

Docente: Filipa Monteiro; Responsável pela UC: Maria Manuel Romeiras

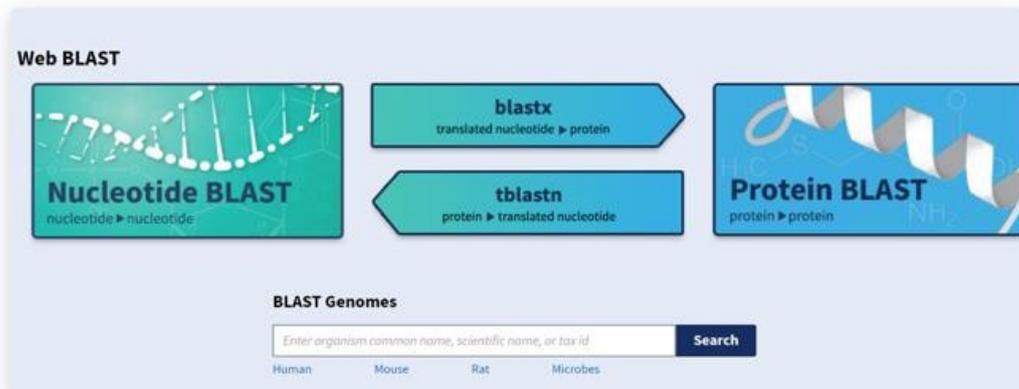
## Exercício: Identificação de espécies vegetais usando o BLAST

### Contextualização

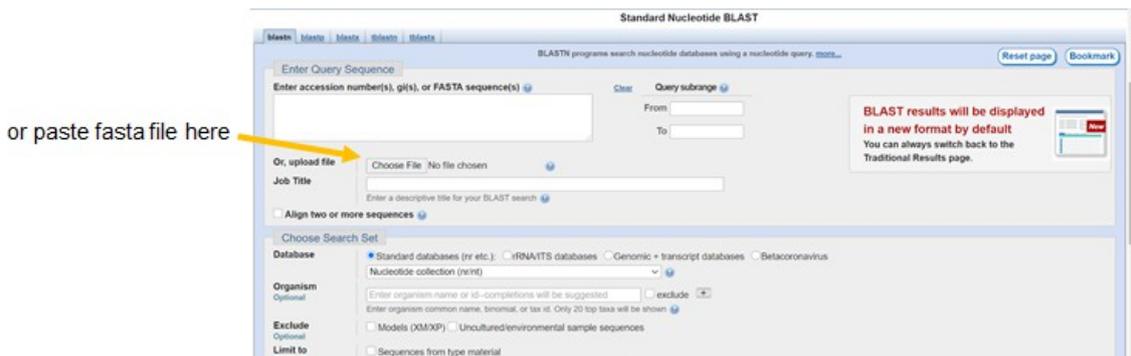
**BLAST ( Basic Local Alignment Search Tool ) é uma ferramenta** de busca online fornecida pelo NCBI (National Center for Biotechnology Information). Ela permite “encontrar regiões de similaridade entre sequências” (nucleotídeos ou proteína). O NCBI mantém um enorme banco de dados de sequências biológicas, que compara as sequências de consulta para encontrar as mais semelhantes com o disponível na base de dados. Usando o BLAST, pode-se inserir uma sequência de DNA de interesse e pesquisar bibliotecas moleculares inteiras para determinação de sequências idênticas ou semelhantes.

Tipo Blast	Sequência de consulta	Banco de dados	Alinhamento	Usar
Blastn	nucleotídeo	nucleotídeo	nucleotídeo	sequência identidade, útil para todas as categorias de taxa
blastx	nucleotídeo (traduzido para proteína)	proteína	proteína	Identificar proteínas codificadas, detecção de novos vírus
pblast	proteína	proteína	proteína	sequência EU IA e busca por similaridade
tblastx	nucleotídeo (traduzido para proteína)	nucleotídeo (traduzido para proteína)	proteína	Sequências de nucleotídeos de ID com codificação regiões semelhantes à consulta
tblastn	proteína	nucleotídeo (traduzido para proteína)	proteína	Sequências de banco de dados de ID que codificam proteínas semelhante para a consulta

1. Escolher o tipo de BLAST baseado no objetivo. BLAST é geralmente suficiente para confirmar um taxon.



2. Colar a sua sequência na caixa identificada caixa ou fazer upload de um ficheiro fsta com várias sequências.



### FASTA formatar

O formato FASTA é usado para representar sequências de nucleotídeos ou peptídeos. A primeira linha é um comentário, começando com ">" e descrevendo a sequência. Todas as linhas seguintes são a sequência, em texto simples.

### Exemplo sequência de DNA no formato FASTA :

>gi|23423|ref|NM\_23542.0| Proteína de Homo sapiens

ATGAATCGATACGATAGCTAGCTATCGATGCAGATCAGAGAGGGGCTTAGCTAGCTAAGCTAG

### Exemplo proteína sequência em FASTA formatar:

>MCHU - Calmodulina - Humano, coelho, bovino, rato, e frango

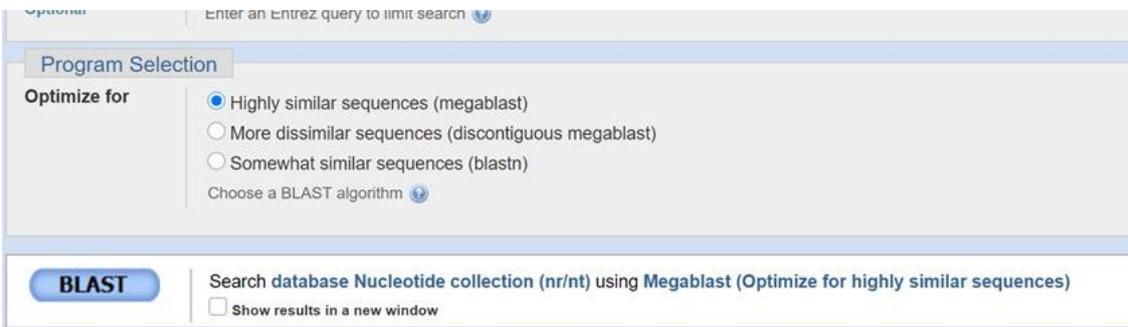
ADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRS LGQNPTAE LQDMINEVDADGNGTID  
FPEFLTMMARKMKD TDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGKLTDEEVD EMIREA  
DIDGDGQVNYEEFVQMMTAK\*

3. Mais abaixo na página escolher o programa para otimizar a busca. Navegar para a Seção “Seleção de Programas”. Para o blastn, pode escolher entre:

1. Megablast - lata ser usado para encontrar o melhor correspondência sequêcia.
2. Descontíguo megablast - usado para encontrar mais diferente sequências.
3. Blastn - usado para encontrar relacionado sequências de outro organismos.

Megablast é geralmente usado para o blast.

#### 4. Clique BLAST.

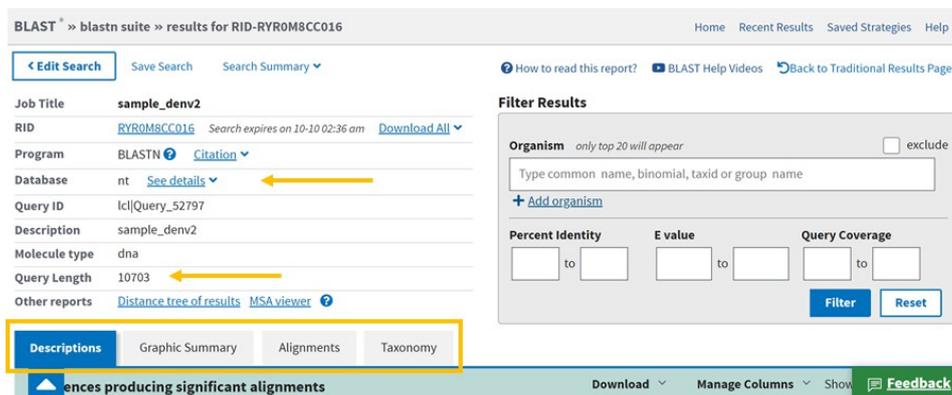


#### 5. Interpretação dos resultados do BLAST Resultado.

Os resultados do BLAST mostram todos os taxa disponíveis na base de dados com sequência semelhante sequência de consulta. As métricas e os gráficos podem ajudar a determinar a qualidade do resultado obtido

O resultado da página vai mostrar o seguinte resumo de um BLAST:

1. O banco de dados usado para a busca.
2. O comprimento da nossa sequência.
3. Resultados (Descrições, Gráfico Resumo, Alinhamentos, Taxonomia)



5.1. Navegar para o “ **Descrições** ” aba que tem métricas que pode ajuda você determinar a qualidade do golpe.

**Max score:** a maior pontuação de hits calculada a partir de correspondências e incompatibilidades de alinhamento. Quanto maior a pontuação, melhor o alinhamento.

**Total score:** a soma do alinhamento de todos os segmentos da sequência. Quanto maior a pontuação, melhor o alinhamento.

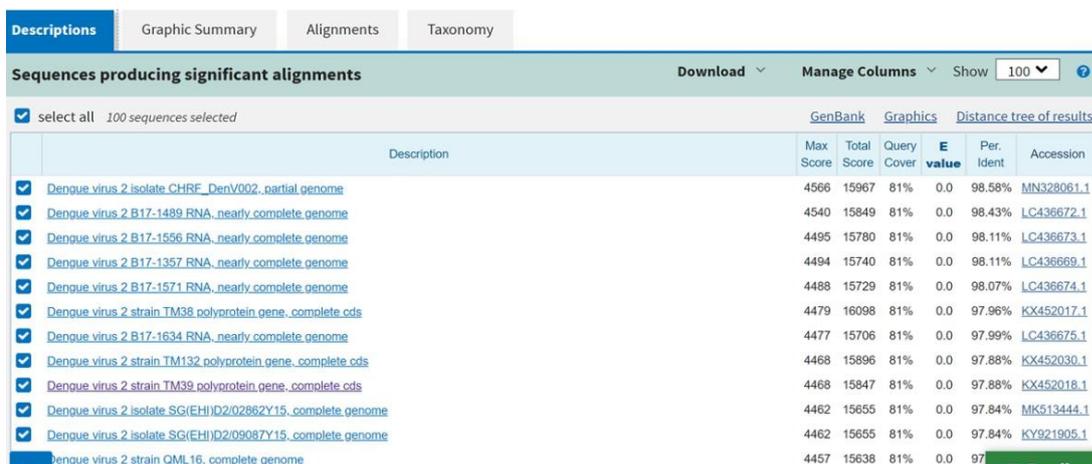
**Query coverage:** % do comprimento contíguo que alinha com o NCBI. Uma pequena porcentagem de cobertura de consulta significa que apenas uma pequena parte da sequência está alinhada. Se houver um alinhamento com 100% de identidade e 5% de cobertura de consulta, a sequência provavelmente não pertence a esse taxon.

**Valor E :** o número de sucessos esperado para ser visto por acaso. Quanto mais perto para 0, o melhorar. O sucessos são automaticamente classificados por E valor (melhor para pior). Essa métrica é extremamente útil para identificar acertos reais.

- E valor  $1e^{-50}$**  pequeno E valor: baixo número de acessos, mas de alta qualidade.
- Valor E 0.01:** BLAST com valor E menor que 0,01 ainda podem ser considerados bons acertos para correspondências de homologia.
- Valor E 10.** Valor E grande: muitos acertos, alguns de baixa qualidade. Um valor E menor que dez incluirá acertos que não podem ser considerados tão significativos quanto um valor E baixo.

**Percent identity:** a porcentagem de bases que são idênticas ao genoma de referência

**Accession [number]:** um identificador exclusivo atribuído para registros na base de dados do NCBI.



Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 isolate CHR_F_DenV002...partial genome</a>	4566	15967	81%	0.0	98.58%	<a href="#">MN328061.1</a>
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 B17-1489 RNA...nearly complete genome</a>	4540	15849	81%	0.0	98.43%	<a href="#">LC436672.1</a>
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 B17-1556 RNA...nearly complete genome</a>	4495	15780	81%	0.0	98.11%	<a href="#">LC436673.1</a>
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 B17-1357 RNA...nearly complete genome</a>	4494	15740	81%	0.0	98.11%	<a href="#">LC436669.1</a>
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 B17-1571 RNA...nearly complete genome</a>	4488	15729	81%	0.0	98.07%	<a href="#">LC436674.1</a>
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 strain TM38 polyprotein gene...complete cds</a>	4479	16098	81%	0.0	97.96%	<a href="#">KX452017.1</a>
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 B17-1634 RNA...nearly complete genome</a>	4477	15706	81%	0.0	97.99%	<a href="#">LC436675.1</a>
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 strain TM132 polyprotein gene...complete cds</a>	4468	15896	81%	0.0	97.88%	<a href="#">KX452030.1</a>
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 strain TM39 polyprotein gene...complete cds</a>	4468	15847	81%	0.0	97.88%	<a href="#">KX452018.1</a>
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 isolate SG(EHI)D2/02862Y15...complete genome</a>	4462	15655	81%	0.0	97.84%	<a href="#">MK513444.1</a>
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 isolate SG(EHI)D2/09087Y15...complete genome</a>	4462	15655	81%	0.0	97.84%	<a href="#">KY921905.1</a>
<input checked="" type="checkbox"/> <a href="#">Denque virus 2 strain QML16...complete genome</a>	4457	15638	81%	0.0	97.84%	<a href="#">KY921905.1</a>

5.2. Clique sobre a opção “Alinhamentos”.

- a) Observação da sequência submetida à base de dados em azul entre o principal.
- b) A localização e o comprimento dos alinhamentos de sequência são representados abaixo. Cada linha representa um táxon.
- c) A cor do alinhamento representa a qualidade do alinhamento, baseado na pontuação do alinhamento. Vermelho representa os alinhamentos com a pontuação mais alta (melhor alinhamento), enquanto preto representa a pior pontuação e não é necessariamente confiável.
- d) As linhas horizontais cinzentas representam lacunas no alinhamento.

100 sequences selected

Alignment view: Pairwise  CDS feature [Restore defaults](#)

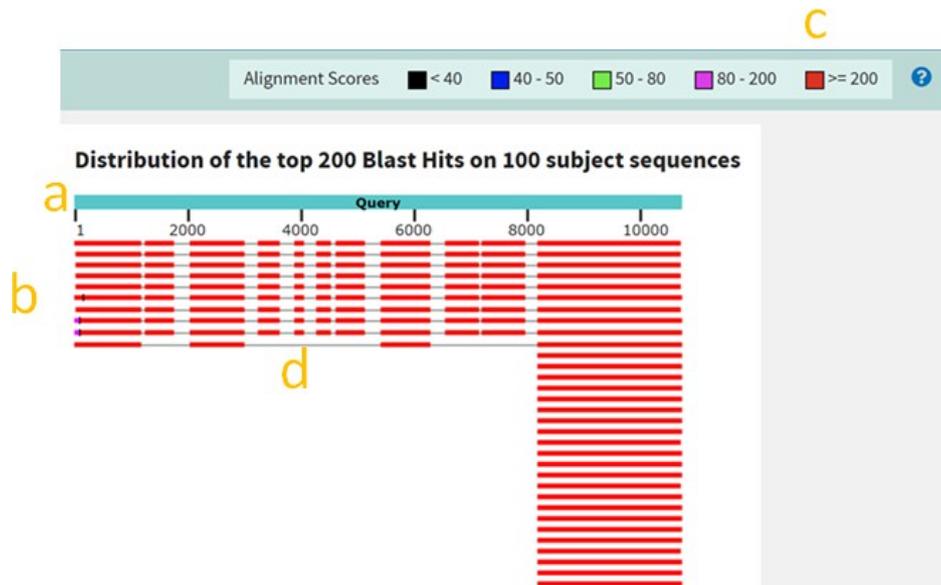
[Download](#) [GenBank](#) [Graphics](#) Sort by: E value

**Dengue virus 2 isolate CHRF\_DenV002, partial genome**  
Sequence ID: [MN328061.1](#) Length: 10699 Number of Matches: 11

Range 1: 8153 to 10696 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
4566 bits(2472)	0.0	2508/2544(99%)	0/2544(0%)	Plus/Plus
Query 8149	CTACAAAKGAAATACGGAGGAGCTTTGGTGAGGAATCCACTCTCACGAAATCCACACAC	8208		
Sbjct 8153	CTACAAAGGAAATACGGAGGAGCTTTGGTGAGGAATCCACTCTCACGAAATCCACACAC	8212		
Query 8209	GAGATGTACTGGGTATCCAATGCTTCCGGGAACATAGTGTATCAGTGAACATGATTTCA	8268		
Sbjct 8213	GAGATGTACTGGGTATCCAATGCTTCCGGGAACATAGTGTATCAGTGAACATGATTTCA	8272		
Query 8269	AGAATGTTGATTAACAGATTACAATGAGACACAAGAAGGCCACATACGAGCCGGATGTT	8328		
Sbjct 8273	AGAATGTTGATTAACAGATTACAATGAGACACAAGAAGGCCACATACGAGCCGGATGTT	8332		
Query 8329	GATCTCGGAAGTGGAACCCGCAACATCGGAATTGAAAGTGAAGTACCAAATCTAGACATA	8388		
Sbjct 8333	GATCTCGGAAGTGGAACCCGCAACATCGGAATTGAAAGTGAAGTACCAAATCTAGACATA	8392		

Clique no alinhamento para visualizar o alinhamento da sequência.



5.3. Pode ser feito o **download do alinhamento ou das sequências** ao clicar download e escolhendo o arquivo de interesse no menu.

GenBank FASIA [Link To This View](#) [Feedback](#)

500 K 1 M 1,500 K 2 M 2,500 K 3 M 3,500 K 4 M 4,500 K 5,087,117

CP026491.1 Find:  Tools Tracks Download

2,331 K 2,331,500 2,332 K 2,332,500 2,333 K 2,333,500 2,334 K

Sequence

Genes

C9K24\_J2145

rrf

rRNA-SS ribosomal R...

(U) BLAST Results for: Nucleotide Sequence Query\_45959

(U) Cleaned Alignments - BLAST Results for: Nucleotide Seq... Query\_45959

2,331 K 2,331,500 2,332 K 2,332,500 2,333 K 2,333,500 2,334 K

CP026491.1: 2.3M..2.3M (3,361 nt) Tracks shown: 4/9

**6. Registe as seguintes métricas:**

- accession number of the best hit (usually the top hit).
- E value
- query coverage

**7. O taxon foi identificado! Usar esses resultados para responder o questões abaixo.**

---

**Questões**

**1.** Na secção Descrições, verificar o principal resultado, qual deve ser o resultado com a pontuação mais alta. Anote as informações sobre a melhor combinação:

Descrição

E value

Identity

Query cover

**2.** Na secção Alinhamentos, observe o alinhamento entre a sua sequência e a sequência de referência.

Fazer você ver alguma incompatibilidade?

**3.** É possível identificar a sequência ao nível da espécie? Por favor elaborar.

**4.** Identificar o estado de conservação da espécie identificada.