ISSN 1679-1150

150

Circular Técnica

Sete Lagoas, MG
Setembro, 2010

Autores

Roberto Willians Noda
Biólogo, Ph.D. em Bioinformática, Pesquisador da Embrapa Milho e Sorgo, Sete Lagoas, MG, roberto.noda@cnpms.embrapa.br

Cynthia Maria Borges Damasceno
Bióloga, Ph.D. em Biologia Molecular, Pesquisadora da Embrapa Milho e Sorgo, Sete Lagoas, MG, cynthia@cnpms.embrapa.br

Sylvia Morais de Sousa
Bióloga, Ph.D. em Biologia Molecular, Pesquisadora da Embrapa Milho e Sorgo, Sete Lagoas, MG, smsousa@cnpms.embrapa.br

Embrapa

Anotação Funcional de Sequências com BLAST2GO

Introdução

Há duas maneiras de se obter as sequências dos genes. Uma maneira seria sequenciar todo o genoma e depois utilizar softwares que predizem as regiões gênicas. A outra, seria sequenciar o transcriptoma, ou seja, os genes transcritos.

Resumidamente, para sequenciar o genoma completo, deve-se obter milhões de cópias dele, fragmentar as cópias (os equipamentos de sequenciamento atuais possuem uma capacidade limitada de leitura, por isso o genoma precisa ser fragmentado), efetuar a leitura e montar as peças (sequências), sobrepondo-as, recompondo a sequência original. Já para sequenciar o transcriptoma, obtém-se os genes transcritos, representados pelo conjunto do ácido ribonucléico mensageiro (mRNA), converte-os em ácido desoxirribonucléico complementar (cDNA) com a enzima transcriptase reversa e faz-se a leitura no sequenciador.

A anotação pode ser definida como o processo de descoberta de componentes importantes do genoma, principalmente genes e seus produtos, adicionando a eles análises e interpretações necessárias para extrair sua importância biológica e colocando-os no contexto dos processos biológicos (STEIN, 2001).

Cada vez mais, os profissionais de Bioinformática têm buscado desenvolver softwares e ferramentas com interface gráfica e amigável para facilitar análises da área, diminuindo a necessidade de conhecimentos avançados de Tecnologia da Informação (TI) para os usuários não bioinformatas.

O BLAST2GO (CONESA et al., 2005), disponível em http://www.blast2go.org/, é uma ferramenta web com interface Java (Java é uma linguagem de programação e uma plataforma de computação), que funciona em quaisquer sistemas operacionais (Windows, Linux e outros), para análise funcional de sequências (Figura 1).

O sistema de ontologias utilizado pelo BLAST2GO é o Gene Ontology (THE GENE ONTOLOGY CONSORTIUM, 2000), disponível em http://www.geneontology.org/, uma tentativa de padronizar a representação dos genes e seus produtos para todos os sistemas biológicos, subdividindo-os em três categorias: processo biológico – refere-se à atividade biológica com o qual o gene ou seu produto contribui; função molecular – atividade bioquímica do gene ou produto; componente celular – local na célula onde o gene ou seu produto é ativo.

A versão atual (2.4.5) do BLAST2GO permite fazer buscas *online* com BLAST – busca por sequências similares (ALTSCHUL, 1990), InterProScan – busca por assinaturas proteicas similares (ZDOBNOV; APWEILER, 2001), GO-Slim – um sub-conjunto dos termos do Gene Ontology, há GO-Slim para plantas, levedura e outros organismos e grupos (THE GENE ONTOLOGY CONSORTIUM, 2000), Enzime Code – busca do código da enzima (http://www.chem.qmul.ac.uk/iubmb/enzyme/) e KEGG - Kyoto Encyclopedia of Genes and Genomes – visualização dos mapas metabólicos onde os genes e/ou seus produtos atuam (KANEHISA; GOTO,

2000) para, a partir do arquivo FASTA contendo as sequências (para saber mais sobre o formato FASTA veja: http://en.wikipedia.org/wiki/FASTA_format), determinar suas possíveis funções e ontologias.

Como Utilizar o BLAST2GO

Os passos para a anotação automática com o BLAST2GO são:

1. Verificar Java.
   Antes de instalar o BLAST2GO é conveniente verificar a versão do Java. Clique no botão "Faça o download do software Java agora", no site http://www.java.com/pt_BR/download/installed.jsp e, caso seja necessário, faça a atualização do Java.

2. Instalar o BLAST2GO.
   Entrar no site http://www.blast2go.org/start_blast2go e, de acordo com a quantidade de memória RAM do seu computador, selecione, na seção "Download BLAST2GO", a opção 512, 1024, 1500 ou 2048 MB.

3. Entrar com os dados das sequências no formato FASTA.
   Para isso, basta clicar, na barra de Menu, em "File" e selecionar "Load FASTA file" (Figura 2) ou digitar "Alt + O" (conjuntamente nas teclas "Alt" e "letra O") e selecionar o arquivo FASTA. O formato FASTA é texto plano, ou seja, um texto sem formatação (negrito, parágrafo duplo, etc). A extensão de texto plano é o ".txt" e os softwares mais recomendados para a sua edição são o Bloco

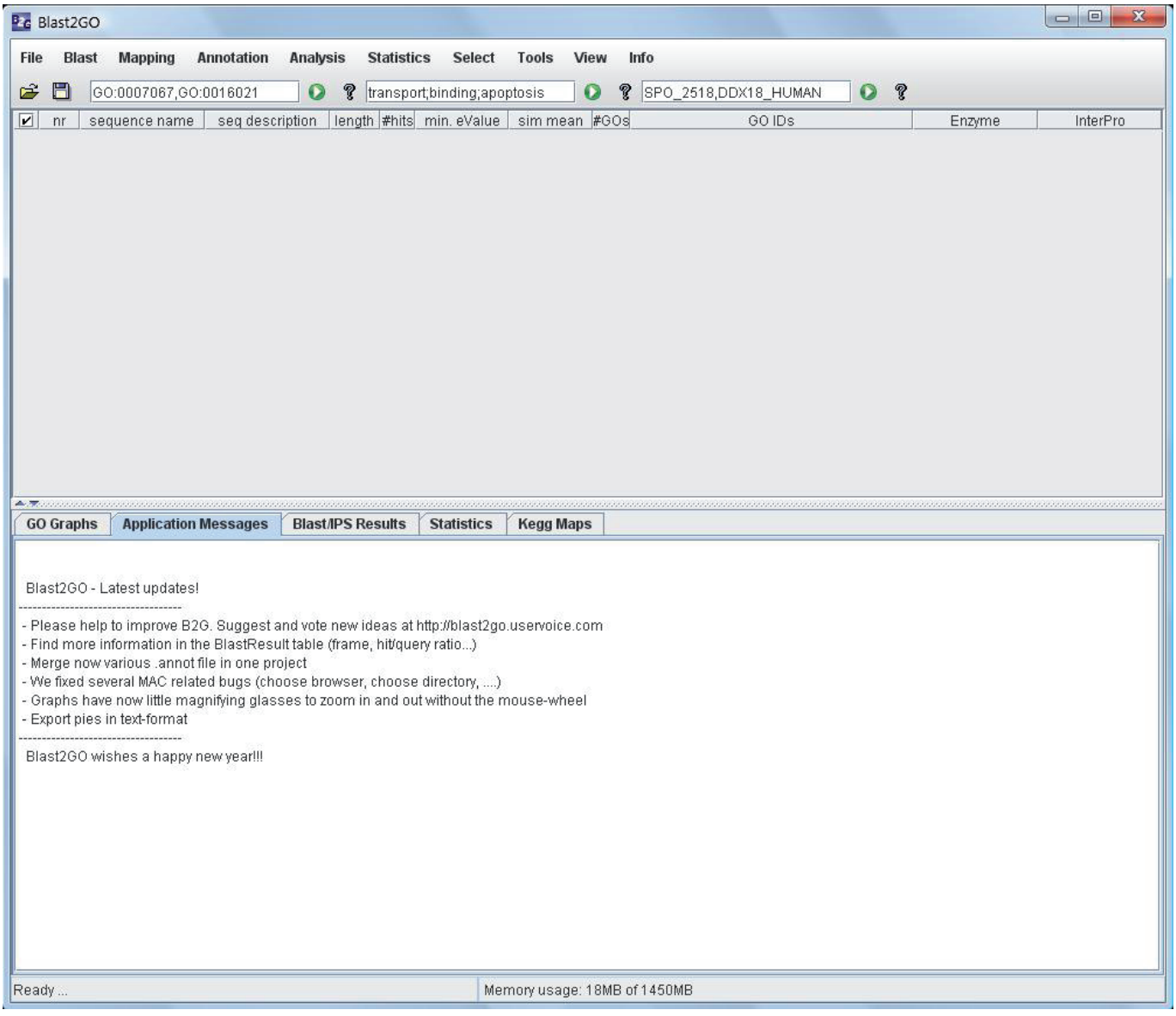

Figura 1. Tela do BLAST2GO, uma ferramenta web com interface Java para análise funcional de sequências genômicas.

de notas (Notepad), o gedit, o vi, o vim e outros. Mudando a extensão de um arquivo ".txt" para ".fasta", será produzido um arquivo FASTA, desde que seja respeitado o formato FASTA dentro do arquivo (http://en.wikipedia.org/wiki/FASTA_format).

Observação: os projetos de anotação do BLAST2GO podem ser salvos em arquivo ".dat". Caso tenha o arquivo ".dat", ele pode ser carregado digitando "Alt + Z" ("File/Load B2G-Project (.dat)") e selecionando-o. Veja como salvar o projeto de anotação no item 15.

Dica: vá salvando o projeto ao longo do processo de anotação para não perder dados de análises já realizadas, em caso de problemas com o computador, a energia e/ou a internet.

4. Executar o BLAST para encontrar sequências similares.
   Digitar "Alt + B" ("Blast/Run BLAST Step"), escolher os parâmetros e iniciar as buscas.

   Observação: Caso tenha dúvida sobre o parâmetros das análises, utilize os parâmetros pré-estabelecidos ou busque informações dos parâmetros para cada análise. Por exemplo, sobre o uso do BLAST, busque informações no http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs.

5. Mapear os termos Gene Ontology (GO) associados aos hits encontrados com o BLAST.
   Digitar "Alt + M" ("Mapping/Run GO-Mapping Step") e iniciar o mapeamento.

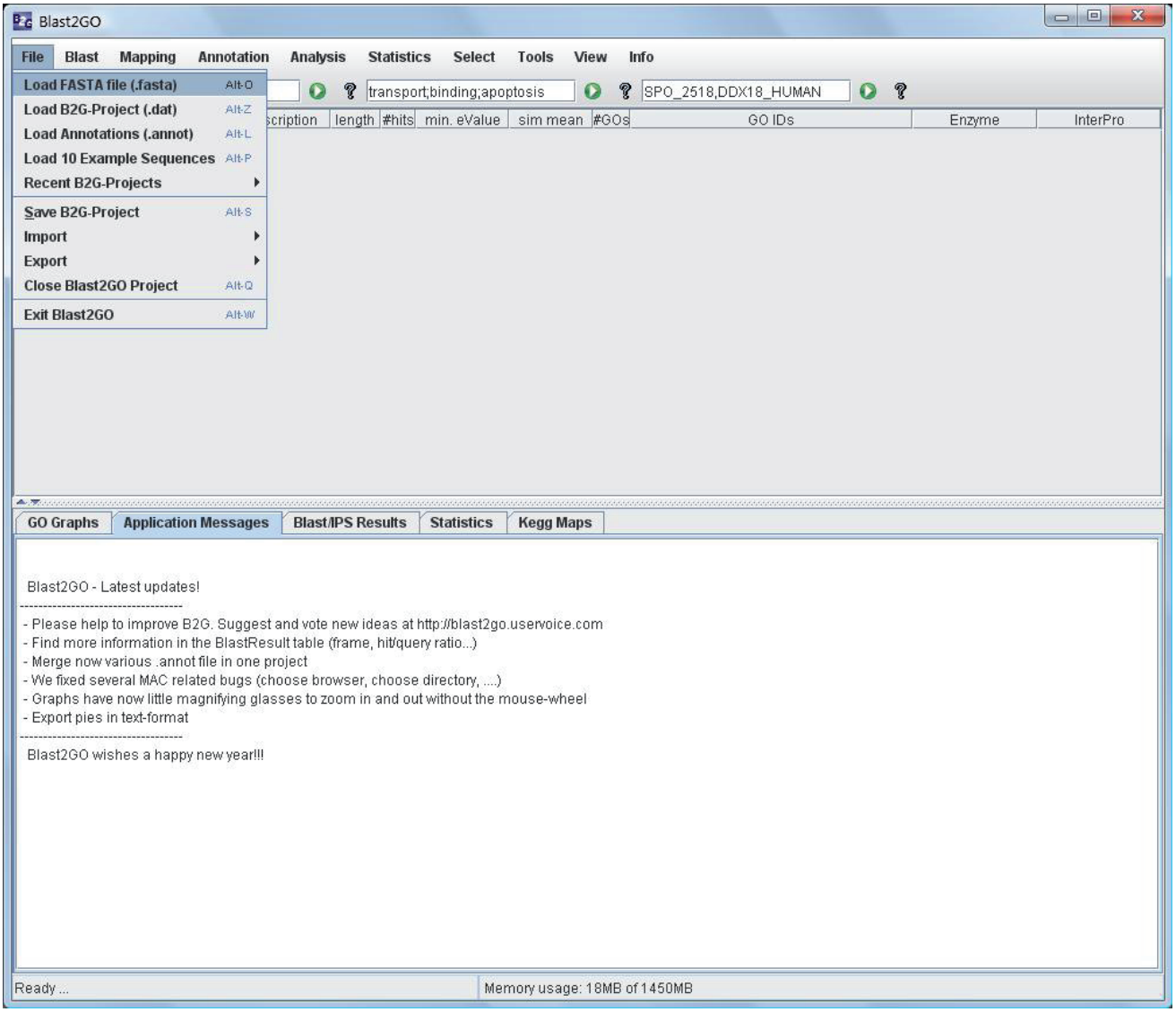

Figura 2. Tela do BLAST2GO, com a opção de carregar um arquivo FASTA selecionada, na barra de menu.

6. Anotar as sequências.
   Digitar "Alt + A" ("Annotation/Run Annotation Step") e iniciar a anotação.

7. Anotar motivos/domínios das sequências com InterProScan.
   Digitar "Alt + I" ("Annotation/InterProScan/Run InterProScan (online)"), selecionar as aplicações a serem utilizadas e iniciar a anotação. E digitar "Alt + J" ("Annotation/InterProScan/Merge InterProScan GOs to Annotation") para adicionar os termos GO obtidos com o InterProScan aos termos GO obtidos com o BLAST.

8. Aumentar a quantidade de sequências anotadas.
   Digitar "Alt + N" (Annotation/Run ANNEX (Annotation Augmentation)") para executar o ANNEX (Annotation Expander). O ANNEX é composto pelas relações, manualmente curadas, entre os termos GO para função molecular envolvidos em "processos biológicos" e atuando em "componentes celulares", e essas relações permitem inferir os termos GO para processos biológicos e componentes celulares para sequências com termos GO para função molecular, aumentando os termos GO associados às sequências.

9. Opcionalmente, pode-se resumir a anotação baseando o mapeamento em termos GO-Slim direcionados, por exemplo, para plantas ou para leveduras.
   Digitar "Alt + G" ("Annotation/GO-Slim/Run GO-Slim (online)"), selecionar o GO slim desejado e iniciar a anotação.

10. Corrigir anotação manualmente.
    Clicar com o botão direito do mouse sobre a sequência, selecionar "Change Annotation and Description" (Figura 3), fazer as alterações necessárias, marcar "Mark manual Annotation" e confirmar a correção (clicar no botão "play", no canto superior esquerdo da janela de Correção).

11. Validar anotação.
    Clicar em "Annotation", na barra de Menu e selecionar "Validate Annotations".

12. Determinar o código da enzima (EC number).
    Digitar "Alt + E" ("Annotation/Enzyme Code and KEGG/Run GO-Enzyme Code Mapping") para iniciar a busca do EC number baseado nos termos GO das sequências.

13. Obter os mapas metabólicos do KEGG em que as sequências atuam.
    Digitar "Alt + H" ("Annotation/Enzyme Code and KEGG/Load Pathways-Maps from KEGG (online)") e selecionar o diretório (pasta) onde as imagens dos mapas metabólicos serão salvos para iniciar o download.

14. Visualizar resultados e realizar análises de enriquecimento.
    Os menus "Analysis" e "Statistics" disponibilizam os resultados possíveis de serem visualizados e a aba "Kegg Maps" permite visualizar os mapas metabólicos obtidos no item anterior (Figura 4). No menu "Analysis", também é possível fazer análises de enriquecimento, como Teste Exato de Fisher.

Salvar o projeto de anotação.

Digitar "Alt + S" ("File/Save B2G-Project") para selecionar o nome do arquivo e diretório (pasta) onde o arquivo ".dat" será salvo.

15. Exportar resultados.

No menu "File/Export" há sete opções para exportar os dados das sequências e das suas anotações. Como o "File/Export/Export Sequence Table" ou "Alt + T" (Figura 5), que produz um arquivo TXT tabulado, que pode ser aberto em um software de planilha (como o Excel ou Calc) ou convertido em tabela em um software de edição de texto (como o Word ou Write), e possui as colunas apresentadas na Tabela 1.

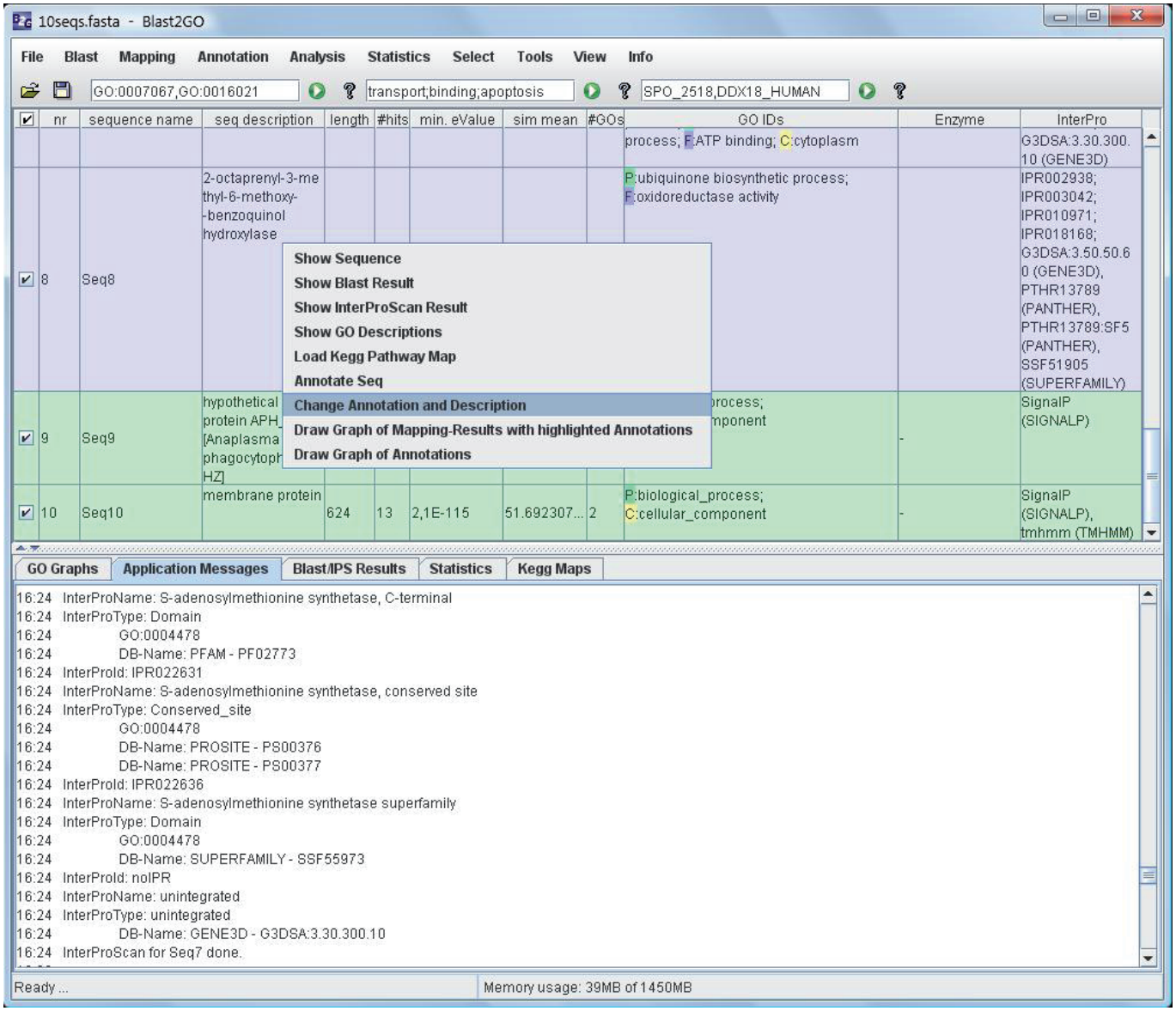

Figura 3. Tela do BLAST2GO, com a opção de corrigir a anotação e a descrição da sequência selecionada, na barra de menu.

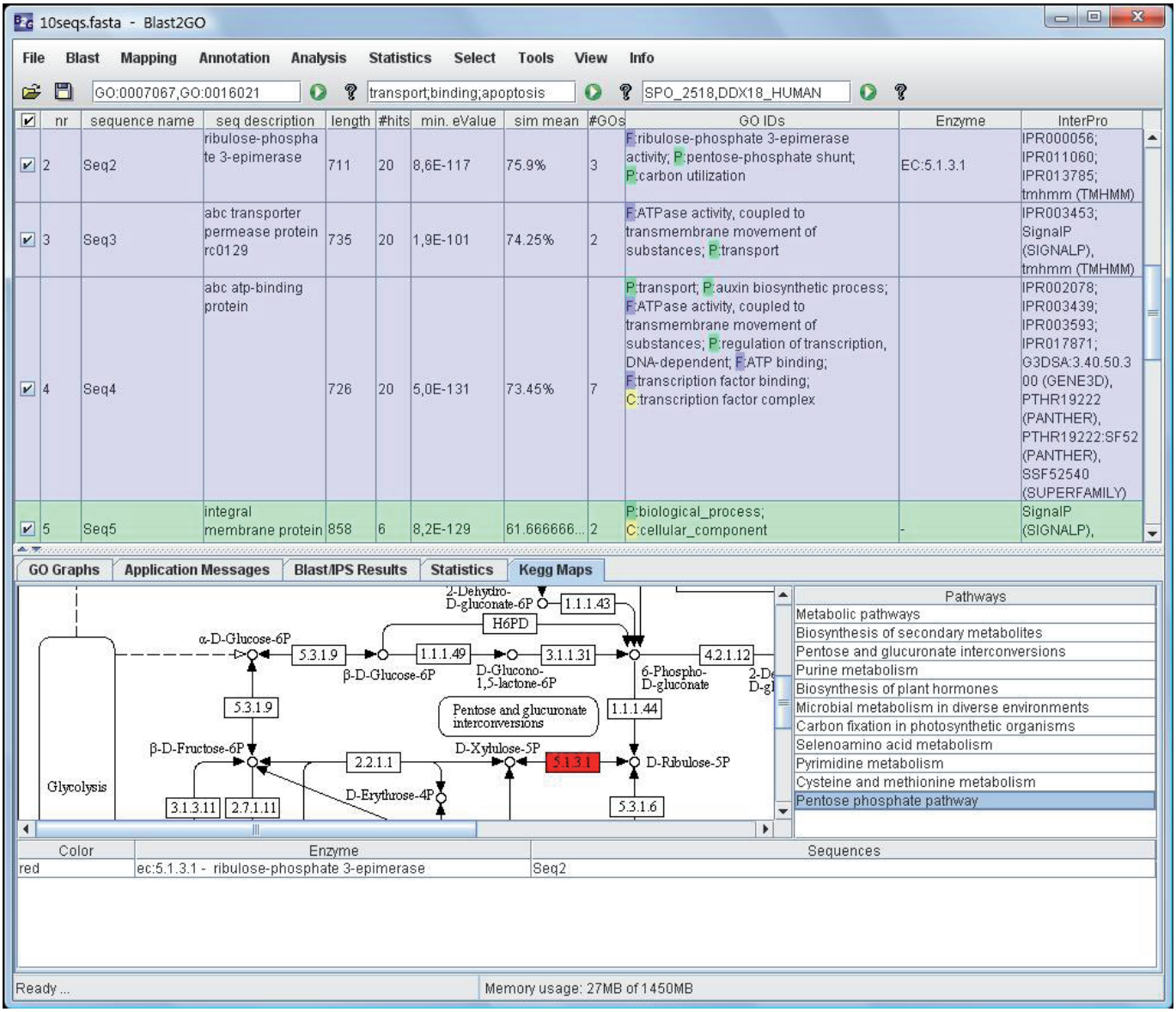

Figura 4. Tela do BLAST2GO, mostrando a aba "Kegg Maps", onde é possível visualizar as vias metabólicas do KEGG (http://www.genome.jp/kegg/) em que a sequência ou seu produto atua. A(s) sequência(s) ou seu(s) produto(s) presente(s) na via metabólica é (são) destacada(s) com cor(es).

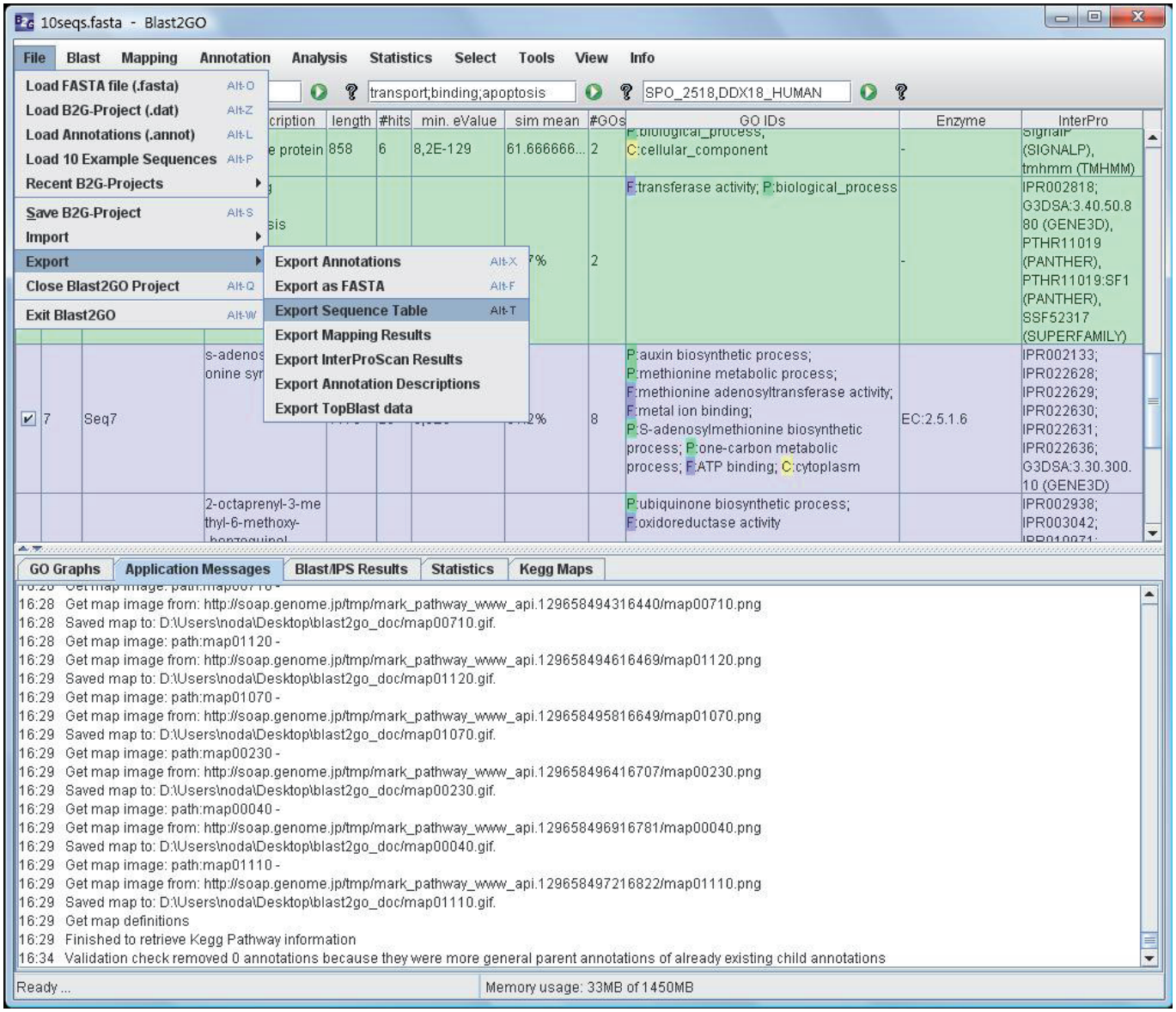

Figura 5. Tela do BLAST2GO, com a opção de exportar os dados das sequências, na barra de menu.

Tabela 1. Dados exportados na opção "File/Export/Export Sequence Table" ou "Alt + T" e convertidos em tabela num software de edição de textos.

Seq. Name	Seq. Description	Seq. Length	#Hits	min. eValue	mean Similarity	#GOs	GOs	Enzyme Codes	InterProScan
Seq1	dna polymerase i	2562	20	0.0	71.65%	5	F:DNA-directed DNA polymerase activity; P:DNA replication; F:DNA binding; F:5'-3' exonuclease activity; C:DNA polymerase complex	EC:2.7.7.7	IPR001098; IPR002298; IPR002421; IPR008918; IPR012337; IPR018320; IPR019760; IPR020045; IPR020046; IPR020047; G3DSA:1.10.150.20 (GENE3D), G3DSA:1.20.1060.10 (GENE3D), G3DSA:3.30.420.10 (GENE3D), G3DSA:3.30.70.370 (GENE3D), G3DSA:3.40.50.1010 (GENE3D), PTHR10133 (PANTHER), tmhmm (TMHMM), SSF56672 (SUPERFAMILY), SSF88723 (SUPERFAMILY)
Seq2	ribulose-phosphate 3-epimerase	711	20	8,43E-112	75.9%	3	F:ribulose-phosphate 3-epimerase activity; P:pentose-phosphate shunt; P:carbon utilization	EC:5.1.3.1	IPR000056; IPR011060; IPR013785; tmhmm (TMHMM)
Seq3	abc transporter permease protein rc0129	735	20	1,84E-96	74.25%	2	F:ATPase activity, coupled to transmembrane movement of substances; P:transport		IPR003453; SignalP (SIGNALP), tmhmm (TMHMM)

Conclusão

A quantidade de dados gerados nos laboratórios é cada vez maior e as análises *in silico* desses dados necessitam ser cada vez mais automatizadas, municiando de informações quem precisa decidir quais análises, *in silico* ou laboratoriais, devem ser executadas posteriormente, para tanto a Bioinformática precisa produzir softwares que atinjam esse objetivo, mas que também sejam mais amigáveis e fáceis para os usuários com menores conhecimentos em Tecnologia da Informação, Linux e/ou Bioinformática. O BLAST2GO é um exemplo de ferramenta gráfica, com interface intuitiva e multiplataforma que permite fazer as análises mais usuais para anotação de sequências.

Os resultados da anotação automática (comparações automáticas das sequências com bancos de dados) devem passar por uma curadoria (verificação manual), onde as informações possam ser confirmadas ou corrigidas. Entretanto, cabe ressaltar que a verdadeira validação dos resultados de análises *in silico* deve ser biologicamente realizada. Os resultados de anotação automática são muito importantes, pois auxiliam na descoberta da importância biológica da sequência dentro do contexto em que ela foi obtida.

Referências

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, P. Basic local alignment search tool. **Journal of Molecular Biology**, London, v. 215, p. 403-410, 1990.

CONESA, A.; GÖTZ, S.; GARCÍA-GÓMEZ, J. M.; TEROL, J.; TALÓN, M.; ROBLES, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, v. 21, n. 18, p. 3674-3676, 2005.

KANEHISA, M.; GOTO, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. **Nucleic Acids Research**, London, v. 28, p. 27-30, 2000.

STEIN, L. Genome annotation: from sequence to biology. **Nature Reviews. Genetics**, London, v. 2, n. 7, p. 493-503, Jul. 2001.

THE GENE ONTOLOGY CONSORTIUM. Gene ontology: tool for the unification of biology. **Nature Genetics**, New York, v. 25, n. 1, p. 25-29, May 2000.

ZDOBNOV, E. M.; APWEILER, R. "InterProScan - an integration platform for the signature-recognition methods in InterPro." **Bioinformatics**, v. 17, n. 9, p. 847-848, 2001.

Circular Técnica, 150

Exemplares desta edição podem ser adquiridos na:
Embrapa Milho e Sorgo
Endereço: Rod. MG 424 km 45 Caixa Postal 151
CEP 35701-970 Sete Lagoas, MG
Fone: (31) 3027 1100
Fax: (31) 3027 1188
E-mail: sac@cnpms.embrapa.br
1ª edição
1ª impressão (2010): on line

Ministério da Agricultura, Pecuária e Abastecimento
BRASIL UM PAÍS DE TODOS GOVERNO FEDERAL