Como desvendar enigmas genéticos a partir da comparação de sequências*

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

[1] Bolsista CAPES no Programa de Pós-Graduação em Genética e Biologia Molecular, ICB, Universidade Federal de Goiás, Goiânia, Campus II, ICB I.
[2] Bolsista PQ1D-CNPq, Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Campus II, ICB I.

Autor para correspondência: adrianaantunesbio@gmail.com

*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.

A principal função da atividade é demonstrar como funciona uma das etapas para a análise dos dados de sequenciamento de primeira e/ou segunda geração, utilizando uma ferramenta da Bioinformática conhecida como "BLAST" (*Basic Local Alignment Search Tool*). Ao resolver os enigmas propostos, que simulam situações reais, graduandos ou pós-graduandos manipularão as ferramentas em busca de respostas e consolidarão seus conhecimentos sobre a análise de dados de sequenciamento e as múltiplas possibilidades de uso e aplicações em diferentes áreas.

A principal função da atividade proposta é mostrar uma das etapas para a análise dos dados de sequenciamento, utilizando uma ferramenta da Bioinformática conhecida como "BLAST" (*Basic Local Alignment Search Tool*). As metodologias de sequenciamento geram uma grande quantidade de dados que podem ser organizadas e avaliadas pelas ferramentas analíticas da Bioinformática, uma ciência que usa a matemática e a computação para armazenar, comparar e analisar sequências do DNA, RNA ou de proteínas. Por meio do BLAST é possível inferir relações funcionais e evolutivas entre sequências. A atividade proposta apresenta três sequências de nucleotídeos, denominadas enigma 1, enigma 3 e enigma 5 e duas sequências de aminoácidos, denominadas enigma 2 e enigma 4 que deverão ser analisadas usando a ferramenta "BLAST" para analisar a similaridade entre as sequências que estão sendo analisadas, em comparação com outras sequências armazenadas nos bancos de dados públicos. A atividade permitirá aos estudantes consolidar o conhecimento sobre a análise de dados de sequenciamento e, assim, ampliar a visão e as possibilidades de análises no nível molecular, uma área importante da genética que tem interface com as diferentes áreas das ciências.

PROBLEMA PROPOSTO

O DNA (ácido desoxirribonucleico) e o RNA (ácido ribonucleico) armazenam e transmitem a informação genética nos organismos. Tanto o DNA quanto o RNA são compostos por unidades químicas (monômeros) chamadas de nucleotídeos, unidos em longas cadeias não ramificadas. Por outro lado, a estrutura de uma proteína é comumente definida em termos de quatro níveis hierárquicos. A estrutura primária diz respeito à sequência de resíduos de aminoácidos. A estrutura secundária inclui resíduos de aminoácidos arranjados em conformações particularmente estabilizadas por pontes de hidrogênio, que originam padrões regulares e repetitivos. A estrutura terciária inclui todos os aspectos do padrão de enovelamento tridimensional de uma proteína. Proteínas que apresentam duas ou mais subunidades possuem ainda a estrutura quaternária, que descreve como as várias subunidades estão dispostas no espaço. As metodologias e o equipamento disponíveis atualmente nas áreas de Genética e Biologia Molecular possibilitam sequenciar o DNA, o RNA e as proteínas, permitindo investigar questões biológicas importantes em diversas áreas das ciências.

As metodologias de sequenciamento permitem identificar a ordem das bases nitrogenadas no DNA ou de aminoácidos nas proteínas. Com o passar dos anos, essas metodologias têm aprimorado e possibilitado a geração de um volume de dados cada vez maior. O **método de Sanger** foi proposto primeiramente para analisar sequências de nucleotídeos e durante vários anos foi o único método disponível. Na última década surgiram novas tecnologias de sequenciamento de DNA, denominados de "**segunda geração**" ou, como conhecido em inglês, *Next Generation Sequencing* (NGS), que têm permitido sequenciar genomas inteiros e gerar um volume maior de sequências, em um curtíssimo espaço de tempo, e com custo reduzido. O sequenciamento de proteínas também utiliza novas tecnologias, tais como a **espectrometria de massa**, que também tem permitido uma aceleração na disponibilização de sequências de aminoácidos de proteínas descobertas por grupos de pesquisas em todo o mundo. Um importante banco de dados de sequências é o GenBank (http://www.ncbi.nlm.nih.gov/genbank/) que armazena um grande conjunto de dados que é disponibilizado de forma gratuita ao público.

Uma das maneiras de comparar sequências de nucleotídeos ou aminoácidos é a partir da utilização da ferramenta "BLAST", um algoritmo que analisa comparativamente a sequência de interesse com outras sequências conhecidas e armazenadas nos banco de dados, de diversos organismos. Existem diferentes tipos de BLAST (Tabela 1) que podem ser acessados a partir do *web site* do NCBI (*National Center for Biotechnology Information*) (http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi?CMD=Web&PAGE_TYPE=BLASTHome).

Método de Sanger, também conhecido como método dideoxi, requer a síntese enzimática de uma fita de DNA complementar à fita analisada usando dideoxinucleotídeos, nucleotídeos que quando inseridos interrompem a síntese do DNA, pois não possuem o grupo hidroxila no carbono 3' que é necessário para formação da próxima ligação fosfodiéster. São gerados fragmentos de diferentes tamanhos que são analisados em gel de eletroforese. A análise conjunta dos fragmentos no gel permite ler a sequência de nucleotídeos do fragmento.

Existem várias tecnologias de **sequenciamento de segunda geração** (NGS) comercialmente disponíveis, entre elas a Roche454, a Illumina Solexa, a Life Technologies (SOLiD) e a Helicos BioSciences. Cada uma delas possui características particulares, mas no geral usam métodos baseados em amplificação ou de molécula única na preparação de reações. Para o sequenciamento existem vários métodos como ciclo de terminação reversível, sequenciamento por ligação ou piro sequenciamento.

Para a identificação por **espectrometria de massa**, as proteínas são separadas por eletroforese ou cromatografia e em seguida analisadas por ionização a laser, técnica através da qual são gerados íons e estes analisados através da relação massa-carga.

Tipo	Análise realizada
BLASTn	Procura uma sequência de nucleotídeos em um banco de dados de sequências de DNA
BLASTp	Procura uma sequência de aminoácidos em um banco de dados de proteínas
BLASTx	Procura uma sequência de nucleotídeos em um banco de dados de proteínas
tBLASTn	Procura uma sequência de aminoácidos em um banco de dados de nucleotídeo
tBLASTx	Procura uma sequência de nucleotídeos traduzidos em um banco de dados de nucleotídeos traduzidos

Tabela 1.
Análises realizadas com cada tipo de BLAST.

Alinhamento de sequência – O BLAST faz o alinhamento simples e local. O alinhamento simples é aquele realizado entre duas sequências. O alinhamento local é aquele no qual apenas pedaços ou trechos das sequências são comparados.

No BLAST, a análise de similaridade entre sequências é realizada por meio do **alinhamento da sequência** submetida ao programa (chamada de "*Query*"), com sequências do banco de dados (denominada "*Subject*"). Alinhar duas sequências consiste em comparar a correspondência de nucleotídeos entre elas. Dessa forma, o alinhamento organiza e compara sequências visando identificar regiões de similaridade. No alinhamento, os nucleotídeos nos quais as bases são correspondentes ou iguais são chamados de "match" e nas bases em que não são correspondentes de "*mismatch*". No alinhamento também podem existir regiões com inserção e/ou deleção de um ou mais nucleotídeos na sequência (*indel*), onde se pode inserir um hífen, indicando um "***gap***" (Figura 1).

Gap são espaços introduzidos entre os nucleotídeos ou aminoácidos a fim de obter o melhor alinhamento possível.

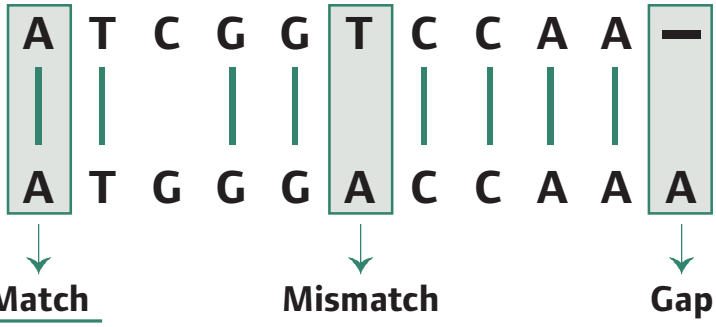


Figura 1.
Exemplo de alinhamento de sequência de nucleotídeos do DNA.

Quando todas as bases do *Query* pareiam com as bases da sequência *Subject*, o alinhamento recebe o nome de "Full-Lenght". No entanto, nem sempre as sequências são alinhadas por completo. Em muitos casos, apenas trechos conservados entre as sequências são correspondentes no alinhamento. Nesse sentido, durante a análise no BLAST a sequência de interesse (*Query*) é alinhada com várias sequências do banco de dados (*Subject*) e pode ser escolhido o melhor alinhamento com base em alguns parâmetros como, por exemplo, o valor de *score* e o *E-value*.

O valor de *score* para cada alinhamento é calculado por meio da atribuição de valores negativos de penalidade para os gaps e *mismatchs*. O melhor alinhamento é aquele que possui maior valore de *score*. O *E-value* também é um parâmetro de confiança para análise de alinhamentos e corresponde ao número esperado de sequências com similaridade tão alta quanto a encontrada por acaso. Quanto maior o *E-value*, menor a confiança de que as correspondências sejam reais, e também maior a chance da similaridade entre as sequências serem devido ao acaso. Geralmente um alinhamento só é considerado real se o *E-value* for menor que 10^{-20}, valor utilizado como ponto de corte.

Quando uma sequência é analisada no BLAST, é importante verificar o valor de *score* e o *E-value* para decidir qual o melhor alinhamento. O melhor resultado da análise no BLAST é chamado de "*Best-Hit*" e, na maioria das vezes, ele é o primeiro resultado da tabela de resultados gerada no BLAST. Se a sequência em análise não tiver similaridade significativa com nenhuma sequencia do banco de dados, tem-se um "*No hit*". Neste caso o *E-value* possui um valor alto.

Nesta atividade está sendo proposta uma análise por meio do BLAST de cinco sequências denominadas "enigma". As sequências do enigma 1, do enigma 3 e do enigma 5 são compostas de nucleotídeos, enquanto as sequências do enigma 2 e do enigma 4 são compostas de aminoácidos. Para a análise das sequências "enigmas" serão usados os cinco tipos de BLAST. O primeiro desafio da atividade é escolher o tipo de BLAST adequado para solucionar o questionamento de cada enigma. Ao submeter as sequências para análise no BLAST, é preciso anotar os valores de *score* e *E-value* e definir o alinhamento *Best-Hit* para responder às questões complementares.

Sequência enigma 1

- enigma 1

TGCAAAAGCAAAATTGATTACTCTGTTTTGGTTTTGGATCAGATCAACCAAAGGAAGCGAAGAAGAAGAAGACTTTATACAC
TCTCGACGCTATTGATTTCGTCTGTGCAGTTTTGTTTCTCTATCAGTTAGTTTTAAAGATGAGTTCTTCAGAGAGTGTGGAAAAC
GAGTGCATGTGTTGGGCTGCAAGAGATCCATCTGGTCTTCTTTCTCCTCATACTATCACTCGCAGGTCTGTTACAACTGACGATG
TTTCACTCACAATCACTCATTGTGGAGTGTGTTACGCTGATGTTATCTGGAGTAGAAACCAACATGGAGACTCCAAGTACCCTT
TGGTTCCTGGGCATGAGATTGCTGGAATAGTGACTAAGGTTGGGCCTAATGTTCAACGATTCAAAGTTGGAGACCATGTTGGTG
TTGGAACGTATGTTAACTCCTGCAGGGAGTGTGAATATTGTAATGAAGGACAAGAAGTTAATTGTGCGAAAGGAGTTTTTACT
TTCAATGGCATTGATCATGATGGCTCTGTTACTAAAGGAGGCTACTCTAGTCACATTGTTGTTCATGAAAGGTACTGCTACAAGA
TACCTGTGGACTATCCCTTGGAATCAGCTGCACCATTACTCTGTGCTGGAATCACGGTTTATGCTCCTATGATGCGTCACAATAT
GAATCAACCTGGTAAATCTCTTGGGGTGATCGGGCTAGGTGGTCTTGGACACATGGCGGTTAAGTTTGGCAAGGCTTTTGGACT
TAGTGTTACGGTTTTTAGCACCAGCATTTCCAAGAAAGAAGAAGCTTTGAATCTGCTAGGAGCTGAGAATTTCGTTATCTCATCT
GACCATGACCAGATGAAGGCACTAGAGAAATCTCTAGACTTTCTAGTTGACACAGCATCTGGTGATCACGCGTTTGATCCTTACA
TGTCTCTCTTGAAGATTGCTGGAACTTATGTATTGGTTGGTTTCCCAAGTGAAATTAAAATCAGTCCTGCCAATCTCAATCTTGG
TATGAGAATGCTCGCTGGAAGTGTAACCGGGGGGACCAAAATAACACAGCAAATGTTAGATTTCTGTGCAGCTCATAAGATTTA
TCCAAACATAGAGGTGATTCCCATTCAAAAGATAAACGAAGCTCTCGAAAGAGTGGTGAAGAAGGACATCAAGTACCGTTTCG
TGATTGACATCAAGAACTCCCTCAAATAGATGTTGCTCAAAGGAAGGAATAATGGAGTCTGTAATAAGAGAATAATACTCACTGC
TACAATCTTTATTACGTATTTTCTCGTTTTTCATTAGTAAAGCAAATAAATTAAGACT

Sequência enigma 2

- enigma 2

VSSCSGVSPGVAGPRQGSQGWVELGFSLGGSGTNPLVLVGKPRPVPARCVLWEERGLGALGWSAGQQRSPEWERPSRAVSPS
SLPQGTEVHDRRGEAVRALYQELLCSGRPASPVSPRPAPPLGLRRAWSRPSPLSHRPGCLDRSHPHLPQCFSKQKFGSLALA
GPVSAV

Sequência enigma 3

- enigma 3

GGTTCCTCTTTTTCTATCTTCCTTCTAGCCCTTCTCTCTTGCTTGACCGTGCCCGCGTCAGCCTACCAAGTACGCAACTCCTCGGG
CCTCTACCATGTCACCAATGATTGCCCTAACTCGAGTATTGTGTACGAGACGGCCGATACCATACTACACTCTCCGGGGTGTG
TCCCTTGCGTTCGCGAGGGTAACAACTCGAGGTGTTGGGTGGCGGTGGCCCCCACAGTCGCCACCAGGGACGGCAAACTCCC
CACAACGCAGCTTCGACGTCATATCGATCTGCTTGTCGGGAGCGCCACCCTTTGCTCGGCCCTCTATGTGGGGGACTTGTG
CGGGTCTGTCTTTCTTGTCGGCCAACTGTTCACCTTCTCCCCCAGGCGCCACTGGACAACGCAAGACTGCAACTGCTCTATC
TACCCCGGCCATATAACGGGTCACCGAATGGCATGGGATATGATGATGAACTGGTCCCCTACAACAGCGCTGGTAGTAGCTCAGC
TGCTCAGGGTCCCGCAAGCCATCGTGGATATGATCGCTGGTGCCCACTGGGGAGTCCTAGCGGGCATAGCGTATTTCTCCATGG
TAGGGAACTGGGCGAAGGTCCTGGTAGTGCTGTTGCTGTTTGCCGGCGTCGATGCCGAGACCTACACCACTGGGGGGAGTGC
TGCCAGGATCACGACCGGACTCGTCAGTCTTTTCAGTCCGGGCGCCAAGCAGAATATCCAGCTGATGAACACCAACGGCAGT
TGGCACATCAATCGCACGGCCCTGAACTGTAATGCGAGCCTCGACACCGGCTGGGTGGCGGGGCTCTTCTACCACCACAAATT
CAACTCCTCGGGCTGCCCCGAGAGGATGGCCAGCTGTAGACCCCTTGCCGATTTTGACCAGGGCTGGAGCCCTATCACCCACGC
CAACGGAAGTGGCCCCGAACATCGCCCCTACTGCTGGCACTACCCCCCAAAGCCCTGTGGTATCGTGCCAGCACGGAACGTGTG
TGGCCCAGTGTATTGTTTCACTCCTAGCCCCGTGGTGGTGGGAACGACCGACGTGCTGGGCGTGCCTACCTACACCTGGGGTGG
TAATGATACGGACTTCTTCGTCCTTAACAACACCAGGCCACCGTTGGGCAATTGGTTTGGTTGCACCTGGATGAACTCGTCTG
GATTTACCAAAGTGTGCGGAGCGCCTCCTTGCGTCATCGGAGGGGTGGGCAACAACACCTTGCACTGCCCTACTGACTGTTTC
CGCAAGCATCCAGAAGCCACATACTCTCGGTGTGGCTCCGGTCCCTGGATCACGCCCAGGTGCCTGGTCCACTATCCTTATAGG
CTTTGGCACTACCCTTGCACCGTCAACTACACCCTGTTCAAGGTCAGGATGTACGTAGGAGGGGTCGAGCACAGGCTGGAGGT
TGCTTGCAACTGGACGCGGGGCGAGCGTTGTGATCTGGACGACAGGGACAGGTCCGAGCTCAGCCCGCTGCTGCTGTCCACCA
CGCAGTGGCAGGTCCTTCCGTGCTCCTTCACGACCTTGCCAGCCTTGACCACTGGCCTCATCCATCTCCACCGGAACATCGTGGA
CGTGCAATATTTGTACGGGGTGGGGTCAAGCATTGTGTCCTGGGCCATCAAGTGGGAATACGCCATTCTCTTATTTCTCCTGCT
TGCAGACGCGCGCATCTGCTCTTGCTTGTGGATGATGTTACTCATATCCCAAGCGGAGGCG

Sequência enigma 4

- enigma 4

```
MVPQTETKAGAGFKAGVKDYRLTYYTPDYMVKDTDILAAFRMTPQPGVPPEECGAAVAAESSTGTWTTVWTDGLTSLDRYK
GRCYDIEPVAGEDNQYIAYVAYPIDLFEEGSVTNLFTSIVGNVFGFKALRALRLEDLRIPPAYAKTFQGPPHGIQVERDKINKY
GRGLLGCTIKPKLGLSAKNYGRAVYECLRGGLDFTKDDENVNSQPFMRWRDRFTFVAEAIYKSQAETGEIKGHYLNVTAATSE
EMMKRAECAKDLGVPIIMHDYLTAGLTANTSLAHYCRDTGLLLHIHRAMHAVIDRQRNHGIHFRVLAKALRLSGGDHLHSG
TVVGKLEGEREVTLGFVDLMRDDYIEKDRSRGIYFTQDWCSLPGVMPVASGGIHVWHMPALVEIFGDDACLQFGGGTLGHPWG
NAPGAAANRVALEACTQARNAGVDLARKGGDVIRAACKWSPELAAACEVWKEIKFEFETIDKL
```

Sequência enigma 5

- enigma 5

```
GGAACCATTATGCACTCTTCAATAGTTTTGGCCACCGTGCTCTTTGTAGCGATTGCTTCAGCATCAAAAACGCGAGAGCTATGCA
TGAAATCGCTCGAGCATGCCAAGGTTGGCACCAGCAAGGAGGCGAAGCAGGACGGCATCGACCTCTACAAACATATGTTCGAG
CACTATCCAGCAATGAAGAAATACTTCAAGCATCGTGAAAATTATACACCGGCCGATGTCCAAAAGGATCCCTTCTTTATTAAACA
AGGTCAAAATATCTTGCTCGCCTGTCACGTTTTGTGCGCCACATACGACGATCGTGAGACATTCGACGCGTACGTTGGTGAGCT
GATGGCACGACACGAGCGGGACCATGTTAAAGTACCGAATGATGTTTGGAATCACTTCTGGGAACATTTCATCGAGTTTCTGG
GAAGTAAGACCACGTTGGACGAGCCAACCAAGCACGCATGGCAAGAGATCGGTAAAGAATTCTCACATGAAATCAGCCACCA
CGGTCGACATTCGGTTCGCGACCATTGCATGAACTCGTTGGAGTATATCGCGATCGGCGATAAGGAACATCAAAAGCAGAATGG
CATTGACCTTTACAAGCATATGTTCGAGCATTATCCACATATGAGAAAGGCATTCAAGGGACGCGAAAACTTCACGAAAGAAGA
CGTTCAAAAGGACGCATTCTTCGTTAACAAGGACACAAGATTCTGTTGGCCCTTCGTATGCTGTGACTCCTCATACGATGACGAG
CCAACATTCGACTATTTTGTTGATGCCCTAATGGATCGTCATATCAAAGATGATATTCATCTACCTCAGGAACAATGGCATGAGTTC
TGGAAATTGTTTGCCGAATATTTGAACGAAAAGAGTCACCAACATTTGACAGAAGCCGAGAAACATGCATGGAGTACAATAGGT
GAGGACTTCGCGCATGAGGCCGATAAGCATGCAAAGGCCGAAAAAGACCATCATGAAGGAGAGCACAAAGAGGAACACCAC
TGAACCAACCCGTCGTCGTTCAACTTAAGCCTTCAGCTTAAGCTCGAGCTAAAGCCTCAGCTTGAGCTCAATCTTATGTCCTCAGG
CCTAAACTTGAATTTTAAAAGCATTTTGTTGAAGCAGTGCTAGCCAATCTCTTATCTTATCGGTGCTATTATCAATTTACTCTATGC
CACCCCCCCCCCCCTCTCTCTGTTCTCTATTTGATATTCTGTTCTTTTAGTGCCAGATGTTAGTACCAGATGTTATTTTCTGCATAAT
TTTCTTCTCTTTACTTCGTTATTTTTTCGTTCTTCTATTTTTATGGCCAATTTTTGTGATGTCGAAGTCAATAAAACCATTTTT
```

INSTRUÇÕES PARA O PROFESSOR

1. É recomendável que o professor aplique esta atividade em turmas que já possuem conhecimentos sobre genética molecular, assim o público alvo mais adequado são alunos de Graduação ou Pós Graduação na área de Biológicas. Nesse sentido, é importante o conhecimento prévio acerca da estrutura e função do DNA e das proteínas. Além disso, é preciso que os alunos tenham conhecimentos das tecnologias de análise das macromoléculas incluindo métodos de sequenciamento, como o Sanger, e espectrometria de massa, metodologias que geram os dados das sequências que serão analisados a partir do BLAST.

2. A atividade será realizada individualmente por cada estudante que receberá do professor os problemas propostos. Cada estudante deverá executar a análise no BLAST e ao final entregar um relatório. A atividade pode ser realizada como aula prática em laboratório de informática ou como exercício de casa.

3. O professor deverá fornecer aos estudantes, por meio eletrônico, dois arquivos. O primeiro arquivo em **formato FASTA** contendo as cinco sequências a serem consultadas e as questões enigmas. O segundo arquivo pode ser enviado em PDF contendo as orientações para os estudantes e as questões adicionais (A a D) que devem ser respondidas.

O **formato FASTA** permite representar sequências de nucleotídeos e aminoácidos. Na primeira linha do arquivo há o símbolo maior (>) seguido pelo título da sequência e, nas linhas seguintes, a sequência representada pelo código de única letra.

PROCEDIMENTOS PARA OS ESTUDANTES

1. Acessar o site do NCBI (http://www.ncbi.nlm.nih.gov/). No menu lateral direito, acessar a ferramenta BLAST. Em "Basic BLAST" acessar o tipo de BLAST que considerar mais adequado para responder cada um dos cinco enigmas. Na caixa de texto copiar a sequência enigma em formato FASTA, ou ainda, escolher o arquivo a ser submetido. Escolher a base de dados a ser consultada. Dar o comando de execução do BLAST e resolver os 5 enigmas abaixo descritos.

- **Enigma 1.** Um pesquisador que trabalha com um determinado gene utilizou um par de ***primers*** que amplifica a região do genoma que corresponde a esse gene específico. Após obter o sequenciamento desta região do genoma, o pesquisador precisa confirmar se houve a amplificação e o sequenciamento da região alvo. Para isso, ele decidiu utilizar o BLAST como ferramenta de bioinformática. Qual versão do BLAST ele precisará utilizar e por quê? Qual o nome do gene com o qual o pesquisador trabalha?
- **Enigma 2.** Foi isolado o material biológico de um indivíduo do qual foi extraída uma dada proteína. Após o sequenciamento, a sequência de aminoácidos está disponível, mas não se sabe de que proteína se trata. Qual BLAST você usaria para identificar a proteína e por quê? Qual proteína foi extraída?
- **Enigma 3.** Você trabalha com sequências de cDNA e quer saber se uma delas codifica ou não alguma proteína no organismo do qual você isolou a amostra. A ferramenta BLAST pode ser usada para responder esta pergunta. Que tipo de BLAST você usaria e por quê? A sequência de cDNA do enigma 3 codifica alguma proteína? Se sim, qual?
- **Enigma 4.** Para investigar uma dada proteína depois de isolá-la e sequenciá-la, você pode usar o BLAST para chegar até a sequência correspondente no genoma? Que BLAST você usaria e por quê?
- **Enigma 5.** Você trabalha com amostras de pacientes com deficiências hematológicas e ainda não sabe exatamente o que causa a anomalia. O sequenciamento dos genes das vias hemolíticas destes pacientes é realizado e você decide usar o BLAST como uma primeira ferramenta para tentar identificar a similaridade dessas sequências com outras sequências nos bancos de dados. Que BLAST você usaria? Houve similaridade com algum gene? Qual? Isso faz sentido com a sua pergunta biológica?

2. Interprete os resultados gerados em cada questão enigma e responda também às seguintes perguntas:

- **Questão A.** Em "Graphic Summary" em que cor as sequências inseridas aparecem e o que isso significa na interpretação dos resultados?
- **Questão B.** Em "Descriptions", se houver, qual foi o resultado do *Best-Hit*, seu *E-value* e *Score* total para as sequências. E o que significa "Complete cds"?
- **Questão C.** Com base nos resultados obtidos no BLAST para as sequências, responda se foi possível identificar a que organismo pertence e também qual é a sua função.
- **Questão D.** Qual a importância da utilização do BLAST no estudo de espécies não modelo, ou que não possuem genoma descrito e de espécies que já têm informações genéticas bem conhecidas?

Primer é uma sequência curta de nucleotídeos que serve como iniciador para a replicação do DNA.

RESPOSTAS PARA OS ENIGMAS

Enigma 1. *Nucleotide BLAST*, pois trata-se de uma sequência de nucleotídeos que será alinhada contra sequências de banco de dados de nucleotídeos. O BLAST indicará em que região do genoma a sequência foi alinhada com maior similaridade, que deve ser exatamente o gene alvo do estudo em questão. A sequência enigma 1 corresponde ao gene "cinnamyl-alcohol dehydrogenase mRNA".

Enigma 2. *Protein BLAST*, pois trata-se de uma sequência de aminoácidos que será alinhada contra sequências de um banco de dados de aminoácidos. O BLAST mostrará em seu *Best Hit* a qual proteína e de qual organismo a sequência *Query* é mais similar, possibilitando descobrir qual a proteína putativa proveniente da amostra em estudo. A sequência enigma 2 corresponde à proteína "Cytochrome b-245, alpha polypeptide, isoform CRA".

Enigma 3. *BLASTx*, a sequência *Query* é composta por nucleotídeos (DNA codificado a partir de RNA) e o resultado deste BLAST será em aminoácidos caso se trate de uma região codificadora de proteína. Também será possível descobrir a proteína putativa e o organismo com maior similaridade para aquela região que contenha sequência depositada no GeneBank. A sequência do enigma 3 corresponde ao Gene do envelope do vírus da hepatite C.

Enigma 4. *T BLASTn*, a sequência *Query* é composta por aminoácidos e o resultado retornaria uma sequência de nucleotídeos, o que permitiria acesso a uma provável sequência genômica para aquela proteína.

Enigma 5. TBLASTx, trata-se de uma sequência de nucleotídeos que é traduzida sendo alinhada contra um banco de dados de nucleotídeos traduzidos. Sim, houve similaridade com um gene mutado da hemoglobina. Isso faz sentido de acordo com a pergunta biológica e pode ajudar a explicar o quadro clínico dos pacientes.

Questão A. O resultado gerado pela comparação de sequências (Figura 2) mostra primeiramente um gráfico sumário, contendo os *scores* associados à correspondência encontrada entre as sequências e quanto maiores os *scores*, mais alta a correspondência. Essa informação também é evidenciada nos *scores* com que as sequências inseridas aparecem. No resultado da sequência enigma 1, o gráfico exibe o alinhamento em rosa e vermelho (*score* de 80 a >=200), que indica alta similaridade. Para a sequência enigma 2, o gráfico exibe *score* de <40 a >=200. Para a sequência enigma 3, o gráfico exibe *score* de >=200. Para a sequência enigma 4, o gráfico exibe *score* >=200. Para a sequência enigma 5, o gráfico exibe *score* de <40 a >=200.

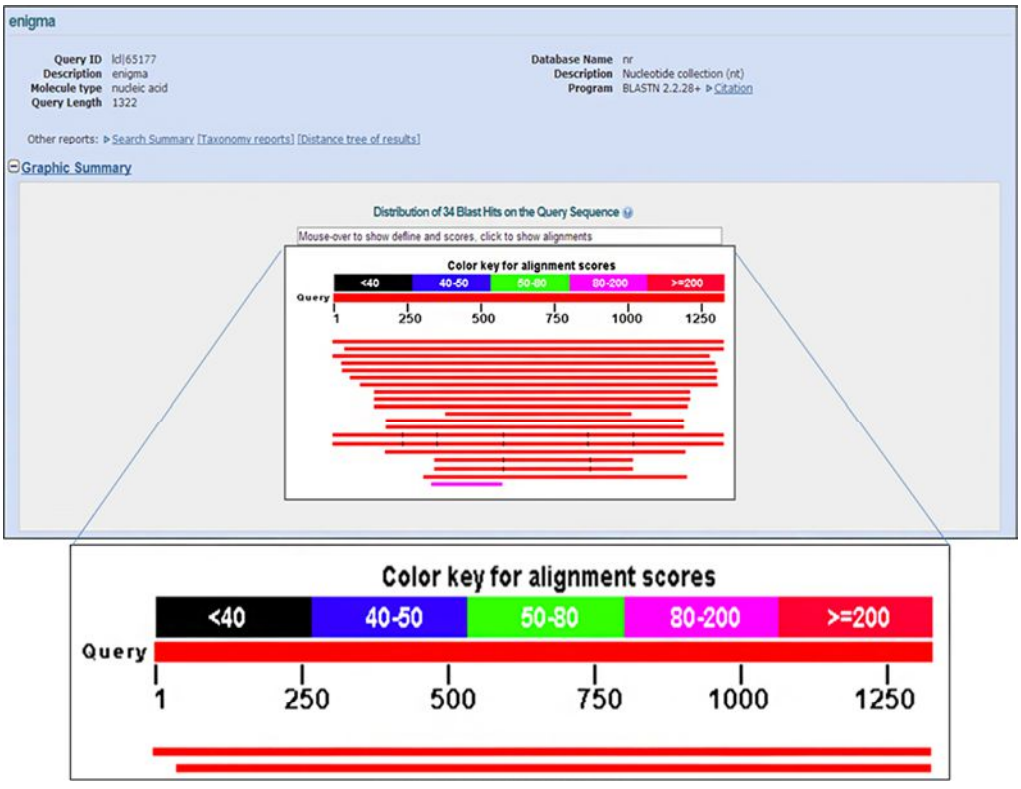
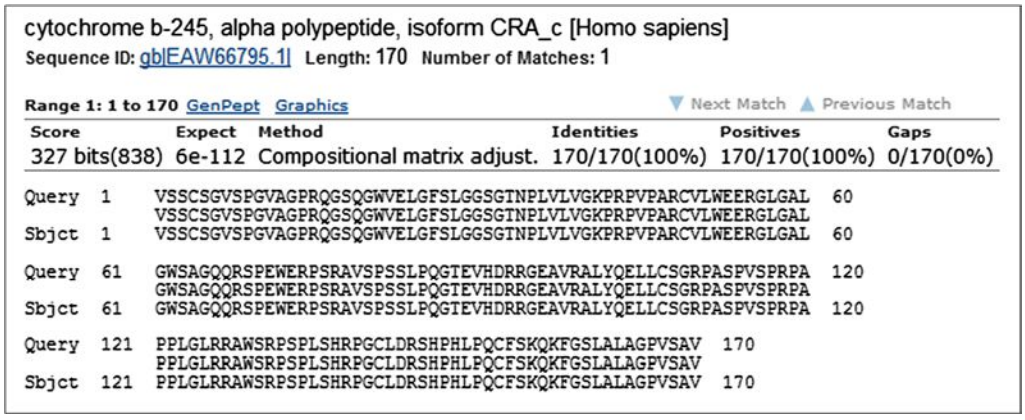


Figura 2.
Resultado gráfico do BLAST para a sequência enigma 1 evidenciando a chave de cores para *scores* de alinhamento.

Questão B. Em geral, o melhor resultado é o "*Best-Hit*", ou seja, o primeiro resultado da tabela de resultados do BLAST. Acessando esse resultado com um clique, aparecem as sequências comparadas e parâmetros onde *Query* é a sequência fornecida para a busca e *Subject* é a sequência conhecida. Foram encontrados *Best-Hits* para as cinco sequências fornecidas (Tabela 2). Quando se tem um "Complete cds", significa que todas as bases informadas no *Query* pareiam com as bases da sequência *Subject*.

Enigma	Best-hit	Espécie	Scores	E-value
1	cinnamyl-alcohol dehydrogenase mRNA, complete cds	*Arabidopsis thaliana*	2442	0,0
2	Cytochrome b-245, alpha polypeptide, isoform CRA_c	*Homo sapiens*	327	6^{e-112}
3	Hepatitis C virus isolate 1A14.32/MN-2/C5/22.05.1995 envelope gene	Virus hepatite c	1157	0,0
4	Large subunit of Rubisco	*Stigeoclonium helveticum*	949	0,0
5	Pseudoterranova decipiens (frameshift mutated) hemoglobin mRNA, complete cds	*Homo sapiens*	797	0,0

Tabela 2.
Best-hit para cada uma das cinco sequências enigmas.


cytochrome b-245, alpha polypeptide, isoform CRA_c [Homo sapiens]
Sequence ID: gb|EAW66795.1| Length: 170 Number of Matches: 1

Range 1: 1 to 170 GenPept Graphics ▼ Next Match ▲ Previous Match

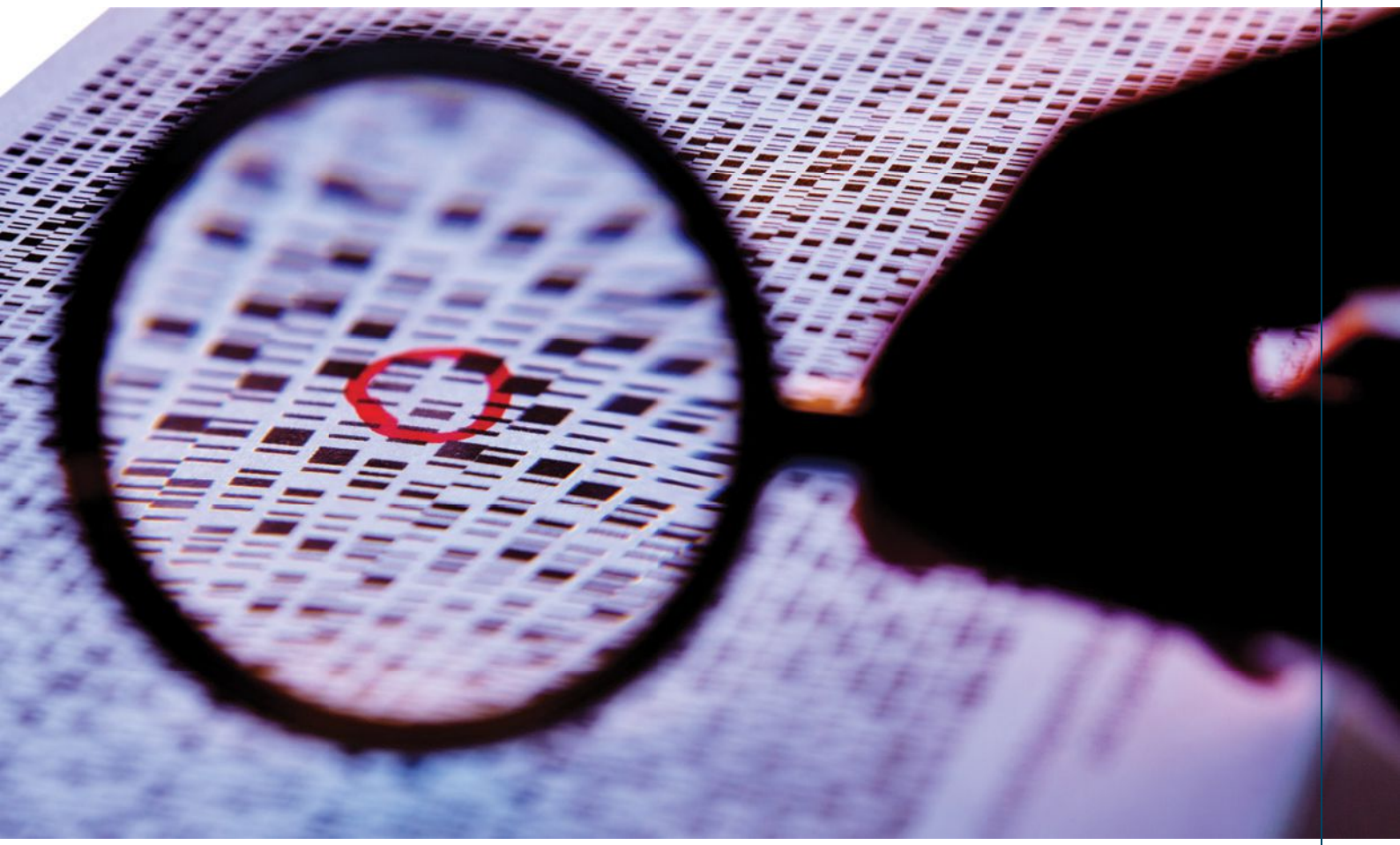
Score	Expect	Method	Identities	Positives	Gaps
327 bits(838)	6e-112	Compositional matrix adjust.	170/170(100%)	170/170(100%)	0/170(0%)

```
Query  1    VSSCSGVSPGVAGPRQGSQGWVELGFSLGGSGTNPLVLVGKPRPVPARCVLWEERGLGAL  60
            VSSCSGVSPGVAGPRQGSQGWVELGFSLGGSGTNPLVLVGKPRPVPARCVLWEERGLGAL
Sbjct  1    VSSCSGVSPGVAGPRQGSQGWVELGFSLGGSGTNPLVLVGKPRPVPARCVLWEERGLGAL  60

Query  61   GWSAGQQRSPEWERPSRAVSPSSLPQGTEVHDRRGEAVRALYQELLCSGRPASPVSPRPA  120
            GWSAGQQRSPEWERPSRAVSPSSLPQGTEVHDRRGEAVRALYQELLCSGRPASPVSPRPA
Sbjct  61   GWSAGQQRSPEWERPSRAVSPSSLPQGTEVHDRRGEAVRALYQELLCSGRPASPVSPRPA  120

Query  121  PPLGLRRAWSRPSPLSHRPGCLDRSHPHLPQCFSKQKFGSLALAGPVSAV  170
            PPLGLRRAWSRPSPLSHRPGCLDRSHPHLPQCFSKQKFGSLALAGPVSAV
Sbjct  121  PPLGLRRAWSRPSPLSHRPGCLDRSHPHLPQCFSKQKFGSLALAGPVSAV  170
```

Figura 3.
Resultado do alinhamento da sequência enigma 2 com a sequência *subject* do banco de dados.



Questão C. Sim, foi possível identificar o organismo e a função de cada uma das sequências informadas, assim como descrito abaixo:

- **Sequência enigma 1:** É o mRNA de uma enzima que catalisa a reação química, cinamil álcool-desidrogenase, de *Arabidopsis thaliana*.
- **Sequência enigma 2:** É um Alpha Citocromo B-245, isoforma CRA de *Homo sapiens*.
- **Sequência enigma 3:** É a sequência de uma proteína "e", de envelope celular de um isolado de Vírus de Hepatite C.
- **Sequência enigma 4:** É um genoma parcial de cloroplasto, gene da proteína Rubisco da alga *Stigeoclonium helveticum*.
- **Sequência enigma 5:** É um RNA ribossomal parcial de gene, região 16 mitocondrial de *Hynobius leechii*.

Questão D. Uma pesquisa no BLAST permite a um investigador comparar uma dada sequência fornecida em uma consulta com a maior biblioteca de base de Dados de Sequências do mundo e identificar sequências que se assemelham à investigada (*Query*) e que estejam acima de um nível mínimo de semelhança. Quando se trabalha com espécies das quais se tem pouco conhecimento de informações genéticas, o BLAST pode ser uma ferramenta muito importante, pois pode ser utilizado, entre outros, para descobrir se as sequências novas em estudo contêm algum gene homólogo já descrito para outra espécie ou ainda se uma destas novas sequências, se sabidamente genes, já foram descritos para outros organismos. Trabalhando com organismos para os quais já se tem informações genéticas disponíveis, por exemplo, organismos modelo, é possível utilizar novos dados adquiridos para esses organismos e compará-los aos dados já publicados no banco de dados do NCBI, podendo inclusive ser verificado o nível de similaridade entre as duas sequências, o que pode fornecer importante informações.