Genética na Escola – ISSN: 1980-3540

Ser ou não ser, eis a questão: predizendo a função de sequências de DNA usando o Blast2GO*

Adriana Maria Antunes[1], Lays Karolina Soares da Cruz[1], Mariana Pires de Campos Telles[2]

[1] Pós-graduanda em Genética e Biologia Molecular, ICB, Universidade Federal de Goiás, Goiânia, Campus II, ICB I, Brasil
[2] Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas I, Universidade Federal de Goiás, Campus II; Escola de Ciências Agrárias e Biológicas, Pontifícia Universidade Católica de Goiás

Autor para correspondência: tellesmpc@gmail.com

Palavras-chave: alinhamento, bioinformática, DNA, expressão gênica, genômica, similaridade

* Material didático desenvolvido como parte da atividade de Estágio de Docência na disciplina Genética Molecular, coordenado pela Profa. Mariana Pires de Campos Telles, no curso de graduação em Ciências Biológicas, Instituto de Ciências Biológicas, Universidade Federal de Goiás. Apoiado pelo projeto "Desenvolvimento de marcadores, genotipagem e caracterização genômica de espécies do Cerrado (GENPAC 02)" - CNPq 563839/2010-4.

Genética na Escola – ISSN: 1980-3540

O Blast2GO é uma ferramenta de Bioinformática que usa o Blast (*Basic Local Alignment Search Tool*) e o GO (*Gene Ontology*) para identificar e anotar funcionalmente genes por meio da busca de similaridade significativa com sequências conhecidas armazenadas em bancos de dados. A atividade tem como objetivo apresentar uma estratégia para predizer genes a partir da análise de sequências de DNA de diferentes organismos, a fim de estudá-los do ponto de vista funcional. A atividade é indicada para estudantes de graduação e pós-graduação que já cursaram ou que estejam cursando Genética e pretende propiciar ao estudante a oportunidade de ampliar o conhecimento sobre as estratégias de análises genômicas.

FUNÇÃO PEDAGÓGICA

A atividade proposta permite mostrar uma das possibilidades de anotação de sequências de ácidos nucleicos. É a partir da **anotação** que é realizada a identificação e análise dos elementos genéticos que constituem a estrutura dos **genomas**. Um dos primeiros e mais importantes passos na anotação é a identificação de genes, uma vez que os estudos das regiões codificadoras fornecem muitas informações sobre o funcionamento dos organismos. O Blast2GO é uma ferramenta de **Bioinformática** que usa o Blast (*Basic Local Alignment Search Tool*) e o GO (*Gene Ontology*) para identificar e anotar funcionalmente genes em sequências de ácidos nucleicos por meio da busca de similaridade significativa com sequências de genes conhecidos e armazenados em bancos de dados.

A atividade tem como objetivo apresentar uma estratégia para predizer genes a partir da análise de sequências de DNA, utilizando o programa Blast2GO, em trechos do genoma das seguintes espécies: *Escherichia coli*, *Arabidopsis thaliana*, *Mus musculus* e *Homo sapiens*. A atividade pretende despertar o interesse pelas áreas de genética molecular e genômica, a fim de consolidar conhecimentos sobre estratégias de investigação nessas áreas da ciência.

PROBLEMA PROPOSTO

O DNA ou ácido desoxirribonucleico é uma macromolécula constituída por unidades chamadas de nucleotídeos, cada nucleotídeo é formado por um grupo fosfato, uma desoxirribose e uma base nitrogenada. É a ordem das bases nitrogenadas nas sequências do DNA que controla o funcionamento dos organismos por meio da **expressão gênica**, seguindo os padrões definidos no **código genético**. Os genes codificam as proteínas e as proteínas atuam sozinhas ou conjuntamente com outras macromoléculas para desempenhar alguma função celular. Assim, os milhares de genes expressos em uma célula, em particular, determinam as funções biológicas ativas nessa célula. Além disso, cada etapa do fluxo de informações do DNA para o RNA e para a proteína fornece a célula pontos em potencial para regulação dos processos de transcrição e tradução, ajustando a quantidade e o tipo de proteínas que serão fabricadas.

A identificação e a anotação funcional de genes em sequências de ácidos nucleicos permite compreender a estrutura e o funcionamento do genoma. O processo de anotação do genoma requer uma etapa de identificação das sequências do gene predito, bem como a indicação da função biológica dessas sequências codificadoras. A identificação de genes em organismos procariotos é mais simples devido ao tamanho e à estrutura dos seus genomas. Em procariotos, os genomas são menores e as regiões gênicas (que codificam para alguma proteína) são colineares em seus produtos gênicos. Por outro lado, em organismos eucariotos os genes estão localizados em meio a uma enorme quantidade de sequências não codificadoras (intergênicas). Além disso, os genes de eucariotos normalmente apresentam regiões não codificadoras (íntrons) intragênicas. Apesar das dificuldades em estudar genes em genomas grandes e complexos como os eucariotos, essas investigações tornaram-se possíveis a partir do desenvolvimento de diversas ferramentas de análise de dados da bioinformática.

A identificação de genes pode ser realizada utilizando-se características intrínsecas do código genético, tais como os **códons** de início e de parada. Além disso, é possível usar métodos baseados em similaridade de sequências, uma vez que os genes são identificados em função da quantidade de similaridade com sequências gênicas bem caracterizadas e armazenadas em **bancos de dados genômicos**. Para genomas eucariotos, a anotação de genes usando métodos de similaridade é mais eficiente, uma vez que os métodos baseados em características intrínsecas são mais propensos a erros em genomas mais complexos.

Uma importante ferramenta para anotação do genoma é o Blast2GO (CONESA *et al.*, 2005) que permite realizar a identificação e a anotação funcional de sequências gênicas. A identificação dos genes é realizada pelo Blast (do inglês *Basic Local Alignment Search Tool*) (Veja mais em ANTUNES *et al.*, 2014), um algoritmo que compara a sequência de ácidos nucleicos de interesse com outras sequ-

Anotação - identificação na sequência de DNA de elementos funcionais para a expressão gênica, através de ferramentas de bioinformática ou experimentais.

Genoma - conteúdo inteiro de material genético em um conjunto haploide de cromossomos de uma espécie.

Bioinformática - conjunto de sistemas computadorizados de informação e métodos analíticos aplicados a problemas biológicos.

Códon - sequência de três nucleotídeos do RNA mensageiro que codifica um único aminoácido.

Bancos de dados genômicos - bancos de dados que disponibilizam as sequências de DNA de diversos organismos já estudados em todo o mundo. Geralmente os dados são disponibilizados via web.

Expressão gênica - processo no qual a informação contida em um ou mais genes é usada na síntese de um produto funcional.

Código genético - conjunto de correspondências entre trincas de nucleotídeos no RNA mensageiro e aminoácidos na proteína.

ências conhecidas e armazenadas em banco de dados para diversos organismos. A comparação ou análise de similaridade entre as sequências é feita por meio do alinhamento. O alinhamento permite comparar a correspondência de nucleotídeos entre duas ou mais sequências. Em um alinhamento, as bases correspondentes entre as sequências recebem o nome de *match*, as bases não correspondentes recebem o nome de *mismatch* e as regiões onde ocorre inserção ou deleção de bases recebem o nome de *gap*.

Existem várias possibilidades de alinhamento entre duas sequências e o melhor alinhamento é escolhido com base nos parâmetros valor de score e E-valor. O valor de *score* é calculado com base nas pontuações dos *matchs*, *mismatch* e *gaps*. Os *matchs*, que representam regiões de similaridades entre as sequências, geralmente recebem pontuações positivas, enquanto, os *mismatchs* e os *gaps*, que representam regiões de diferenças entre as sequências, recebem pontuações negativas. O melhor alinhamento é o que possui maior valor de *score* e menor *E-valor*. O *E-valor* corresponde ao número esperado de sequências com similaridade tão alta como a encontrada por razão do acaso. Quanto maior o *E-valor*, maior a chance das correspondências entre as sequências não serem verdadeiras, dessa forma busca-se um *E-valor* menor que 10^{-20}, valor considerado como ponto de corte. Durante o Blast, a similaridade de sequências é investida considerando os sentidos 3'- 5' e 5'- 3' da fita de DNA.

O Blast permite a análise de similaridade entre a sequência de interesse e sequências armazenadas em bancos de dados. Nesse sentido, existem vários tipos de Blast. O Blastn busca sequência de nucleotídeos em banco de dados de DNA. O Blastp busca sequências de aminoácidos em banco de dados de proteínas. O Blastx busca sequências de nucleotídeo em banco de dados de proteínas. O tBlastn busca sequências de aminoácidos em banco de dados de nucleotídeo. E o tBlastx busca sequência de nucleotídeos traduzidos em banco de dados de nucleotídeo traduzidos. A identificação de genes em sequências de ácidos nucleicos pode ser realizada usando, por exemplo, o Blastp, para identificação de genes codificadores de proteínas.

Após a identificação das sequências dos genes, é feita a anotação funcional usando o *Gene Ontology*, uma iniciativa que visa relacionar genes e produtos de genes para diversas espécies. O GO tem como objetivo classificar e representar os genes em três categorias funcionais:

a) componentes celulares - o termo "componente celular" indica a localização do produto do gene na célula, ou seja, indica se está localizado em organelas, no citoplasma, núcleo ou parede celular.

b) função molecular - o termo "função molecular" indica a função bioquímica do produto de expressão (atividade enzimática, catalítica, regulatória ou outros).

c) processos biológicos - o termo "processo biológico" indica processos mais complexos como vias metabólicas (respiração celular ou metabolismo de carboidratos).

O mesmo gene pode ser anotado em um ou nos três termos do *Gene Ontology*. As anotações funcionais e as relações estabelecidas entre os genes baseadas no GO são inferidas a partir de semelhanças entre as sequências. A anotação funcional é realizada por meio da análise dos produtos da expressão do gene, ou seja, das proteínas. Existe a possibilidade porque a função das proteínas é associada à presença de domínios, que são unidades estruturais, funcionais e evolutivas. Os domínios são responsáveis pelo dobramento das cadeias de aminoácidos e determinam o papel global das proteínas. O alinhamento da sequência de interesse com banco de dados de domínios permite identificar as regiões funcionais das proteínas e assim classificar as famílias de proteínas. Por exemplo, os sítios ativos das enzimas são bastante conservados e permitem identificar genes com função catalítica. O compartilhamento de domínios é resultado de duplicações gênicas ao longo da evolução (genes ortologos). Nesse contexto, a anotação usando o Blast2GO faz uso do GO e permite a anotação funcional do genoma em larga escala.

O Blast2GO é uma ferramenta desenvolvida em linguagem JAVA que pode ser instalada em servidores ou mesmo no computador pessoal (é uma ferramenta computacional automatizada que permite a análise de um

grande número de sequências com uma precisão aceitável e uma velocidade suficiente). A versão básica é oferecida gratuitamente para a comunidade científica para ser usado sem fins lucrativos e a versão PRO, a mais rápida, é oferecida livremente por uma semana para teste. O Blast2GO permite a realização de anotação funcional *in silico*, que inclui análises estatísticas e visualizações gráficas dos resultados. Nesta atividade será realizada uma análise utilizando o Blast2GO versão PRO a partir de análise de três **arquivos FASTA** contendo partes de sequências do genoma de diferentes espécies eucariotas e procariotas. Este arquivo de entrada, obtido anteriormente a partir de bancos de dados públicos, será usado para a identificação e investigação funcional das sequências. Os estudantes terão que rodar as análises no Blast2GO e interpretar os resultados. Os resultados obtidos representam inferências sobre a sequência e função dos genes, porem são confiáveis visto que os resultados são estatisticamente significativos. No entanto, os resultados poderão, posteriormente, ser testados experimentalmente por pesquisadores visando a confirmação das informações obtidas.

INSTRUÇÕES PARA O PROFESSOR

1 - É recomendável que o professor aplique esta atividade em turmas de Graduação que possuem conhecimentos de Genética Molecular, tais como estrutura de DNA, mecanismos de expressão gênica e métodos de sequenciamento, incluindo os métodos clássicos e os de nova geração.

2 - A proposta de atividade pode ser realizada em aula prática em laboratório de informática ou como atividade extraclasse. Recomenda-se que cada estudante entregue um relatório ao final das atividades.

PROCEDIMENTOS PARA OS ESTUDANTES

1 - Após receber do professor o roteiro de atividades e os arquivos em formato ".fasta", o estudante deve acessar o site do Blast2GO, disponível em https://www.blast2go.com/. Em "*Request a Free PRO Trial*", checar se o computador utilizado cumpre os requisitos necessários para a execução do software. Em seguida, o estudante deve acessar a página de download em "https://www.blast2go.com/blast2go-pro/b2g-register-basic" e registrar-se. O estudante receberá um e-mail com o código de acesso que deverá ser inserido na janela inicial do software e ativado.

Arquivo FASTA - formato de arquivo em que na primeira linha há o símbolo maior (>), seguido pelo título da sequência e nas linhas seguintes apresentam as sequências de bases nitrogenadas do DNA ou de aminoácidos de polipeptídios.

2 - Em seguida, é preciso acessar "Download Blast2GO"; logo após deve-se acessar *show more download options* e selecionar a opção adequada ao computador em uso para que o download seja iniciado. Basta executar a instalação nos padrões de recomendação do próprio software; ao terminar, clicar em *finish* e um ícone terá sido criado em sua área de trabalho. Esta versão será disponibilizada gratuitamente por uma semana. Ao inicializar o software, é preciso concordar com os termos de uso e, em seguida, inserir o código enviado para o e-mail do estudante no momento em que se registrou.

3 - Após a ativação do código de acesso, o software está liberado para uso. Na parte superior direita serão exibidos ícones que representam os passos de análise. O estudante deve começar pelo primeiro *start* quando serão inseridas as sequências. Clicando em 'start', acessar a opção *load sequences* e uma nova janela será aberta. Nesta nova janela é preciso clicar em *Browse* para escolher o arquivo a ser carregado. Não é necessário alterar os demais parâmetros. A sequência fornecida nesta atividade é de nucleotídeos, o que a torna compatível com o default do B2GO. Ao final, a sequência carregada aparecerá em verde no canto inferior esquerdo da tela. As análises deverão ser realizadas, separadamente, para cada sequência.

4 - Para iniciar as análises, é preciso clicar no ícone *blast*, escolher a opção *run blast* e será exibida a opção de banco de dados a ser utilizado. Recomenda-se manter o default do software que

Genética na Escola – ISSN: 1980-3540

utiliza o banco de dados do NCBI. É preciso fornecer um e-mail válido para registro no NCBI.

5 - Avançando a execução será necessário escolher um formato de arquivo de resultados (output) a ser gerado no fim da etapa. Recomenda-se que os estudantes selecionem o formato html para visualização gráfica dos resultados. Selecionar a pasta de destino dos outputs e clicar em *run*. O Blast será iniciado e uma barra de progressão aparecerá no local onde a sequência foi carregada.

6 - Ao final da execução, aparecerá uma mensagem informando que o estudante deve prosseguir para as próximas etapas. Neste momento é preciso clicar no ícone *interpro* e escolher a opção *run InterProScan*. Com todas as opções de resultados selecionadas, prosseguir igualmente com todas as análises até *chart*. Nesta opção é necessário gerar as análises gráficas de cada etapa separadamente. Acessando chart, comece pela primeira opção *Project Statistics* e repita a ação até o fim. Ao fim de cada análise estatística, salvar os gráficos gerados em formato pdf. Em *graphs*, gerar o gráfico combinado com todas as opções de resultados selecionadas e o GO *graphs*.

A partir da análise dos resultados gerados, responder às questões propostas na atividade.

MATERIAL DIDÁTICO

O professor deve fornecer aos estudantes: o roteiro contendo as informações da seção "procedimento para os estudantes", os arquivos .fasta contendo as sequências a serem analisadas; o questionário.

QUESTIONÁRIO

1) Com base na tabela de resultados para cada uma das sequências informadas responder:

   a) Qual foi o *E-value* de cada uma das sequências?

   b) Quantas e quais foram as classificações GO para cada uma das sequências?

   c) Comparando os três *E-values* informar que sequência teve menor chance de erro tipo 1 e explicar.

2) Com base no gráfico combinado para a sequência de *Escherichia coli*, fazer uma análise descritiva da anotação feita pelo Blast2GO.

3) Supor que é um pesquisador que trabalha com pacientes que estão se submetendo a testes de medicamentos. Em uma dada amostra foi identificada a superexpressão de um gene. A partir da sequência deste gene fazer a anotação usando o Blast2GO como uma análise complementar a fim de identificar a provável função do gene afetado pelo medicamento, que poderá

auxiliar os estudos de efeitos adversos. De acordo com a anotação utilizando o Blast2GO e o gráfico (GO graphs) obtido

a) Qual é o gene?

b) Qual o processo biológico relacionado ao gene e qual é sua descrição?

c) Qual a função molecular relacionada ao gene e qual é sua descrição?

4) Supor que o pesquisador está estudando diferentes condições de desenvolvimento em plantas, utilizando o organismo modelo *Arabidopsis thaliana*. Anotar sequências de DNA de indivíduos que foram submetidos à condições de estresse hídrico durante seu desenvolvimento. Para melhor entender os efeitos da privação de água em plantas.

a) Algum gene apresentou resposta ao tratamento de estresse hídrico?

b) Fazer uma análise do provável motivo do gene responder ao estresse.

ENTENDENDO A ATIVIDADE (RESPOSTAS)

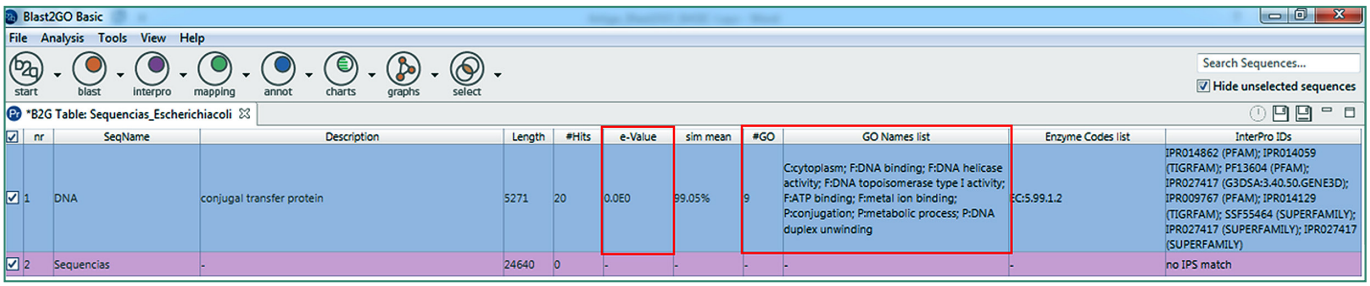
1) Com base na tabela de resultados para cada uma das sequências informadas responder:

a) Qual foi o *E-value* de cada uma das sequências?

Resposta: Para a sequência de *E. coli* foi obtido um *e-value* de 0,0. Para a sequência de *Homo sapiens* foi obtido *e-value* de 4,8E-75. E, para a sequência de *A. thaliana*, foi obtido *e-value* de 2,2E-39.

b) Quantas e quais foram as classificações GO para cada uma das sequências?

Resposta: Para *H. sapiens* foram obtidas 4 anotações nas categorias GO e, para *A. thaliana*, foram obtidas 6. Para *E. coli*, conforme exemplificado na figura 1, foram obtidas 9 classificações nas categorias GO:



nr	SeqName	Description	Length	#Hits	e-Value	sim mean	#GO	GO Names list	Enzyme Codes list	InterPro IDs
1	DNA	conjugal transfer protein	5271	20	0.0E0	99.05%	9	C:cytoplasm; F:DNA binding; F:DNA helicase activity; F:DNA topoisomerase type I activity; F:ATP binding; F:metal ion binding; P:conjugation; P:metabolic process; P:DNA duplex unwinding	EC:5.99.1.2	IPR014862 (PFAM); IPR014059 (TIGRFAM); PF13604 (PFAM); IPR027417 (G3DSA:3.40.50.GENE3D); IPR009767 (PFAM); IPR014129 (TIGRFAM); SSF55464 (SUPERFAMILY); IPR027417 (SUPERFAMILY); IPR027417 (SUPERFAMILY)
2	Sequencias	-	24640	0	-	-	-	-	-	no IPS match

Figura 1. Tabela de resultados do Blast2GO exibindo *e-value*, número de categorias GO encontradas para a sequência de *E. coli* e o nome de cada categoria.

c) Comparando os três *E-values* informar que sequência teve menor chance de erro tipo 1 e explicar.

Resposta: Para a sequência de *E. coli* foi obtido um *e-value* de 0,0. Para a sequência de *H. sapiens* foi obtido *e-value* de 4,8E-75. E, para a sequência de *A. thaliana*, foi obtido *e-value* de 2,2E-39. Comparativamente, a sequência de *E. coli* teve o melhor *e-value*, pois foi o mais próximo de zero. Quanto maior o *E-value*, menor a confiança de que as correspondências sejam reais, e maior a chance da similaridade entre as sequências serem devido ao acaso. Ou seja, para esta sequência, a probabilidade de atribuição de falso positivo, ou cometimento do erro tipo 1 é 0.

2) Com base no gráfico combinado para a sequência de *Escherichia coli* fazer uma análise descritiva da anotação feita pelo Blast2GO.

Resposta: Foram identificados 14 termos GO de processo biológico e 4 de componente celular. O processo biológico do gene foi identificado como organização da biogênese, divisão nuclear mitótica e processo celular entre outros. O componente celular formado pelo produto do gene está relacionado à membrana. Descrição do processo biológico: um processo do ciclo celular, compreendendo os passos pelos quais o núcleo de uma célula eucariótica divide; o processo envolve a condensação de DNA cromossômico numa

forma altamente compactada. Produz dois núcleos filhos, cujo complemento cromossômico é idêntico ao da célula-mãe. Descrição do componente celular: componente de uma membrana que consiste em produtos de genes e complexos de proteínas que têm pelo menos uma parte da sua sequência de peptídeo incorporado na região hidrofóbica da membrana.

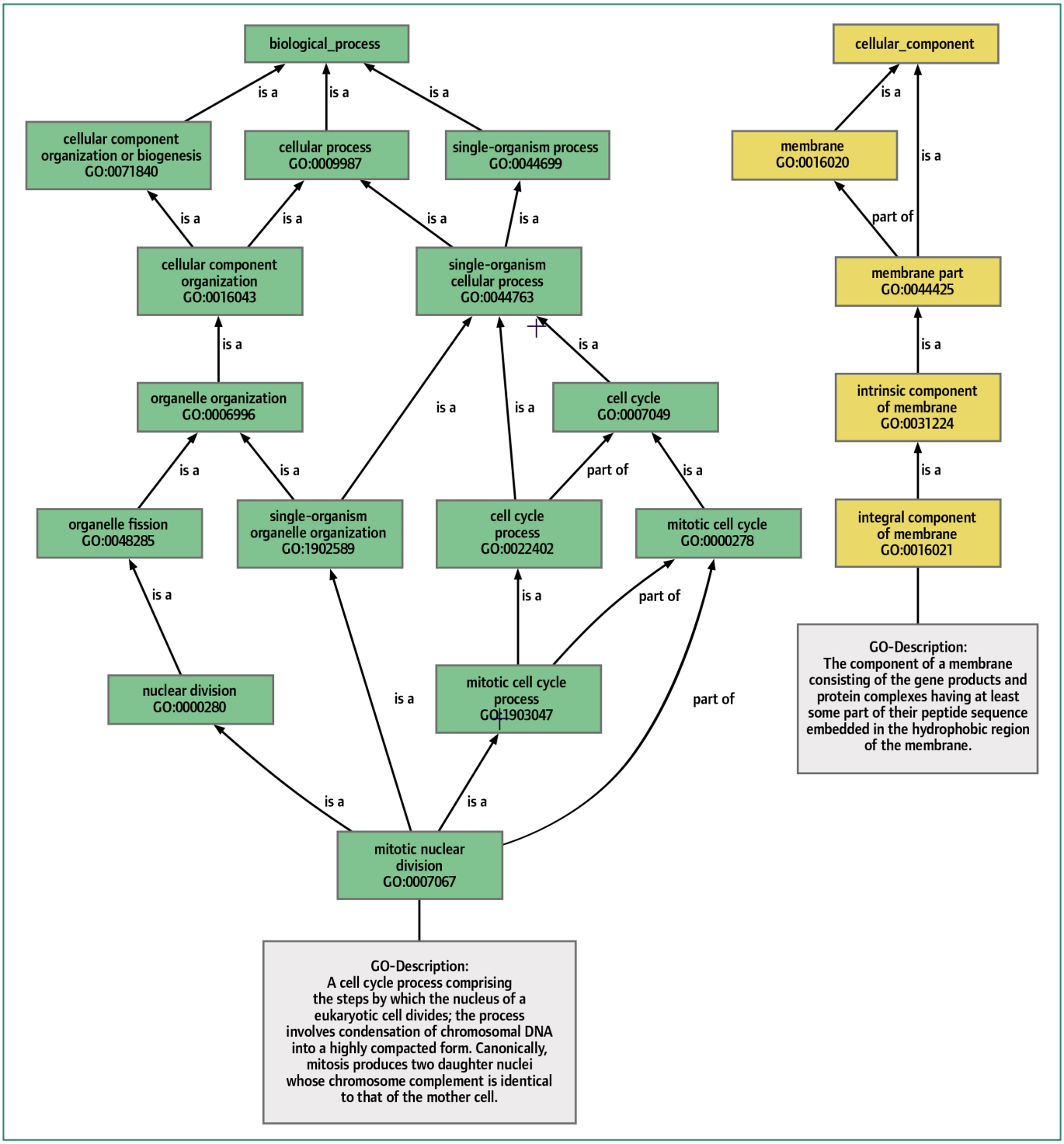


Figura 2.
Gráfico combinado gerado a partir da sequência de *E. coli*.

3) Supor o seguinte: um pesquisador que trabalha com pacientes que estão se submetendo a testes de medicamentos. Em uma dada amostra, foi identificada a superexpressão de um gene. A partir da sequência deste gene, fazer a anotação usando o Blast2GO como uma análise complementar a fim de identificar a provável função do gene afetado pelo medicamento, que poderá auxiliar os estudos de efeitos adversos. De acordo com a anotação utilizando o Blast2GO e o gráfico (GO *graphs*) obtido:

a) Qual é o gene?

Resposta: Gene de NADH dehydrogenase.

b) Qual o processo biológico relacionado ao gene e sua descrição?

Resposta: Transporte mitocondrial de elétrons. Envolvido na transferência de elétrons do NADH para a ubiquinona, que ocorre durante a fosforilação oxidativa, mediada pela enzima de múltiplas subunidades conhecidas como complexo I.

c) Qual a função molecular relacionada ao gene e sua descrição?

Resposta: A, não-covalente seletiva, muitas vezes estequiométrica, interação de uma molécula com um ou mais locais específicos sobre outra molécula.

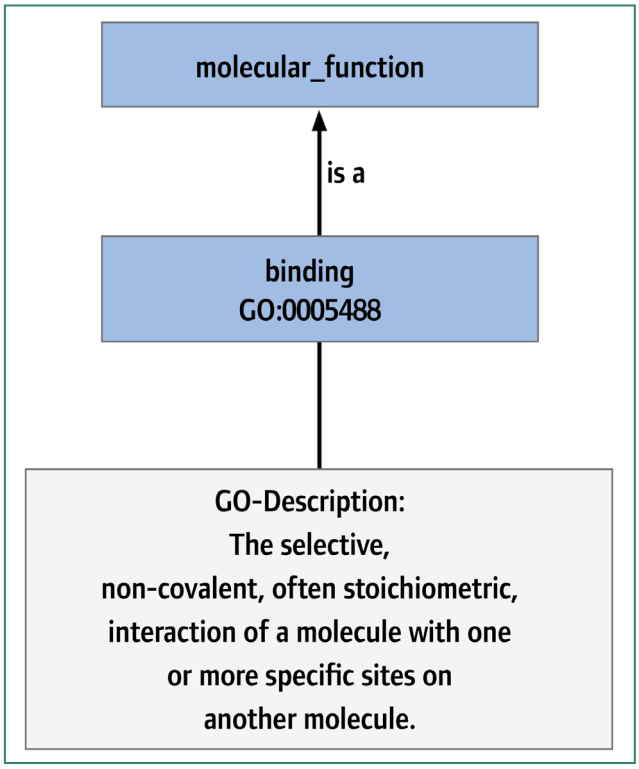


Figura 3.
Resultado obtido de função molecular para a sequência de *H. sapiens*.

4) Supor o seguinte: um pesquisador estudando diferentes condições de desenvolvimento em plantas, utilizando o organismo modelo *Arabidopsis thaliana*. Anotar sequências de DNA de indivíduos que foram submetidos a condições de estresse hídrico durante seu desenvolvimento. Para melhor entender os efeitos da privação de água em plantas.

a) Algum gene apresentou resposta ao tratamento de estresse hídrico?

Resposta: Fator 3 de crescimento meristemático em raiz.

b) Fazer uma análise do provável motivo do gene responder ao estresse.

Resposta: A expressão diferencial de um gene que induz o crescimento de raiz em um indivíduo que está sob estresse hídrico possivelmente está relacionada ao tratamento, uma vez que algumas plantas têm como estratégia aprofundar suas raízes para atingir níveis mais profundos do solo em busca de água quando expostas a condições desfavoráveis de abastecimento.

REFERÊNCIAS

ANTUNES, A. M.; CRUZ, K. S.; TELLES, M. P. C. Como desvendar enigmas genéticos a partir da comparação de sequências. *Revista Genética na Escola* v. 9, n. 2, p. 136-145, 2014.

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 Applications Note*, v.21, n.18, 2005.

GRIFFITHS, A. J. F.; WESSLER, S. R.; CARROLL, S. B.; LEWONTIN, R. C. *Introdução à Genética* – Rio de Janeiro:Guanabara, 10ª Ed., 2013.

The Gene Ontology Consortium. "The Gene Ontology project in 2008". *Nucleic Acids Res.* 36 (Database issue): D440–4. doi:10.1093/nar/gkm883. (January 2008).

VERLI, H. *Bioinformática: da Biologia à Flexibilidade Molecular*. 1ª edição, Porto Alegre, Brasil, 2014 (disponível para download gratuitamente na internet)