Genoma Funcional do Fruto do Guaranazeiro: Terminada a Fase de Anotação Automática

Paula Cristina da Silva Angelo[1]; Marcelo de Macedo Brígido[2]; Jorge Rebelo Porto[3]; Marcos José Salgado Vital[4]; Jean Charles da Cunha Peixoto[5]; Maria Paula Cruz Schneider[6]; Horácio Schneider[7]; Wilsea Fernandez[8]; Emygdia R. L. R. B. P. L. Mesquita[9]; Márcio Antônio da Silveira[10]; Luiz Hildebrando Pereira Silva[11]; Margarida Lima Carvalho[12]; Spartaco Astolfi Filho[13], Dra. Elionor R. Almeida[14]

Por se tratar de produto nativo da Amazônia, utilizado na alimentação, com cadeia produtiva instalada e pronta para crescer, e por ter potencial para utilização pela indústria de fitofármacos e fitocosméticos, o guaranazeiro (*Paullinia cupana* var. *sorbilis*) foi eleito como objeto de estudo da recém-formada Rede da Amazônia Legal de Pesquisas Genômicas - REALGENE. Esta Rede é coordenada pela Universidade Federal do Amazonas e dela fazem parte instituições de pesquisa e ensino da Amazônia Legal, sendo Embrapa Amazônia Ocidental, o INPA, o IPEPATRO em Rondônia, Universidade Federal do Acre, Universidade Federal de Amapá, Universidade Federal do Maranhão, Universidade Federal de Pará (três grupos), Universidade Federal de Roraima, Universidade Federal de Tocantins/UNITINS. São diretrizes do projeto de formação da Rede implantar e melhorar a infra-estrutura laboratorial em biologia molecular/genética genômica/bioinformática em Instituições de Ensino, Pesquisa e Desenvolvimento de diferentes Estados da Amazônia Legal e contribuir para a sofisticação da cadeia produtiva do guaraná a partir da aplicação dos resultados obtidos ao final da execução do "Projeto Genoma Funcional do Guaranazeiro". O objetivo desta contribuição ao I Seminário sobre Pesquisas com Guaranazeiro é divulgar que parte das metas já foi cumprida e registrar que a Genômica do guaranazeiro já dispõe de resultados.

A Embrapa Amazônia Ocidental forneceu, mediante a assinatura de acordo de cooperação técnica entre os participantes da REALGENE, o material vegetal - partes de plantas do clone registrado na Secretaria Nacional de Proteção de Cultivares, sob a denominação comercial BRS-Amazonas - para a geração das bibliotecas de cDNA e seqüenciamento do genoma funcional. Para tal, foram coletados frutos com sementes em três estádios de maturação e o RNA (Figura 1) foi extraído na Embrapa Recursos

[1]Pesquisadora III Embrapa Amazônia Ocidental. C.P. 319, CEP. 69011-970 Manaus, Amazonas. paula@cpaa.embrapa.br
[2]Professor, Dr., Universidade de Brasília; [3]Pesquisador, Instituto de Pesquisas da Amazônia; [4]Professor, Universidade Federal de Roraima; [5]Professor, Universidade Federal do Amapá; Professora, Universidade Federal do Pará; [7]Professor, Universidade Federal do Pará; [8]Professora, Universidade Federal do Pará; [9]Professora, Universidade Federal do Maranhão; [10]Professor, UNITINS; [11]Professor, Universidade Federal de Rondônia; [12]Professora, Universidade Federal do Acre; [13]Professor, Universidade Federal do Amazonas, [14]Pesquisadora III da Embrapa Recursos Genéticos e Biotecnologia.

Paula Cristina da Silva Angelo; Marcelo de Macedo Brígido; Jorge Rebelo Porto; Marcos José Salgado Vital; Jean Charles da Cunha Peixoto; Maria Paula Cruz Schneider; Horácio Schneider; Wilsea Fernandez; Emygdia R. L. R. B. P. L. Mesquita; Márcio Antônio da Silveira[0]; Luiz Hildebrando Pereira Silva[1]; Margarida Lima Carvalho[2]; Spartaco Astolfi Filho[3], Dra. Elionor R. Almeida

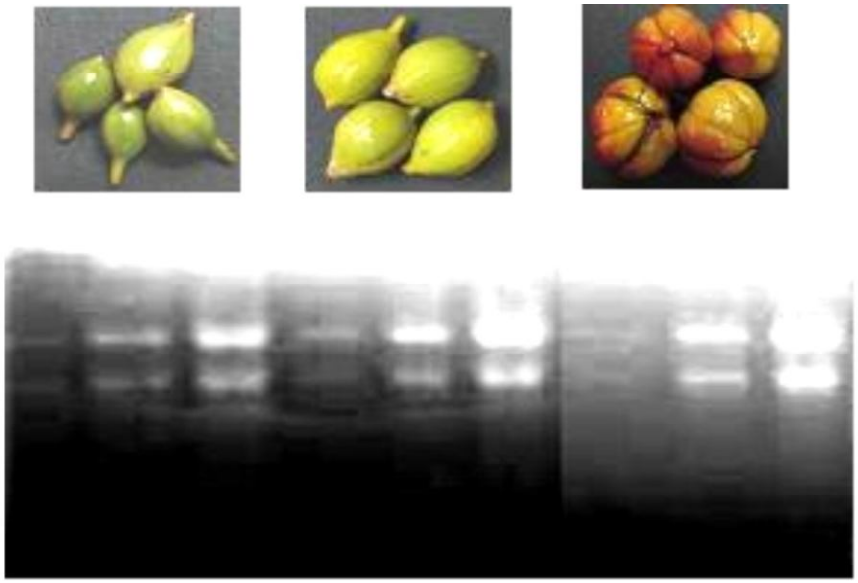

Figura 1. Frutos do guaranazeiro nos três estádios de desenvolvimento e o RNA total extraído para a construção das bibliotecas de cDNA e geração de ESTs.

Genéticos e Biotecnologia, utilizando o "Concert Plant Reagent" (Invitrogen). A síntese do cDNA foi realizada utilizando o "kit Super Script" (Invitrogen). O cDNA foi fracionado por filtração para seleção de fragmentos entre 600 e 1.000 pares de bases e clonado no vetor plasmidial pSPORT6 (Invitrogen). Os insertos foram sequenciados a partir da extremidade 5' e os arquivos de ESTs ("expressed sequence tags") foram depositados no servidor da Universidade de Brasília, podendo ser acessados, mediante senha, através da página https://www.biomol.unb.br/GR/. As seqüências de cDNA clonadas são mantidas na Universidade Federal do Amazonas.

Compõem o banco de dados do projeto 15.387 seqüências, que foram aceitas por apresentarem pelo menos 200 bases com qualidade acima de 20 (índice de qualidade do aplicativo PHRED para qualificar os picos de fluorescência dos eletroferogramas gerados pelo sequenciador automático). Após a eliminação da contaminação por pares de bases do vetor, 9.418 seqüências foram organizadas em "contigs" (grupos de seqüências que apresentam alta identidade) com o programa CAP3. Portanto, constam do banco de dados 2.628 "contigs" e 5.969 "singlets" (seqüências que foram encontradas apenas uma vez entre os clones de cDNA seqüenciados), num total de 8.597 grupos (Figura 2) ou cDNAs representados. Cada "contig" tem, em média, 760 bases.

Seguiu a esta manipulação das seqüências, a busca automática por similaridade em bancos de genes e proteínas, como o GenBank, do NCBI (National Center for Biotechnology Information - http://www.ncbi.nih.gov/), o SwissProt, do Instituto Suíço de Bioinformática (http://ca.expasy.org/sprot/) e o GO (GeneOntology http://www.godatabase.org/dev/database/).

Paula Cristina da Silva Angelo; Marcelo de Macedo Brígido; Jorge Rebelo Porto; Marcos José Salgado Vital; Jean Charles da Cunha Peixoto; Maria Paula Cruz Schneider; Horácio Schneider; Wilsea Fernandez; Emygdia R. L. R. B. P. L. Mesquita; Márcio Antônio da Silveira[0]; Luiz Hildebrando Pereira Silva[1]; Margarida Lima Carvalho[2]; Spartaco Astolfi Filho[3], Dra. Elionor R. Almeida

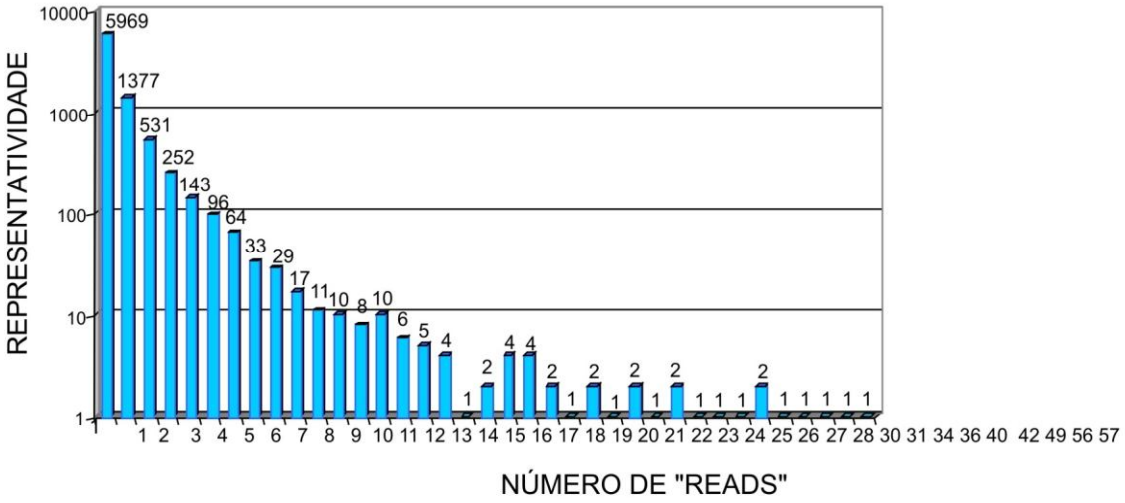

Figura 2. distribuição e número de "reads" (ESTs aceitas). A freqüência de distribuição das ESTs foi obtida após a análise com o aplicativo CAP3. A maior parte (91,6%) dos grupos têm de 1 ("singlets") a 3 ESTs ("contigs"). Existem 5.969 "singlets" e existem poucos grupos constituídos por mais de 18 ESTs.

A busca por identidade em bancos de proteínas é feita depois de terem sido deduzidas as seqüências protéicas codificadas pelas ESTs, sendo que existem seis possibilidades de tradução dos nucleotídeos em aminoácidos para cada EST. Este processo foi realizado pelo aplicativo Blast X.

O resultado da busca descrita acima foi, sempre que a similaridade apresentou-se estatisticamente bastante significativa, a anotação automática da identidade mais provável de cada "contig" e "singlet", do nome do produto gênico, da categoria funcional segundo classificação do KOG e identificação e diponibilização do número EC para as enzimas. Esta anotação automática, armazenada no servidor da UnB, vem acompanhada de "links" para os arquivos de referência que estão nos bancos consultados. Estes arquivos incluem o valor estatístico da similaridade para cada par de seqüências (seqüência de referência x seqüência de guaraná anotada), porque são testadas mais de um milhão de possibilidades de similaridade durante a busca e nem todas as anotações automática apresentam valor aceitável. Quanto maior o banco consultado mais demorada a busca e, geralmente, mais resultados com valor estatístico bom são encontrados. Quanto melhor o valor estatístico da similaridade, mais completa fica a anotação automática. Todo esse material fica à disposição do revisor humano que conclui a anotação manualmente.

Para algumas ESTs do guaraná não foram encontrados similares nos bancos de sequências. Existem, também, ESTs que apresentam mais de uma possibilidade de tradução para um mesmo grupo de nucleotídeos e, portanto, mais de uma seqüência protéica é anotada automaticamente, algumas vezes com valores estatísticos aproximados. Para

Paula Cristina da Silva Angelo; Marcelo de Macedo Brígido; Jorge Rebelo Porto; Marcos José Salgado Vital; Jean Charles da Cunha Peixoto; Maria Paula Cruz Schneider; Horácio Schneider; Wilsea Fernandez; Emygdia R. L. R. B. P. L. Mesquita; Márcio Antônio da Silveira[0]; Luiz Hildebrando Pereira Silva[1]; Margarida Lima Carvalho[2]; Spartaco Astolfi Filho[3], Dra. Elionor R. Almeida

outras, foi registrada similaridade com seqüências depositadas que ainda não apresentam função definida. Isto ocorre porque o depósito de seqüências cuja função ainda não foi desvendada tornou-se comum, desde que passaram a ser seqüenciados genomas inteiros de muitos organismos. Estas são ESTs que precisam ser analisadas com cuidado e que terão suas funções definidas, possivelmente, pelos grupos que estão vinculados a cada projeto genoma, o que contribuirá para Genômica, de maneira geral. Pelo menos uma seqüência apresentou similaridade com aquelas expressas em espécies dos mais diferentes filos, foi, portanto, aparentemente conservada ao longo do processo evolutivo e, ainda assim, sua função não está claramente definida em nenhuma espécie.

A anotação automática está, durante os meses de setembro, outubro e novembro de 2005, sendo revista e aprimorada por anotação manual. Ao longo do processo de anotação manual é possível complementar o registro existente para cada seqüência, indicando, por exemplo, que, apesar de não ter sido identificado automaticamente, um "contig" ou "singlet" contém ou codifica regiões ou "motivos", por exemplo os sítios ativos das enzimas, que por vezes são bem conhecidos e conservados e tornam-se importantes para a indicação da função da EST. Durante esta fase é também possível registrar quais seqüências têm potencial para gerar patentes. Sempre que possível, a classificação KOG (definição da classe a que pertence a proteína, se é uma proteína que participa do metabolismo de aminoácidos, ou da síntese de ácidos nucléicos, ou do transporte de elétrons e geração de energia e etc.) e o nome pelo qual a EST será reconhecida dentro dos registros estão sendo anotados manualmente. Quando não existe registro automático de classe enzimática, a referência à classe enzimática mais provável pode ser anotada manualmente.

Entre as seqüências identificadas está pelo menos uma sintetase da cafeína, o que é interessante, já que a expressão de sintetase de cafeína foi relatada poucas vezes em endosperma e os dados publicados são, na maioria, sobre a expressão destes genes em folhas. Foram também identificados possíveis carreadores de ferro para a semente, proteínas que têm afinidade por metais pesados e enzimas que são parte da via de síntese de metabólitos secundários que não a cafeína. Existem enzimas cuja estrutura é mais conservada ao longo da evolução e que são mais comuns à maioria das plantas como a RUBISCO, aquelas que participam dos processos de geração de ATP e de transporte de elétrons. Outras são comumente encontradas em eucariotos, como aquelas que participam do "splicing" do RNA. Estas últimas ficaram, geralmente, com registros automáticos muito completos. Estão anotadas manualmente cerca de 3.300 seqüências.

Com a análise que será realizada após o processo de anotação manual, será possível ter uma visão ampla dos processos metabólicos que ocorrem no fruto do guaranazeiro.

Agradecimentos

Ao CNPq/MCT, pelo financiamento do projeto.

O número de pessoas que têm contribuído para a geração dos dados é muito maior do que a listagem dos autores, que são os Coordenadores de Polos da REALGENE, listados no sentido horário da localização das instituições a que estão vinculados, a partir de Manaus, com exceção do Coordenador Geral do Projeto, que está listado como 13° autor. Agradecemos especialmente à Enedina N. Assunção, técnica do Laboratório de Tecnologia do DNA da Universidade Federal do Amazonas.