# Exercise 2: Pairwise alignments

Program is available

DOTLET

<http://myhits.isb-sib.ch/cgi-bin/dotlet>

Alignment based on dot representation (use NetScape)

https://dotlet.vital-it.ch/

Use the following sequences in FastA format:

***DATASET 1 Two terminal oxidases from the same family***

>gi|13449404|ref|NP\_085587.1| cytochrome c oxidase subunit 1 [Arabidopsis thaliana]

MKNLVRWLFSTNHKDIGTLYFIFGAIAGVMGTCFSVLIRMELARPGDQILGGNHQLYNVLITAHAFLMIFFMVMPAMIGGFGNWFVPILIGAPDMAFPRLNNISFWLLPPSLLLLLSSALVEVGSGTGWTVYPPLSGITSHSGGAVDLAIFSLHLSGVSSILGSINFITTIFNMRGPGMTMHRLPLFVWSVLVTAFLLLLSLPVLAGAITMLLTDRNFNTTFFDPAGGGDPILYQHLFWFFGHPEVYILILPGFGIISHIVSTFSGKPVFGYLGMVYAMISIGVLGFLVWAHHMFTVGLDVDTRAYFTAATMIIAVPTGIKIFSWIATMWGGSIQYKTPMLFAVGFIFLFTIGGLTGIVLANSGLDIALHDTYYVVAHFHYVLSMGAVFALFAGFYYWVGKIFGRTYPETLGQIHFWITFFGVNLTFFPMHFLGLSGMPRRIPDYPDAYAGWNALSSFGSYISVVGICCFFVVVTITLSSGNNKRCAPSPWALELNSTTLEWMVQSPPAFHTFGELPAIKETKSYVK

>gi|461786|sp|P33517|COX1\_RHOSH Cytochrome c oxidase polypeptide I (Cytochrome AA3 subunit 1)

MADAAIHGHEHDRRGFFTRWFMSTNHKDIGVLYLFTGGLVGLISVAFTVYMRMELMAPGVQFMCAEHLESGLVKGFFQSLWPSAVENCTPNGHLWNVMITGHGILMMFFVVIPALFGGFGNYFMPLHIGAPDMAFPRMNNLSYWLYVAGTSLAVASLFAPGGNGQLGSGIGWVLYPPLSTSESGYSTDLAIFAVHLSGASSILGAINMITTFLNMRAPGMTMHKVPLFAWSIFVTAWLILLALPVLAGAITMLLTDRNFGTTFFQPSGGGDPVLYQHILWFFGHPEVYIIVLPAFGIVSHVIATFAKKPIFGYLPMVYAMVAIGVLGFVVWAHHMYTAGLSLTQQSYFMMATMVIAVPTGIKIFSWIATMWGGSIELKTPMLWALGFLFLFTVGGVTGIVLSQASVDRYYHDTYYVVAHFHYVMSLGAVFGIFAGSTSGIGKMSGRQYPEWAGKLHFWMMFVGANLTFFPQHFLGRQGMPRRYIDYPEAFATWNFVSSLGAFLSFASFLFFLGVIFYSLSGARVTANNYWNEHADTLEWTLTSPPPEHTFEQLPKREDERAPAH

***DATASET 2 Two terminal oxidases from a different family***

>gi|13449404|ref|NP\_085587.1| cytochrome c oxidase subunit 1 [Arabidopsis thaliana]

MKNLVRWLFSTNHKDIGTLYFIFGAIAGVMGTCFSVLIRMELARPGDQILGGNHQLYNVLITAHAFLMIFFMVMPAMIGGFGNWFVPILIGAPDMAFPRLNNISFWLLPPSLLLLLSSALVEVGSGTGWTVYPPLSGITSHSGGAVDLAIFSLHLSGVSSILGSINFITTIFNMRGPGMTMHRLPLFVWSVLVTAFLLLLSLPVLAGAITMLLTDRNFNTTFFDPAGGGDPILYQHLFWFFGHPEVYILILPGFGIISHIVSTFSGKPVFGYLGMVYAMISIGVLGFLVWAHHMFTVGLDVDTRAYFTAATMIIAVPTGIKIFSWIATMWGGSIQYKTPMLFAVGFIFLFTIGGLTGIVLANSGLDIALHDTYYVVAHFHYVLSMGAVFALFAGFYYWVGKIFGRTYPETLGQIHFWITFFGVNLTFFPMHFLGLSGMPRRIPDYPDAYAGWNALSSFGSYISVVGICCFFVVVTITLSSGNNKRCAPSPWALELNSTTLEWMVQSPPAFHTFGELPAIKETKSYVK

>gi|2114418|gb|AAB58264.1| cbb3-type cytochrome oxidase component FixN [Rhizobium leguminosarum bv. viciae]

MNYTTETMVIAVAAFLALLVAAFAHDHLFAVHMGILCLCLVMGAVLMVRKVDFSPAGQQRNVDRSGYFDEVIRYGLIATVFWGVVGFLVGVIIALQLAFPDLNIAPYLNFGRLRPVHTSAVIFAFGGNALIMTSFYVVQRTCRARLFGGNLAWFVFWGYQLFIVMAATGYVLGITQGREYAEPEWYVDLWLTIVWVAYLAVYLGTILKRKEPHIYVANWFYLSFIVTIAMLHVVNNLAVPASFLGSKSYSVSSGVQDALTQWWYGHNAVGFFLTAGFLGMMYYFVPKQANRPVYSYRLSIIHFWALIFMYIWAGPHHLHYTALPDWAQTLGMVFSIMLWMPSWGGMINGLMTLSGAWDKIRTDPIIRMMIVAIAFYGMSTFEGPMMSVKTVNSLSHYTEWTIGHVHSGALGWVGMITFGAIYYLTPKLWGRERLYSLRMVNWHFWLATFGIVVYAAVLWVAGIQQGLMWREYNSQGFLVYSFAETVAAMFPYYVLRAVGGTLYLAGGLVMAWNVFMTIRGHLRDEAAIPTTFVPQAQPAE

***DATASET 3 Random sequences***

>gi|13449404|ref|NP\_085587.1| cytochrome c oxidase subunit 1 [Arabidopsis thaliana]

MKNLVRWLFSTNHKDIGTLYFIFGAIAGVMGTCFSVLIRMELARPGDQILGGNHQLYNVLITAHAFLMIFFMVMPAMIGGFGNWFVPILIGAPDMAFPRLNNISFWLLPPSLLLLLSSALVEVGSGTGWTVYPPLSGITSHSGGAVDLAIFSLHLSGVSSILGSINFITTIFNMRGPGMTMHRLPLFVWSVLVTAFLLLLSLPVLAGAITMLLTDRNFNTTFFDPAGGGDPILYQHLFWFFGHPEVYILILPGFGIISHIVSTFSGKPVFGYLGMVYAMISIGVLGFLVWAHHMFTVGLDVDTRAYFTAATMIIAVPTGIKIFSWIATMWGGSIQYKTPMLFAVGFIFLFTIGGLTGIVLANSGLDIALHDTYYVVAHFHYVLSMGAVFALFAGFYYWVGKIFGRTYPETLGQIHFWITFFGVNLTFFPMHFLGLSGMPRRIPDYPDAYAGWNALSSFGSYISVVGICCFFVVVTITLSSGNNKRCAPSPWALELNSTTLEWMVQSPPAFHTFGELPAIKETKSYVK

>gi|16121653|ref|NP\_404966.1| transport ATP-binding protein [Yersinia pestis]

MQTSHLMNKTRQYELIRWLKKQSAPAQRWLRLSMLLGLLSGLLIIAQAWLLATLLQSLIIDKLPRATLTTEFSLLAGAFALRAVISWLRERVGFICGMRVRQQIRKVVLDRLEQLGPSWVKGKPAGSWATIILEQIEDMQEYYSRYLPQMYLAVFIPVLILIAVFPINWAAGLILFVTAPLIPIFMILVGMGAADANRRNFVALARLSGNFLDRLRGLDTLRLFNRAKAETDQIRDSSEDFRSRTMEVLRMAFLSSAVLEFFAAISIAVVAVYFGFSYLGELNFGSYGLGVTLFAGFLVLILAPEFFQPLRDLGTFYHAKAQAVGAAESLVTFLSSEGEAIGQGEKQLDGKEAIALEANELEILAPNGTRLAGPLNFSLPAGKRVAIVGQSGAGKSSLLNLLLGFLPYRGSLKVNGIELRELEPQVWRSQLSWVGQNPHLPEQTLATNILLRQPDASEHQLQQAVERAYINEFLKDLPQGLNTEIGDHSARLSVGQAQRIAVARALLNPCRLLLLDEPTASLDAHSEQLVMKALEEASRAQSTLLVTHQLEDTLGYDQIWVMDNGRLIQQGDYSTLSQSAGSFANLLSQRNEEL

### Make an alignment using dotlet

<https://dotlet.vital-it.ch/>

Import the sequences

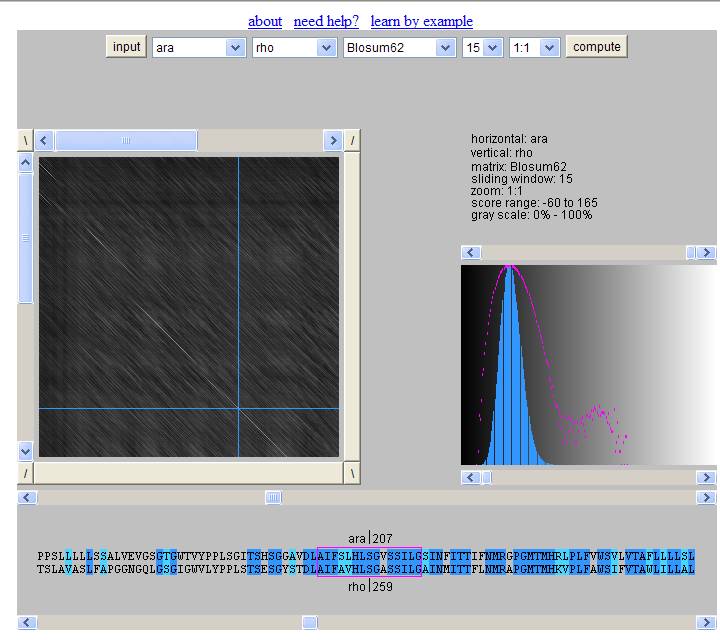
1) Compare a sequence to itself e.g. Arabidopsis to Arabidopsis

2) Align the sequence of Rhodobacter (COX1\_RHOS) with the one of Arabidopsis (Dataset 1)

3) Align the sequence of Arabidopsis to the one of Rhizobium (dataset 2)

4) Align the sequence of Arabidopsis to the sequence from dataset 3

Try out the effect of different substitution matrices and of the sequence length used to put a dot.



### Perform a global alignment (Needleman Wunsch)

http://www.ebi.ac.uk/Tools/emboss/align/index.html

a) Use the Default settings (EMBOSS global alignment)

Write down the score for

Use dataset 1: 1715

Use dataset 2: 169

Use dataset 3: 31

And visually inspect the alignments

**Question: Can you visually make a distinction between the relevant and spurious alignments?**

a

**Question: How is the score derived? Can it help you making a distinction between the relevant and spurious alignments?**

a

**Question: Is there much difference between the scores of the different datasets?**

A

**Question: Can you compare the score between the different datasets?**

a

In principle you cannot because you are comparing different sequence sets that might not have exactly the same length. However, because the sequence lengths between the datasets here are quite comparable the score level gives a hint.

b) Try different gap parameter on the second dataset

Lower the gap opening penalty term: Increase the gap extension parameter.

What changes do you expect? Can you see them in the alignment? Use dataset2 as here the changes in parameter setting will most severely affect the results

Dataset2 13

Gap open 1

Gap extend 10

Dataset 2 656

Dataset 3 437

(many small gaps)

Gap open 100

Gap extend 10

Dataset 2 6

Dataset 3 2

(one big gap)

Question: Explain why you find each time you use different parameter setting a different score.

Because if you use a different scoring scheme different scores are obtained and these scores can not be compared mutually. As soon as you change the gap scoring systems alignment scores become incomparable (and a higher score does not reflect as better alignment)

c) change the substitution scoring matrices:

Question: From which score matrices can you choose?

Answer: a whole series of PAM and BLosum matrices

Try on dataset2 a different scoring scheme (lower blosum matrix 30).

Question: Why would this scoring scheme be more appropriate? –

Answer: Because we deal with evolutionary distant sequences (646.5)

Question: How is the score influenced by choosing a lower numbered BLOSUM matrix?

Answer: The score should improve (# Score: 643.5 instead of 169)! Note that this setting in which you have a global alignment, use the same gap penalty parameter and substitution scoring system (except for the numbers) and the same sequence set, the obtained alignment scores can be compared.

### Perform a local alignment (Smith Waterman)

Use first the smith waterman procedure of Emboss

http://www.ebi.ac.uk/Tools/psa/emboss\_water/

Default parameter settings used: (Blosum 60)

Write down the score?

Dataset 1: 1725

Dataset 2: 189

Dataset 3: 46

Question: For which dataset would you use the local alignments instead of the global ones? Why?

Answer: dataset 2 as this contain evolutionary quite distinct sequences that cannot well be aligned globally but that must contain local stretches that are conserved during evolution (in this case those that are involved in the binding of heme groups that are of relevance to the functioning of the protein)

Conclusion:

* Scores of alignments are length dependent and dependent on the scoring system so they can never be compared when trying to align different sequences (so no comparison of scores between datasets is possible)
* Scores of datasets of the same length (sequences to be aligned have the same length) can be compared as the used scoring system is the same (the gap penalty scoring and the substitution matrices).
* Scores obtained with different PAMs or BLOSUMs on the same pairwise alignment will differ and here it is assumed that the highest score is the result of the matrix being more adapted to the phylogenetic distance of the sequences to be compared. This does not tell anything about whether the alignment obtained with the respective different scoring systems will be the same.

Having a statistical meaning of the scores would help us in distinguishing between good and bad alignments

### Statistical significance of local alignments

Now we will as a local alignment procedure, the procedure

http://www.ebi.ac.uk/Tools/services/web\_lalign/

Use the BLOSUM62:

Default gap parameters are then: -7 open, -1 extension

>>gb|AAB58264.1| cbb3-type cytochrome oxidase (540 aa)

Waterman-Eggert score: 198; 29.7 bits; E(1) < 0.00033

24.6% identity (49.7% similar) in 455 aa overlap (23-423:74-451)

30 40 50 60

ref|NP FGAIA----GVMGTCFSVLIRMELARPGDQI-----LGGNHQLY-NVLITA---HAFLMI

.: :: ::.: .:.: ..:: : .: .: . .. ...: : .:..:

gb|AAB YGLIATVFWGVVGFLVGVIIALQLAFPDLNIAPYLNFGRLRPVHTSAVIFAFGGNALIMT

80 90 100 110 120 130

70 80 90 100 110 120

ref|NP FFMVMP----AMIGGFGN--WFVPILIGAPDMAFPRLNNISFWLLPPSLLLLLSSALVEV

: :. : . : :: ::: :: . : :

gb|AAB SFYVVQRTCRARLFG-GNLAWFV------------------FW----------GYQLFIV

140 150 160

130 140 150 160 170

The advantage of this approach is that it calculates the significance of the alignments by shuffling.

Write down the scores:

1) Dataset 1: Waterman-Eggert score: 1720; 195.3 bits; E(1) < 5e-54

2)Dataset 2 Waterman-Eggert score: 181; 36.6 bits; E(1) < 2.8e-06

3) Dataset 3: Waterman-Eggert score: 47; 17.6 bits; E(1) < 0.8

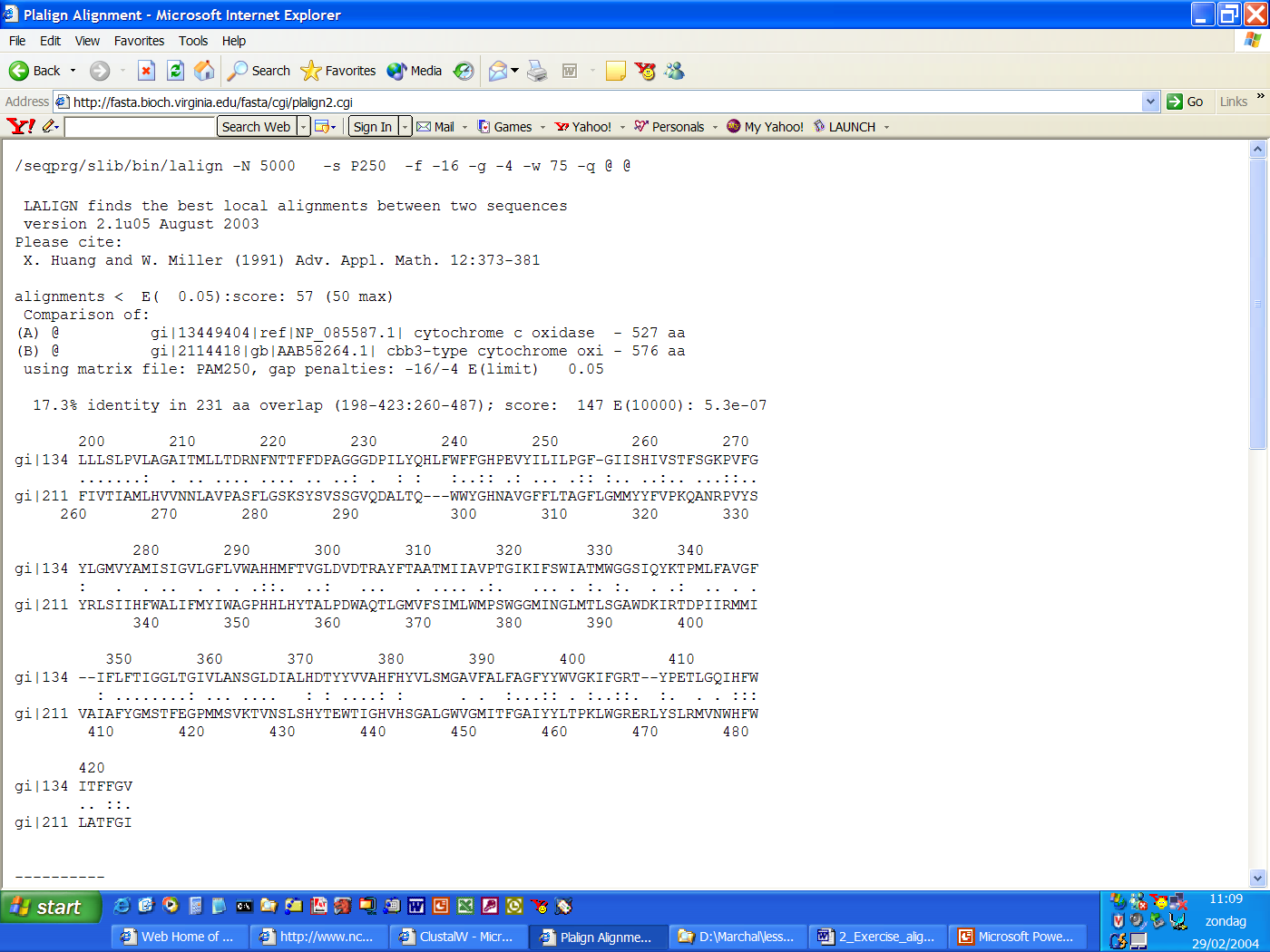
What is the meaning of the scores?

Can you compare the E-values between alignments?

**Especially for 2 sequences that are not very similar anymore it is difficult to assess whether the alignment is still biologically true. Introducing more sequences and making use of a multiple sequence alignment can increase the information (see exercise 2)**



Tuning the parameters of the pairwise alignment allows you to get a better alignment of the relevant residues in case of data set 2 (see below for the parameters). The aminoacids required for the enzymatic reaction get better aligned.



Dataset 2

### Database search using FastA

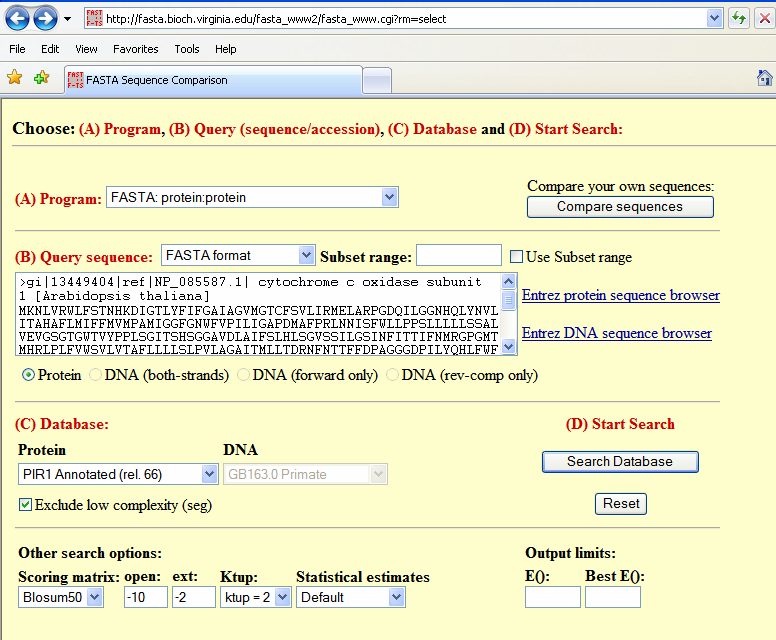
Take a protein sequence from dataset 1

Use the FASTA program: <http://fasta.bioch.virginia.edu/fasta/cgi/searchx.cgi?pgm=fa>

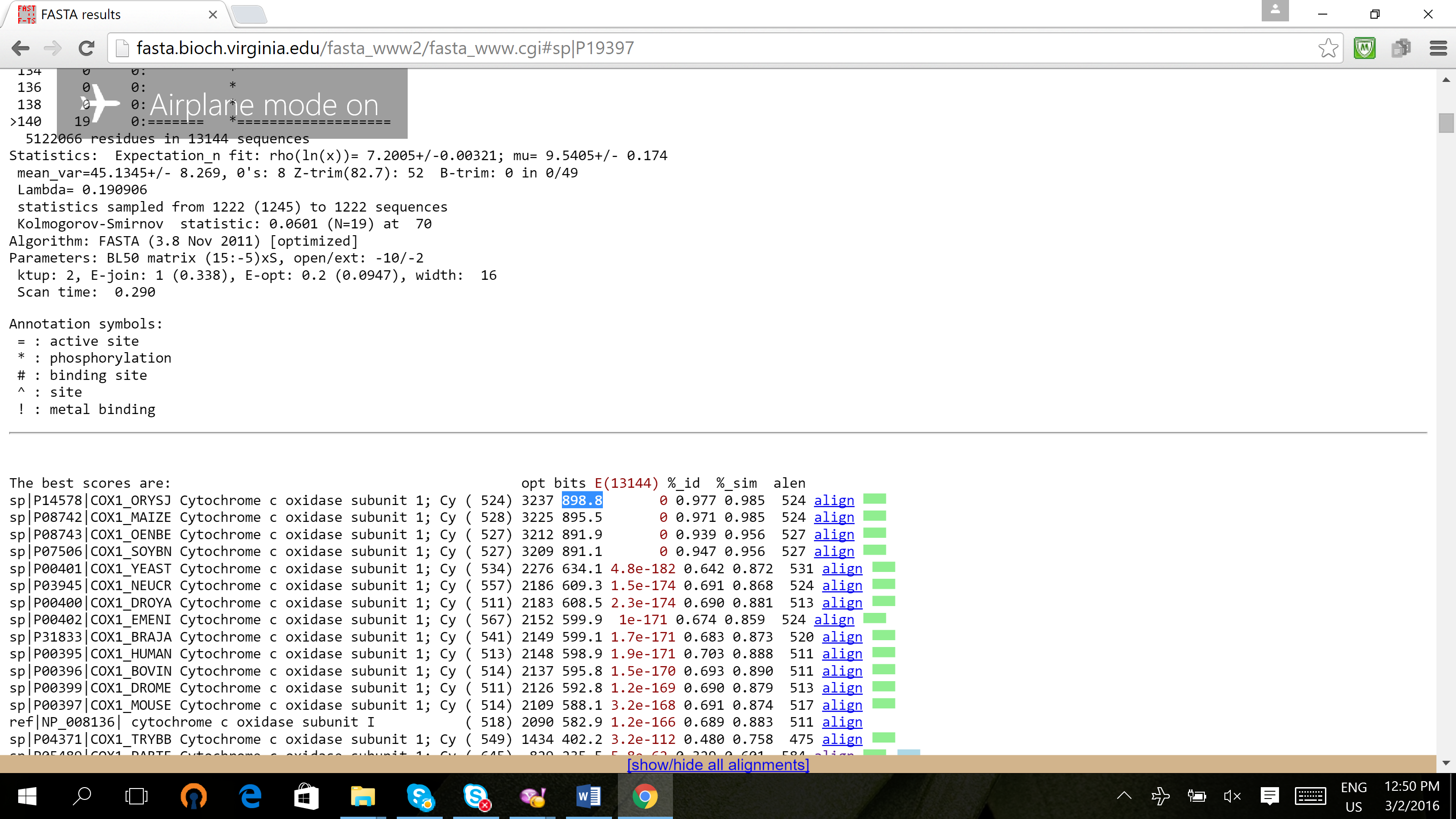
What does this program do?

What is the meaning of ktup?

Calculate the statistical significance of the alignment.



Interpret the results?



Raw

score

Bit

score

E value

(length target seq)

(length match)

# fasta36 -p -q -w 80 -m 9i -m 6 -H -f -10 -V "!./annot/ann\_feats2ipr.pl --neg --acc\_comment" -S -g -2 TMP.q A 2

FASTA searches a protein or DNA sequence data bank

version 36.3.7a Jan, 2015(preload9)

Please cite:

W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

Query: TMP.q

1>>>gi|13449404|ref|NP\_085587.1| cytochrome c oxidase subunit 1 [Arabidopsis thaliana] - 527 aa

Library: PIR1 Annotated (rel. 66)

5121825 residues in 13143 sequences

opt E()

< 40 8 0:===

42 0 0: one = represents 3 library sequences

44 0 0:

46 0 0:

48 6 1:\*=

50 5 5:=\*

52 22 15:====\*===

54 26 32:========= \*

56 67 55:==================\*====

58 73 79:========================= \*

# fasta36 -p -q -w 80 -m 9I -m 6 -m 9I -m 6 -H -f -10 -V "!./annot/ann\_feats2ipr.pl --neg --acc\_comment" -S -g -2 TMP.q A 2

FASTA searches a protein or DNA sequence data bank

version 36.3.8c Dec, 2015(preload9)

Please cite:

W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

Query: TMP.q

1>>>gi|13449404|ref|NP\_085587.1| cytochrome c oxidase subunit 1 [Arabidopsis thaliana] - 527 aa

Library: PIR1 Annotated (13K)

5122066 residues in 13144 sequences

opt E()

< 40 8 0:===

42 0 0: one = represents 3 library sequences

44 0 0:

46 0 0:

48 7 1:\*==

50 6 5:=\*

52 18 15:====\*=

54 28 32:==========\*

56 64 55:==================\*===

58 71 79:======================== \*

60 48 99:================ \*

62 87 111:============================= \*

64 113 115:======================================\*

66 121 111:====================================\*====

68 92 102:=============================== \*

70 89 91:==============================\*

72 87 78:=========================\*===

74 77 65:=====================\*====

76 52 54:=================\*

78 35 44:============ \*

80 35 35:===========\*

82 33 28:=========\*=

84 24 22:=======\*

86 33 17:=====\*=====

88 26 14:====\*====

90 11 11:===\*

92 8 8:==\*

94 13 6:=\*===

96 10 5:=\*==

98 7 4:=\*=

100 3 3:\*

102 0 2:\*

104 1 2:\*

106 6 1:\*=

108 1 1:\* inset = represents 1 library sequences

110 1 1:\*

112 2 1:\* :\*=

114 0 1:\* :\*

116 2 0:= \*==

118 3 0:= \*===

120 1 0:= \*=

122 1 0:= \*=

124 0 0: \*

126 1 0:= \*=

128 0 0: \*

130 1 0:= \*=

132 0 0: \*

134 0 0: \*

136 0 0: \*

138 0 0: \*

>140 19 0:======= \*===================

5122066 residues in 13144 sequences

Statistics: Expectation\_n fit: rho(ln(x))= 7.2005+/-0.00321; mu= 9.5405+/- 0.174

mean\_var=45.1345+/- 8.269, 0's: 8 Z-trim(82.7): 52 B-trim: 0 in 0/49

Lambda= 0.190906

statistics sampled from 1222 (1245) to 1222 sequences

Kolmogorov-Smirnov statistic: 0.0601 (N=19) at 70

Algorithm: FASTA (3.8 Nov 2011) [optimized]

Parameters: BL50 matrix (15:-5)xS, open/ext: -10/-2

ktup: 2, E-join: 1 (0.338), E-opt: 0.2 (0.0947), width: 16

Scan time: 0.290

The best scores are: opt bits E(13143) %\_id %\_sim alen

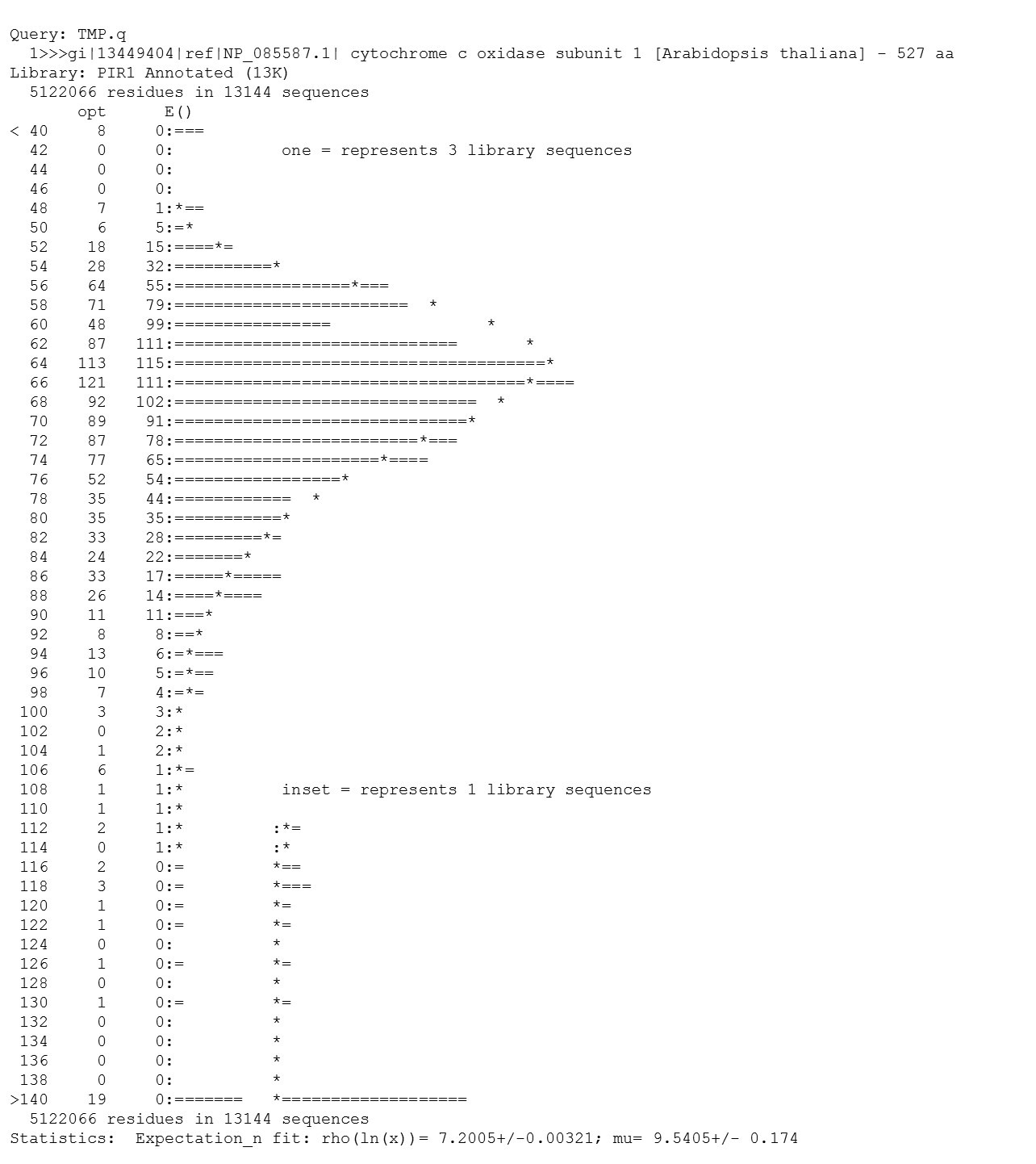
sp|P14578|COX1\_ORYSJ Cytochrome c oxidase subunit 1; Cy ( 524) 3237 898.8 0 0.977 0.985 524 [align](http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi#sp|P14578)

sp|P08742|COX1\_MAIZE Cytochrome c oxidase subunit 1; Cy ( 528) 3225 895.5 0 0.971 0.985 524 [align](http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi#sp|P08742)

sp|P08743|COX1\_OENBE Cytochrome c oxidase subunit 1; Cy ( 527) 3212 891.9 0 0.939 0.956 527 [align](http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi#sp|P08743)

sp|P07506|COX1\_SOYBN Cytochrome c oxidase subunit 1; Cy ( 527) 3209 891.1 0 0.947 0.956 527 [align](http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi#sp|P07506)

sp|P00401|COX1\_YEAST Cytochrome c o



Alignment score (bins)

Observed number of target sequences with a given match score

19 sequences have a match with the query with a score higher than 140

Expected number of target sequences with a given match =score

### Database search using Blast

Go to the Blast homepage at NCBI

What is the difference between blastn, blastp,blastX

What does nr (on redundant mean)

Blast the unknown sequence

Interpret the result:

Explain E-value, bitscore, Identities, Positives, Query Subject

What does E-value = 0 mean?

What is the best match in the database?

Do you have a clue on the function of the protein?

Go to the GenBank file of the best hit: (click on the link)

What information can you find in the GenBank file?

Suppose in the lab you sequenced the following sequence.

To learn more about the function of this gene, search for a homologue in the protein database.

>gene 1

ATGACATCAGCGACTCTGACGCCAGGGGCCGCCCTGGGCAGCCAGCGGGTGTCGGAAAATGTGCGTTACTACGAAGACGCCGTCCGACTCTTCGTCATCGCTGCAGTGTTCTGGGGCGTCGTCGGCTTCCTCGCCGGCGTCTTCATCGCGCTGCAGCTGGCTTTTCCGGCGCTGAATCTCGGCCTTGAGTGGACGAGCTTCGGGCGCCTGCGGCCGGTCCACACCTCGGCCGTGATCTTCGCGTTTGGCGGCAACGTCCTGTTCGCCACCTCGCTCTACTCCGTGCAGCGCACCAGCCGCCAGTTCCTGTTCGGCGGCGAGGGCCTCGCGAAGTTCGTCTTCTGGAACTACAACATCTTCATCGTCCTGGCGGCGCTCAGCTACGTGCTCGGCTACACCCAGGGCAAGGAGTATGCAGAGCCGGAGTGGATCCTCGACCTCTACCTGACGGTCATCTGGGTCCTCTACGCCATCCAGTTCGTCGGCACGGTGATGACCCGCAAGGAGTCGCACATCTACGTCGCCAACTGGTTCTTCATGGCGTTCATCCTGACCGTCGCGATCCTCCACATCGGCAACAACGTCAACGTCCCGGTGTCGCTGACCGGGATGAAGTCCTACCCGTTCGTCTCGGGCGTGCAGAGCGCCATGGTGCAGTGGTGGTACGGCCACAACGCGGTCGGCTTCTTCCTGACCGCCGGCTTCCTCGGCATCATGTCTACTTCGTTCCGAAGCGCGCGGAGCGGCCGGTCTATTCGTACCGCCTGTCGATCGTGCACTTCTGGACGCTGATCTTCCTCTACATCTGGGCCGGCCCGCACCACCTGCACTACACGGCCCTGCCGGATTGGGCGCAGACGCTGGGCATGACCTTCTCGGTCATGCTGTGGATGCCGTCCTGGGGCGGCATGATCAACGGCATCATGACCCTGTCGGGTGCCTGGGACAAGCTGCGCACCGACCCGGTCCTGCGCTTCCTCGTGACGTCGGTGGCCTTCTACGGCATGTCGACCTTCGAGGGCCCGCTGATGTCGGTGAAGCCGGTCAACGCCCTGTCGCACTACACCGACTGGACGATCGGCCACGTGCACTCCGGTGCGCTCGGCTGGGTGGCCTTCATCTCCTTCGGCGCGATCTACTATCTGGTCCCGGTCCTGTGGAAGCGCTCGCAGCTCTACAGCCTGCGTCTGGTCAGCTACCACTTCTGGACCGCCACCATCGGCATCGTGCTCTACATCACCGCCATGTGGGTGTCGGGCATCATGCAGGGCCTGATGTGGCGCGCCTACGACAACCTCGGCTTCCTCCAGTACTCGTTCGTCGAGACGGTCGCGGCCATGCATCCCTTCTACGTGATCCGTGCGCTGGGCGGCGTCCTGTTCCTGGCTGGTGCCCTGATCATGGTCTACAACCTGTGGCGCACGGCCAAGGGTGACGTCCGCATCGAGAAGCCCTATGCCTCCGCCCCGCACAAGGCGGCGGTCGGTGCGGCCTGA

Blast the following sequence to the non redundant protein database.

Which sequence has the highest score.

How do you explain the different HSPs observed for the same sequence?

What does the E-value tell you?

