UM *SOFTWARE* PARA EXTRAÇÃO DE ESTs, *CONTIGS* E *SINGLETS* DO CAFEST[1]

Rafael Luciano Guerra[2]; Samuel Mazzinghy Alvarenga[3]; Eveline Teixeira Caixeta[4], Carlos de Castro Goulart[5]

[1]Trabalho financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).
[2]Graduando em Ciência da Computação, Universidade Federal de Viçosa, Bolsista de Iniciação Científica PIBIC/CNPq, faelwar@gmail.com
[3]Doutorando em Genética e Melhoramento, Universidade Federal de Viçosa, Viçosa-MG, samalvarenga@gmail.com
[4]Pesquisadora, D.Sc., Embrapa Café, Brasília-DF, eveline.caixeta@embrapa.br
[5]Professor, D.Sc., Universidade Federal de Viçosa, goulart@dpi.ufv.br, autor para correspondência

RESUMO: Uma das dificuldades encontrada atualmente para os usuários da base de dados do Projeto Brasileiro do Genoma Café (CafEST) é a dificuldade em extrair manualmente todas as sequências armazenadas em seus projetos, no sistema *Gene Projects*. A partir dessa necessidade, foi desenvolvido um *software* para fazer a automatização do processo. Desta forma, o usuário do CafEST passa a dispor da comodidade de apenas informar ao programa quais projetos ele deseja que tenham as ESTs, *contigs* ou *singlets* extraídos. O programa consiste em dois *scripts*, um para fazer a extração das ESTs e outro para fazer a extração dos *contigs* e dos *singlets*. Os dois *scripts* retiram do CafEST apenas as ESTs, os *contigs* e *singlets* e salvam cada um desses em um arquivo do tipo FASTA. Isto facilita o uso dos mesmos em outras bases de dados e/ou ferramentas de bioinformática que normalmente trabalham com esse tipo de arquivo. Os *scripts* podem ser obtidos mediante solicitação por email.

Palavras-Chave: Programação *Script*, Genoma Café, ESTs.

A SOFTWARE FOR EXTRACTION OF ESTs, CONTIGS AND SINGLETS FROM CAFEST

ABSTRACT: A major problem currently faced by the Brazilian Coffee Genome Project Database (CafEST) users is the difficulty in manually extract all the sequences stored in their projects from Gene Projects System. Based on this demand, we have developed a computer software to automate the sequence extraction process. Thus, the CafEST user has now the convenience of just informing the program from which project the ESTs, contigs or singlets are supposed to be extracted. The program consists of two scripts, one for extracting ESTs and the other to extract contigs and singlets. Both scripts withdraw the ESTs, contigs and singlets from CafEST and save each one of them in a FASTA file. It makes them easy to be used in other databases and/or bioinformatics tools which commonly work with this sort of file. The scripts can be obtained by email request.

Key words: Script Programming, Coffee Genome, ESTs.

INTRODUÇÃO

Atualmente a pesquisa na área genômica do café, levou a criação de uma enorme base de dados no Brasil, utilizada para armazenamento de sequências, realização de busca por similaridade via BLAST e clusterização de *reads*. Essa base é conhecida como CafEST e está disponível em http://www.lge.ibi.unicamp.br/cafe/ (Vieira et al., 2006).

A base de dados é dotada de diferentes aplicativos, mas os usuários não têm a opção de fazer o *download* das suas ESTs, *contigs* e *singlets* mineradas pelo sistema *Gene Projects* (Carazzolle et al., 2007). Desta forma, não podem realizar qualquer tipo de análise de bioinformática em outras bases de dados ou utilizar outras ferramentas computacionais. A única forma de fazer isso era acessando a página referente a cada EST, *contig* ou *singlet*. Essa atividade, na maioria dos casos, requer muito tempo, pois, para visualizar os *contigs* e os *singlets* de um projeto do sistema *Gene Projects* é necessário carregar todas as sequências. Dependendo da conexão, isto poderia levar muito tempo e ainda ser necessária a presença do usuário em frente ao computador, acompanhando todo o processo.

Outra dificuldade muitas vezes encontrada pelos usuários ao tentar armazenar as informações em seus computadores é o erro humano, como o simples erro de "pular" uma ou mais sequências, esquecendo-se de armazená-las. Assim o usuário ficaria sem as sequências esquecidas em seu projeto.

O objetivo desse trabalho foi facilitar a extração e o armazenamento de sequências, economizando o tempo, gerando arquivos organizados, criando uma aplicação de fácil uso e que possua o código aberto. Com isso, disponibilizamos um *software* que qualquer usuário que tenha acesso ao CafEST possa utilizar e ajustar seu funcionamento para melhor atendê-lo.

MATERIAL E MÉTODOS

Para elaboração do programa, a melhor forma encontrada para uma primeira versão foi dividir o programa em dois, um para fazer a extração das ESTs do sistema *Gene Projects* (GCES) e uma segunda para fazer a extração dos *contigs* e dos *singlets* (GCECS), gerados pela clusterização desse sistema. Como a única forma encontrada para acessar a base de dados é por meio de páginas da *web*, foi necessário a utilização de uma linguagem de programação que desse suporte a algum tipo de emulação de um navegador.

Considerando esse aspecto da base de dados, a linguagem Perl (Deitel et al., 2002) demonstrou ser a mais indicada por possuir um módulo para emulação de paginas web, o WWW::Mechanize. Usando este módulo ganhamos tempo na execução do programa, considerando que o módulo está implementado de forma eficiente.

Uma dificuldade encontrada durante a criação do *software* foram as diferentes formas de armazenamento de arquivos entre diferentes sistemas operacionais. Uma alternativa encontrada para solucionar esse problema foi a criação de duas versões da aplicação, evitando, assim, possíveis erros no armazenamento das sequências feito pelo programa. Desta forma, o programa tem versões disponíveis para ser executado nos sistemas operacionais Linux (GCES_L e GCECS_L) e Windows (GCES_W e GCECS_W). Para a execução do *software* é necessário ter um interpretador de Perl instalado. Para Windows é recomendado o uso do ActivePerl 5.12 ou superior, que pode ser encontrado em <http://www.activestate.com/activeperl>. Este já possui o módulo WWW::Mechanize incluído. Em Linux é necessário instalar o programa via terminal usando os seguintes comandos: em modo de superusuário, *"apt-get install perl"* que irá instalar o interpretador da linguagem e em seguida entrar com o comando *"apt-get install libwww-mechanize-perl"*, que irá instalar o módulo WWW::Mechanize necessário para execução dos *scripts*.

Para executar a aplicação em Windows, depois de instalar o interpretador e o módulo, basta clicar no arquivo que contém o *script*, que o *prompt* de comando será inicializado automaticamente. Em Linux, o usuário deve acessar o terminal que pode ser encontrado no menu na área de trabalho. No terminal, dentro da pasta onde se encontra os *scripts*, basta digitar o comando "*perl GCES_L.pl*" para o *script* que faz a extrações das ESTs ou digitar "*perl GCECS_L.pl*" para iniciar o *script* que faz a extração dos *contigs* e *singlets*. Posteriormente, basta seguir as informações que aparecem na tela. Após o tempo de execução o usuário poderá verificar as sequências salvas em arquivos *.FASTA na pasta de destino.

No programa para extração de ESTs, primeiramente, o *script* de execução pede a identificação do usuário por meio de *login* e senha. Em sequência são mostrados todos os projetos existentes para aquele usuário. Após a escolha do projeto a ser extraído, o script abre o projeto e as ESTs nele contidas. Após conferir se alguma EST já fora extraída em outro momento, o *script* inicia a extração das ESTs. Após a extração, o *script* finaliza a execução (Figura 1).

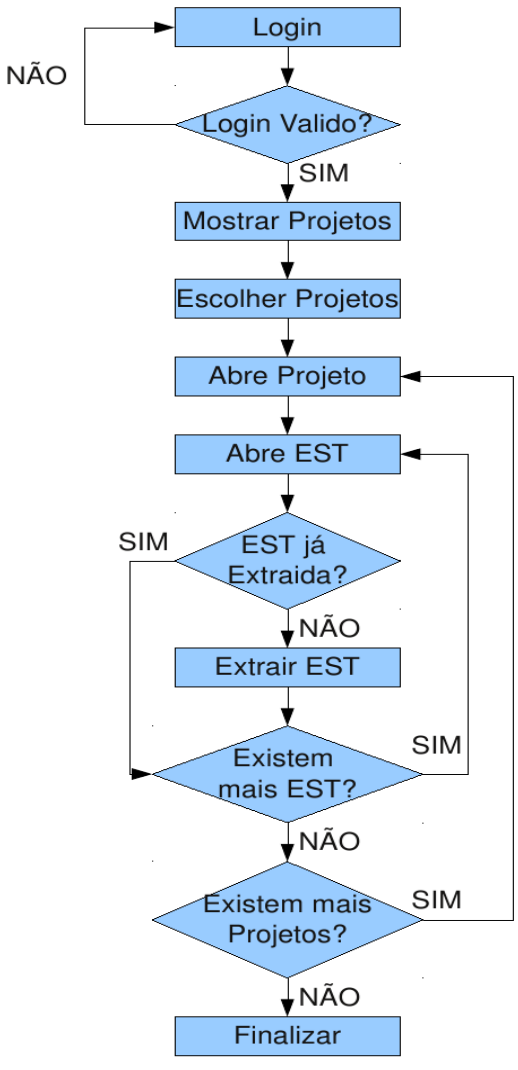

Figura 1 – *Workflow* do *script* para extração de ESTs do Sistema *Gene Projects* do CafEST.

No programa para extração de *contigs* e *singlets*,

após abrir o projeto escolhido pelo usuário, o *script* abre a página dos resultados de clusterização do sistema *Gene Projects*. Após conferir se algum *contig* ou *singlet* já fora extraído em outro momento, o *script* inicia a extração das sequências.

RESULTADOS E DISCUSSÃO

Para o caso de extração das ESTs, o *script* consiste em um sistema de *login* de usuário onde qualquer usuário cadastrado no CafEST pode acessar. Após o *login*, será apresentada ao usuário uma lista de todos os seus projetos do sistema *Gene Projects* que o programa encontrou. Assim o usuário poderá informar em quais projetos ele deseja que a extração das ESTs seja feita. Logo após esta escolha, o programa inicia a extração das ESTs, sempre informando ao usuário em qual projeto a extração está sendo feita naquele momento. Terminado o processo de extração, a janela do script fechará automaticamente. Nesse momento o usuário poderá verificar na pasta de destino as ESTs salvas em arquivos *.FASTA.

Para a extração de *contigs* e *singlets*, o procedimento é o mesmo, com a diferença que quando o usuário escolhe o projeto, serão extraídos os *contigs* e *singlets* gerados pela clusterização do sistema *Gene Projects* (Figura 2), conforme mostrado na Figura 2. Na ilustração são detalhados os passos que o usuário deve seguir: no *passo 1* o usuário deve entrar com o *login* com o qual está cadastrado no sistema *Gene Projects* do CafEST; no *passo 2*, digitar a senha (ambas as informações foram omitidas, por razões óbvias); no *passo 3*, o programa mostrará uma lista contendo todos os projetos criados por aquele usuário; finalmente, no *passo 4*, o usuário deve escolher de qual(is) o(s) projeto(s) deseja ter suas ESTs extraídas.

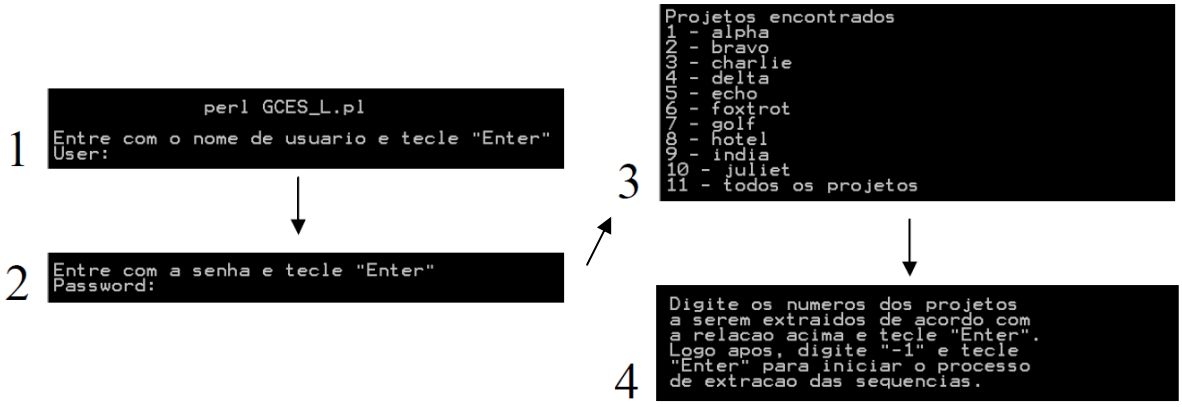

Figura 2 - Exemplo de execução da aplicação para extração de ESTs em sistema operacional Linux

Em caso de queda de energia, falha de conexão com a internet ou qualquer falha do computador, as ESTs que já foram extraídas não serão perdidas. Elas ficarão salvas no disco rígido do usuário e se este tentar extrair o projeto uma segunda vez após a falha, o programa verificará quais ESTs já foram extraídas e não necessitará realizar a extração novamente.

O *software* também informa qual EST está sendo extraída no momento e se esta se trata de uma EST que já se encontra armazenada em seu computador, considerando que o *script* será sempre executado na mesma pasta. Isto porque o *script* verifica apenas se existe uma EST de igual nome dentro da pasta criada por ele com o nome do projeto. Quando a extração de sequências de um projeto é finalizada o programa informa ao usuário e inicia o próximo projeto, se houver. Caso não haja mais projetos para extração, a aplicação é finalizada.

A relação do tempo médio de extração de cada EST para projetos de diferentes tamanhos pode ser verificada na Tabela 1. Lembrando que esses valores podem variar de acordo com a velocidade da conexão com a internet e com a configuração de *hardware* do computador.

Com esses valores podemos observar a principal vantagem de se usar o *software*: o tempo para a extração de ESTs. Esse tempo tende a reduzir com o aumento do número de ESTs armazenadas em um projeto. O tempo médio observado foi de 2,03 s para a extração de cada EST, algo que manualmente não é possível de ser feito.

Além do ganho de tempo, há também o ganho em comodidade para o usuário, que não precisa ficar em frente ao computador realizando a extração, possivelmente durante algumas horas para que o trabalho seja finalizado. Ao executar o programa, o usuário pode se tomar de outras atividades, enquanto o *script* pode extrair cerca de 1.000 ESTs em um pouco mais de 27 minutos (Tabela1).

Para o segundo caso, a extração dos *contigs* e dos *singlets* gerados a partir da clusterização do sistema *Gene Projects*, a interface da aplicação é mesma utilizada para o primeiro caso, a extração das ESTs. A diferença é a forma de armazenamento das informações retiradas da base de dados. O *software* cria uma pasta com o nome do projeto e dentro dela cria outras duas pastas, uma para colocar os *contigs* e uma segunda para guardar os *singlets*.

O programa armazena todos os *singlets* e os dois arquivos referentes a cada *contig* (*contig e icontig*), sempre verificando se o arquivo existe. Se existe significa que o *contig* ou *singlet* já foi extraído anteriormente e não é necessário realizar o procedimento de extração novamente.

Tabela 1 – Relação segundos por EST extraída e de tempo total gasto para um determinado número de ESTs, em média.

Quantidade de ESTs	Tempo por EST (em segundos)	Tempo total para extração das ESTs (em segundos)
10	4,40	44
20	4,15	83
50	2,50	125
100	1,73	173
150	1,72	259
200	1,57	314
250	1,77	443
300	1,75	527
400	1,67	670
500	1,64	819
650	1,65	1074
700	1,64	1145
1000	1,63	1631
1300	1,63	2124
1500	1,66	2486
1850	1,65	3048
2500	1,67	4171

Para realização dos testes de desempenho do programa, foram considerados dois valores importantes. O primeiro, o tempo gasto para extrair um *contig* ou *singlet* e o tempo total gasto para extrair um projeto inteiro. A Tabela 2 mostra o tempo gasto para extrair um *contig* ou um *singlet* e o tempo total gasto para extrair um projeto completamente.

Tabela 2 – Relação de tempo de extração por *contig* ou *singlet* e tempo total para extração dos *contigs* e *singlets* de um projeto pela quantidade de *contigs* e *singlets.*

Quantidade de *contigs* e *singlets*	Tempo para extração de um *contig* ou *singlet* (em segundos)	Tempo para extração de completa de um projeto (em segundos)
10	0,80	8
25	0,72	18
50	0,42	21
100	0,45	45
150	0,61	92
250	0,44	111
300	0,53	160
350	0,56	196
400	0,46	183
450	0,57	257
500	0,54	269
550	0,48	264
700	0,64	446
750	0,54	406
900	0,62	561
1800	0,53	952
3600	0,54	1957

Para os valores medidos nota-se que, na média, o tempo necessário para a extração de um *contig* ou *singlet* foi de 0,54 segundos. O tempo varia próximo a esse valor devido a quantidade de ESTs armazenadas em um projeto, pois o algoritmo é forçado a esperar que a página que contém todas as ESTs tenha sido completamente carregada para poder acessar a página que contém os *contigs* e *singlets*, e assim poder abrir cada um deles e fazer a extração.

Vale ressaltar outra vez que esses valores podem variar conforme a conexão do usuário, configuração de *hardware* e, para esse caso, o número de sequências de cada projeto pode causar variações no tempo de extração. Outra

informação que também pode influenciar na variação de tempo é a quantidade de *contigs* em um projeto. Como a extração deles é feita criando 2 arquivos diferentes, quanto maior a quantidade de *contigs* existentes, maior o tempo de execução total do programa.

Devido a eficiência do *software* é possível fazer a extração e armazenamento de aproximadamente 3600 *contigs* e *singlets* em pouco mais de 30 minutos (Tabela 2), algo que não seria possível de ser feito manualmente.

Os *scripts* podem ser obtidos mediante solicitação por email.

CONCLUSÕES

Com o resultado mostrado nos testes apresentados nas tabelas 1 e 2 para execução dos dois *scripts*, podemos notar a eficiência deles para extração das informações do CafEST. Para extração de ESTs, é possível extrair cerca 1 sequência a cada 1,7 segundos, e para *contigs* e *singlets* tem-se, em média, 1 a cada 0,55 segundos. Definitivamente, esses são valores que não são possíveis de serem alcançados ao fazer-se a extração manualmente.

Apesar do uso do *prompt* de comando para execução do *software*, o que para alguns usuários pode causar alguma dificuldade, o programa tenta ser o mais claro possível nas informações que são pedidas ao usuário. Uma característica que ainda pode ser melhorada numa próxima versão da aplicação.

REFERÊNCIAS BIBLIOGRÁFICAS

CARAZZOLLE, M. F., FORMIGHIERI, E. F., DIGIAMPIETRI, L. A., ARAUJO, M. R. R., COSTA, G. G. L., PEREIRA, G. A. G. Gene projects: A genome Web tool for ongoing mining and annotation applied to CitEST. **Genetics and Molecular Biology**, v.30, n.3, p.1030-1036. 2007.

DEITEL, H. M., DEITEL, P. J., NIETO, T. R. MCPHIE, D. C. **Perl: Como Programar**. 3ª Edição. Porto Alegre: BOOKMAN, 2002. 952 p.

VIEIRA, L. G. E., ANDRADE, A. C., COLOMBO, C. A., MORAES, A. H. A., METHA, A., OLIVEIRA, A. C., LABATE, C. A., MARINO, C. L. E C. B. MONTEIRO-VITORELLO, MONTE, D. C., GIGLIOTI, E., KIMURA, E. T., ROMANO, E., KURAMAE, E. E., LEMOS, E. G. M., ALMEIDA, E. R. P., JORGE, E. C., ALBUQUERQUE, E. V. S., SILVA, F. R., VINECKY, F., SAWAZAKI, H. E., DORRY, H. F. A., CARRER, H., ABREU, I. N., BATISTA, J. A. N., TEIXEIRA, J. B., KITAJIMA, J. P., XAVIER, K. G., LIMA, L. M., CAMARGO, L. E. A., PEREIRA, L. F. P., COUTINHO, L. H., LEMOS, M. V. F., ROMANO, M. R., MACHADO, M. A., COSTA, M. M. C., SÁ, M. F. G., GOLDMAN, M. H. S., FERRO, M. I. T., TINOCO, M. L. P., OLIVEIRA, M. V., SLUYS, M. V., SHIMIZU, M. M., MALUF, M. P., EIRA, M. T. S., FILHO, O. G., ARRUDA, P., MAZZAFERA, P., MARIANI, P. D. S. C., OLIVEIRA, R. L. B. C., HARAKAVA, R., BALBAO, S. F., TSAI, S. M., MAURO, S. M. Z., SANTOS, S. N., SIQUEIRA, W. J., COSTA, G. G. L., FORMIGHIERI, E. F., CARAZZOLLE, M. F., PEREIRA, G. A. G. 2006. Brazilian coffee genome project: an EST-based genomic resource. **Brazilian Journal of Plant Physiology**, v.18, n.1, p.95-108. 2006.