CEN0485 – ACH5555

Bioinformática

maio 12 de 2022

Prova 1 – Parte 2

Prof. Dr. Diego M. Riaño Pachón

Aluno: Rafaela Prado Graciano

No USP: 10318285

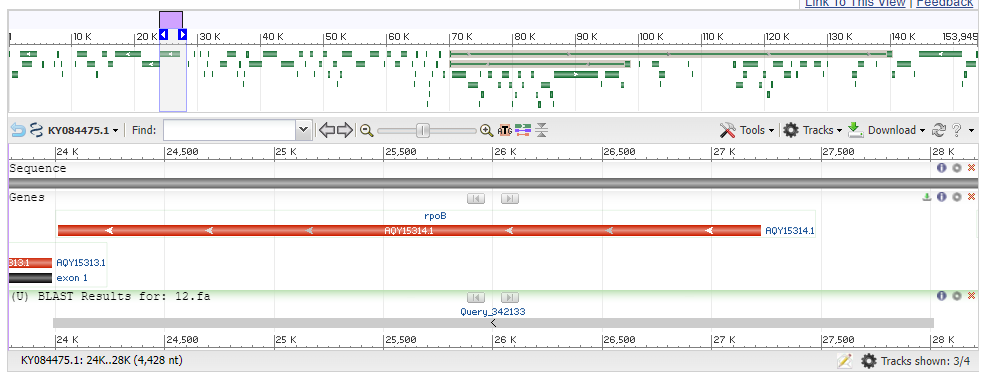
Vocês receberam pelo e-mail da USP uma sequência de ácidos nucleicos. Usando essa sequência e ferramentas de bioinformática responder o seguinte:

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

O organismo mais provável do qual a sequência foi obtida é da planta *Cannabis Sativa,* mais precisamente do cloroplasto desse organismo. Para chegar a essa conclusão, eu utilizei a ferramenta blast, a qual faz um alinhamento local com sua base de dados, no caso eu utilizei o blastn, que considera a base de dados de nucleotídeos. Então carreguei o arquivo que eu possuía, e usei as configurações padrões de buscas. Como resultado, o cloroplasto da *Cannabis Sativa*  foi o mais provável, com 100% de identidade.

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

Uma forma de tentar responder essa pergunta seria analisando o resultado da sequência no blast.



Como podemos observar, ao alinhar nossa sequência, representada como Query\_342133, ela coincide com o gene representado em vermelho, que codifica a uma subunidade beta da RNA polimerase do cloroplasto, representada abaixo, então podemos afirmar que nossa sequência pode ser traduzida.



Caso não tivéssemos encontrado nenhum resultado satisfatório no blast, poderíamos tentar responder essa perguntar utilizando o Emboss, com o comando ‘getorf’, para encontrar e emitir as sequências de quadros de leitura aberta (ORFs) em sequência de nucleotídeos, na qual ORF é definida como uma região de um tamanho mínimo (cujo padrão é 30 bases) especificados entre dois códons STOP (ou entre um START e STOP). As ORFs podem ser emitidas como a sequência de nucleotídeos ou como a tradução de proteínas. A saída é um arquivo de sequência contendo quadros de leitura abertos previstos maiores do que o tamanho mínimo.

Realizando esse comando com a sequência em questão resulta em um arquivo com 154 orfs, na qual a maior delas é representada abaixo:

>12.fa\_95 [787 - 4002]

KKMLGDGNGKMSTIPGFNQIQFEGFCRFIDQGLTEELYKFPKMEDTDQEIEFQLFVETYQLAEPLIKERDAVYESLTYSSELYVSAGLIWKTSRDMQEQTIFIGNIPLMNSLGTSIVNGIYRIVINQILQSPGIYYRSELDHNGISVYTGTIISDWGGRLELEIDRKARIWARVSRKQKISILVLSSAMGSNLREILENVCYPEIFLSFLNDKEKKKIKSKENAILEFYQQFACVGGDPVFSESLCKELQKKFFQQRCELGRIGRRNMNRRLNLDIPQNNTFLLPRDILAAADHLIGMKFGMGILDDMNHLKNKRIRSVADLLQDQFGLALVRLENMVRGTMSGAIRHKLIPTPQNLVTSTTLTTTFESFFGLHPLSQVLDRTNPLTQIVHGRKLSYLGPGGLTGRTASFRIRDIHPSHYGRICPIDTSEGINVGLIGSLSIHARIGPWGSLESPFYEISERSKKVRMLYLSPSRDEYYMVAAGNSLALNRGSQEEQVVPARYRQEFLTIEWEQVHLRNIFPFQYFSIGASLIPFIEHNDANRALMSSNMQRQAVPLSRSEKCIVGTGLESQVALDSGVPAIAEHEGKIVYTDTDKIILSGNGDTLSIPLVRYQRSNKNTCMHQKPQVARGKCIKKGQILADGAATVGGELSLGKNVLVAYMPWEGYNFEDAVLINERLVYGDIYTSFHIRKYEIQTHVTSHGPERITNEIPHLEAHLLRNLDKKGVVMLGSWVETGDILVGKLTPQMAKESSYAPEDRLLRAILGIQISTSKETCLKLPIGSRGRVIDVRWIQKKGGSSYNPETIRVYISQKREIKVGDKVAGRHGNKGIVSKILPRQDMPYLQDGRPVDMVFNPLGVPSRMNVGQIFECSLGLAGELLDRHYRIAPFDERYEQEASRKLVFSELYEASKQTANPWVFEPEYPGKSRIFDGRTGDPFEQPVIIGKPYILKLIHQVDDKIHGRCSGHYALVTQQPLRGRAKQGGQRVGEMEVWALEGFGVAHILQEMLTYKSDHIRARQEVLGTTIIGGPIPKPEDAPESFRLLVRELRSLALELNHFLVSEKNFQINRKDA

Então caso não soubéssemos nada sobre a sequência, levando em conta códons de start e stop, poderíamos dizer que essa sequência tem a possibilidade de codificar proteínas, mas que precisaria ser investigado.

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

contagem frequência

A 1321 0.3281172

C 639 0.1587183

G 830 0.2061600

T 1236 0.3070045

* 1. Composição de dímeros

Contagem Frequência

AA 464 0.1152795

AC 170 0.0422360

AG 238 0.0591304

AT 449 0.1115528

CA 200 0.0496894

CC 125 0.0310559

CG 116 0.0288199

CT 198 0.0491925

GA 328 0.0814907

GC 107 0.0265839

GG 229 0.0568944

GT 166 0.0412422

TA 328 0.0814907

TC 237 0.0588820

TG 247 0.0613665

TT 423 0.1050932

* 1. Tamanho da sequência

4026

1. Qual é a função biológica dessa sequência?

A função biológica da sequência é codificar o gene que dará origem a subunidade beta da RNA polimerase, a qual irá ajudar no processo de síntese de RNA, como a transcisão.

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

Para responder a pergunta 2 foi realizado o procedimento da tradução no Emboss, como descrito na resposta. Para a próxima questão será usado a sequência de proteínas com a fase de leitura mais comprida, presente na questão 2 é representada abaixo

>12.fa\_95 [787 - 4002]

KKMLGDGNGKMSTIPGFNQIQFEGFCRFIDQGLTEELYKFPKMEDTDQEIEFQLFVETYQLAEPLIKERDAVYESLTYSSELYVSAGLIWKTSRDMQEQTIFIGNIPLMNSLGTSIVNGIYRIVINQILQSPGIYYRSELDHNGISVYTGTIISDWGGRLELEIDRKARIWARVSRKQKISILVLSSAMGSNLREILENVCYPEIFLSFLNDKEKKKIKSKENAILEFYQQFACVGGDPVFSESLCKELQKKFFQQRCELGRIGRRNMNRRLNLDIPQNNTFLLPRDILAAADHLIGMKFGMGILDDMNHLKNKRIRSVADLLQDQFGLALVRLENMVRGTMSGAIRHKLIPTPQNLVTSTTLTTTFESFFGLHPLSQVLDRTNPLTQIVHGRKLSYLGPGGLTGRTASFRIRDIHPSHYGRICPIDTSEGINVGLIGSLSIHARIGPWGSLESPFYEISERSKKVRMLYLSPSRDEYYMVAAGNSLALNRGSQEEQVVPARYRQEFLTIEWEQVHLRNIFPFQYFSIGASLIPFIEHNDANRALMSSNMQRQAVPLSRSEKCIVGTGLESQVALDSGVPAIAEHEGKIVYTDTDKIILSGNGDTLSIPLVRYQRSNKNTCMHQKPQVARGKCIKKGQILADGAATVGGELSLGKNVLVAYMPWEGYNFEDAVLINERLVYGDIYTSFHIRKYEIQTHVTSHGPERITNEIPHLEAHLLRNLDKKGVVMLGSWVETGDILVGKLTPQMAKESSYAPEDRLLRAILGIQISTSKETCLKLPIGSRGRVIDVRWIQKKGGSSYNPETIRVYISQKREIKVGDKVAGRHGNKGIVSKILPRQDMPYLQDGRPVDMVFNPLGVPSRMNVGQIFECSLGLAGELLDRHYRIAPFDERYEQEASRKLVFSELYEASKQTANPWVFEPEYPGKSRIFDGRTGDPFEQPVIIGKPYILKLIHQVDDKIHGRCSGHYALVTQQPLRGRAKQGGQRVGEMEVWALEGFGVAHILQEMLTYKSDHIRARQEVLGTTIIGGPIPKPEDAPESFRLLVRELRSLALELNHFLVSEKNFQINRKDA

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.

Para encontrar a segunda sequência, como foi utilizado uma sequência de aminoácidos, a busca foi realizada no BLASTP. Então foi colocada a sequência ">12.fa\_ 95 [787 - 4002]” na busca do BLASTP, utilizando as configurações padrões. Em seguida e como esperado, as primeiras sequências mais similares pertenciam a espécie *Cannabis Sativa,* então foi selecionada a sequência de outra espécie entre os primeiros resultados, pertencente a especie *Pteroceltis tatarinowii*

* 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?

O código de acesso é YP\_009530027.1, o E value é 0 e a porcentagem de identidade é 97.76%.

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?

Para obter o alinhamento global foi utilizado o Rstudio ( como não moro em Piracicaba não consegui utilizar os computadores do CENA). Então foi utilizado o pacote Biostrings e a função pairwiseAlignment(), a qual faz o alinhamento global utilizando o algoritmo de Needleman-Wunsch.

Seguindo os parâmetros descritos acima foi criada a matriz de substituição, representada abaixo:

A R N D C Q E G H I L K M F P S T W Y V B J Z X \*

A 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

R -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

N -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

D -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

C -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

Q -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

E -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

G -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

H -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

I -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

L -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

K -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

M -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

F -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

P -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2

S -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2 -2

T -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2 -2

W -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2 -2

Y -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2 -2

V -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2 -2

B -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2 -2

J -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2 -2

Z -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2 -2

X -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5 -2

\* -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 5

Então, foi utilizado o seguinte código para obter o alinhamento, sendo X a sequência de proteínas com a fase de leitura mais comprida da sequência da pergunta 1 e Y a sequência encontrada na pergunta 6.

X<-'KKMLGDGNGKMSTIPGFNQIQFEGFCRFIDQGLTEELYKFPKMEDTDQEIEFQLFVETYQLAEPLIKERDAVYESLTYSSELYVSAGLIWKTSRDMQEQTIFIGNIPLMNSLGTSIVNGIYRIVINQILQSPGIYYRSELDHNGISVYTGTIISDWGGRLELEIDRKARIWARVSRKQKISILVLSSAMGSNLREILENVCYPEIFLSFLNDKEKKKIKSKENAILEFYQQFACVGGDPVFSESLCKELQKKFFQQRCELGRIGRRNMNRRLNLDIPQNNTFLLPRDILAAADHLIGMKFGMGILDDMNHLKNKRIRSVADLLQDQFGLALVRLENMVRGTMSGAIRHKLIPTPQNLVTSTTLTTTFESFFGLHPLSQVLDRTNPLTQIVHGRKLSYLGPGGLTGRTASFRIRDIHPSHYGRICPIDTSEGINVGLIGSLSIHARIGPWGSLESPFYEISERSKKVRMLYLSPSRDEYYMVAAGNSLALNRGSQEEQVVPARYRQEFLTIEWEQVHLRNIFPFQYFSIGASLIPFIEHNDANRALMSSNMQRQAVPLSRSEKCIVGTGLESQVALDSGVPAIAEHEGKIVYTDTDKIILSGNGDTLSIPLVRYQRSNKNTCMHQKPQVARGKCIKKGQILADGAATVGGELSLGKNVLVAYMPWEGYNFEDAVLINERLVYGDIYTSFHIRKYEIQTHVTSHGPERITNEIPHLEAHLLRNLDKKGVVMLGSWVETGDILVGKLTPQMAKESSYAPEDRLLRAILGIQISTSKETCLKLPIGSRGRVIDVRWIQKKGGSSYNPETIRVYISQKREIKVGDKVAGRHGNKGIVSKILPRQDMPYLQDGRPVDMVFNPLGVPSRMNVGQIFECSLGLAGELLDRHYRIAPFDERYEQEASRKLVFSELYEASKQTANPWVFEPEYPGKSRIFDGRTGDPFEQPVIIGKPYILKLIHQVDDKIHGRCSGHYALVTQQPLRGRAKQGGQRVGEMEVWALEGFGVAHILQEMLTYKSDHIRARQEVLGTTIIGGPIPKPEDAPESFRLLVRELRSLALELNHFLVSEKNFQINRKDA'

Y <-'MLGGGNGKMSTIPGFNQIQFEGFCRFIDQGLTEELYKFPKIEDTDQEIEFQLFVETYQLVEPLIKERDAVYESLTYSSELYVSAGLIWKTSRDMQEQIIFIGNIPLMNSLGTSIVNGIYRIVINQILQSPGIYYRSELDHNGISVYTGTVISDWGGRLELEIDRKARIWARVSRKQKISILVLSSAMGSNLREILENVCYPEIFLSFLNDKEKKKIKSKENAILEFYQQFACVGGDPVFSESLCKELQKKFFKQRCELGRIGRRNMNRRLNLDIPQNNTFLLPRDILAAADHLIGMKFGMGILDDMNHLKNKRIRSVADLVQDQFGLALVRLENMVRGTMCGAIRHKLIPTPQNLVTSTTLTTTYESFFGLHPLSQVLDRTNPLTQIVHGRKSSYLGPGGLTGRTASFRVRDIHPSHYGRICPIDTSEGINVGLIGSLAIHARIGHWGSLESPFYEISERSKKIRMLYLSPSRDEYYMVAAGNSLALNRGSQEEQVVPARYRQEFLTIEWEQVHFRNIYPFQYFSIGASLIPFIEHNDANRALMSSNMQRQAVPLSRSEKCIVGTGLESQVALDSGVPAIAEHEGKIIYTDTDKIILSGNGDTLSIPLVIYQRSNKNTCMHQKPQVARGKCIKKGQILADGAATVDGELSLGKNVLVAYMPWEGYNSEDAVLINERLVYGDIYTSFHIRKYEIQTHVTSHGPERITNEIPHLEAHLLRNLDKKGVVMLGSWVETGDILVGKLTPQMAKESSYAPEDRLLRAILGIQISTSKETCLKLPIGSRGRVIDVRWIQKRGGSSYNPETIRVYISQKREIKVGDKVAGRHGNKGIVSKILPRQDMPYLQDGRPVDMVFNPLGVPSRMNVGQIFECSLGLAGGLLDRHYRLAPFDERYEQEASRKLVFSELYEASKQTANPWVFEPEYPGKSRIFDGRTGDPFEQPVIIGKPYILKLIHQVDDKIHGRCSGHYALVTQQPLRGRAKQGGQRVGEMEVWALEGFGVAHILQEMLTYKSDHIRARQEVLGATIIGGPIPKPEDAPESFRLLVRELRSLALELNHFLVSEKNFQINRKDA'

globalAlign<- pairwiseAlignment(X, Y, type= "global",

substitutionMatrix = Matriz,

gapOpening = -6,

gapExtension = -6,

scoreOnly = FALSE)

> globalAlign

Global PairwiseAlignmentsSingleSubject (1 of 1)

pattern: KKMLGDGNGKMSTIPGFNQIQFEGFCRFIDQGLTEELYKFPKMEDTDQEIEFQ...VLGTTIIGGPIPKPEDAPESFRLLVRELRSLALELNHFLVSEKNFQINRKDA

subject: --MLGGGNGKMSTIPGFNQIQFEGFCRFIDQGLTEELYKFPKIEDTDQEIEFQ...VLGATIIGGPIPKPEDAPESFRLLVRELRSLALELNHFLVSEKNFQINRKDA

score: 5164

Acima podemos observar o começo e o final do alinhamento, e também o score.

* 1. Qual a pontuação do alinhamento?

5164