CEN0485 – ACH5555

Bioinformática

maio 12 de 2022

Prova 1 – Parte 2

Prof. Dr. Diego M. Riaño Pachón

Aluno: Ana Julia Bernardi de Souza

No USP: 10755620

Vocês receberam pelo e-mail da USP uma sequência de ácidos nucleicos. Usando essa sequência e ferramentas de bioinformática responder o seguinte:

1. Identifique o organismo mais provável do qual essa sequência foi obtida. Descreva o procedimento, brevemente, que usou.

O procedimento foi realizado usando o BLAST, uma vez que 1.fa é uma única sequência de nucleotídeos e pequena.

Para descobrir qual organismo a sequência se refere, entrei no site e no campo que se referia a entrada e saída de uma sequência de nucleotídeos. Dessa forma, procurei com qual organismo a sequência utilizada tinha maior porcentagem de identidade (também foram observados (outros parâmetros também foram considerados).

Outra forma é utilizar o BLAST na linha de comando Unix para saber a qual organismo essa sequência se refere. Utilizando o database de nucleotídeos **conhecidos** (o BLAST exige que cada banco de dados de assunto seja pré-formatado para uso; esta é uma forma de agilizar certos tipos de buscas) o primeiro comando é “makeblastdb -in db\_no formato.fna -dbtype nucl -out db\_no\_formato.fna.db”.

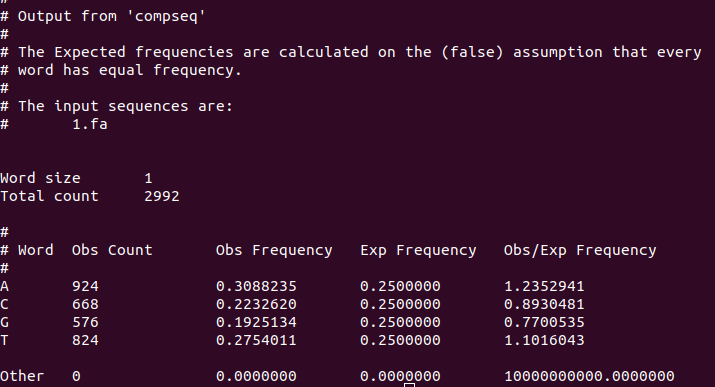
Em seguida,usamos o comando com um arquivo que tenha a sua consulta (query) para rodar o BLAST “blastn - db\_no\_formato.fna.db. query 1.fa”. Após isso você pode simplesmente olhar o output ou salvá-lo.

Dessa forma, o organismo mais provável é *Arabidopsis thaliana.*

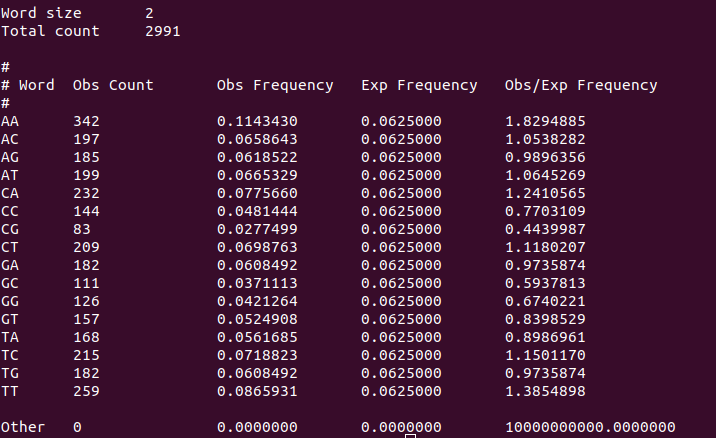
1. Essa sequência pode dar origem a uma proteína, i.e., pode ser traduzida? Por quê? Justifique sua resposta.

Sim para ambos, uma vez que essa sequência de nucleotídeos corresponde a uma fração do gene 1 do organismo, podendo ser traduzido em uma proteína. Além disso, foi utilizado o comando ‘tcode’ para identificar regiões codificadoras de proteínas usando a estatística Fickett Testcode (e o resultado mostrou que haviam regiões codificadoras). Por fim, foi utilizado a própria sequência de aminoácidos no BLAST para ver se a sequência tinha similaridade com alguma proteína.

1. Calcule estadísticas básica da sequência do ponto 1:
   1. Composição de monómeros



* 1. Composição de dímeros



* 1. Tamanho da sequência

2992

1. Qual é a função biológica dessa sequência?

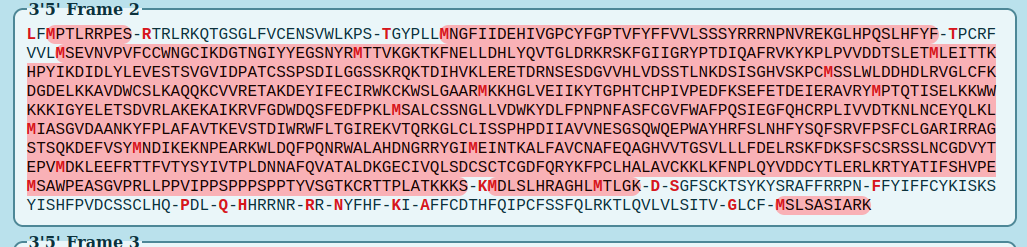
Ela corresponde a uma sequência que traduz uma transposase da família MuDR; em que sua função é a ligação do íon zinco; segundo o NCBI.

1. Caso seja possível traduzir a sequência numa proteína, descreva o procedimento que seguiu para estabelecer isto.

Utilizei o comando ‘transeq 1.fa -frame=6’ para traduzir a sequência em todos os seis frames (para frente e reverso). Gerou um novo arquivo com a sequência traduzida (de seis formas diferentes).

1. Faz uma busca de BLAST com a sequência original (se sua sequência pode ser traduzida use a sequência de proteínas com a fase de leitura mais comprida) e descarregue a sequência mais semelhante de uma outra espécie. Por exemplo, se no ponto 1 você identifica que sua sequência é de *Homo sapiens*, aqui você tem que procurar uma sequência semelhante dentro do resultado do BLAST de uma outra espécie diferente de *Homo sapiens*.
   1. Descreva brevemente o procedimento para achar essa segunda sequência.

Utilizando a fase de leitura mais comprida (a parte em rosa):



Foi copiado e colado no BLAST essa sequência e excluído da busca *Arabidopsis thaliana* e na busca o melhor resultado foi MULE transposase domain [Arabidopsis suecica].

* 1. Qual o código de acesso (Accession Number) no banco de dados do NCBI, para essa sequência? O E value, e a porcentagem de identidade calculado pelo BLAST?

Número de acesso: KAG7658385

E value: 0.0

Porcentagem de identidade: 97.33%



1. Faz um alinhamento global exato com o algoritmo de Needleman & Wunsch entre a sequência original e a sequência do ponto 6, usando os seguintes parâmetros: MATCH=5; MISMATCH=-2; GAPS=-6
   1. Descreva o procedimento que usou para obter o alinhamento?

Utilizando o comando needle e a matrix correta (BLOSUM62), com os parâmetros definidos utilizando as sequências de interesse.

* 1. Qual a pontuação do alinhamento?

# Length: 3557

# Identity: 1280/3557 (36.0%)

# Similarity: 1280/3557 (36.0%)

# Gaps: 1887/3557 (53.1%)

# Score: 1637.5