DETERMINAÇÃO DO UNIGENE DO PROJETO GENOMA CAFÉ

Raphael M. O. B. SALES[1] e-mail: raphael.melo21@gmail.com, Alan C. ANDRADE[1] e Felipe R. DA SILVA[1]

[1]Embrapa Recursos Genéticos e Biotecnologia, Brasilia, DF.

Resumo:

O Projeto Genoma Café gerou seqüências parciais de mais de duzentos mil clones de EST (*Expressed Sequence Tag*). Essa estratégia gera dados redundantes. Nesse trabalho, selecionamos o conjunto mínimo de clones que representam todos os transcritos encontrado no projeto. Para tanto, as 213.157 seqüências geradas pelo projeto, apos um processo criterioso que resultou em 145.507 seqüências limpas, foram agrupadas por similaridade dando origem a 32.958 possíveis transcritos, aqui chamados de Unigenes. Para cada Unigene, determinamos o clone correspondente à extremidade 5' o que, pela metodologia empregada na construção das bibliotecas, deve corresponder ao clone de maior extensão. Todos os resultados obtidos foram centralizados e organizados em uma base de dados relacional, de forma a facilitar sua utilização em posteriores aplicações de diferentes plataformas e linguagens. O SGDB usado foi o PostgreSQL. Desenvolvemos uma interface Web usando as linguagens PHP e Perl rodando sobre o Apache para permitir a usuários acesso aos dados de maneira simplificada e rápida. Escolhemos essas ferramentas por serem todas de código livre, permitindo personalizações, se necessárias, e por não agregarem nenhum vinculo de licença.

Palavras-chave: Bioinformática, EST, Genômica, *Coffea*, Bancos de Dados.

Abstract:

The Coffee Genome Project has generated partial sequences in excess of two hundred thousand EST (*Expressed Sequence Tag*) clones. This approach generates redundant data. In this work, we have devised the minimal clone set that represents all transcripts found in the project. The 213,157 sequences generated by the project were submitted to an elaborated cleaning process that resulted in 145,507 trimmed sequences. Those trimmed sequences where grouped by similarity in 32,958 putative transcripts, here called Unigenes. For each Unigene, we have picked the clone in the 5' edge, which should correspond to the one with the largest insert, due to the methodology used in library construction. All data was organized in a single relational database, allowing its use by future applications in diverse platforms and languages. The RDBMS in this work is PostgreSQL. An Web interface, using PHP and Perl over Apache was developed, allowing users fast and simple access to the data. We have chosen to work with open source tools because it allows us to make customizations, if necessary, and due to its free license and distribution policy.

Key words: Bionformatics, EST, Genomics, *Coffea*, Data Bank.

Introdução

O desenvolvimento de tecnologias aplicadas à biologia e genética molecular vem propiciando uma produção enorme de informações na área de genomas. O seqüenciamento do genoma total de organismos superiores é tecnicamente complicado devido não somente à sua maior complexidade, mas também ao seu tamanho. Em vegetais superiores a tarefa pode ser ainda mais complicada devido às porções de DNA repetitivo, que pode corresponder a 70 % do total de DNA genômico. O EST (*Expressed Sequence Tags*) é uma tecnologia de seqüenciamento genético, rápido e incompleto, onde apenas os genes expressos pelo organismo são seqüenciados. Internacionalmente, um grande número de ESTs de plantas de importância agronômica como a soja, o milho, o arroz, a alfafa, o tomate, vem sendo registrados, além da planta modelo *Arabidopsis* (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).

O Projeto Genoma Café, executado pela rede AEG-Fapesp e a Embrapa – Recursos Genéticos e Biotecnologia, tem por objetivo o seqüenciamento e a anotação de 20.000 genes de *Coffea spp.*, a partir da análise de 200.000 sequências de ESTs, obtidas de várias 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.

Projetos genomas com estratégia de seqüenciamento de EST tem como vantagem um custo menor do que os de seqüenciamento completo e seus dados são mais interessantes para projetos focados na seqüência codificadora, pois não ocorre seqüenciamento de áreas inter-genicas e das regiões de íntrons. No entanto, cada gene tem um nível de atividade diferente, sendo uns mais expressos que outros. Isso faz com que o genes muito ativos sejam clonados varias vezes, em contraste com genes expressos com menor freqüência. Um dos desafios de um projeto com essa estratégia é evitar essa redundância para aprimorar o uso dos dados obtidos pelo mesmo.

Neste trabalho, descrevemos a metodologia empregada na identificação do número transcritos presentes no conjunto total dos dados gerados pelo Projeto Genoma Café, assim como o processo de seleção do clone que melhor representa cada um dos transcritos.

Métodos:

Os cromatogramas referentes ao seqüenciamento automático dos clones do Projeto Genoma Café, obtidos junto ao Laboratório de Bioinformática da Embrapa Recursos Geneticos e Biotecnologia, 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.

Criamos um esquema de banco de dados para guardar as seqüências e representar os relacionamentos entre elas. Para representar os elementos envolvidos na camada de aplicação, definimos as seguintes entidades no sistema: Cluster (grupo), HSP, Biblioteca, Laboratório, Placa, Clone e Espécie. Cluster representa um resultado de agrupamento. HSP representa a comparação de um Cluster ou um Clone contra o banco protéico do Genbank. Biblioteca identifica o conjunto de clones de mesma origem. Placas são a oganização física dos clones, em grupos de 96, ou 384, identificados por coordenadas. Laboratório identifica o laboratório onde foi seqüenciada a placa. A estrutura das tabelas do banco de dados está representada na Figura 1.

As seqüências limpas de cromatogramas, as seqüências consenso dos Unigenes, a localização relativa de cada clone em um Unigene 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.

Para determinar o clone mais representativo de um Unigene, os seguintes casos foram tipificados sequencialmente:

a) Se o Unigene é formado por apenas um clone, este clone é seu representante.
b) Se o Unigene apresenta similaridade com proteínas nos quadros de leitura positivos e houver um clone com orientação direta na extremidade 5', este clone é seu representante.
c) Se o Unigene apresenta similaridade com proteínas nos quadros de leitura negativos e houver um clone com orientação reversa na extremidade 3', este clone é seu representante.
d) Se o Unigene apresenta similaridade com proteínas nos quadros de leitura positivos, é formado por clones nas 2 orientações e não houver um clone com orientação direta na extremidade 5', o clone com orientação direta mais próximo da extremidade 5' é seu representante.
e) Se o Unigene apresenta similaridade com proteínas nos quadros de leitura positivos, e todos os clones que o formam apresentam orientação reversa, o clone da extremidade 3' é seu representante.
f) Se o Unigene apresenta similaridade com proteínas nos quadros de leitura negativos, é formado por clones nas 2 orientações e não houver um clone com orientação reversa na extremidade 3', o clone com orientação reversa mais próximo da extremidade 3' é seu representante.
g) Se o Unigene apresenta similaridade com proteínas nos quadros de leitura negativos, e todos os clones que o formam apresentam orientação direta, o clone da extremidade 5' é seu representante.
h) Se o Unigene não apresenta similaridade com proteínas, e todos os clones que o formam apresentam orientação direta, o clone mais próximo da extremidade 5' é seu representante.
i) Se o Unigene não apresenta similaridade com proteínas, e todos os clones que o formam apresentam orientação reversa, o clone mais próximo da extremidade 3' é seu representante.
j) Se o Unigene não apresenta similaridade com proteínas, é formado por clones nas 2 orientações, e nas duas extremidades encontram-se clones com orientação direta, o clone da extremidade 5' é seu representante.
k) Se o Unigene não apresenta similaridade com proteínas, é formado por clones nas 2 orientações, e nas duas extremidades encontram-se clones com orientação reversa, o clone da extremidade 3' é seu representante.
l) Se o Unigene não apresenta similaridade com proteínas, é formado por clones nas 2 orientações, na extremidade 5' há um clone com orientação direta, e na extremidade 3' há um clone com orientação reversa, ambos se tornam representantes deste Unigene.
m) Se o Unigene não apresenta similaridade com proteínas, é formado por clones nas 2 orientações, na extremidade 5' há um clone com orientação reversa, e na extremidade 3' há um clone com orientação direta, o clone de orientação direta mais próximo da extremidade 5' e o de orientação reversa mais próximo da extremidade 3' tornam-se representantes deste Unigene.

O processo seqüencial pode ser melhor compreendido na Figura 2.

Resultados:

No final do processo de limpeza, 67.650 seqüências foram descartadas. As 145.507 seqüências mantidas foram agrupadas em 32.958 Unigenes, dos quais 13.928 são formados por apenas um clone (caso (a)). A existência de 151 casos dos tipos (l) e (m) gerou um total de 33.109 clones representando os 32.958 Unigenes do projeto. O total de clones escolhido em cada um dos casos tipificados encontra-se na Figura 2.

Com o banco populado, construímos uma interface Web simples, usando a linguagem PHP, que permite ao usuário realizar buscas diversas. A partir do nome de um clone qualquer do projeto, por exemplo, é possível se chegar ao Unigene que o representa. A partir do nome do Unigene, pode-se saber que clones o compõem. Pode, ainda, encontrar todos os unigenes que contenham palavras presentes nas HSPs, inclusive com operações booleanas de "E" e "OU" e uso de parênteses para combinar palavras. Finalmente, é possível, rapidamente, encontrar todos os Unigenes formados por qualquer combinação de bibliotecas, *i.e.*, aqueles formados exclusivamente por clones provenientes de uma única biblioteca, ou por qualquer combinação delas.

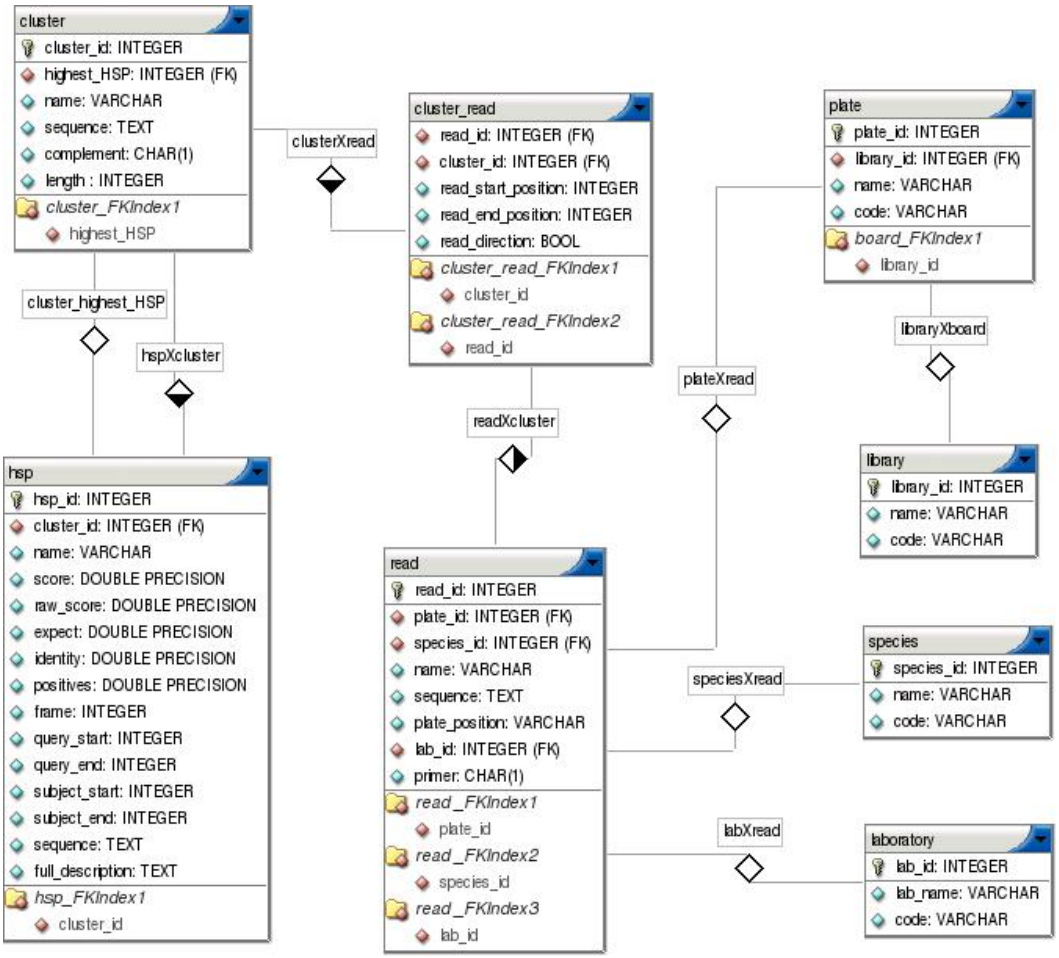

Figura 1. O esquema de banco de dados relacional empregado na determinação do clones que compõem o Unigene do Projeto Genoma Café.

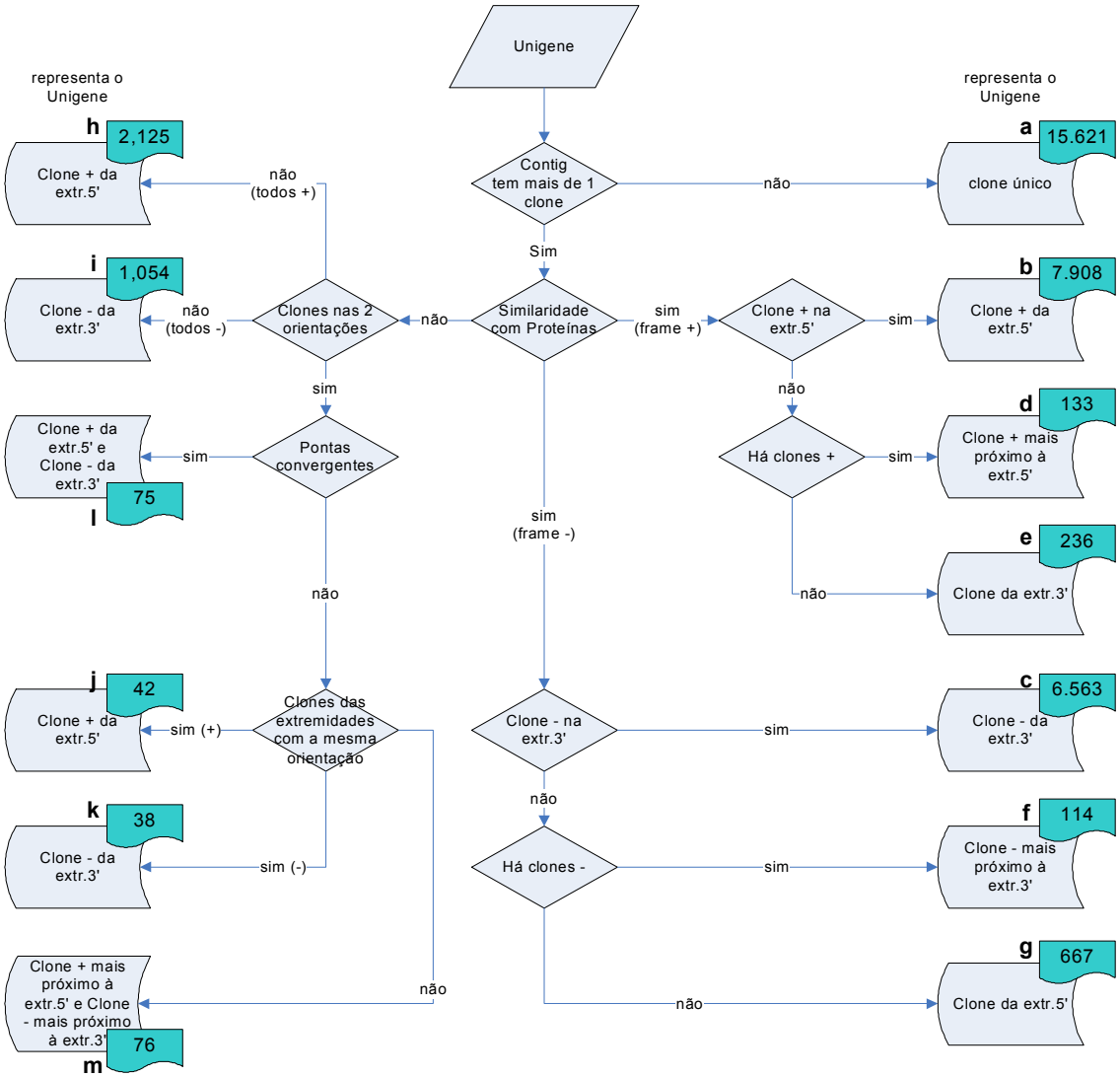

Figura 2. Fluxo decisório de escolha do clone mais representativo de um Unigene. As letras indicam os casos tipificados em *Métodos*. Os números indicam a quantidade de clones selecionados por cada um dos critérios.

Referências bibliográficas

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.

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

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