TRABALHO DE FORMATURA

Projeto de Iniciação Científica

ANOTAÇÃO DE GENES ASSOCIADOS COM O CONTROLE DA PROLIFERAÇÃO CELULAR E ORIGEM DE TUMORES

Aluno: André Fujita

Orientadora: Prof[a] Dr[a] Mari Cleide Sogayar (IQ-USP)

Supervisores: Prof. Dr. Alan Mitchell Durham (IME-USP)

Katlin Brauer Massirer (IQ-USP)

Instituto de Química - USP

Departamento de Bioquímica

Julho-Dezembro 2002

RESUMO

Este projeto visa implementar, no Laboratório de Biologia Celular e Molecular do Instituto de Química da USP, um sistema de organização e anotação de seqüências gênicas de interesse para a equipe do laboratório.

A linha de pesquisa deste laboratório enfoca, principalmente, o isolamento e a caracterização estrutural e funcional de genes diferencialmente expressos em processos relacionados com a proliferação celular e a origem dos tumores. As principais linhas de estudo incluem o efeito de agentes anti-tumorais em células de gliomas de rato, efeito de agentes mitogênicos em células beta de ilhotas pancreáticas humanas e, mais recentemente, a participação nos projetos Genoma do Câncer brasileiro (HGCP), Transcriptoma (TFI) e Genoma Funcional, nos quais se busca, entre outras coisas, a comparação entre tumores e tecidos normais de alguns órgãos humanos.

Com o avanço de projetos envolvendo seqüenciamento de DNA e análise de genes e proteínas, uma enorme quantidade de dados vem sendo gerada, criando a necessidade de avaliar e integrar estas informações através do desenvolvimento de ferramentas de Bioinformática. O objetivo deste projeto foi criar um sistema computacional de anotação e armazenamento de seqüências expressas de DNA (ESTs), envolvidas no controle da proliferação celular e na origem de tumores. O sistema é composto por um ambiente gráfico servidor-cliente (*web*) apoiado por um banco de dados em MySQL. A opção por uma interface *web* propicia a combinação de diversas ferramentas para uma análise mais completa, com resultados homogêneos. Inicialmente, as ESTs são identificadas e localizadas, de maneira automática, em relação a um genoma, pelos programas de alinhamento BLAST e BLAT. Em seguida, são avaliadas informações relativas às funções dos produtos gênicos, através de buscas em diversos bancos de dados públicos. Seqüências já descritas são confirmadas. Para as demais, são extrapoladas características referentes à seqüência de maior similaridade. Além da anotação automática, o ambiente gráfico permite que os usuários adicionem informações referentes aos genes. As informações obtidas na anotação são armazenadas no banco de dados. A facilidade de acesso aos dados e a aceleração na interpretação dos mesmos propicia maior agilidade na publicação dos resultados e na escolha de alvos para experimentos subseqüentes.

ABSTRACT

The advancement of projects involving DNA sequencing and analysis of genes and proteins, has been generating a large amount of data, creating the necessity for evaluating and integrating these informations through the development of Bioinformatic tools. The objective of this project was the development of a computational gene annotation system for the expressed sequence tags (ESTs) involved in the control of celular proliferation and the origin of tumors. This system is composed by a graphic client-server (web) interface supported by a MySQL database. The use of a web interface aims to integrate the numerous publicly available genomic tools, and thus obtain a complete analysis of the ESTs, in an homogenous manner. Firstly, the ESTs are identified and located automatically in a genome using the BLAST and BLAT programs. Then, the information associated to functions of gene products are evaluated through search in public databases. Known sequences are confirmed. For unknown sequences, the characteristics of the best hit are extrapolated. In addition, the graphic interface allows the annotator to add information related to these genes. The information obtained from the annotation are stored in the database.

This annotation system decreases repetitive effort tasks in numerous searches in databases. The results obtained using this system can easily be used for future consultation and to help the design of new experiments.

1 – INTRODUÇÃO

Recentemente, com o avanço de projetos envolvendo seqüenciamento de DNA e análise de genes e proteínas, uma enorme quantidade de dados vem sendo gerada, criando a necessidade de avaliar e integrar estas informações através do desenvolvimento de ferramentas de Bioinformática.

As tecnologias de Genômica, Proteômica e Bioinformática estão sendo utilizadas por nosso grupo para compreender os mecanismos moleculares dos processos normais e regulados de proliferação e diferenciação celular e do desvio deste controle, que ocorre em células que dão origem a tumores, utilizando modelos de transgênese em sistemas celulares e em animais.

Para a produção de proteínas, grupos de diferentes genes são ativados ou reprimidos em resposta a sinais internos ou externos. Isto significa que as células de um organismo tem o mesmo DNA, mas diferentes genes contidos neste DNA são transcritos em RNA e posteriormente traduzidos em proteínas, as quais irão exercer funções específicas. O conjunto de genes que estão ativados ou reprimidos num determinado momento compõe o que denominamos de perfil de expressão gênica (Alberts, 1994).

Em nosso laboratório, busca-se o isolamento e a caracterização estrutural de genes diferencialmente expressos na presença de agentes mitogênicos, indutores de diferenciação, vírus tumorais, agentes anti-tumorais/anti-proliferativos em modelos celulares e em tumores humanos (Sasahara *et al*, 1995, Armelin *et al*, 1996, Valentini & Armelin, 1996, Vedoy *et al*, 1999, Flatchart & Sogayar, 1999, Vedoy & Sogayar, 2001, Bengtson *et al*, 2002). De forma geral, são utilizadas metodologias experimentais que permitem o isolamento de genes regulados em processos relacionados à proliferação celular e em condições específicas de tratamentos.

O principal modelo de estudo são células ST1 provenientes de tumores de glia de rato, nas quais procura-se isolar os genes associados com a ação anti-tumoral de glicocorticóides e de retinóides sobre estas células. Mais recentemente, participando dos projetos de Genoma e Transcriptoma humanos do câncer, passou-se a estudar os genes diferencialmente expressos entre tecidos normais e tumorais. Os tecidos atualmente estudados com este enfoque são tecidos de cérebro, próstata e pâncreas.

Estudando os tecidos de próstata, o grupo também participa do projeto denominado CAGE (“Cooperation for Analysis of Gene Expression”), realizado em conjunto com o Instituto de Matemática e Estatística (IME) da USP. Este projeto emprega análises de

microarranjos de DNA (*microarray*) feitos em lâminas de vidro, através de robótica, para análise da expressão gênica diferencial em larga escala (análises de aproximadamente 4.000 genes por experimento).

Na Unidade de Ilhotas Pancreáticas Humanas, que é um anexo do Laboratório de Biologia Celular e Molecular, são estudados genes isolados de ilhotas de pâncreas humanos tratadas com agentes mitogênicos (glicose e/ou prolactina), visando auxiliar na elucidação do diabetes mellitus (DM) tipo I e no desenvolvimento de novas terapias para esta doença. Genes diferencialmente expressos entre células beta de pâncreas normais e tumorais (insulinomas) também estão sendo isolados e caracterizados.

Após o isolamento dos genes envolvidos nos referidos processos, são realizados estudos funcionais para caracterizar estes genes e seus produtos protéicos. As informações que se tem sobre os genes envolvidos nestes processos de proliferação, diferenciação e malignidade são ainda incompletas e o uso de técnicas experimentais modernas aliadas às análises de Bioinformática, constituem ferramentas potentes para a elucidação destes processos.

Com o aumento da quantidade de seqüências codificadoras obtidas (ESTs – *Expressed Sequence Tags*) e a caracterização experimental das respectivas proteínas codificadas por estes genes, a predição de funções de proteínas, utilizando ferramentas computacionais, se torna cada vez mais importante (Bork *et al*, 1998).

O objetivo do presente projeto foi o desenvolvimento de um sistema de anotação para as seqüências de ESTs geradas no laboratório. Esta anotação tem por finalidade identificar os genes representados por uma determinada EST ou grupo de ESTs e descrever as funções à elas relacionadas.

O sistema foi composto por um ambiente gráfico servidor-cliente (*web*) apoiado por um banco de dados em MySQL. O ambiente gráfico contém uma página de submissão de seqüências gênicas e uma de ferramentas de anotação dos genes para que os usuários possam tanto visualizar os dados quanto adicionar novas informações. Estas informações são armazenadas de forma organizada no banco de dados de suporte.

A opção por uma interface *web* propicia a combinação de diversas ferramentas para se obter uma análise completa, com resultados homogêneos, de fácil uso e acesso.

A anotação de ESTs foi dividida, basicamente, em duas fases (Stein, 2001). A primeira consistiu na identificação e localização da EST em relação a um genoma. Nesta

etapa utiliza-se programas de alinhamento de seqüências. Para isto, é realizada a anotação automática, preliminar, de seqüências, através dos programas BLAST e BLAT, obtendo-se informações catalogadas sobre ORFs (*Open Reading Frames*) passadas como entrada. Os dados da seqüência do banco que apresentam o melhor *hit* são atribuídos à seqüência.

Na fase seguinte, são avaliadas as informações relativas às funções dos genes, por meio de buscas em diversos bancos de dados públicos disponíveis, contendo, inclusive, proteínas já conhecidas (LocusLink, CGAP, PFAM, Gene Ontology). Seqüências conhecidas são confirmadas, enquanto para as seqüências não totalmente conhecidas, serão extrapoladas características referentes à seqüência de maior similaridade. Desta forma são armazenadas as ESTs devidamente caracterizadas, com fácil acesso aos usuários.

Na etapa final, a correção de anotações funcionais exige um enorme esforço já que esta fase ainda precisa ser revisada manualmente. Isto se deve, principalmente, à dificuldade de se desenvolver sistemas que selecionem palavras-chaves e informações funcionais de artigos científicos (Bork *et al*, 1998).

Com esta organização e homogeneização dos dados gerados por diferentes projetos do laboratório, a integração dos resultados é facilitada e os mecanismos envolvidos nos processos estudados podem ser elucidados mais rápida e eficientemente.

2 – OBJETIVOS

Criar um sistema de anotação e armazenamento de ESTs, as quais estão envolvidas no controle da proliferação celular e na origem de neoplasias.

2.1 – Objetivos específicos

Analisar e caracterizar ESTs isoladas experimentalmente, através de uma interface de anotação do tipo *web*, contendo ferramentas publicamente disponíveis.

Identificar, organizar e armazenar dados relativos aos genes de interesse, em um banco de dados MySQL, de fácil acesso e atualização.

3 - VIABILIDADE E RELEVÂNCIA

O projeto está sendo desenvolvido no Laboratório de Biologia Celular e Molecular (IQ-USP) sob orientação da Prof[a] Dr[a] Mari Cleide Sogayar e supervisão da MSc. Katlin Brauer Massirer. O Prof. Dr. Alan Mitchell Durham do IME-USP é o supervisor da parte computacional.

O laboratório conta com uma estação de trabalho Alpha, que dispõe do sistema operacional Tru64Unix (Compaq) na qual está sendo desenvolvido todo o sistema de anotação e a qual também funciona como servidor de página para a Internet.

As ferramentas computacionais já existentes, necessárias para o desenvolvimento do projeto estão disponíveis gratuitamente na Internet para o uso acadêmico. As demais ferramentas, para as necessidades específicas do grupo, são desenvolvidas pela equipe de Bioinformática.

Os alunos do laboratório (maioria doutorandos) fornecem, a partir de seus experimentos, as seqüências a serem anotadas e contribuem nas discussões das ferramentas necessárias para as análises, bem como no esclarecimento de dúvidas referentes à Biologia Molecular.

O laboratório é composto por cerca de 20 alunos entre Iniciação Científica e pós-graduação (lista da Equipe em anexo) os quais podem ser beneficiados com o desenvolvimento deste projeto.

A inter-relação dos dados obtidos experimentalmente no laboratório, seguida das análises de Bioinformática irão proporcionar maior integração dos diferentes projetos, e, ainda, permitir que os alunos iniciantes acessem, rapidamente, os dados anteriormente obtidos. A facilidade de acesso aos dados e a aceleração na interpretação dos mesmos, poderá contribuir para atingir maior agilidade na publicação dos resultados e na escolha de alvos para experimentos subseqüentes.

4– MÉTODOS E FERRAMENTAS COMPUTACIONAIS

A princípio, foi feito um levantamento de dados, juntamente com a supervisora Katlin Brauer Massirer, para avaliar quais seriam as ferramentas de anotação necessárias para suprir as informações de interesse para a equipe do laboratório. Baseado nisto, foram criados dois protótipos de telas: a interface de submissão e a de anotação. Apresentando estas telas aos alunos e recebendo suas sugestões, foi possível completar um nível razoável de refinamento para o início da implementação.

4.1 – *Pipeline* de Anotação

Os mecanismos utilizados para realizar as transações entre anotador e banco de dados foram implementados nos próprios programas encarregados da submissão e visualização das seqüências.

Os programas de submissão, anotação automática, visualização e *parsers* (programas de filtragem de dados) formam juntos o *pipeline* de anotação de ESTs como descrito no esquema abaixo.

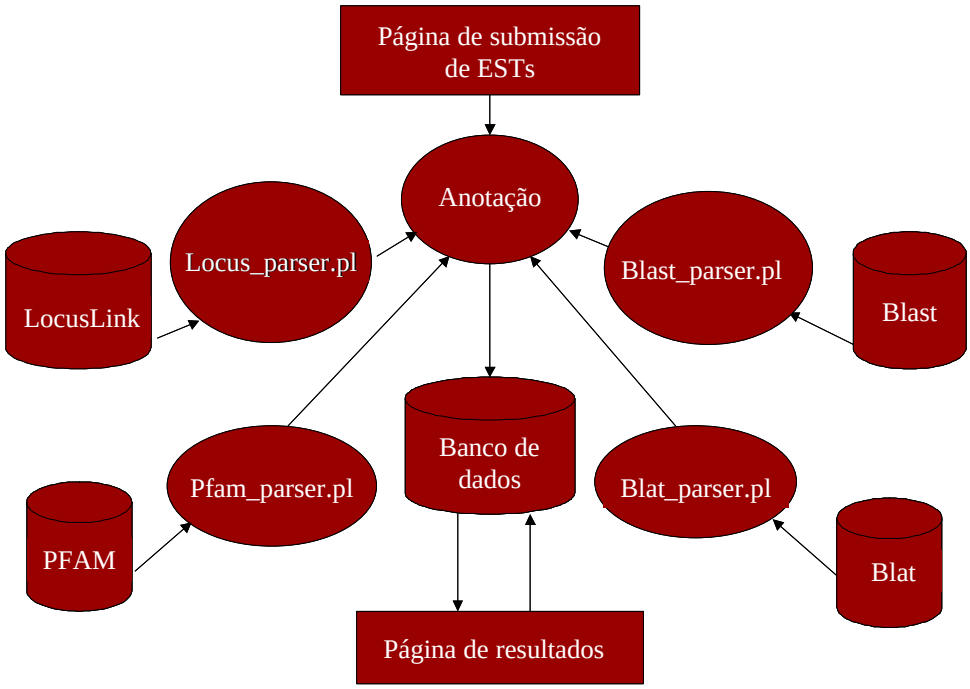

Figura 1. Representação esquemática do *pipeline* do sistema de anotação.

4.2 – Interfaces

A fim de facilitar a submissão das seqüências a serem anotadas e o acesso ao banco de dados e agilizar a anotação manual, foram criadas algumas interfaces gráficas em HTML.

Para a construção dessas interfaces, foi utilizado o aplicativo FrontPage 2000 da Microsoft por oferecer maior agilidade no seu desenvolvimento.

As interfaces estão divididas em dois tipos: a de submissão e a de anotação manual.

A tela de submissão contém opções de múltipla escolha. São escolhidas as informações quanto ao nome do anotador, modelo experimental, estado de malignidade das células, tratamento, tempo de tratamento e agente do tratamento. Em relação à expressão gênica são submetidos os dados de nível de expressão e a razão entre a expressão de cada uma das duas condições.

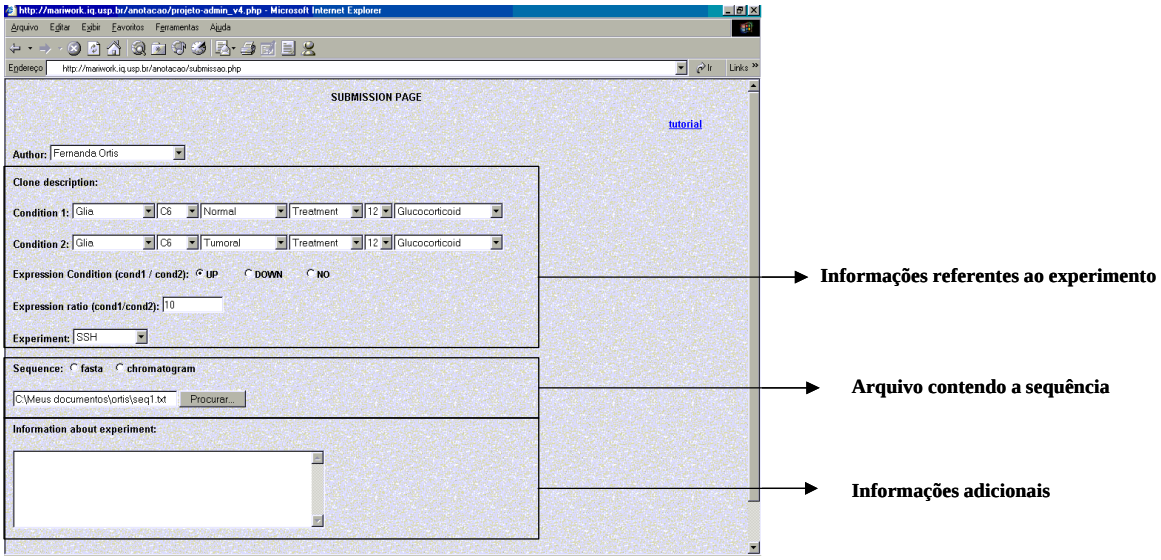

Figura 2. Página de submissão.

A submissão da seqüência do gene pode ser realizada diretamente através de um arquivo em formato FASTA que consiste em um arquivo texto com a primeira linha começando com o caracter “>” seguido pela descrição da seqüência, e nas linhas abaixo se encontra a seqüência propriamente dita.

Exemplo de um arquivo em formato FASTA:

> gi|21312785|ref|NM_028991.1|[21312785]

ttagtatgaactacaatgcaatgcacttgtaactcatggaaataaatgtacatctttatttacacccatgataagattcagtgttgat

tttctctggattggtgtgtcctaagtaggcactcataatcaatttatggcttgtgcttcagacaaaaatgttcatgggccttactcta

ctccccactccaccctacccccccatgcactgcccctcacagcagtttacgtatatgggaaaggtccttttcagctgcacatggt
gccatgcatcgttaatcccagcattcagtcagaggcaggtggatctctgaatggaagcaggcctgatttgcatagggaggtc
caagctggaactctataggtcctgtctcaaaaaaaacagagtcctccccgtctgcctct

As informações adicionais ou comentários quanto ao experimento podem ser inseridas em uma área de texto disponível no canto inferior da tela.

Na tela de visualização e anotação são listados todos os resultados obtidos pela anotação automática que estão armazenados no banco de dados. Nesta página também estão disponíveis alguns *links* para as ferramentas e bancos de dados públicos, no caso do anotador desejar obter maiores informações sobre o gene ou também para realizar consultas mais atualizadas. Há também uma área de texto para a inserção desses novos comentários.

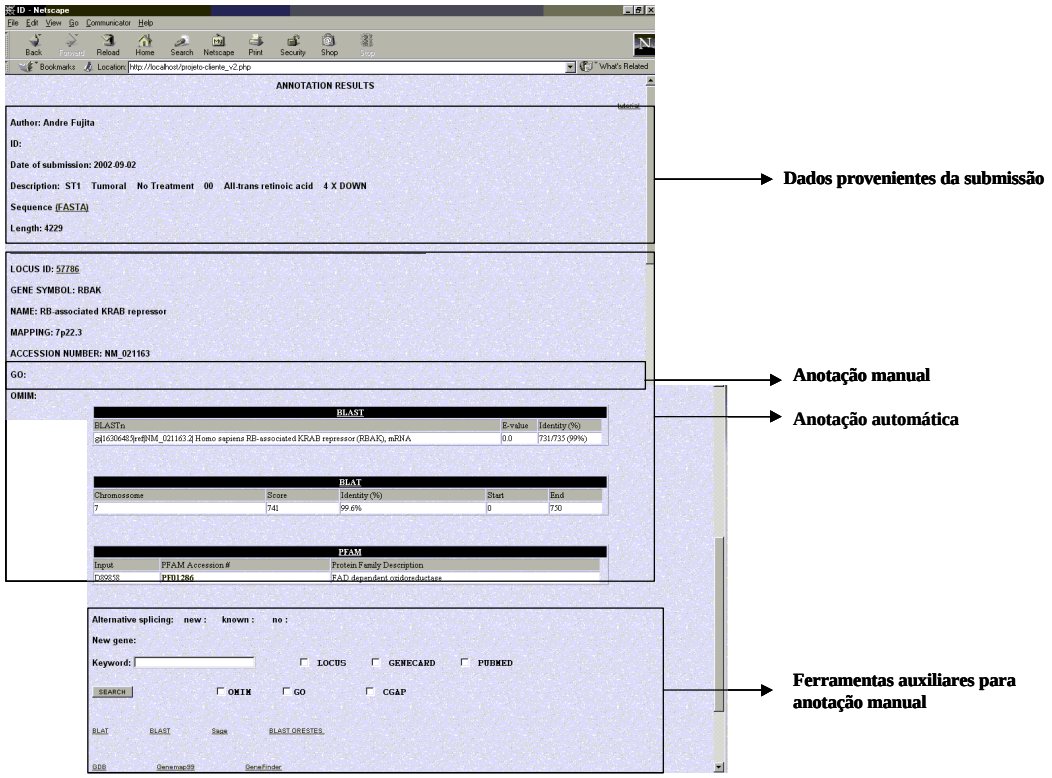

Figura 3. Página de anotação.

4.3 – Ferramentas de localização e caracterização gênica

BLAST (*Basic Local Alignment Search Tool*): é uma ferramenta de alinhamento de seqüências desenvolvido pelo *National Center for Biotechnology Information* (NCBI). Dada uma seqüência de consulta, o programa apresenta, como resultado, uma lista com os *hits*

seguida da extensão e qualidade da similaridade da seqüência, bem como a disposição de *gaps* no alinhamento (Wheeler *et al*., 2001). Em cada um dos alinhamentos encontrados é listada uma descrição da seqüência, a pontuação do alinhamento obtido que é normalizada, permitindo comparações entre consultas realizadas em diferentes bancos de seqüências, e o *E value* que é o valor que indica a significância estatística do *hit*, ou seja, o número estimado de alinhamentos com pontuação igual ou maior ao do *hit* que poderia ter sido encontrado ao acaso.

O BLAST possui algumas variantes dentre as quais, a variante usada para a anotação foi o blastn, um programa que compara uma seqüência de nucleotídeos contra todas as seqüências de nucleotídeos contidas no banco.

BLAT é um programa desenvolvido pela Universidade da Califórnia, Santa Cruz. É similar ao BLAST, porém, realiza o alinhamento apenas contra o genoma humano e de camundongo. O método de indexação de todos os K-*mers* sem sobreposição torna-o mais rápido e sensível que o Blast e cerca de 500 vezes mais rápido que outras ferramentas populares de alinhamentos de mRNA/DNA (Kent, 2002). Inclui também outras ferramentas úteis como programas de predição gênica e visualização de ESTs, permite a visualização de *splicings* alternativos e fornece a identificação e localização gênica com *links* para várias outras ferramentas de caracterização.

LOCUSLINK é um banco de dados público mantido pelo *National Center for Biotechnology Information* (NCBI) e que contém informações descritivas sobre o *loci* dos genes incluindo nomenclatura, identificador no banco de dados (ID), doenças associadas, posição cromossômica e acesso às seqüências, centralizando os dados relativos aos genes de mosca de frutas, camundongo, rato, humano, e lebiste (Pruitt & Maglott, 2001).

4.4 – Ferramentas de caracterização funcional

GO (*Gene Ontology*) é um banco de dados independente dos demais conhecidos que fornece descrição das funções moleculares, processos biológicos e componentes celulares de produtos gênicos.

PFAM (*Protein Families Database of Alignments and HMMs*) é um banco de dados representado por famílias de proteínas, cada qual composta por seqüências similares e analisadas pelo modelo de *hidden Markov*, sendo utilizado na avaliação de domínios de proteína. Este banco contem anotações de cada família em forma de texto e, também, *links*

para as referências bibliográficas (Bateman *et al*, 2001). É mais sensível que alguns outros bancos que são amplamente utilizados, como PROSITE, pois contém dados interligados de outros bancos (Bork *et al*, 1998).

CGAP (*Cancer Genome Anatomy Project*) abrange bancos de dados com enfoque em alterações moleculares que ocorrem em câncer. Inclui bancos de ESTs (expressed sequence tags), padrões de expressão gênica, SNPs (single nuclear polymorphisms) e informações citogenéticas e ainda, um diretório de genes supressores de tumores e oncogenes.

SAGE (*Serial Analysis of Gene Expression*) é um componente do CGAP que permite depositar, restaurar e analisar dados de expressão gênica humana (Lal *et al*, 1999). Atualmente é composto por 100 bibliotecas de cDNAs de tecidos (normal ou tumoral) ou condições de tratamento diferentes.

Permite a busca de dados de genes individuais ou de listas de genes diferencialmente expressos entre diferentes bibliotecas de SAGE.

4.5 – Anotação automática

Os *parsers* para a busca de informações relevantes nas ferramentas de anotação foram desenvolvidos utilizando-se a linguagem de programação Perl e as transações de dados baseadas em scripts PHP.

Primeiramente, o programa compara a seqüência submetida pelo anotador com o banco de seqüências de nucleotídeos nt do *GenBank* usando o programa BLAST. O resultado do BLAST fornece a identificação da seqüência em diferentes organismos e a localização genômica, devolvendo, como resultado, o nome da seqüência do *hit* mais significativo acompanhado do *score*, da identidade, do *accession number* e do *E value*.

Esta mesma seqüência submetida é comparada com os bancos do genoma humano e do camundongo pelo programa BLAT. Este, por sua vez, retorna a localização no cromossomo, o *score*, a identidade, o início e o fim do alinhamento do melhor *hit*.

No entanto, há casos em que o melhor alinhamento, ou seja, o que obteve o melhor *hit* não identifica a seqüência de interesse corretamente. Para estes casos, a página de visualização permite que o anotador realize a correção manualmente.

A partir do *accession number* obtido pelo BLAST, é possível obter o locus id (chave do banco de dados do Locuslink), fazendo-se uma busca no banco de dados do CGAP

instalado localmente. Este banco de dados é formado por dois arquivos texto onde estão inseridas inúmeras informações quanto ao genoma humano e ao de camundongo. A fim de facilitar a busca neste banco, pois grande parte da informação era desnecessária para a anotação, foi criada uma tabela contendo apenas os valores do locus id e os seus respectivos *accession numbers*. Feita a associação do *accession number* com o locus id, é realizada a busca no LocusLink que retorna a descrição do gene, a posição específica no cromossomo e o símbolo.

Após a identificação da seqüência, é necessário, ainda, caracterizar a função bioquímica da proteína codificada. Uma alternativa é realizar buscas em bancos de proteínas já conhecidas e encontrar alguma seqüência similar para a qual é possível extrapolar a função. Caso alguma seqüência seja encontrada, a hipótese que esta nova seqüência tem a mesma função do gene do banco (Gaasterland & Oprea, 2001). Para isto, são realizadas buscas no banco de dados do PFAM.

4.6 – Anotação manual

Concluída a anotação automática, é necessário, ainda, uma verificação manual dos resultados obtidos, pois algumas vezes o gene de interesse não é localizado automaticamente, de maneira correta, pela anotação. A anotação manual consiste em verificar possíveis erros, corrigí-los, caso necessário, e, ainda, adicionar informações que não se pode obter computacionalmente como a relação com as doenças como o câncer, por exemplo.

No caso deste projeto, foi deixado para que o próprio anotador obtivesse manualmente o resultado do *Gene Ontology*, por não ser possível desenvolver uma boa heurística de procura, pelo fato de o anotador procurar diferentes classificações de acordo com o modelo de estudo.

4.7 – Montagem do banco de dados

Os dados de anotação são armazenados e organizados em um banco de dados MySQL (gerenciador de sistema de banco de dados). O MySQL tem, como característica principal, ser do tipo relacional, ou seja, separa os dados em tabelas, permitindo maior flexibilidade e velocidade no acesso. As tabelas são interligadas através de relações que tornam possível a combinação de dados de tabelas distintas (Gibas, 2001).

Este banco é formado basicamente por uma tabela que armazena os dados do anotador, as datas das anotações e os resultados obtidos na busca em cada um dos bancos ou na execução de ferramentas auxiliares. Inicialmente, o banco está sendo formado por oito tabelas com as suas respectivas relações mostradas na figura abaixo.

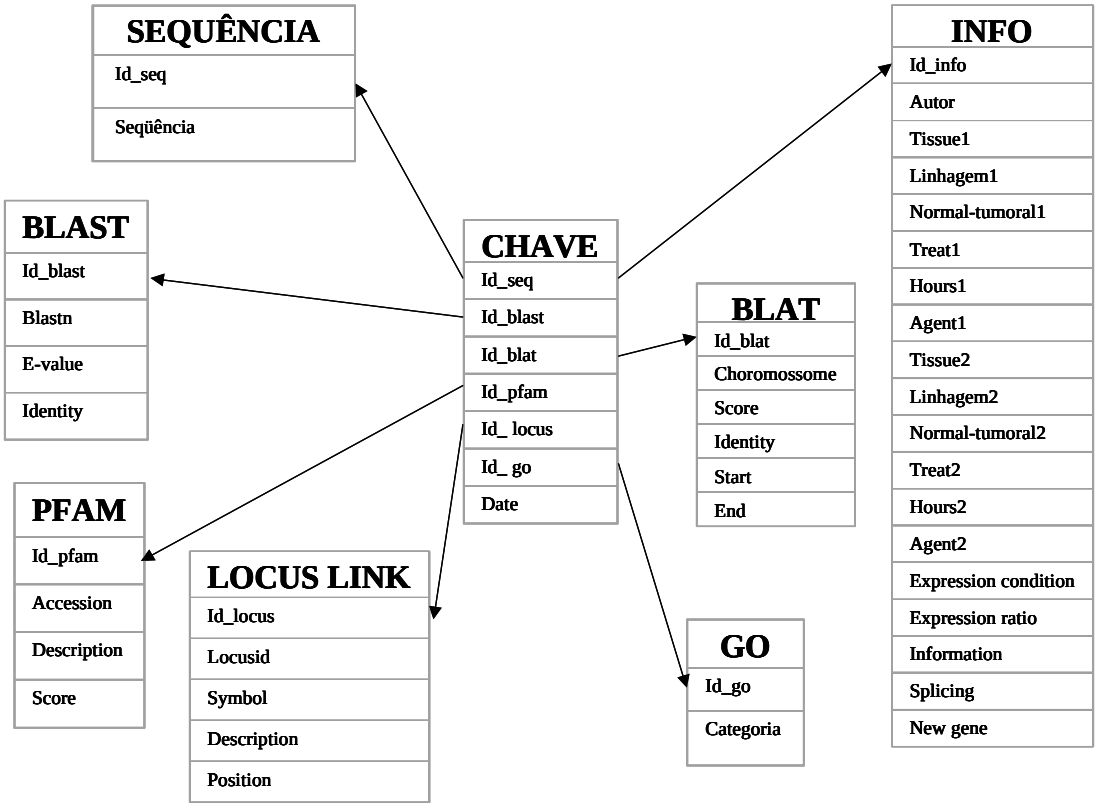

Figura 4. Estrutura do banco de dados.

Este banco contém a identificação dos genes e as características funcionais anotadas como *Gene Bank Accession Number*, a seqüência de nucleotídeos, a identificação do clone físico, a localização cromossômica, a classificação funcional (*Gene Ontology*), presença de domínios protéicos, polimorfismos, formas alternativas de transcritos, tecidos e níveis de expressão diferencial.

O banco de dados implementado possibilita o armazenamento de todo o histórico de anotação, ou seja, armazena cada versão da atualização do banco. Com isto, o sistema passa a adotar, como referência, a data da última anotação, além de manter as anotações anteriores.

Outras características deste banco são:

- Permite armazenar os resultados obtidos por apenas uma ou mais ferramentas sem a necessidade de se executar todas as outras;
- Cada uma das tabelas que armazenam os resultados de uma ferramenta específica podem ser alteradas sem afetar as outras;
- Facilita a inserção de novas ferramentas para anotação conforme as necessidades do laboratório.

5 – RESULTADOS E DISCUSSÃO

Apesar de já existir uma ferramenta de filtro (Bioperl) para as informações geradas pelo BLAST e outros programas, optou-se por criar *parsers* próprios pela necessidade de se criar filtros também para os resultados gerados pelos módulos do Bioperl.

Uma das principais preocupações na implementação das interfaces tanto de submissão quanto de anotação manual e visualização da anotação automática foi de projetá-las de tal forma que minimizasse o erro operacional humano, fornecendo ao usuário todas as opções para que este apenas realize a escolha, diminuindo a quantidade de digitação na entrada dos dados. Isto garante resultados mais confiáveis e minimiza o tempo gasto pelo anotador. O ambiente gráfico também oferece maior conforto, diferente dos programas convencionais de Bioinformática que são operados via linhas de comando, o que dificulta muito o manuseio e aprendizado por usuários inexperientes.

Nos alinhamentos realizados usando os programas BLAST e BLAT, ocorre que, na maioria dos casos, o primeiro *hit* traz a resposta correta, mas em algumas exceções, a resposta correta não é o melhor *hit*, havendo a necessidade de corrigi-lo manualmente. No entanto, esta anotação automática já auxilia muito, minimizando o esforço, trabalho e tempo do anotador no caso em que o primeiro *hit* é o mais relevante.

O banco de dados do BLAT é atualizado a cada quatro meses e, juntamente, a URL de acesso. Como o acesso ao BLAT a partir do sistema de anotação é realizado via URL, há uma necessidade de se alterar uma linha do código da anotação a cada momento em que o banco do BLAT é atualizado.

Para cada ferramenta utilizada, foi criado um filtro para selecionar as informações relevantes para o grupo do laboratório. Isto acarreta em constante manutenção do sistema, já que a cada nova versão destas ferramentas, a disposição dos dados pode ser alterada levando à uma seleção errônea de dados.

Um dos principais problemas ao se realizar uma anotação é o fato de algumas das informações se tornarem desatualizadas com a descoberta e o depósito de novas seqüências

nos bancos de dados públicos (Gaasterland & Oprea, 2001). Para superar esse problema, a página de anotação manual disponibiliza *links* de visualização de alinhamentos mais recentes do BLAST e BLAT. Além disso, o próprio banco de dados foi estruturado e projetado para armazenar inúmeras versões da anotação.

6 – CONCLUSÕES

Com o sistema de anotação automática é possível diminuir o esforço em tarefas repetitivas do anotador como buscas em inúmeros bancos de dados.

Os resultados obtidos pelo sistema de anotação servem como fonte de consulta de resultados além orientar análises experimentais posteriores.

7 - REFERÊNCIAS BIBLIOGRÁFICAS

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Introduction to the cell. In: **Molecular Biology of the Cell**. cap1, p. 4-39, 1994

Armelin MC, Sasahara RM, Flatschart R, Vedoy C. Use of cDNA cloning to study the mechanism of action of glucocorticoid hormones at the molecular level. Braz J Med Biol Res. v.29, n.12, p.1751-7. Review, 1996.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res v.30, p. 276-80, 2002.

Bengtson MH, Maria-Engler, SS Rodrigues, LO, Hasegawa, AP, Colin C & Sogayar, MC Celular and molecular characterization of retinoic acid effects on ST1 rat glioma cells. Artigo submetido Glia, Nov 2002.

Bork P, Dandekar, T Diaz-Lazcoz, Y, Eisenhaber F, Huynen M, Yuan Y. Predicting Function: From Genes to Genomes and Back. J.Mol.Biol., v.283, p.707-725, 1998.

Flatschart RB, Sogayar MC. Functional analysis of newly discovered growth control genes: experimental approaches. Braz J Med Biol Res. v.32, n.7, p.867-75. Review, 1999.

Gaasterland T, Oprea M. Whole-genome analysis: annotations and updates. Current Opinion in Structural Biology, v.11, n.3, p.377-381, 2001.

Gibas C, Jambeck, P. Building Biological Database. In: **Developing Bioinformatics and Computer Skills**. Sebastopol: O'Reilly, 2001. cap.13, p.350-382.

Kent WJ, BLAT-The BLAST-Like Alignment Tool v.12, n.4, p. 656-664, 2002.

Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, Marra MA, Prange C, Morin PJ, Polyak K, Papadopoulos N, Vogelstein B, Kinzler KW, Strausberg RL, Riggins GJ. A public database for gene expression in human cancers. Cancer Res., v.1, n.59(21), p.5403-5407, 1999.

Mount DW. **Bioinformatics. Sequence and Genome Analysis**. 1.ed. NY: CSHL Press, 2000. 564p.

Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res., 1;29(1) p.137-40, 2001.

Sasahara RM, Valentini SR & Armelin MCS. Molecular basis of glucocorticoid action in transformed to normal phenotypic reversion. *Anais do Simpósio Nipo-Brasileiro de Ciência e Tecnologia*, 94- 95, 1995.

Stein L. Genome annotation: from sequence to biology. Nat. Rev. Genet., v.2, n.7, p.493-503, 2001.

Valentini SR, Armelin MC. Cloning of glucocorticoid-regulated genes in C6/ST1 rat glioma phenotypic reversion.J Endocrinol. v.148, n.1, p.11-7, 1996.

Vedoy CG, Bengtson MH, Sogayar MC. Hunting for differentially expressed genes. Braz J Med Biol Res. v.32, n.7, p.877-84, 1999.

Vedoy CG, Sogayar MC - Isolation and characterization of genes associated with the anti-tumor activity of glucocorticoids. Mol. Brain Res., 2002 (*in press*).

Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res., v.1, n.30(1), p.13-16, 2002.

Páginas da internet

BLAST Basic Local Alignment Search Tool.

Disponível em http://www.ncbi.nlm.nih.gov/BLAST/ Acesso em: 27 junho 2002.

BLAT Search Genome.

Disponível em http://genome.ucsc.edu/cgi-bin/hg Blat?Command =start Acesso em: 03 julho 2002.

CGAP Cancer Genome Anatomy Project.

Disponível em http://cgap.nci.nih.gov/ Acesso em: 27 junho 2002.

GeneMark.

Disponível em http://www.ebi.ac.uk/genemark/ Acesso em: 25 junho 2002.

Gene Ontology Consortium

Disponível em http://www.geneontology.org/ Acesso em: 25 outubro 2002.

GEO Gene Expression Omnibus.

Disponível em http://www.ncbi.nlm.nih.gov/geo/ Acesso em: 03 julho 2002.

MySQL.

Disponível em http://www.mysql.com/doc/W/h/What-is.html Acesso em: 03 julho 2002.

PFAM Protein families database of alignments and HMMs.

Disponível em http://www.sanger.ac.uk/Software/Pfam/ Acesso em: 27 junho 2002.

PHP Manual

Disponível em http://www.zend.com/manual/

SWISS-PROT knowledgebase.

Disponível em http://www.expasy.ch/sprot/ Acesso em: 27 junho 2002.

8 - DESAFIOS E FRUSTRAÇÕES ENCONTRADOS

Durante toda a Iniciação Científica fui exposto a inúmeros desafios. tanto pela minha Orientadora quanto pela minha Supervisora. Dentre eles, citarei alguns deles abaixo:

- Comunicar de forma eficaz com profissionais das Ciências Biológicas;
- Apresentar seminários semestrais em inglês para o grupo do laboratório;
- Apresentar o seminário sobre o tema “O que é Bioinformática” para o curso “Ferramentas de engenharia genética aplicadas à biotecnologia” ministrado para alunos e professores do ensino médio do Colégio Bandeirantes e da Escola de Aplicação;
- Apresentar o seminário de introdução ao uso de ferramentas de Bioinformática para o curso “Biologia Molecular da Transformação Maligna” ministrado para os estudantes da pós-graduação;
- Cursar a disciplina QBQ0126 – Biologia Molecular do Gene.

Dentre todos os desafios, o único fato frustrante, foi sem dúvida, o seminário apresentado para os estudantes da pós-graduação. Não que o seminário tenha sido ruim, mas, acredito que, por falta de experiência ou maturidade, não pude apresentar um seminário do nível que eu esperava apresentar. Provavelmente, o fato de ter como público estudantes de pós me fez sentir um pouco de nervosismo e insegurança.

9 - LISTA DAS DISCIPLINAS CURSADAS NO BCC MAIS RELEVANTES

Devido ao fato de ter realizado uma Iniciação Científica num ramo multidisciplinar, a Bioinformática e, além de tudo, no Departamento de Bioquímica do Instituto de Química, acredito que muitas disciplinas foram extremamente importantes para o bom desenvolvimento do projeto. Dentre elas, merecem destaque:

- MAC0426-Sistemas de Bancos de Dados e MAC0439-Laboratório de Bancos de Dados, pois meu projeto estava diretamente relacionado a implementação de um banco para anotação;
- MAC0332-Engenharia de Software e MAC0446-Princípios de Interação Homem-Computador tanto para a estratégia de desenvolvimento do projeto quanto para a

criação de uma interface amigável ao usuário, já que a ferramenta desenvolvida é direcionada aos biólogos;

- MAC0414-Autômatos e Linguagens Formais para a construção de expressões regulares capazes de filtrar arquivos gerados pelas diversas ferramentas auxiliares.
- MAC0422-Sistemas Operacionais para a administração de uma máquina Unix;

Além dessas disciplinas, a possibilidade de realizar disciplinas extracurriculares como QBQ0126-Biologia Molecular do Gene me permitiu compreender melhor as dificuldades e o modo de raciocínio dos profissionais das ciências biológicas, em particular, dos biólogos moleculares permitindo uma comunicação mais precisa.

10 - INTERAÇÃO COM MEMBROS DA EQUIPE QUE TENHAM AGIDO COMO MENTORES DO TRABALHO

A interação foi bem próxima e amigável, isto é, minha supervisora, Katlin Brauer Massirer esteve praticamente todo o tempo auxiliando na resolução de dúvidas relacionadas aos problemas biológicos, opinando sobre a relevância e auxiliando no desenvolvimento de todas as partes do projeto.

Já a Professora Mari Cleide Sogayar sempre me incentivou a estudar Biologia Molecular e se mostrou estar bem entusiasmada tanto com o projeto quanto com essa nova área, a Bioinformática. Mostrou-se sempre disposta a ler e corrigir os resumos, relatórios e projetos mesmo nos períodos mais difíceis.

11 - DIFERENÇAS NOTADAS ENTRE A FORMA DE COOPERAÇÃO COM COLEGAS DO BCC NAS TAREFAS EM GRUPO DAS DISCIPLINAS E A FORMA DE TRABALHO CONJUNTO NO LABORATÓRIO

No BCC, todos nós trabalhamos em grupo sobre temas similares, enquanto no laboratório, os trabalhos, pelo menos relacionados ao meu projeto, foram complementares, ou seja, grande parte do conhecimento biológico foi fornecido pelos alunos pós-graduandos do laboratório enquanto eu fornecia o lado computacional.

12 - OBSERVAÇÕES SOBRE APLICAÇÃO DE CONCEITOS ESTUDADOS NOS CURSOS NO CONTEXTO PRÁTICO DE APLICAÇÕES REAIS

A grande parte das disciplinas cursadas não tiveram uma aplicação direta no desenvolvimento do projeto, porém, essas ditas como “teóricas” me ajudaram a desenvolver um raciocínio e um poder de abstração, suficiente para que eu pudesse aprender novas tecnologias e novas áreas do conhecimento.

13 - SE O ALUNO FOSSE CONTINUAR ATUANDO NA ÁREA EM QUE EXERCEU A IC, QUE PASSOS TOMARIA PARA APRIMORAR OS CONHECIMENTOS TÉCNICOS/METODOLÓGICOS/CIENTÍFICOS/ETC RELEVANTES PARA ESTA ATIVIDADE?

Para um maior desenvolvimento, estudaria disciplinas relacionadas às Ciências Biológicas como Biologia Celular, Bioquímica, disciplinas com enfoque mais prático, ou seja, com experimentos de bancada como, por exemplo, o curso “Biologia Molecular da Transformação Maligna”. Creio que a participação em congressos e seminários relacionados a Bioinformática também auxiliariam muito para ganhar maturidade e experiência nesta área mais acadêmica.