O BANCO DE DADOS DO "PROJETO GENOMA CAFÉ" NA EMBRAPA RECURSOS GENÉTICOS E BIOTECNOLOGIA (CENARGEN)

Raphael M. de O. B. SALES[1,2]; Alex Vicente BARBOSA[1,2]; Felipe Rodrigues DA SILVA[1], E-mail: felipes@cenargen.embrapa.br

[1]Embrapa Recursos Genéticos e Biotecnologia, Brasilia, DF, [2]Universidade de Brasília, Brasilia, DF.

Resumo:

Projetos Genoma EST são uma forma relativamente barata de inventariar os genes de um organismo. O grande volume de dados gerado por eles, entretanto, torna sua análise complicada. O Projeto Genoma Café brasileiro gerou 218.150 seqüências EST de *C. arabica, C. canephora* e *C. racemosa*. A análise realizada na Embrapa Recursos Genéticos e Biotecnologia descartou 63.380 seqüências e agrupou as 154.770 restantes em 45.366 grupos (UniGenes). Este trabalho apresenta o ferramental de análise disponibilizado até o momento aos usuários do projeto. Acessando o endereço https://alanine.cenargen.embrapa.br/cafEST/ e utilizando a senha designada pela coordenação do projeto, o usuário pode localizar rapidamente os dados de interesse partindo de diversos parâmetros, incluindo o nome do UniGene, o nome do clone seqüenciado, ou ainda palavras presentes na descrição de proteínas homólogas às do projeto. Buscas mais elaboradas incluem listar UniGenes específicos ou preferencialmente expressos em um conjunto de bibliotecas.

Palavras-chave: genoma, EST, bioinformática, banco de dados de DNA, *Coffea*.

COFFEE GENOME PROJECT DATABASE AT EMBRAPA GENETIC RESOURCES AND BIOTECHNOLOGY (CENARGEN)

Abstract:

EST Genome Projects are a relatively inexpensive way to describe genes. The overwhelming amount of data coming from such a project complicates analyzing it. The Brazilian Coffee Genome Project generated 218,150 EST sequences from *C. arabica, C. canephora* e *C. racemosa*. The analysis done at Embrapa Genetic Resources and Biotechnology discarded 63,380 sequences and grouped the 154,770 remaining sequences on 45,366 clusters (UniGenes). This work describes the tools currently available to the project users. Going to https://alanine.cenargen.embrapa.br/cafEST/ with the username and password assigned by the project coordinators, one can promptly find the sequence of interesting by imputing data such as the UniGene name, sequenced clone name or even words used in the description of homologous protein. Powerful searched includes finding library specific or preferentially expressed UniGenes.

Key words: genome, EST, bioinformatics, DNA database, *Coffea*.

Introdução

Projetos Genoma com estratégia de seqüenciamento EST (*Expressed Sequence Tags*) são baseados no seqüenciamento rápido e incompleto apenas dos genes expressos de um organismo. Internacionalmente, já foi gerado um grande número de ESTs de plantas de importância agronômica como a soja, o milho, o arroz, a alfafa e o tomate, além da planta modelo Arabidopsis (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Projetos genomas do tipo EST têm como vantagem um custo menor do que os projetos de seqüenciamento completo e seus dados são focados na seqüência codificadora (e, por conseguinte, na proteína), evitando-se o seqüenciamento de regiões inter-genicas e íntrons. Entretanto, como o nível de expressão dos genes de um organismo tem uma amplitude grande, sendo uns mais expressos que outros, em projetos deste tipo os genes muito ativos são clonados e seqüenciados varias vezes, em contraste com genes expressos com menor intensidade.

O Projeto Genoma Café, executado pela rede AEG-Fapesp e a Embrapa Recursos Genéticos e Biotecnologia, gerou 218.150 seqüências EST de *Coffea arabica, C. canephora* e *C. racemosa*. Estas seqüências são oriundas de clones de 49 bibliotecas de cDNA provenientes de diferentes tecidos ou órgãos das plantas de café (folhas, raízes, flores, sementes, frutos etc) em diferentes estádios de desenvolvimento e/ou sob estresses bióticos e abióticos. De sorte a permitir que os usuários do projeto consigam identificar rapidamente genes de interesse nesta montanha de dados, este trabalho apresenta as ferramentas já implementadas na base de dados hospedada na Embrapa Recursos Genéticos e Biotecnologia.

Material e Métodos

Os cromatogramas referentes ao seqüenciamento automático dos clones do Projeto Genoma Café foram analisados utilizando-se o software phred (EWING et al., 1998). As seqüências geradas foram processadas, com remoção de seqüências provenientes de rRNA e eliminação de porções contendo seqüências de vetor, adaptador, caudas de poliA e

bases de baixa qualidade, e agrupadas com uso do programa CAP3 (HUANG e MADAN, 1999) para a formação do conjunto UniGene de Café, conforme metodologia estabelecida por Telles e da Silva (2001), com algumas modificações. As seqüências consenso de cada UniGene foram comparados com seqüências protéicas presentes no GenBank utilizando-se o software BlastX (ALTSCHUL *et al.*, 1997). Apenas semelhanças com *e-values* melhores que 0.00001 (10^{-5}) foram consideradas. Para cada UniGene, um máximo de 20 descrições (*i.e.*, o nome da proteína com a qual a seqüência se parece) e até 5 alinhamentos foram armazenados no banco.

O modelo de banco de dados relacional esquematizado na Figura 1 foi implementado no Sistema Gerenciador de Banco de Dados PostgreSQL (v 8.1.4) instalado em um servidor Sun V890 com 4 processadores UltraSPARC IV+ e 24 GB de RAM empregado exclusivamente para análises do Projeto Genoma Café.

As seqüências limpas de cromatogramas, as seqüências consenso dos Unigenes, a localização relativa de cada seqüência em um UniGene, as qualidades das seqüências individuais e das seqüências consenso e os resultados de Blast, originalmente arquivos no formato texto, foram importados para o banco de dados, através de scripts desenvolvidos usando a linguagem Perl. A interface Web foi desenvolvida nas linguagens PHP e Perl rodando sobre o Apache para permitir a usuários acesso aos dados de maneira simplificada e rápida. A escolha destas ferramentas se deve ao fato de serem todas de código livre, permitindo personalizações, se necessárias, e por não agregarem nenhum vinculo de licença.

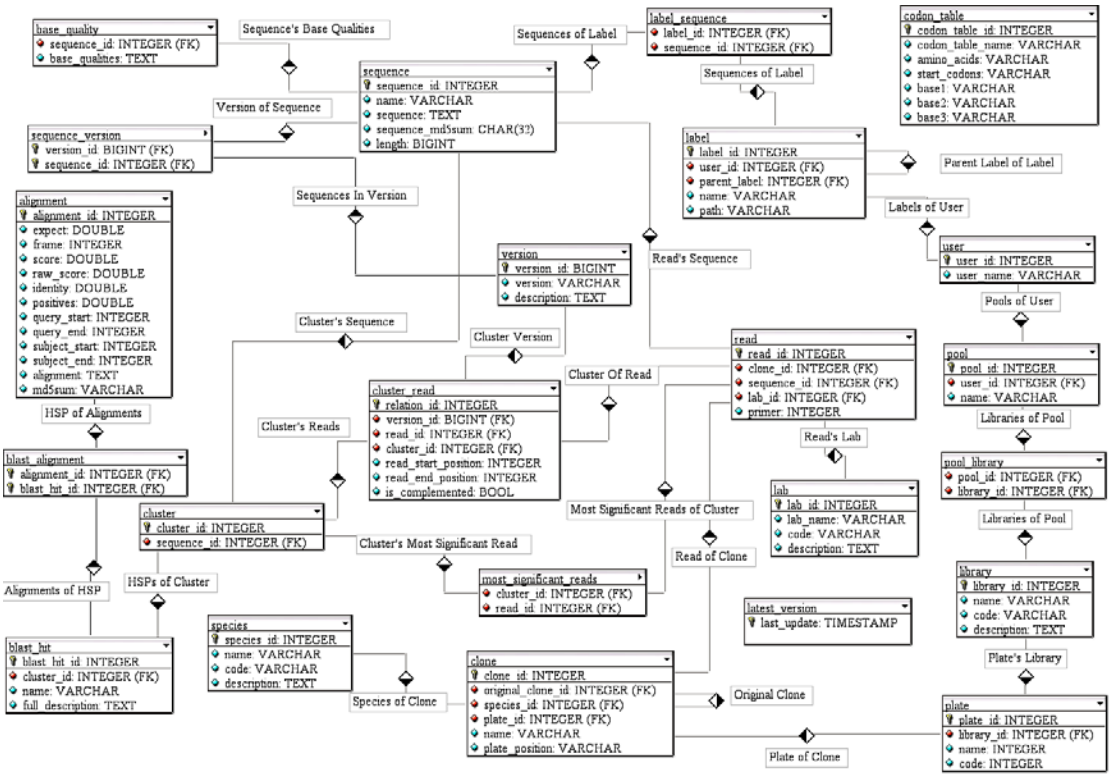

Figura 1- Esquema do banco de dados relacional implementado para a armazenagem dos dados.

Resultados e Discussão

Desde o dia 7 de março de 2007 os 67 usuários (de 8 instituições em 5 estados brasileiros) cadastrados para acesso aos dados gerados pelo Projeto Genoma Café dispõem de um novo sistema para análise de dados. O servidor, instalado em novembro no Laboratório de Bioinformática da Embrapa Recursos Geneticos e Biotecnologia, permite acesso às seguintes funcionalidades.

Sistemas de busca

Na parte superior direita da tela há um campo para busca de dados no banco (Figura 2, seta A). Abaixo dele, um menu *pull-dow* (Figura 2, seta B) que permite escolher 4 categorias de busca:Cluster, Read, Palavra-Chave e Biblioteca-Específica.

Cluster

Escolhendo-se Cluster neste menu, pode-se encontrar UniGenes pelo nome. Se um valor numérico for inserido ele procurará pelo UniGene com nome Contig<o número inserido>. Isso é, buscar por 25 mostrará o Contig25.

Read

Escolhendo-se Read efetua-se a procura por nome (ou parte do nome) de Reads. Se for procurado SB7-A0, o resultado será uma lista contendo todos os Reads que tenham no nome a palavra digitada. Para cada Read encontrado é mostrado e também o UniGene do qual ele faz parte. Caso o número de Reads seja grande demais, o resultado é dividido em páginas, com 15 resultados por página. A partir da tela de resultado, é possível ir para outra tela onde é mostrada a seqüência do Read com as qualidades de base e também a tradução nos 6 frames dessa seqüência.

Palavra-Chave

Escolhendo-se Palavra-Chave pode-se buscar palavras presentes na descrição das proteínas que apresentaram similaridade com UniGenes, ou seja, procurar no resultado do BLASTX das seqüências do projeto. Cada UniGene do projeto já foi comparado com as proteínas presentes no GenBank usando o programa BLAST. O resultado dessas comparações é guardado no banco de dados.

É possível empregar expressões booleanas na busca, utilizando-se os operadores AND, OR e parênteses. Desta forma, a busca pela palavra "photolyase" retorna todos os UniGenes que apresentaram similaridade com uma proteína que contenha esta palavra na descrição. Já a busca por "photolyase AND Dunaliella" retorna apenas os contigs que têm hit com as fotoliases da espécie Dunaliella (ou que por qualquer outro motivo contenha estas duas palavras na descrição). Finalmente, a busca "photolyase AND (Dunaliella OR Cucumis)" retorna o conjunto de UniGenes que apresentam hits com fotoliases de Dunaliella somados aos que apresentam hits com fotoliases de Cucumis.

Note que as aspas empregadas aqui no texto não podem ser digitadas no campo de busca e que um conjunto de palavras separadas por espaços é tratado como uma frase. Ou seja, buscar "Photosystem I reaction center subunit V" retorna apenas os ss que têm hit com uma proteína que apresenta esta frase (as mesmas palavras, na mesma ordem), o que é diferente de buscar "Photosystem AND I AND reaction AND center AND subunit AND V" que vai trazer hits que contenham estas palavras em qualquer ordem.

Biblioteca-Específica

Finalmente, a busca por Biblioteca-Específica busca UniGenes por Biblioteca(s). Ele recebe o nome de 1 ou mais Bibliotecas e mostra todos os UniGenes que tem pelo menos um Read de cada uma das Bibliotecas procuradas não possuindo Reads de qualquer outra. Por exemplo, em um projeto hipotético formado por apenas três bibliotecas, A , B e C, se o usuário procurar "B" (sem aspas), o sistema mostrará todos os UniGenes formados exclusivamente por reads da biblioteca B, descartando quaisquer UniGenes contendo reads das demais bibliotecas. Se a busca for feita por "A B" (sem aspas), o sistema mostrará todos os UniGenes que têm ao mesmo tempo Reads proveniente da biblioteca A e da biblioteca B, mas que não possuem nenhum read proveniente da biblioteca C. Isso permite fazer testes básicos de expressão, como quais UniGenes só são expressos em uma determinada biblioteca, ou fazer qualquer tipo de combinação de bibliotecas para verificar quais UniGenes são expressos em todas elas.

Note que outras buscas são possíveis. Para encontrar todos os UniGenes que contém pelo menos um Read de uma biblioteca (independente da inclusão de reads de outras bibliotecas), basta escolher Listar Bibliotecas Disponíveis no menu e clicar na biblioteca de interesse. A página seguinte mostrará os Contigs desta biblioteca organizados por número de reads da biblioteca. Desta forma, é possível que um UniGene apontado como tendo um read quando se olha a lista de uma biblioteca seja formado por 40 reads: 1 da biblioteca em questão e 39 de outras bibliotecas. Outra busca possível é a por todos os UniGenes que contém reads de um conjunto de bibliotecas (a união, ao invés de intersecção). Para tanto, crie um grupo no Teste de Fischer (veja adiante) e escolha o ícone da lupa.

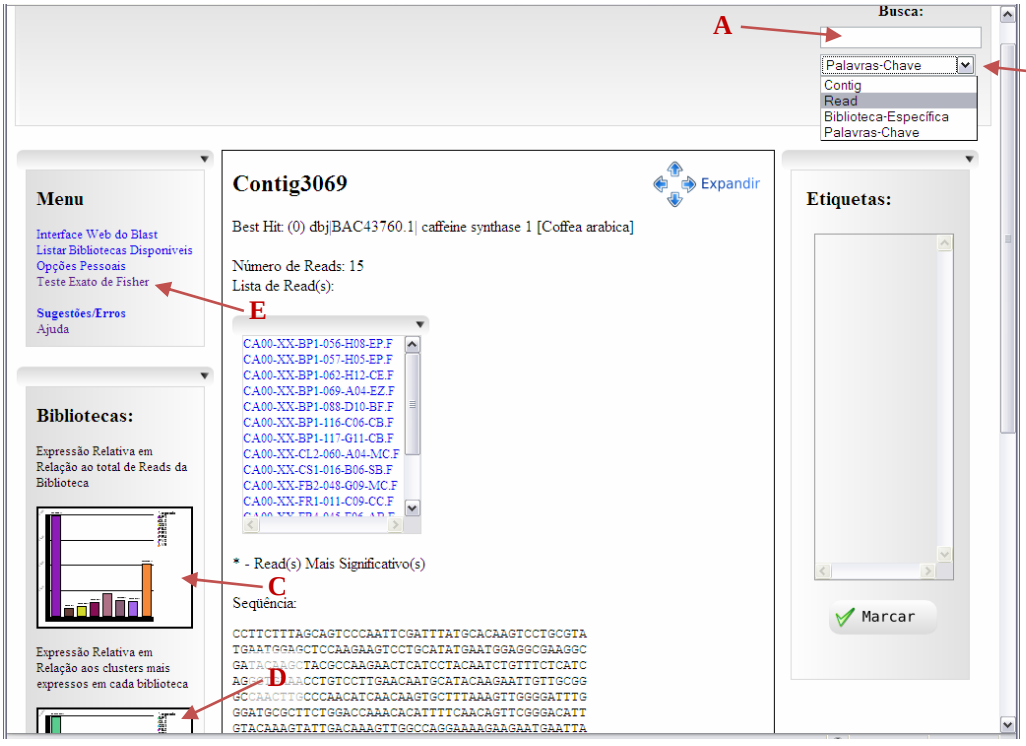

Figura 2- Exemplo da interface gráfica de busca no banco de dados.

Sistema de gráficos

Apenas listar os Reads que compõem o UniGene não responde muito sobre a expressão relativa do UniGene. Se um UniGene possuir 5 Reads da Biblioteca A e 5 da Biblioteca B pode-se ter a impressão errada de que ele tem expressão

similar em ambas as Bibliotecas. Mas se a Biblioteca A for formada por 1000 Reads e a B por 500, pode-se inferir que o transcrito representado pelo UniGene é mais expresso na Biblioteca B. Os gráficos permitem que a informação de expressão relativa seja mostrada de maneira rápida e intuitiva.

A página que mostra o UniGene apresenta 2 gráficos à esquerda (no formato de visualização "Reduzido"). O primeiro gráfico (Expressão Relativa em Relação ao total de Reads da Biblioteca, Figura 2, seta C) mostra a razão entre o número de Reads da Biblioteca presentes no UniGene em relação ao número total de Reads da Biblioteca. O segundo (Expressão Relativa em Relação aos UniGenes mais expressos em cada biblioteca, Figura 2, seta D, visível apenas parcialmente na figura) calcula a razão em relação não ao total de reads da biblioteca, mas em relação ao "transcrito mais expresso" da biblioteca. O "transcrito mais expresso" é o UniGene que contém mais reads daquela biblioteca. Desta forma, cada gráfico emprega uma normalização diferente. O primeiro nos dá uma idéia da abundância do transcrito em relação à biblioteca como um todo e o segundo de como ele se comporta com relação ao transcrito mais expresso desta mesma biblioteca.

Teste Exato de Fisher (Northern Blot eletrônico)

Apesar dos gráficos de expressão relativa serem úteis para se ter uma idéia da diferença de expressão de um UniGene nas Bibliotecas, eles podem mostrar diferenças que não tem significância estatística, ou seja, para o tamanho da amostragem, aquela diferença pode ser apenas um artefato de amostragem, e a expressão das seqüências ser na verdade muito próxima. Para lidar com estes desvios de amostragem, inerentes a este tipo de projeto, foi desenvolvida uma forma de aplicar o Teste Exato de Fisher (Fisher, 1922) para determinar quais diferenças são significativas.

Além disso, tratar diversas bibliotecas como se fossem uma só pode ser extremamente vantajoso para responder algumas perguntas biológicas. Por exemplo, o projeto genoma café tem 9 bibliotecas de *C. arabica* feitas a partir de folhas. Algumas induzidas por hormônios, outras não. Esta implementação do Teste Exato de Fisher permite que diversas bibliotecas sejam agrupadas. Desta forma, além de encontrar expressões significativamente distintas entre 2 bibliotecas, é possível encontrar expressões distintas entre 2 grupos de bibliotecas.

Ao ser acessada pela primeira vez (Figura 2, seta E), a página do Teste Exato de Fisher mostra uma lista com todas as bibliotecas do projeto. Para aplicar o Teste, é necessário antes criar 2 grupos (algo como controle e tratamento). Para criar o primeiro grupo, seleciona-se uma ou mais bibliotecas da lista, digita-se um nome para o grupo no campo no topo da lista e clica-se no botão "Criar Grupo" no final da lista. O nome do grupo recém criado será mostrado no quadro do lado esquerdo. O mesmo procedimento deve ser efetuado para a criação de cada grupo. É possível a criação de vários grupos, mas apenas 2 podem ser comparados de cada vez. Os ícones à direita do nome do grupo permitem (i) visualizar o grupo (listando as bibliotecas incluídas e gerando um relatório como se o grupo todo fosse uma única biblioteca), (ii) editar o grupo (alterando seu nome ou as bibliotecas que o compõem) e (iii) descartar o grupo (apagar este grupo de sua lista de grupos).

Para realizar o Teste Exato de Fisher, clica-se "Comparar Grupos" no quadro da esquerda e seleciona-se dois dentre os grupos criados anteriormente para serem comparados. O campo "Divisor:" permite escolher a significância da comparação. Quanto menor o número, mais signficativa a diferença. Após selecionados os 2 grupos um clique no botão "Comparar Grupos" inicia efetivamente o Teste.

O resultado do Teste são os UniGene nos quais a distribuição dos reads dos 2 grupos foge da "regra geral". Ou, no jargão biológico, quais genes estão super-expressos em um grupo quando comparado com o outro.

Os grupos criados, modificados ou apagados nesta ferramenta devem ser salvos antes do encerramento da sessão (antes de fechar o navegador), ou serão perdidos. Para tanto, utiliza-se a opção Gravar Grupos e confirme sua opção na nova página.

Busca por similaridade de seqüência (Blast local)

Clicando-se em Interface Web do Blast no menu à esquerda (Figura 2, seta E) pode-se partir de uma seqüência conhecida e localizar todas as seqüências de UniGenes similares presentes no banco de dados. Podem ser feitas buscas de blastN (nucleotídeo/nucleotídeo), TblastN (proteína/nucleotídeo traduzida) ou TblastX (nucleotídeo traduzida/nucleotídeo traduzida). O resultado, apresentado no formato tradicional de buscas de Blast via web, têm os resultados ligados diretamente ao banco de dados. Desta forma, ao se encontrar a seqüência desejada, com mais um clique se está na página dela.

Sistema de Etiquetas

Apesar dos sistemas de buscas serem eficientes, tornou-se evidente que refazer várias vezes as mesmas buscas não é a melhor opção. Para evitar a repetição do trabalho, é dada ao pesquisador a opção de criar Etiquetas para organizar as seqüências da maneira que achar melhor. As etiquetas têm comportamento similar à estrutura de diretórios encontrada em vários sistemas operacionais. Diferentemente do sistema de pastas, entretanto, o usuário pode marcar uma mesma seqüência em uma ou mais etiquetas.

Dessa forma, quando o pesquisador desejar visualizar alguma seqüência novamente, basta olhar na etiqueta que ele criou, sem precisar buscar novamente. Para criar etiquetas, segue-se o *link* Opções Pessoais do menu à esquerda (Figura 2, seta E). No instante da criação, além do nome, deve-se indicar se esta será uma etiqueta criada na raiz (*i.e.*, na parte mais alta da hierarquia) ou se ela será uma sub-etiqueta (*i.e.*, uma etiqueta filha de outra). Para criar uma sub-etiqueta selecione a opção Escolher Etiqueta Pai. Ao se clicar no botão Criar, serão apresentadas todas as etiquetas já existentes, e pode-se selecionar a etiqueta-pai da nova etiqueta. Qualquer etiqueta pode se escolhida como pai, permitindo, assim, a criação de quantos níveis hierárquicos forem necessários para sua organização pessoal. As etiquetas de um usuário só são visíveis para ele mesmo.

Etiquetas criadas ou apagadas assim como as seqüências associadas com elas precisam ser explicitamente salvas antes do encerramento da sessão (antes de fechar o navegador), ou serão perdidas. Para tanto, utiliza-se a opção Salvar Etiquetas e confirma-se esta opção na nova página.

Referências Bibliográficas

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25(17):3389-402.

Ewing, B., Hillier, L., Wendl, Mc & Green, P.: (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8: 175-185.

Fisher, R.A. (1922). "On the interpretation of χ^2 from contingency tables, and the calculation of P". *Journal of the Royal Statistical Society* 85(1):87-94.

Huang, X. & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Res* 9 (9): 868-877.

Telles, G.P. & da Silva, F.R. (2001). Trimming and clustering sugarcane ESTs. *Gen. Mol Biol* 24 (1-4): 17-23.