

Procurando agulha no palheiro: predição de genes a partir de sequências genômicas*

Lays Karolina Soares da Cruz¹, Adriana Maria Antunes¹, Mariana Pires de Campos Telles²

¹ Doutoranda no Programa de Pós-Graduação em Genética e Biologia Molecular, ICB, Universidade Federal de Goiás, Goiânia, Campus II, ICB I

² Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas I, Universidade Federal de Goiás, Campus II; Departamento de Biologia, Pontifícia Universidade Católica de Goiás

Autor para correspondência: tellesmpc@gmail.com

Palavras-chave: alinhamento, bioinformática, DNA, 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.



Esta atividade possibilita a consolidação de conhecimentos sobre as estratégias de análise genômicas por meio da predição de regiões gênicas a partir da análise de sequências de DNA e identificação das regiões gênicas preditas por meio da comparação com banco de dados. A atividade é indicada para estudantes de graduação e pós-graduação que já cursaram ou que estejam cursando disciplinas da área de Genética.

Os avanços nas tecnologias de sequenciamento, com o surgimento das plataformas de sequenciamento de nova geração (*Next Generation Sequencing* – NGS), têm facilitado o estudo dos genomas de inúmeras espécies pois as tecnologias NGS permitem gerar um grande volume de dados de sequências de nucleotídeos, em curto espaço de tempo e com custo reduzido. Os dados resultantes do sequenciamento dos genomas são primeiramente utilizados em uma etapa de montagem, a partir do uso de ferramentas de **bioinformática** para a obtenção de sequências consenso, presentes nos genomas. Essas sequências fornecem informações sobre a composição e ordem das bases nitrogenadas ao longo das moléculas de DNA. No entanto, o conhecimento sobre o genoma das espécies depende também da investigação do significado biológico dessas sequências. Nesse contexto, as estratégias de anotação do genoma têm sido aprimoradas, e apresentam grande importância nos estudos genômicos.

Na maioria das espécies, o genoma é constituído em grande parte por sequências não codificantes como, por exemplo, os **elementos repetitivos**, tais como **microssatélites** e **elementos transponíveis**. Em plantas, as sequências de DNA repetitivas podem representar até cerca de 90% do tamanho do genoma e, em humanos, esse percentual ultrapassa 50%. Em meio a essa grande quantidade de sequências estão os genes, sequências de nucleotídeos que contêm a informação capaz de codificar a síntese dos diferentes tipos de RNA, relacionados com o controle da maioria das funções biológicas nos organismos. Nesse sentido, é grande a importância do estudo dos genes que constituem os genomas dos diferentes organismos. Apesar de estarem presentes em pequenas quantidades na maioria dos genomas, as sequências codificantes desempenham papéis sem os quais a existência e a manutenção da vida não seriam possíveis.

Nas metodologias de anotação de genomas, um conjunto de protocolos e estratégias de trabalho podem ser utilizados para delimitar, em uma determinada sequência genômica, possíveis regiões gênicas, e prever a sua função. Nesse sentido, genes podem ser identificados usando características intrínse-

cas dos genomas, como códons de iniciação e parada, ou usando métodos de similaridade, que permitem que a identificação seja realizada por meio da semelhança com genes já conhecidos e armazenados em bancos de dados. O mecanismo de delimitação da sequência gênica depende do organismo cuja sequência genômica foi determinada, uma vez que existem diferenças consideráveis nas estruturas de genes de organismos procarióticos e eucarióticos. Em procariotos, os genes codificantes de proteínas são colineares em seus produtos gênicos, enquanto os genes eucarióticos geralmente possuem intróns, que são sequências não codificantes intragênicas.

Identificar genes em genomas grandes e complexos, como dos eucariotos, não é uma tarefa fácil. O grande volume de sequências obtido a partir dos dados de sequenciamento faz essa tarefa se assemelhar à busca de uma agulha num palheiro. No entanto, a identificação de genes é possível devido ao desenvolvimento de softwares eficientes para predição de sequências com estrutura conservada e que podem ser consideradas um gene em potencial. Entre as diferentes estratégias para discriminar regiões codificantes e não codificantes, a procura de ORFs (ORF = *open reading frame* = matriz de leitura aberta) constitui uma estratégia simples e direta. Existem softwares que identificam genes usando as longas matrizes de leitura abertas (ORFs) pois, no **código genético**, os códons de iniciação (ATG) e de parada (TAA, TAG ou TGA) são bem definidos. Ao identificar uma ORF, encontra-se uma sequência de DNA que tem potencial para codificar um peptídeo, pois possui um códon de início e um códon de parada. As ORFs são identificadas considerando as 6 possibilidades de leitura da dupla fita de DNA. Como o código genético é formado por trinças de nucleotídeo, existem três possibilidades de leitura na fita com **orientação 3'-5'** e três possibilidades de leitura na fita com **orientação 5'-3'**.

Um dos softwares, que permite a realização da anotação de genes por meio da identificação de ORFs, é o *ORFfinder*, disponibilizado via web (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>); com o *ORFfinder* é possível encontrar todos os quadros de leitura aberta

Bioinformática: Conjunto de sistemas computadorizados de informação e métodos analíticos aplicados a problemas biológicos, como a análise genômica.

Elementos repetitivos são sequências de nucleotídeos presentes em múltiplas cópias no genoma.

Microssatélite é uma região do genoma composta de motivos curtos (cerca de 2 a 6 pb) que se repetem em tandem. Alelos diferentes têm números diferentes de repetições do motivo.

Código genético: conjunto de correspondências entre trinças de nucleotídeos no RNA e aminoácidos na proteína.

Elemento transponível é um segmento de DNA que pode se mover de uma posição para outra no genoma.

Orientação 3'-5': orientação da cadeia de polinucleotídeos na fita de DNA *forward*.

Orientação 5'-3': orientação da cadeia de polinucleotídeos na fita de DNA *reverse*.

Montagem de sequência

é a compilação de milhares ou milhões de leituras independentes de sequências de DNA em um conjunto de *contigs*.

Genoma é o complemento inteiro de material genético em um conjunto de cromossomos.

FASTA: Formato de arquivo em que na primeira linha há o símbolo maior (>) seguido pelo título da sequência e nas linhas seguintes as sequências de bases nitrogenadas.

Códon é o trecho do RNA (do tamanho de três nucleotídeos) que codifica um único aminoácido.

Valores de score: no alinhamento, os nucleotídeos correspondentes entre *Query* e *Subject* são chamados de *match* e, aqueles que não são correspondentes, de *mismatch*. Regiões onde existe inserção ou deleção de forma que uma das sequências possui uma base a menos é chamada de *gap*. No cálculo do *score*, *matches* recebem penalidade positivas e, *mismatches* e *gaps*, negativas. O melhor alinhamento é aquele que possui maior valor de *score*, ou seja, maior número de *matches*.

E-valor representa a chance da similaridade entre as sequências ser devido ao acaso. O melhor alinhamento é aquele que apresenta um E-valor baixo, geralmente menor que 10^{-20} .

em arquivos de sequências de nucleotídeos em formato *fasta*. A sequência é inserida no quadro de busca do software e a identificação das ORFs é realizada tomando-se por base o código genético. Na apresentação do resultado é identificado em que fase de leitura a ORF foi encontrada, em que posição (número do nucleotídeo) a sequência codificante começa e termina e qual o seu tamanho total. No entanto, essa ferramenta não indica quais são os genes encontrados, apenas a sua estrutura em termos de sequências de bases nitrogenadas. Para identificar a função das regiões potencialmente gênicas, encontrados pelo *ORFfinder*, é preciso usar a ferramenta *BLAST*.

O *BLAST* é uma ferramenta também disponibilizada via web (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BLASTHome) que permite identificar sequências por meio da comparação por similaridade com sequências armazenadas em bancos de dados. A identificação é realizada por meio do alinhamento da sequência alvo (que recebe o nome de *Query*), com sequências do banco de dados (que são chamadas de *Subject*), visando encontrar regiões de similaridade. O *Query* é alinhado com vários *Subjects* e o melhor alinhamento é definido com base em **valores de score** e **E-valor**, parâmetros de confiança do *BLAST*. Existem vários tipos de *BLAST*, sendo que eles diferem entre si pelo tipo de sequência analisada e tipo de banco de dados consultados. Para estudo dos genes, os mais usados são o *BLAST nucleotide (BLAST n)*, que permite procurar uma sequência de nucleotídeos em um banco de dados de sequências de DNA, e o *BLAST protein (BLASTp)*, que possibilita procurar uma sequência de aminoácidos em um banco de dados de proteínas. Dessa forma, ao analisar a sequência usando o *BLASTn* é possível saber que gene está sendo investigado e, ao usar o *BLASTp*, é possível saber qual a proteína codificada por aquele gene, o que fornece evidências de informações biológicas/funcionais.

FUNÇÃO PEDAGÓGICA

A principal função da atividade proposta é mostrar uma das possibilidades de anota-

ção de sequências de nucleotídeos, obtidas a partir de dados de sequenciamento de ácidos nucleicos. As tecnologias de sequenciamento, principalmente as de nova geração (NGS - *Next Generation Sequencing*), produzem grandes volumes de dados que permitem a **montagem de sequências** consenso do genoma. Nesse contexto, a anotação é o processo que permite extrair o significado biológico das sequências analisadas. Dentre as possibilidades de descrever a estrutura do **genoma**, a anotação de genes vem ganhando cada vez mais importância, uma vez que a expressão gênica determina o funcionamento dos organismos. A atividade proposta tem por objetivo apresentar uma estratégia possível para a predição de genes a partir de sequências do genoma. Assim, as análises serão realizadas com base em arquivos *fasta* com trechos de diferentes genomas e os softwares *ORFfinder* e *BLAST*. O *ORFfinder (Open Reading Frame Finder)* permite identificar regiões que são genes em potencial, com base em características intrínsecas do genoma, tais como os **códons** de iniciação e de parada. O *BLAST (Basic Local Alignment Search Tool)* permite investigar a similaridade dessas sequências com a dos genes que já foram descritos e depositados nos bancos de dados, a partir de metodologias de alinhamento. Com isso, a atividade pretende propiciar aos estudantes a consolidação de seus conhecimentos sobre análises genômicas.

PROBLEMA PROPOSTO

Esta atividade propõe a análise de quatro sequências de nucleotídeos fornecidas em arquivos no formato *fasta*, a fim de realizar a anotação de genes. Inicialmente, os dados de sequências deverão ser submetidos à análise no *ORFfinder* para identificar as sequências potencialmente gênicas. Em seguida, essas sequências serão analisadas, usando o *BLASTn* e o *BLASTx*, para descobrir quais são os genes encontrados e os produtos que eles codificam. Posteriormente, esses genes deverão ser estudados, visando associá-los a um processo biológico específico. Dessa forma, será possível investigar informações funcionais para o conjunto de genes identificados.

INSTRUÇÕES PARA O PROFESSOR

1. É recomendável que essas atividades sejam realizadas com estudantes de Graduação ou Pós-Graduação que possuam conhecimento básico de Genética e Biologia Molecular. Para melhor entendimento das atividades, os estudantes devem ter conhecimentos prévios sobre código genético, processos de transcrição, tradução e expressão gênica, além da estrutura dos genomas de procariotos e eucariotos. É importante também conhecer as metodologias de sequenciamento (clássico e as plataformas de nova geração).
2. As atividades poderão ser realizadas durante a aula, em laboratório de informática, ou como uma atividade extra-classe. No caso da realização em aula prática, o professor pode fornecer as instruções e formar grupos para a realização da atividade. O professor pode solicitar que cada estudante entregue um relatório ao final da atividade, contendo todas as etapas e todos os resultados alcançados.
3. O professor deverá fornecer aos estudantes os quatro arquivos fasta e o PDF com as orientações para realizar a atividade e as questões que devem ser respondidas. As quatro sequências, para serem analisadas, estão disponíveis nos links a seguir: **1**, **2**, **3** e **4**. Elas devem ser enviadas para os e-mails dos estudantes.
4. É recomendável que os estudantes sigam o procedimento abaixo descrito e que respondam as questões incluídas para discussão.

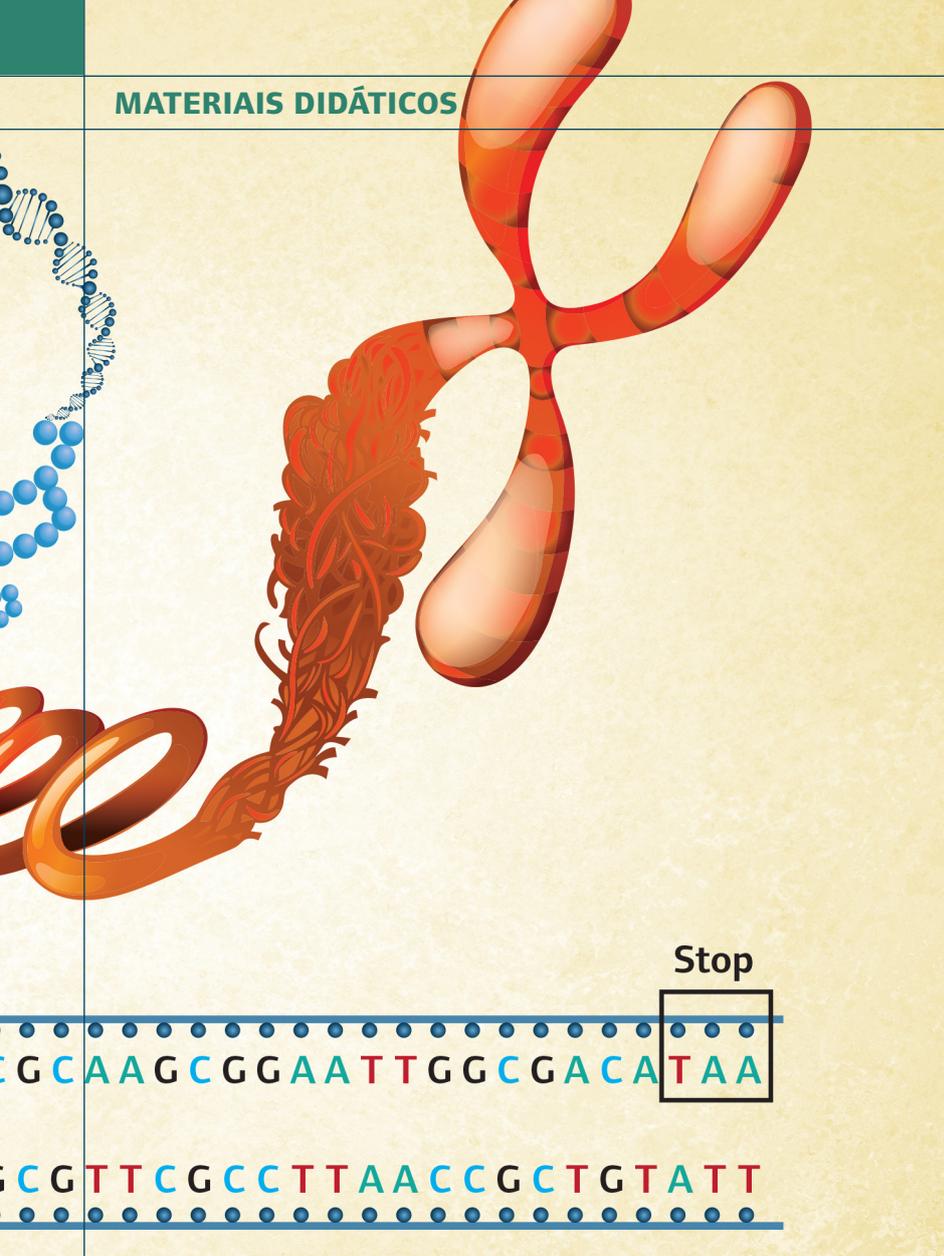
PROCEDIMENTOS PARA OS ESTUDANTES

Acessar o site do NCBI na página em que está disponível a ferramenta *ORFfinder* (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>). Utilizando os arquivos contendo as sequências enviadas pelo professor, dar entrada em cada sequência separadamente na janela “Enter GI or ACCESSION or sequence in FASTA format” do *ORFfinder*. A entrada de dados é realizada copiando a sequência do

arquivo e colando na janela indicada. Clicar no botão “*OrfFinder*” e explorar o resultado como exposto durante a aula para responder as perguntas.

Após a submissão da sequência, a página exibirá uma análise gráfica dos quadros de leitura encontrados. O algoritmo do *ORFfinder* está preparado para lidar com sequências não codificantes que estejam presentes no arquivo como, por exemplo, regiões intrônicas. No gráfico apresentado como resultado são exibidas seis barras de leitura (*Six Frames*), onde as sequências não codificantes são mostradas em branco e, as sequências exônicas, em azul. Além disso, também são exibidos os *frames* dos quadros de leitura putativos identificados. O número correspondente à posição da base de início e término da ORF e o tamanho de cada ORF, em azul, corresponde, proporcionalmente, ao seu tamanho em pares de bases.





Stop

...G C A A G C G G A A T T G G C G A C A T A A

...C G T T C G C C T T A A C C G C T G T A T T

De modo geral, a maior ORF encontrada é o verdadeiro gene contido na sequência do arquivo *fasta*. Acessando essa ORF, ela passará a ser exibida na cor roxa, que indica que as informações visualizadas na página nesse momento são referentes à mesma. Nessa página é possível checar a sequência de nucleotídeos que forma o quadro de leitura putativo. Nesse momento é importante que seja feita uma checagem dos códons de iniciação e parada, que aparecem destacados nas cores azul (códon de iniciação) e rosa (códon de parada). A partir dessa checagem, o usuário precisa assumir que esse é o quadro de leitura correto para prosseguir com a análise. Caso o resultado não seja satisfatório, é possível retornar e tentar uma ORF diferente. Para assumir a ORF como o verdadeiro quadro de leitura do gene e prosseguir, basta clicar em “Accept” e, em seguida a sequência do quadro selecionado aparecerá em verde.

Acessando novamente o quadro de leitura, que está em verde, estará disponível a opção do *BLAST*. Esta opção será exibida em um quadro de ferramentas logo acima da representação gráfica da sequência, onde é possível escolher entre *BLASTn* e *BLASTp*. O usuário deve selecionar uma destas duas opções e clicar em “BLAST” para que o alinhamento seja executado. O estudante deve proceder com as análises das duas opções de *BLAST* para conseguir responder corretamente as perguntas dessa atividade. Assim, é necessário que, após gerar o resultado na primeira opção, o estudante retorne para a página e escolha a segunda opção de alinhamento (*BLASTp*) para obter o novo resultado.

Após a submissão do *BLAST*, será aberta uma nova página que contém informações sobre o *Query* e as opções de resultados e parâmetros. É recomendado que os estudantes mantenham o formato de saída do resultado do *software*, uma vez que ele será exibido em HTML. Nesta etapa é preciso clicar em “View report” e o usuário será guiado à página de resultados. O *BLAST* também gera uma análise gráfica como resultado, sendo que cada cor na imagem representa um *score* de alinhamento, sendo que o maior valor de *score* implica uma maior similaridade do *query* com o *subject*. Os resultados também são apresentados em *hits*, dos quais o primeiro (*Best Hit*) costuma ser o melhor resultado da lista apresentada. O estudante também pode acessar a base de dados “gene” do NCBI para encontrar informações complementares sobre os resultados do *BLAST*.

Nesse momento é importante o usuário checar as informações de qualidade do alinhamento apresentadas nos resultados, como o *score*, no qual o maior valor representa o melhor alinhamento; o *e-value*, que apresenta a chance de cometer erro do tipo I; a porcentagem de identidade entre *query* e *subject*; e o *Accession* que contém informações sobre o gene encontrado. Se as informações geradas no *BLAST* forem estatisticamente relevantes e as informações do gene fizerem sentido com a pergunta biológica, significa que o usuário provavelmente encontrou a ORF verdadeira do gene em questão e o resultado deve ser tratado como putativo até sua con-

firmação por outras técnicas de validação, tais como nocaute gênico, estudos de níveis de expressão gênica e outros.

QUESTÕES

Questão 1. O que é uma ORF (*Open Reading Frame*)?

Questão 2. O que é uma região conservada em uma sequência genômica?

Questão 3. O que significa função putativa?

Questão 4. Para cada uma das sequências fornecidas pelo professor, responder às seguintes questões:

A. Quantas ORFs foram encontradas na sequência e em que fase de leitura estava a maior delas?

B. Qual o nome do gene putativo encontrado pelo *ORFfinder* para a matriz de

leitura, qual a espécie e qual o tamanho da maior matriz de leitura em pares de bases e aminoácidos?

C. Qual o *score* e *E-value* retornado pelo *BLASTn* para o melhor alinhamento da ORF selecionada e submetida?

D. Com que espécie houve maior similaridade entre *query* e *subject*. Essa espécie é um eucarioto ou um procarioto?

Questão 5. Faça uma pesquisa e escreva uma descrição sumarizada da proteína apontada no *Best Hit* do *BLASTp*.

Questão 6. Fazer uma análise comparativa da representação gráfica gerada para os genes entre organismos procariotos e eucariotos quando os genes das quatro sequências são utilizados na busca do banco de dados “gene” do NCBI.



RESPOSTAS DAS QUESTÕES

Questão 1. O que é uma ORF (*Open Reading Frame*)?

ORF, do inglês *Open Reading Frame*, significa matriz de leitura aberta, ou seja, um sítio específico que, quando lido pela maquinaria de transcrição da célula, codifica um RNA.

Questão 2. O que é uma região conservada em uma sequência genômica?

Regiões conservadas do genoma são partes específicas de sequências de nucleotídeos mais preservados entre os organismos, o que aumenta as chances de identidade com sequências da mesma região genômica em outros indivíduos, principalmente indivíduos mais próximos à espécie em questão. Por exemplo, os códons de aminoácidos são regiões conservadas.

Questão 3. O que significa função putativa?

É a função predita para um elemento com base em informações teóricas. Essa

função, no entanto, carece de confirmação fora do âmbito teórico, ou seja, a partir de informações geradas em experimentos, para que seja assumida como uma função verdadeira.

Questão 4. Para cada uma das sequências fornecidas pelo professor, responder às seguintes questões:

A. Quantas ORFs foram encontradas na sequência e em que fase de leitura estava a maior delas?

Sequência 1. Foram encontradas 16 ORFs, e a maior delas estava em *frame* +1 (Figura 1).

Sequência 2. Foram encontradas 27 ORFs, e a maior delas estava em *frame* +1.

Sequência 3. Foram encontradas 27 ORFs, e a maior delas estava em *frame* +1.

Sequência 4. Foram encontradas 53 ORFs, e a maior delas estava em *frame* +3.

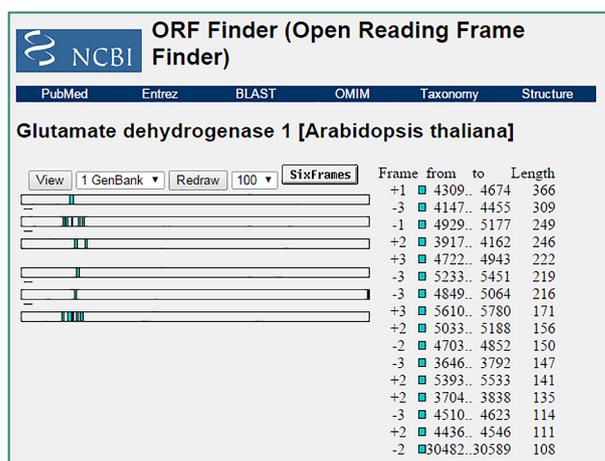


Figura 1.

Representação gráfica das matrizes de leitura abertas encontradas pelo *ORFfinder* para o gene *Glutamate dehydrogenase 1* [*Arabidopsis thaliana*]. NCBI 2015.

B. Qual o nome do gene putativo encontrado pelo *ORFfinder* para a matriz de leitura, qual a espécie e qual o tamanho da maior matriz de leitura em pares de bases e aminoácidos?

Sequência 1. O nome do gene putativo encontrado foi *Glutamate dehydrogenase*

1 [*Arabidopsis thaliana*], com tamanho de 366pb e 122 aminoácidos.

Sequência 2. O nome do gene putativo encontrado foi DNA helicase [*Escherichia coli*], com tamanho de 5271pb e 1757 aminoácidos.

Sequência 3. O nome do gene putativo encontrado foi *Ribosomal protein SA* [*Homo sapiens*], com tamanho de 411pb e 137 aminoácidos.

Sequência 4. O nome do gene putativo encontrado foi *Mitochondria localized glutamic acid rich protein* [*Mus musculus*], com tamanho de 675pb e 225 aminoácidos.

C. Qual o *score* e *E-value* retornado pelo BLASTn para o melhor alinhamento da ORF selecionada e submetida?

Sequência 1. O *score* do melhor alinhamento retornado pelo BLASTn foi 171, e o *E-value* foi 2e-49.

Sequência 2. O *score* do melhor alinhamento retornado pelo BLASTn foi 3564, e o *E-value* foi 0,0.

Sequência 3. O *score* do melhor alinhamento retornado pelo BLASTn foi 111, e o *E-value* foi 1e-28.

Sequência 4. O *score* do melhor alinhamento retornado pelo BLASTn foi 347, e o *E-value* foi 2e-117.

D. Com que espécie houve maior similaridade entre *query* e *subject*. Essa espécie é um eucarioto ou um procarioto?

Sequência 1. Houve maior similaridade com a espécie *Arabidopsis thaliana*, que é um eucarioto.

Sequência 2. Houve maior similaridade com a espécie *Escherichia coli*, que é um procarioto.

Sequência 3. Houve maior similaridade com a espécie *Macaca mulatta*, que é um eucarioto.

Sequência 4. Houve maior similaridade com a espécie *Mus musculus*, que é um eucarioto.

Questão 5. Faça uma pesquisa e escreva uma descrição sumarizada da proteína apontada no *Best Hit* do BLASTp.

Sequência 1. Codifica a subunidade alfa de 43 kDa da desidrogenase do glutamato com um polipeptídeo putativo mi-

tocondrial e trânsito de NAD(H)- e os domínios de ligação de alfa-cetoglutarato. Localização mitocondrial confirmada pelo fracionamento subcelular. Combina com várias relações com a proteína GDH2 (GDH-beta) para formar sete isoenzimas. Catalisa a clivagem de resíduos de glicina. Podem estar envolvidos na assimilação de amônia em condições de excesso de nitrogênio inorgânico. A enzima está quase que exclusivamente nas mitocôndrias da folha.

Sequência 2. DNA helicases são enzimas que catalisam o desenrolamento do DNA dupla fita durante a replicação, recombinação e reparo. Estas enzimas têm sido extensivamente estudadas, no entanto, os detalhes específicos de como qualquer helicase desenrola DNA dupla fita são desconhecidos. O gene que codifica para esta enzima é conservado em chimpanzés, macaco *Rhesus*, cães, peixe-zebra, *C. elegans*, arroz e outros.

Sequência 3. As lamininas, uma família de glicoproteínas da matriz extracelular, são o principal constituinte não colagenosos das membranas basais. Tais não colagenosos têm sido implicados numa grande variedade de processos biológicos incluindo a adesão celular, diferenciação, migração, sinalização, o crescimento de neurites e a metástase. Muitos dos efeitos das lamininas são mediados através de interações com receptores da superfície celular.

Sequência 4. A proteína codificada é uma subunidade de NADH: ubiquinona oxidoreductase (complexo I), o primeiro complexo de enzimas na cadeia de transporte de elétrons localizado na membrana mitocondrial interna, é resultado de *splicing* alternativo em múltiplas variantes de transcritos.

Questão 6. Fazer uma análise comparativa da representação gráfica gerada para os genes entre organismos procariotos e eucariotos quando os genes das quatro sequências são utilizados na busca do banco de dados “gene” do NCBI.

Quando a busca é feita na base gene do NCBI, o resultado apresentado para eucariotos e procariotos é bastante diferente. Em organismos eucariotos, a sequência é representada de maneira interrompida (Figura 2), sendo que as caixas coloridas representam os exons, e a parte pontilhada representa os introns.

No organismo procarioto, a sequência é apresentada de maneira contínua (Figura 3), sem interrupções pontilhadas e formação de caixas, isso acontece pelo fato de, nesses organismos, não serem encontrados introns, sendo somente os exons dispostos lado a lado no genoma.

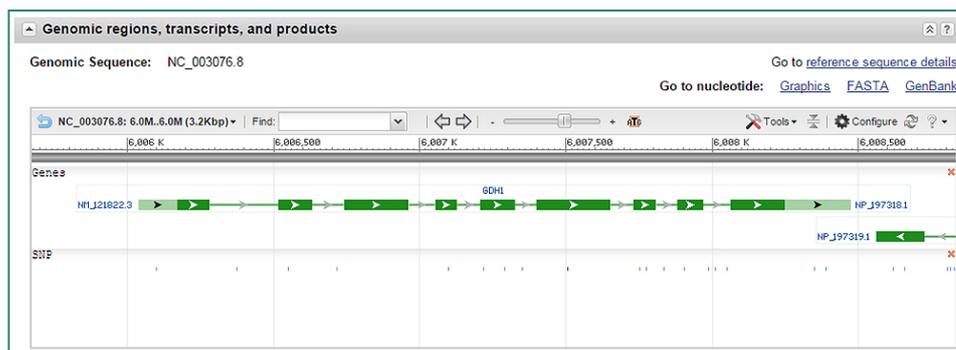


Figura 2.

Representação gráfica do gene *GDH1* glutamate dehydrogenase *1* *Arabidopsis thaliana* (thale cress)]. NCBI 2015.

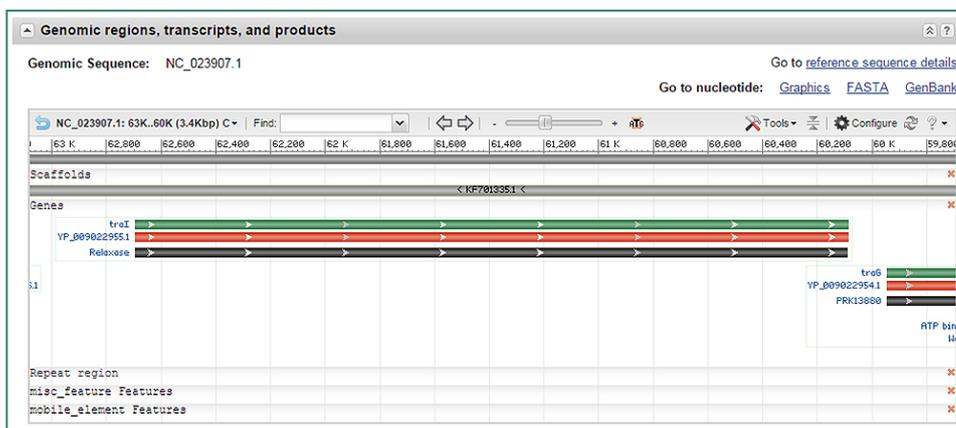


Figura 3.

Representação gráfica do gene tral conjugal transfer nickase/ helicase *TrnI* [*Escherichia coli*]. NCBI 2015.