INSTITUTO FEDERAL Brasília	**Instituto Federal de Brasília - Campus Taguatinga** **Superior em Computação** **Aluna: Danyelle da Silva Oliveira Angelo.**

Ferramentas em Bioinformática 2

Blast

1/2020

Você acabou de receber uma sequência de RNA/DNA coletada em um hospital de sua cidade:

```
>amostra
ATGAATAACCAACGGAAAAAGGCGAAAAACACGCCTTTCAATATGCTGAAACG
CGAGAGAAACCGCGTGT
CGACTGTGCAACAGCTGACAAAGAGATTCTCACTTGGAATGCTGCAGGGACG
AGGACCATTAAAACTGTT
CATGGCCCTGGTGGCGTTCCTTCGTTTCCTAACAATCGCACCAACAGCAGAC
GGATATTGAAGAGATGGA
GGA
```

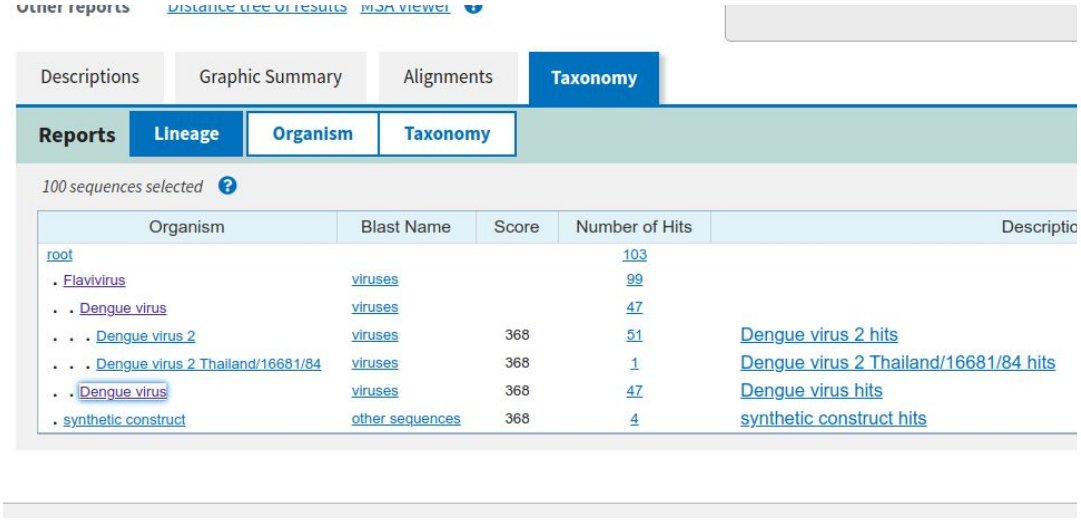
Não se sabe, a priori, de qual organismo e qual função desempenha esta sequência. de forma a analisar melhor que tipo de sequência foi coletada neste hospital, siga o roteiro, respondendo as perguntas que surgirem ao longo dos passos:

1) Acesse o site https://blast.ncbi.nlm.nih.gov/Blast.cgi e entre em **nucleotide blast.**
2) Como *query* adicione a sequência fornecida e utilize como banco de dados o banco *"Nucleotide collection (nr/nt)"*. **Antes de clicar no botão Blast responda:**
    a) Descreva o que é o banco *Nucleotide collection (nr/nt)?* Quantas sequências estão armazenadas nele?
    O nr/nt é uma coleção de nucleotídeos (como o próprio nome diz), estes nucleotídeos são sequências do GenBank, EMBL, DDBJ, PDB e

do RefSeq (ele mescla sequências que podem vir repetidas desses bancos), ao total são 60411854 sequências armazenadas.

3) Pressione o botão **Blast** e aguarde a geração dos resultados.
4) A partir do resultado gerado, **responda**:
    a) Esta sequência veio provavelmente de qual organismo? Justifique interpretando os resultados do blast.

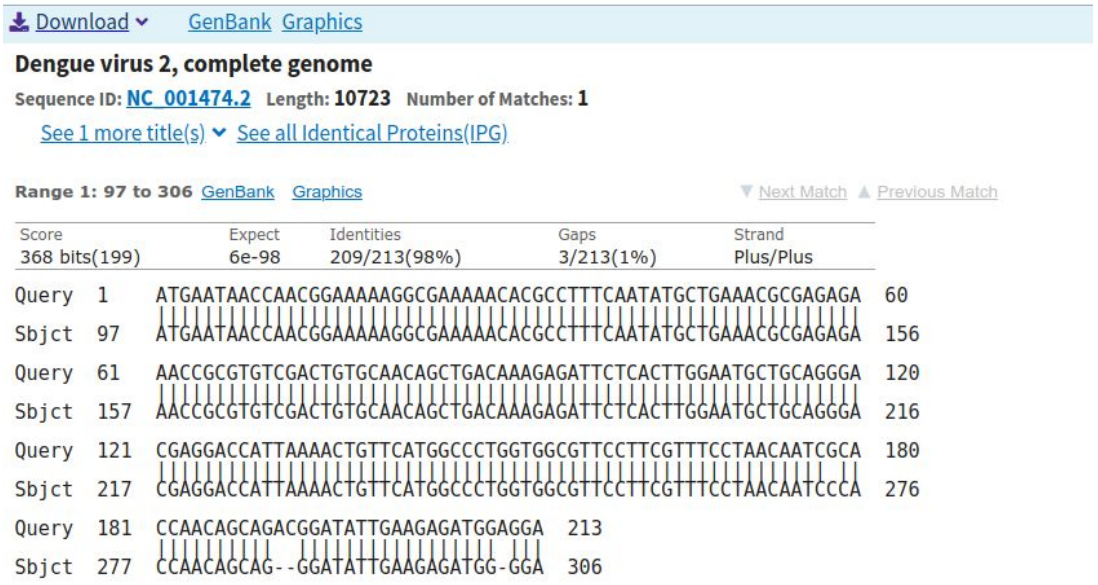
    A sequência é do vírus da dengue é possível deduzir isso através da descrição dos hits (na aba description) e confirmar na aba "Taxanomy", veja a tabela abaixo.



Organism	Blast Name	Score	Number of Hits	Descriptio
root			103	
. Flavivirus	viruses		99	
. . Dengue virus	viruses		47	
. . . Dengue virus 2	viruses	368	51	Dengue virus 2 hits
. . . Dengue virus 2 Thailand/16681/84	viruses	368	1	Dengue virus 2 Thailand/16681/84 hits
. . Dengue virus	viruses	368	47	Dengue virus hits
. synthetic construct	other sequences	368	4	synthetic construct hits

    b) Para que serve a medida *query cover, E-Value* e *Perc. Ident.*?
        - Query cover: a sequência retornada corresponde têm um alinhamento (cobertura) de x% com a sequência pesquisada.
        - E-Value(E): descreve o número de acertos que podemos "esperar" do alinhamento, onde quanto mais baixo for o seu valor, ou mais próximo de zero, melhor será a correspondência. $E = mn\ 2^{-S}$ (onde m e n é o tamanho das sequências e S é o score normalizado).
        - Perc. Ident: quantidade de matches da sequência retornada com a sequência pesquisada.

    c) Selecione algum hit (linhas da tabela "*Sequences producing significant alignments*") e vá para a aba alignments e tire um printscreen do alinhamento realizado (cole o print no seu documento de respostas :).



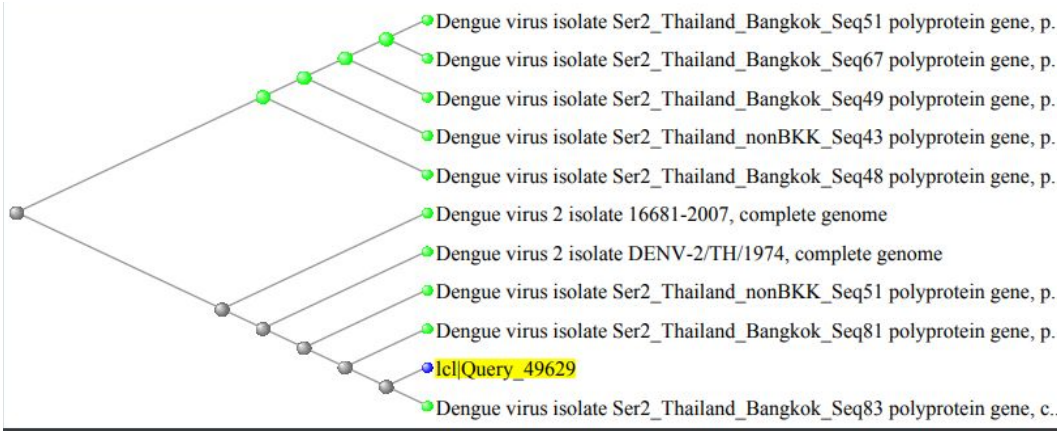
d) A partir do alinhamento anterior você pode concluir que as sequências são similares? Justifique a resposta usando os resultados de E-Value, identidades e gaps

A identidade nos mostra a quantidade de nucleotídeos correspondentes entre duas sequências. Dos 213 nucleotídeos da sequência buscada tivemos uma correspondência com os 209 nucleotídeos da sequência analisada (98%), apenas 3 gaps foram dados (a fim de pararmos um nucleotídeo com outro), e apenas um nucleotídeo não obteve correspondência com o de outra sequência[1].

Além disso o padrão de confiança é extremamente próximo do zero, e através de tudo isso podemos inferir que a similaridade entre as duas sequências é realmente muito alta.

e) Volte para a aba Descriptions e selecione os 5 melhores e 5 piores hits de acordo com o percentual de identidade. Após isso clique em *Distance tree of results* e tire um print da árvore gerada (cole o print no seu documento de respostas :).

[1] ver query 121 e subject 217, ante penúltimos nucleotídeos das duas sequências



f) Para que serve a árvore gerada no exercício anterior? como podemos interpretá-la?

Ela mostra a distância entre as sequências escolhidas, o cálculo dessa distância é feito com base na similaridade/alinhamento das mesmas.

Podemos interpretar ela da seguinte forma:

- O Dengue virus isolate Ser2_Thailand_Bangkok_Seq51 polyprotein gene, partial cds e o Dengue virus isolate Ser2_Thailand_Bangkok_Seq67 polyprotein gene, partial cds são muito próximos entre si.
- Os dois organismos citados no ponto anterior são semelhantes aos organismos seguintes (nível de proximidade vai caindo): Dengue virus isolate Ser2_Thailand_Bangkok_Seq49 polyprotein gene, partial cds , Dengue virus isolate Ser2_Thailand_nonBKK_Seq43 polyprotein gene, partial cds, Dengue virus isolate Ser2_Thailand_Bangkok_Seq48 polyprotein gene, partial cds.
- Os organismos citados nos pontos anteriores possuem similaridades (embora nem tantas) com os organismos: Dengue virus 2 isolate 16681-2007, complete genome, Dengue virus 2 isolate DENV-2/TH/1974, complete genome, Dengue virus isolate Ser2_Thailand_nonBKK_Seq51 polyprotein gene, partial cds, Dengue virus isolate Ser2_Thailand_Bangkok_Seq81 polyprotein gene, partial cds e Dengue virus isolate Ser2_Thailand_Bangkok_Seq83 polyprotein gene, complete cds sendo os apresentados neste ponto (especialmente esse

último) mais próximos da sequência que buscamos no início do exercício *(ponto azul)*.