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

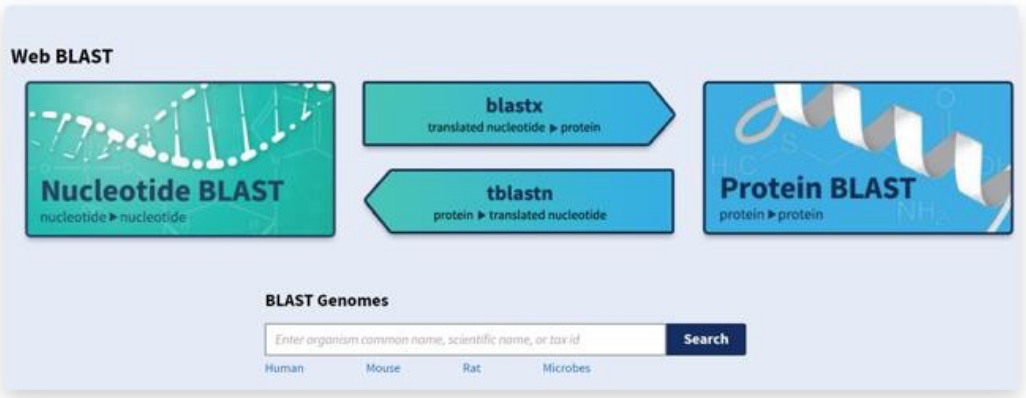
# Exercício: Identificação de espécies vegetais usando o BLAST

## Contextualização

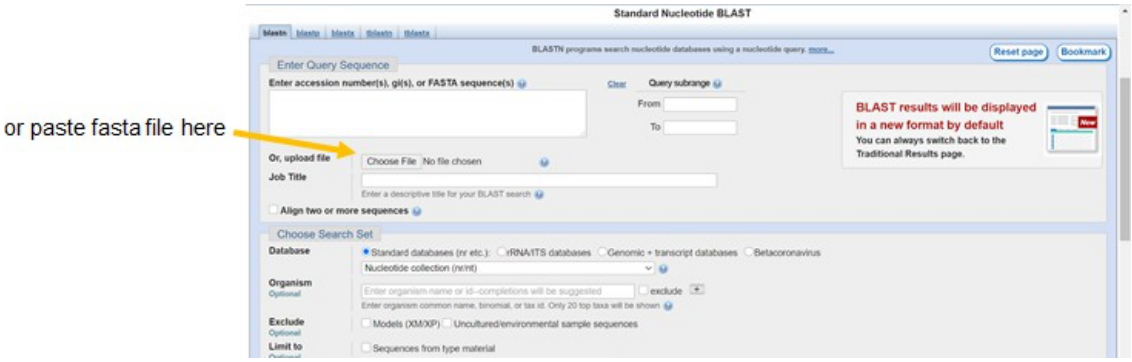
**BLAST** ( **B** asic **Local** Alignment **Search Tool ) é uma ferramenta** de busca online fornecida pelo NCBI (National Center for Biotechnology Information). Ela permite "encontrar regiões de similaridade entre sequências" (nucleotídeos ou proteína). O NCBI mantém um enorme banco de dados de sequências 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

ATGAATCGATACGATAGCTAGCTATCGATGCAGATCAGAGAGGGGCTTTAGCTAGCTAAGCTAG

**Exemplo proteína sequência em FASTA formatar:**

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

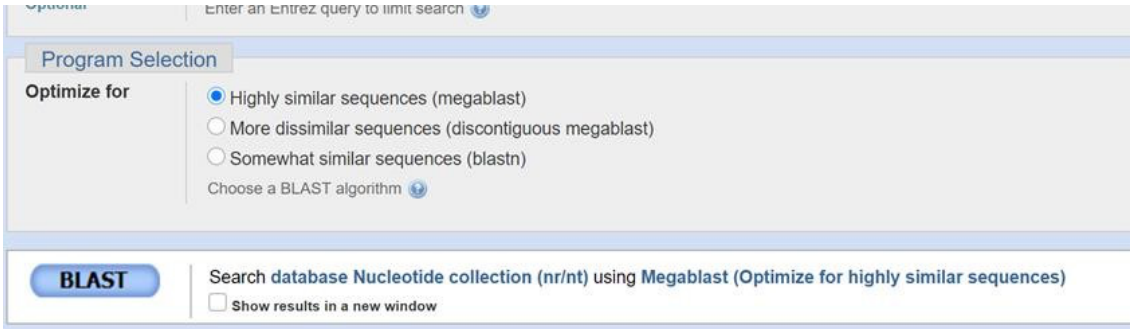
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

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 sequência.
    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.**

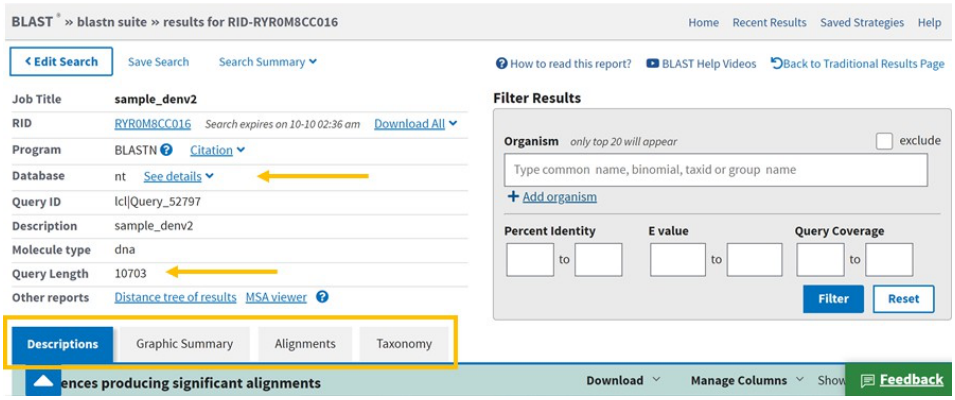



5. **Interpretação dos resultados do BLAST Resultado.**

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 seguindo 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.

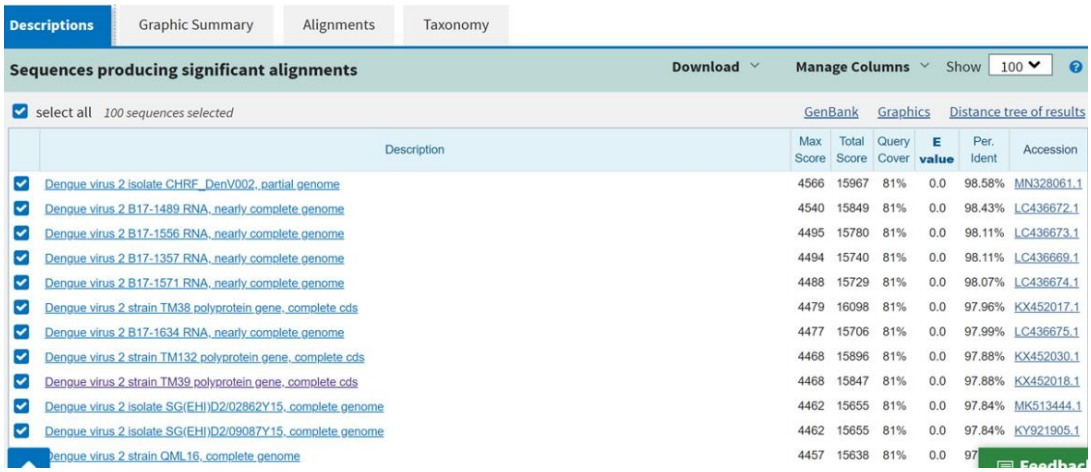
**Query coverage**: % do comprimento contíguo que alinha com o NCBI. Uma pequena porcentagem de cobertura de consulta significa que apenas uma pequena parte da sequência está alinhada. Se houver um alinhamento com 100% de identidade e 5% de cobertura de consulta, a sequência provavelmente não pertence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

a) **E valor 1e$^{-50}$** pequeno E valor: baixo número de acessos, mas de alta qualidade.
b) **Valor E 0.01**: BLAST com valor E menor que 0,01 ainda podem ser considerados bons acertos para correspondências de homologia.
c) **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.



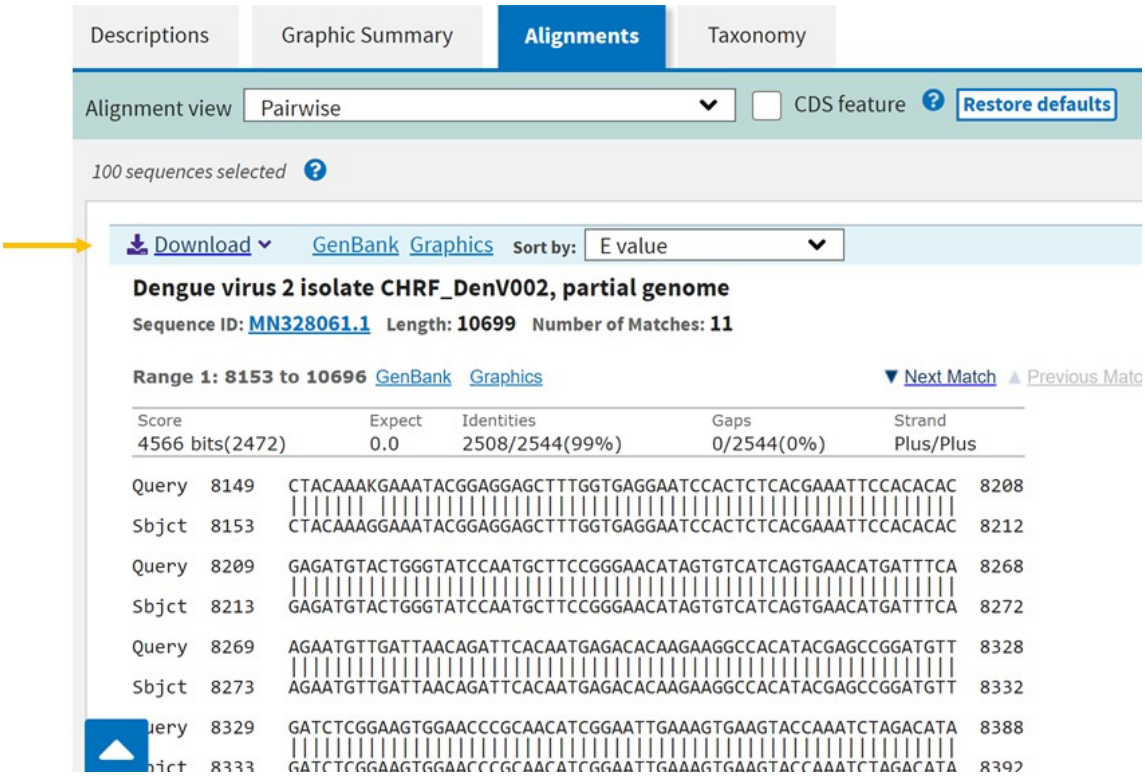
**5.2. Clique sobre a opção "Alinhamentos".**

a) Observação da sequência submetida à base de dados em azul entre o principal.

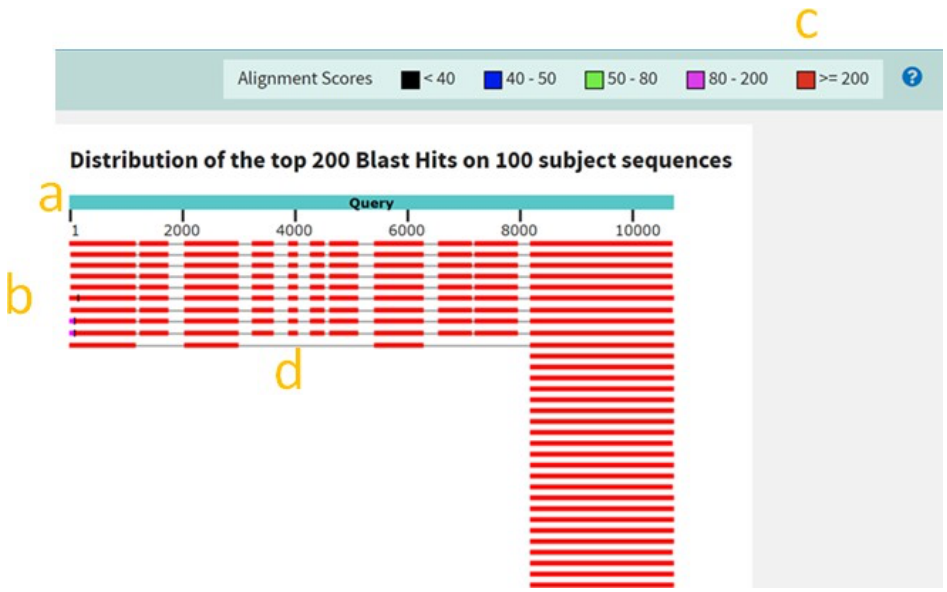
b) A localização e o comprimento dos alinhamentos de sequência são representados abaixo. Cada linha representa um táxon.

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.

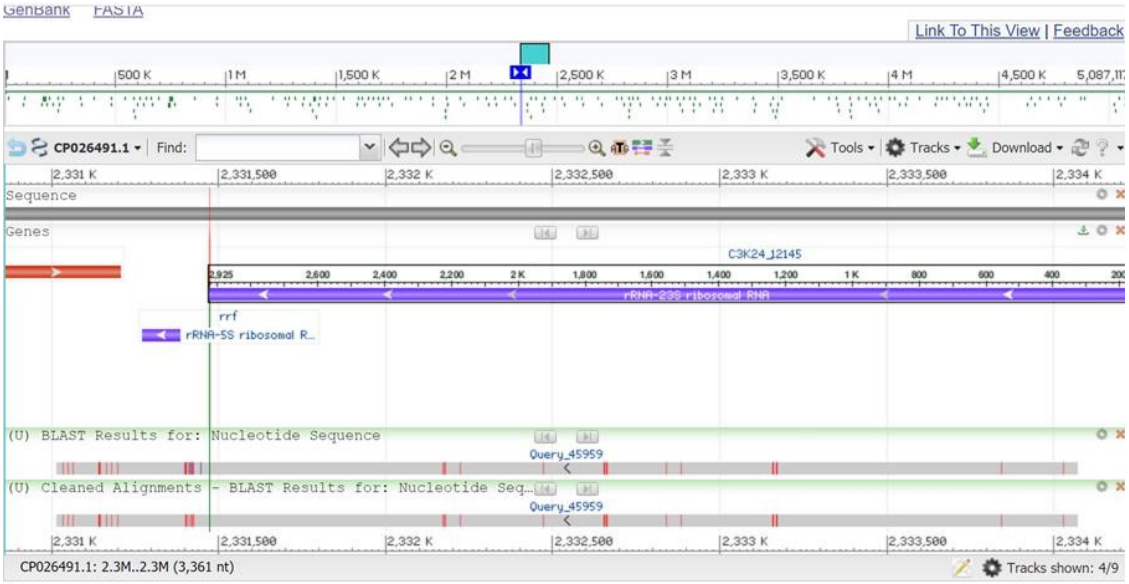
d) As linhas horizontais cinzentas representam lacunas no alinhamento.



**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.



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 sequnência 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.