configurado para ser executado com apenas um único processador. Bealer (2004) demonstra que sistemas como o existente no NCBI, onde são realizadas mais de 100.000 pesquisas por dia, somente podem ser possíveis com a utilização de vários processadores. Em 2004 o NCBI já contava com 280 processadores.

Um teste realizado em um equipamento com 2 processadores Xeon 3.2GHz e 2GB de memória RAM demonstrou que tempos inferiores a 3 minutos podem ser alcançados com o processamento de uma seqüência no *pipeline* definido para o projeto das ESTs de *T. rubrum*.

VI – CONCLUSÃO

6.1 Contribuição

Com o crescente volume de dados e processos em Bioinformática, a descoberta de novas informações biológicas torna-se cada vez mais fácil. Porém, com esta diversidade de formas de análise disponíveis, a demanda por sistemas que auxiliem o trabalho dos pesquisadores é cada vez maior, em especial aos sistemas que auxiliam a combinação dos diversos dados e processos formando o que chamamos nesta dissertação de Sistemas de Gerenciamento e Análise de Dados por Bioinformática.

Neste contexto, esta dissertação apresentou a importância, definiu os requisitos, apresentou um sistema e suas etapas de desenvolvimento baseadas em técnicas de Engenharia de Software e submeteu o sistema desenvolvido a um projeto real de análise de ESTs de um fungo dermatófito denominado *Trichophyton rubrum*.

As etapas de desenvolvimento foram detalhadas com diagramas e modelos conceituais, o que contribui para possíveis extensões.

Ao sistema desenvolvido foi dado o nome de *SGADBio*, sistema este que pode ser instanciado em diversos tipos de pesquisa que envolvem projetos de seqüenciamento. Esta característica foi requisito fundamental no desenvolvimento deste trabalho. Usuários não familiarizados com a criação de *scripts* podem facilmente configurar as ações a que será submetida cada uma de suas seqüências a serem analisadas. Isso se torna possível com o módulo de gerenciamento de projetos.

A preocupação no desenvolvimento deste sistema foi torná-lo capaz de:

- Incluir processos, fontes e recursos normalmente usados em análises de dados de biologia molecular e oferecer mecanismos de extensibilidade;

- Oferecer ferramentas para validação de dados. Verificando se as entradas e saídas geradas possuem coerência;
- Executar o sistema de forma otimizada, de acordo com a arquitetura que está sendo utilizada;
- Controlar usuários, permissões e *logs* como ferramenta de acompanhamento das tarefas executadas;
- Controlar o acesso do que pode ser público ou restrito a um determinado grupo de trabalho;
- Armazenar as informações obtidas nas análises em bancos de dados e criar mecanismos de acesso facilitado;
- Permitir o controle e a execução de programas que estejam em sítios locais ou externos;
- Permitir a comparação entre duas bibliotecas, o chamado Northern Digital;
- Gerar estatísticas sobre os dados analisados, facilitando assim nas tomadas de decisão quanto aos experimentos realizados.

A dissertação incluiu ainda uma descrição sobre cada um dos módulos desenvolvidos e suas funções. Para cada um dos módulos um detalhamento de seus programas, e em alguns casos interfaces foram apresentadas como exemplos.

Por fim, a dissertação apresentou um estudo de caso, onde o sistema implementado foi instalado e testado em um Laboratório de Biologia Molecular no estudo de análise de seqüências oriundas de um projeto de ESTs de *T. rubrum*. O sistema gerou informações importantes que podem ajudar a revelar o padrão de expressão gênica necessário para o estabelecimento da infecção causada por esse fungo e possíveis alvos terapêuticos, tendo em vista as poucas classes químicas de antifúngicos disponíveis na prática médica (Martinez-Rossi et al., 2004).

6.2 Trabalhos Futuros

É possível destacar alguns trabalhos futuros a esta dissertação.

Em primeiro lugar, quanto ao software desenvolvido, é interessante permitir que sejam integradas ao sistema outras estruturas e sistemas existentes com o objetivo de contribuir diretamente para um maior entendimento das análises de bioseqüências. Sistemas como o Biopipe (Hoon et al., 2003) merecem ser estudados, já que este possui uma proposta de criação de um *framework* baseado em um protocolo para definição de *pipelines*.

Outro aspecto que merece atenção é tentar melhorar ainda mais a otimização do software. Para isso, mais testes de desempenho devem ser realizados por tipo e finalidade de aplicação.

Pretende-se também incluir suporte a outros tipos de dados que envolvam estudo de expressão gênica. Dados oriundos de bibliotecas construídas através de metodologia SAGE (*Serial Analysis of Gene Expression*) serão encapsulados no sistema para que o mesmo trabalhe com quantificação da expressão gênica de forma mais robusta.

Técnicas de reconhecimento de padrões utilizando a abordagem estatística denominada Teoria da Decisão ou até mesmo métodos de Inteligência Artificial poderão ser agregados com o objetivo de analisar e identificar regras implícitas aos dados biológicos (Costa, 2004).

Com relação ao estudo de caso utilizado nesta dissertação, novas bibliotecas serão construídas e inseridas no software para análise. O objetivo será identificar o padrão de expressão gênica de *Trichophyton rubrum* durante a interação *in vitro* entre fungo e queratinócitos utilizando o RNA mensageiro de *T. rubrum* obtido em diferentes tempos durante essa interação.

O software ajudará a definir quais são os genes diferencialmente expressos, aqueles mais expressos na interação *T. rubrum* e queratinócitos e aqueles expressos somente por *T. rubrum* na ausência de queratinócitos, através da opção de Northern digital.

Outro importante trabalho será analisar de forma mais cautelosa as seqüências que não foram aceitas pelo padrão de qualidade da análise realizada. Neste caso as seqüências inferiores a 50 bases serão analisadas com critérios mais específicos. Para isso serão estudadas novas técnicas de análise tanto no ponto de vista biológico como computacional.

REFERÊNCIAS

ACCELRYS-SOFTWARE *GCG Wisconsin Package - Homepage.* Disponível em: <http://www.accelrys.com/products/gcg/>. Acesso em: 24 de agosto de 2006, 2006.

AGARWALA, R. *NCBI's Genome Annotation Project - current status.* Disponível em: <http://hgm2001.hgu.mrc.ac.uk/Abstracts/Publish/Workshops/Workshop09/hgm0074.htm>. Acesso em: 30 de agosto de 2006, 2006.

ALTSCHUL, S. F. *Fundamentals of Database Searching.* Trends Guide to Bioinformatics, Elsevier Science, 1998.

ALTSCHUL, S. F., GISH, W., MILLER, W., et al. Basic local alignment search tool. *J Mol Biol*, v. 215, p. 403-410, 1990.

ANDRADE, M. A., BROWN, N. P., LEROY, C., et al. Automated genome sequence analysis and annotation. *Bioinformatics*, v. 15, p. 391-412, 1999.

APWEILER, R., BAIROCH, A., WU, C. H., et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*, v. 32, p. D115-D119, 2004.

BAIROCH, A., APWEILER, R., WU, C. H., et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, v. 33, p. D154-D159, 2005.

BAXEVANIS, A. D. and OUELLETE, B. F. *Bioinformatics: A practical guide to the analysis of gene and proteins*, Second Edition, John Wiley and Sons, 2001.

BEALER, K. *A Fault-Tolerant Parallel Scheduler for Blast.* Disponível em: <ftp://ftp.ncbi.nih.gov/blast/documents/blast-sc2004.pdf>. Acesso em: 03 de setembro de 2006, 2004.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., et al. GenBank. *Nucleic Acids Res*, v. 31, p. 23-27, 2003.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., et al. GenBank. *Nucleic Acids Res*, v. 33, p. D34-D38, 2005.

BOECKMANN, B., BAIROCH, A., APWEILER, R., et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, v. 31, p. 365-370, 2003.

BRAZILIAN-GENOME *Homepage*. Disponível em: <http://www.brgene.lncc.br/>. Acesso em: 05 de agosto de 2006, 2006.

CBRG, COMPUTATIONAL BIOCHEMISTRY RESEARCH GROUP *MultAlign: Multiple Sequence Alignment Tools*. Disponível em: <http://mendel.ethz.ch:8080/Server/MultAlign.html>. Acesso em: 18 de agosto de 2006, 2006.

CERVELATTI, E. P., FERREIRA-NOSAWA, M. S., AQUINO-FERREIRA, R., et al. Electrophoretic molecular karyotype of dermatophyte *Trichophyton rubrum*. *Genetics and Molecular Biology*, v. 27, p. 99-102, 2004.

CORPET, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*, v. 16, p. 10881-10890, 1988.

COSTA, L. DA F. Bioinformatics: perspectives for the future. *Genet Mol Res*, v. 3, p. 564-74, 2004.

COSTA, M., PASSOS, X. S., HASIMOTO E SOUZA, L. K., et al. Epidemiology and etiology of dermatophytosis in Goiania, GO, Brazil. *Rev Soc Bras Med Trop*, v. 35, p. 19-22, 2002.

DOE, US DEPARTMENT OF ENERGY *Homepage*. Disponível em: <http://www.doe.gov>. Acesso em: 04 de agosto de 2006, 2006.

EMMONS, C. W. Dermatophytes: natural grouping based on the form of spores and accessory organs. *Arch. Dermatol. Syphilol.*, v. 30, p. 337-362, 1934.

EWING, B. and GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, v. 8, p. 186-194, 1998.

EWING, B. and GREEN, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet*, v. 25, p. 232-234, 2000.

EWING, B., HILLIER, L., WENDL, M. C., et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, v. 8, p. 175-185, 1998.

FACHIN, A. L., CONTEL, E. P. and MARTINEZ-ROSSI, N. M. Effect of sub-MICs of antimycotics on expression of intracellular esterase of *Trichophyton rubrum*. *Med Mycol*, v. 39, p. 129-133, 2001.

FACHIN, A. L., FERREIRA, S. M., AQUINO, R., et al. Construção de um banco genômico do dermatófito *Trichophyton rubrum*. *Genet. Mol. Biol.*, v. 22, p. 400, 1999.

FALQUET, L., PAGNI, M., BUCHER, P., et al. The PROSITE database, its status in 2002. *Nucleic Acids Res*, v. 30, p. 235-238, 2002.

FELIPE, M. S., ANDRADE, R. V., ARRAES, F. B., et al. Transcriptional profiles of the human pathogenic fungus *Paracoccidioides brasiliensis* in mycelium and yeast cells. *J Biol Chem*, v. 280, p. 24706-14, 2005.

FELIPE, M. S., ANDRADE, R. V., PETROFEZA, S. S., et al. Transcriptome characterization of the dimorphic and pathogenic fungus *Paracoccidioides brasiliensis* by EST analysis. *Yeast*, v. 20, p. 263-271, 2003.

FELSENSTEIN, J. *PHYLIP: the PHYLogeny Inference Package.* Disponível em: <http://evolution.genetics.washington.edu/phylip.html>. Acesso em: 24 de agosto de 2006, 2006.

FOSTER, K. W., GHANNOUM, M. A. and ELEWSKI, B. E. Epidemiologic surveillance of cutaneous fungal infection in the United States from 1999 to 2002. *J Am Acad Dermatol*, v. 50, p. 748-752, 2004.

GALPERIN, M. Y. The Molecular Biology Database Collection: 2006 update. *Nucleic Acids Res*, v. 34, p. D3-D5, 2006.

GREEN, P. *Documentation for Phrap and Cross_Match.* Disponível em: <http://www.phrap.org/phredphrap/phrap.html>. Acesso em: 15 de agosto de 2006, 2006.

GRIFFITHS, A. J. F., MILLER, J. H., SUZUKI, D. T., et al. *An Introduction to Genetic Analysis*, Seventh Edition, WH Freeman & Company, 2000.

HOERSCH, S., LEROY, C., BROWN, N. P., et al. The GeneQuiz web server: protein functional analysis through the Web. *Trends Biochem Sci*, v. 25, p. 33-35, 2000.

HOON, S., RATNAPU, K. K., CHIA, J. M., et al. Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Res*, v. 13, p. 1904-1915, 2003.

HOON, S., RATNAPU, K. K., CHIA, J. M., et al. Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Res*, v. 13, p. 1904-15, 2003.

HUANG, X. and MADAN, A. CAP3: A DNA sequence assembly program. *Genome Res*, v. 9, p. 868-877, 1999.

HWANG, L., HOCKING-MURRAY, D., BAHRAMI, A. K., et al. Identifying phase-specific genes in the fungal pathogen *Histoplasma capsulatum* using a genomic shotgun microarray. *Mol Biol Cell*, v. 14, p. 2314-2326, 2003.

JAMES, P. Protein identification in the post-genome era: the rapid rise of proteomics. *Q Rev Biophys*, v. 30, p. 279-331, 1997.

KANEHISA, M. A database for post-genome analysis. *Trends Genet*, v. 13, p. 375-376, 1997.

KANEHISA, M. and GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, v. 28, p. 27-30, 2000.

KANEHISA, M., GOTO, S., HATTORI, M., et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, v. 34, p. D354-D357, 2006.

KIM, J. Computers Are from Mars, Organisms Are from Venus. *Computer*, v. 35, p. 25-32, 2002.

KULIKOVA, T., ALDEBERT, P., ALTHORPE, N., et al. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*, v. 32, p. D27-D30, 2004.

LBI, LABORATÓRIO DE BIOINFORMÁTICA DO INSTITUTO DE COMPUTAÇÃO DA UNIVERSIDADE DE CAMPINAS *Cancer Annotation Project - Homepage*. Disponível em: <http://cancer.lbi.ic.unicamp.br>. Acesso em: 30 de agosto de 2006, 2006.

LIANG, P. and PARDEE, A. B. Analysing differential gene expression in cancer. *Nat Rev Cancer*, v. 3, p. 869-876, 2003.

MARCHLER-BAUER, A., ANDERSON, J. B., DEWEESE-SCOTT, C., et al. CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res*, v. 31, p. 383-387, 2003.

MARTIN, K. J. and PARDEE, A. B. Identifying expressed genes. *Proc Natl Acad Sci U S A*, v. 97, p. 3789-3791, 2000.

MARTINEZ-ROSSI, N. M., FERREIRA-NOSAWA, M. S., GRAMINHA, M. A. S., et al. *Molecular aspects of dermatophyte-host interactions*. Fungi in Human and Animal Health, Scientific Publishers, 2004.

MASIA CANUTO, M. and GUTIERREZ RODERO, F. Antifungal drug resistance to azoles and polyenes. *Lancet Infect Dis*, v. 2, p. 550-563, 2002.

MOLLER, S., LESER, U., FLEISCHMANN, W., et al. EDITtoTrEMBL: a distributed approach to high-quality automated protein sequence annotation. *Bioinformatics*, v. 15, p. 219-227, 1999.

NCBI, NATIONAL CENTER OF BIOTECHNOLOGY INFORMATION *Homepage*. Disponível em: <http://www.ncbi.nlm.nih.gov>. Acesso em: 17 de agosto de 2006, 2006a.

NCBI, NATIONAL CENTER OF BIOTECHNOLOGY INFORMATION *Open Reading Frame Finder*. Disponível em: <http://www.ncbi.nlm.nih.gov/gorf/gorf.html>. Acesso em: 23 de agosto de 2006, 2006b.

NCBI, NATIONAL CENTER OF BIOTECHNOLOGY INFORMATION *BLAST*. Disponível em: <http://www.ncbi.nlm.nih.gov/BLAST/>. Acesso em: 30 de agosto de 2006, 2006c.

NEEDLEMAN, S. B. and WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, v. 48, p. 443-453, 1970.

NIH, NATIONAL INSTITUTES OF HEALTH *Homepage*. Disponível em: <http://www.nih.gov>. Acesso em: 04 de agosto de 2006, 2006.

ODOM, R. Pathophysiology of dermatophyte infections. *J Am Acad Dermatol*, v. 28, p. S2-S7, 1993.

PEARSON, W. R. and LIPMAN, D. J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, v. 85, p. 2444-2448, 1988.

PEREIRA, M., FACHIN, A. L. and MARTINEZ-ROSSI, N. M. The gene that determines resistance to tioconazole and to acridine derivatives in *Aspergillus nidulans* may have a corresponding gene in *Trichophyton rubrum*. *Mycopathologia*, v. 143, p. 71-75, 1998.

PRESSMAN, R. S. *Software Engineering: A Practitioner's Approach*, Thirth Edition, McGraw-Hill, 1991.

PROSDOCIMI, F., CERQUEIRA, G. C., BINNECK, E., et al. Bioinformática: Manual do Usuário. *Biotecnologia Ciência e Desenvolvimento*, v. 29, p. 18-31, 2002.

RIBICHICH, K. F., SALEM-IZACC, S. M., GEORG, R. C., et al. Gene discovery and expression profile analysis through sequencing of expressed sequence tags from different developmental stages of the chytridiomycete *Blastocladiella emersonii*. *Eukaryot Cell*, v. 4, p. 455-464, 2005.

RICE, P., LONGDEN, I. and BLEASBY, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, v. 16, p. 276-277, 2000.

RICHARDSON, M. D. and WARNOCK, D. W. *Fungal infection: diagnosis and manegement*, Blackwell Scientific Publications, 1993.

ROCHA, E. M., ALMEIDA, C. B. and MARTINEZ-ROSSI, N. M. Identification of genes involved in terbinafine resistance in *Aspergillus nidulans*. *Lett Appl Microbiol*, v. 35, p. 228-232, 2002.

RUTHERFORD, K., PARKHILL, J., CROOK, J., et al. Artemis: sequence visualization and annotation. *Bioinformatics*, v. 16, p. 944-945, 2000.

SALI, A. *MODELLER - A program for protein structure modeling.* Disponível em: <http://salilab.org/modeller/manual/manual.html>. Acesso em: 23 de agosto de 2006, 2006.

SILVA FILHO, A. *Arquitetura de Software*, Editora Campus, 2002.

SMITH, T. F. and WATERMAN, M. S. Identification of common molecular subsequences. *J Mol Biol*, v. 147, p. 195-197, 1981.

SOUSA, M. V. *Gestão da Vida Genoma e Pós-genoma*, Editora Unb, 2001.

STEIN, L. D., MUNGALL, C., SHU, S., et al. The generic genome browser: a building block for a model organism system database. *Genome Res*, v. 12, p. 1599-1610, 2002.

STURTEVANT, J. Applications of differential-display reverse transcription-PCR to molecular pathogenesis and medical mycology. *Clin Microbiol Rev*, v. 13, p. 408-427, 2000.

TATENO, Y., FUKAMI-KOBAYASHI, K., MIYAZAKI, S., et al. DNA Data Bank of Japan at work on genome sequence data. *Nucleic Acids Res*, v. 26, p. 16-20, 1998.

TECHNELYSIUM-PTY *Chromas*. Disponível em: <http://www.technelysium.com.au/chromas.html>. Acesso em: 15 de agosto de 2006, 2006.

TELLES, G. P. and SILVA, F. R. Trimming and clustering sugarcane ESTs. *Genetics and Molecular Biology*, v. 24, p. 1-4, 2001.

TEOREY, T. J., LIGHSTONE, S., NADEAU, T. *Database Modeling and Design*, 4th edition, Morgan Kaufmann Publishers, Inc, San Francisco, 2006.

THOMPSON, J. D., HIGGINS, D. G. and GIBSON, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, v. 22, p. 4673-4680, 1994.

VASCONCELOS, A. M. L. *Introdução à engenharia de software e aos princípios de qualidade*, Lavras - UFLA/FAEPE, 2004.

VELLA ZAHRA, L., GATT, P., BOFFA, M. J., et al. Characteristics of superficial mycoses in Malta. *Int J Dermatol*, v. 42, p. 265-271, 2003.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., et al. The sequence of the human genome. *Science*, v. 291, p. 1304-1351, 2001.

WASINGER, V. Holistic biology of microorganisms: genomics, transcriptomics, and proteomics. *Methods Biochem Anal*, v. 49, p. 3-14, 2006.

WATERMAN, M. S. *Introduction to Computational Biology*, Chapman & Hall, 1996.

WEITZMAN, I. and SUMMERBELL, R. C. The dermatophytes. *Clin Microbiol Rev*, v. 8, p. 240-259, 1995.

WELSH, M. *Running Linux*, Cambridge: O´Reilly, 1999.

WESTBROOK, J., FENG, Z., JAIN, S., et al. The Protein Data Bank: unifying the archive. *Nucleic Acids Res*, v. 30, p. 245-248, 2002.

WU, C. H., HUANG, H., ARMINSKI, L., et al. The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res*, v. 30, p. 35-37, 2002.

ZAHA, A. *Biologia Molecular Básica*, Segunda Edição, Mercado Aberto, 2000.