

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]

```
MKNLVRWLFSTNHKDIGTLYFIFGAIAGVMGTCFSVLIRMELARPGDQILGGNHQLYNVLITAHAFLMIFFMVPAMIGG
FGNWFVPILIGAPDMAFPRLNNISFWLLPPSLLLLSSALVEVGSGTGWTVPPLSGITSHSGGAVDLAIFSLHLSGVSS
ILGSINFITTTIFNMRGPGMTMHRPLPLFVWSVLVTAFLLLLSLPVLAGAITMLLTDNRNFNTTFFDPAGGGDPILYQHLFWF
FGHPEVYIILILPGFGIISHIVSTFSGKPVFGYLGVMYAMISIGVLGFLVWAHMFMTVGLDVDTRAYFTAATMIIAVPTGI
KIFSWIATMWGGSIQYKTPMLFAVGFI FLFTIGGLTGIVLANSGLDIALHDTYYVVAHFHYVLSMGAVFALFAGFYWVG
KIFGRTPETLGQIHFWITFFGVNLTFFPMHFLGLSGMPRRIPDYPDAYAGWNALSSFGSYISVVGICCFVVTITLSS
GNNKRCAPSPWALELNSTTLEWMVQSPPAFHTFGELPAIKETKS YVK
```

>gi|461786|sp|P33517|COX1_RHOSH Cytochrome c oxidase polypeptide I (Cytochrome AA3 subunit 1)

```
MADAAIHGHEHRRGFFTRWFMSTNHKDIGVLYLFTGGVLGLISVAFTVYMRMELMAPGVQFMCAEHLESGLVKGFFQSL
WPSAVENCTPNGLHWNVMITGHGILMMFFVVIPALFGGFGNYFMPLHIGAPDMAFPRMNNLSYWLTVAGTSLAVASLFAP
GGNGQLGSGIGWVLYPPLSTSESGYSTDLAIFAVHLSGASSILGAINMITTFLNMRAPGMTMHKVPLFAWSIFVTAWLIL
LALPVLGAITMLLTDNRNFTTFFQPSGGGDPVLYQHILWFFGHPEVYIIIVLPAFGIVSHVIATFAKKPIFGYLPVMYAM
VAIGVLGFFVWAHMYTAGLSLTQQSYFMMATMVIAPVTGIKIFSWIATMWGGSIELKTPMLWALGFLFLFTVGGVTGIV
LSQASVDRYYHDTYYVVAHFHYVMSLGAVFGIFAGSTSGIGKMSGRQYPEWAGKLHFWMMFVGANLTFFPQHFLGRQGMP
RRYIDYPEAFATWNFVSSLGAFLSFASFLFFLGVI FYSLSGARVTANNYWNEHADTLEWTLTSPPEHTFEQLPKREDER
APAH
```

DATASET 2 Two terminal oxidases from a different family

>gi|13449404|ref|NP_085587.1| cytochrome c oxidase subunit 1 [Arabidopsis thaliana]

```
MKNLVRWLFSTNHKDIGTLYFIFGAIAGVMGTCFSVLIRMELARPGDQILGGNHQLYNVLITAHAFLMIFFMVPAMIGG
FGNWFVPILIGAPDMAFPRLNNISFWLLPPSLLLLSSALVEVGSGTGWTVPPLSGITSHSGGAVDLAIFSLHLSGVSS
ILGSINFITTTIFNMRGPGMTMHRPLPLFVWSVLVTAFLLLLSLPVLAGAITMLLTDNRNFNTTFFDPAGGGDPILYQHLFWF
FGHPEVYIILILPGFGIISHIVSTFSGKPVFGYLGVMYAMISIGVLGFLVWAHMFMTVGLDVDTRAYFTAATMIIAVPTGI
KIFSWIATMWGGSIQYKTPMLFAVGFI FLFTIGGLTGIVLANSGLDIALHDTYYVVAHFHYVLSMGAVFALFAGFYWVG
KIFGRTPETLGQIHFWITFFGVNLTFFPMHFLGLSGMPRRIPDYPDAYAGWNALSSFGSYISVVGICCFVVTITLSS
GNNKRCAPSPWALELNSTTLEWMVQSPPAFHTFGELPAIKETKS YVK
```

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

```
MNYTTETMVIAAVAFLALLVAFAHDHLFAVHMGILCLCLVMGAVLMVRKVDFSPAGQQRNVD RSGYFDEVIRYGLIATV
FWGVVGLVGVIIALQLAFPDLNIAPYLNFGRLRPVHTSAVIFAAGGNALIMTSFYVVQRTCRARLFGGNLAWFVFWGYQ
LFIVMAATGYVLGITQGREYAEPEWYVDLWLTIVVWVAYLAVYLGITILKRKEPHIYVANWFYLSFIVTIAMLVNNLAVP
ASFLGSKSYSVSSGVQDAL TQWWYGHNAVGFLLTAGFLGMMYYFVPKQANRPVYSYRLSIIHF WALIFMYIWAGPHHLHY
TALPDWAQTLGMVFSIMLWMP SWGGMINGLMTLSGAWDKIRTDPIIRMMIVAI AFYGMSTFE GPMMSVKTVNSLSHYTEW
TIGHVHSGALGWVGMITFGAIYYLTPKLWGRERLYSLRMVNWHFWLATFGIVVYA AVLWVAGIQQGLMWREYNSQGFLVY
SFAETVAAMFPYYVLRVGGTLYLAGGLVMAWNVFMTIRGHLRDEAAIPTTFVPQAQPAE
```

DATASET 3 Random sequences

```
>gi|13449404|ref|NP_085587.1| cytochrome c oxidase subunit 1 [Arabidopsis thaliana]
MKNLVRWLFSTNHKDIGTLYFIFGAIAGVMGTCFSVLIRMELARPGDQILGGNHQLYNVLITAHAFMLIFFMVMPAMIGG
FGNWFVPILIGAPDMAFPRLNNISFWLLPPSLLLLLSSALVEVGSGTGWTVPPLSGITSHSGGAVDLAIFSLHLSGVSS
ILGSINFITTTIFNMRGPGMTMHRLLPLFVWSVLVTAFLLLLSLPVLAGAITMLLTDNRNFNTTFFDPAGGGDPILYQHLFWF
FGHPEVYILILPGFGIISHIVSTFSGKPVFGYLGVMYAMISIGVLGFLVWAHMFVGLDQVDTLAYFTAATMIIAVPTGI
KIFSWIATMWGGSIQYKTPMLFAVGFIPLFTIGGLTGIVLANSGLDIALHDTYYVVAHFHYVLSMGAVFALFAGFYWVG
KIFGRTPETLGIHFWITFFGVNLTFFPMHFLGLSGMPRRIPDYPDAYAGWNALSSFGSYISVVGICCFVVTITLSS
GNNKRCAPSPWALELNSTTLEWMVQSPPAFHTFGELPAIKETKSYVK
```

```
>gi|16121653|ref|NP_404966.1| transport ATP-binding protein [Yersinia pestis]
MQTSHLMNKTRQYELIRWLKKQSAPAQRWLRLSMLLGLLSGLLIIAQAWLLATLLQSLIIDKLPRATLTTEFSLLAGAF
LRAVISWLRERVGFICGMRVRQQIRKVVLDRLQLGPSWVKGKPAGSWATIILEQIEDMQEYYSRYLPQMYLAVFIPVLI
LIAVFPINWAAGLILFVTAPLIPIFMILVGMGAADANRRNFVALARLSGNFLDRLRGLDTRLRFNRAKAETDQIRDSED
FRSRTMEVLRMAFLSSAVLEFFAAISIAVVAVYFGFSYLGELNFGSYGLGVTLFAGFLVLILAPEFFQPLRDLGTFFHAK
AQAVGAAESLVTFLSSEGEAIGQGEKQLDGKEAIALEANELEILAPNGTRLAGPLNFSLPAGKRVAIVGQSGAGKSSLLN
LLLGFLPYRGS�KVNIGIELRELEPQVWRSQLSWVGQNPFLPEQTLATNILLRQPDASEHQLQQAVERAYINEFLKDLPOG
LNTEIGDHSARLSVGQAQRIAVARALLNPCRLLLLDEPTASLDAHSEQLVMKALEEASRAQSTLLVTHQLEDTLGYDQIW
VMDNGRLIQQGDYSTLSQSAGSFANLLSQRNEEL
```

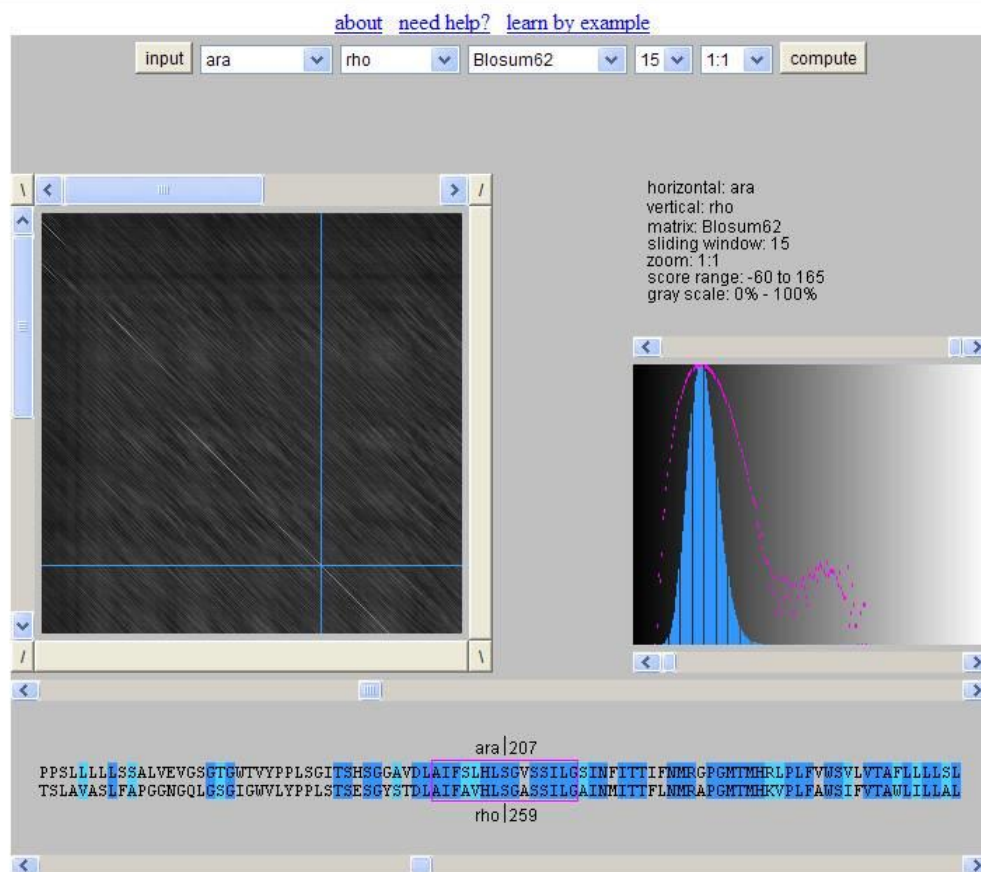
1.1 Make an alignment using dotlet

<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.



1.2 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?

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

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

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

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.

1.3 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

1.4 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---GVMGTCFSVLIRMEIARPGDQI-----LGGNHQLY-NVLITA---HAFLMI
  .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .:
gb|AAB YGLIATVFWGVVGVFLVGVIIALQLAFPDLNIAPYLNFGRLRPVHTSAVIFAFFGGNALIMT
      80      90     100     110     120     130

      70      80      90     100     110     120
ref|NP FFMVMP---AMIGGFGN--WFVPILIGAPDMAFPRLNNISFWLLPPSLLLLSSALVEV
  : .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .:
gb|AAB SFYVVQRTCRARLFG-GNLAWFV-----FW-----GYQLFIV
      140     150
      130     140     150     160     170
```

The advantage of this approach is that it calculates the significance of the alignments by shuffling (use 1000 shuffles).

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)

```
gi|1169145|sp|P98002|CX1B_PARD      FKTP--MLWAFGFLFLFTVGGVTGVVLSQAPLDRVYHDTYYVVAHFHY 416
gi|22956633|ref|ZP_00004383.1|      LKTP--MLWALGFLFLFTVGGVTGIVLSQASVDRYYHDTYYVVAHFHY 424
gi|1169027|sp|P08742|COX1_MAIZ      YKTP--MLFAVGFIPLFTIGGLTGIVLANSGLDIALHDTYYVVAHFHY 383
gi|1169030|sp|P14578|COX1_ORYS      YKTP--MLFAVGFIPLFTIGGLTGIVLANSGLDIALHDTYYVVAHFHY 383
gi|13449404|ref|NP_085587.1|        YKTP--MLFAVGFIPLFTIGGLTGIVLANSGLDIALHDTYYVVAHFHY 383
gi|34555994|emb|CAD33909.1|        WSPA--MLWALGFIPLFTVGGLTGIVLSNSSLDIVLHDTYYVVAHFHY 381
gi|32348033|gb|AAP47921.1|        WSAA--VLWALGFIPLFTVGGLTGIVLANSGLDIVLHDTYYVVAHFHY 381
gi|1352141|sp|Q08855|COX1_RHIL      FATP--MLWALAFIPLFTVGGVTGVVLANSGLDRVLHDTYYVVAHFHY 405
gi|2114418|gb|AAB58264.1|          IRTDPIIRMMIVAIAFYGMSTFEGPMMSVKTVNSLSHYTEWTIGHVHSSA 409
gi|22960507|ref|ZP_00008147.1|      IRTDPIIRMMVVISIGFYGMSTFEGPMMSIKAVNSLSHYTDWTIGHVHSSA 442
gi|3850275|gb|AAC72071.1|          LRTDPLVLRFLVTSVAFYGMSTFEGPLMSVKPVNLSHYTDWTIGHVHSSA 367
```

```
gi|1169145|sp|P98002|CX1B_PARD      GFVVVAHHMKTAGMSLTQQAYFMLATMTIAVPTGIKIFSWIATMWGGSIE 368
gi|22956633|ref|ZP_00004383.1|      GFVVVAHHMKTAGLSLTQQSYFMMATMVAIPTGIKIFSWIATMWGGSIE 376
gi|1169027|sp|P08742|COX1_MAIZ      GFLVVAHHMFTVGLDVTTRAYFTAATMIIAVPTGIKIFSWIATMWGGSIQ 335
gi|1169030|sp|P14578|COX1_ORYS      GFLVVAHHMFTVGLDVTTRAYFTAATMIIAVPTGIKIFSWIATMWGGSIQ 335
gi|13449404|ref|NP_085587.1|        GFLVVAHHMFTVGLDVTTRAYFTAATMIIAVPTGIKIFSWIATMWGGSIQ 335
gi|34555994|emb|CAD33909.1|        GFIVVAHHMFTVGLDVTTRAYFTSATMIIAIPITGVKVFSLATLHGNNIK 333
gi|32348033|gb|AAP47921.1|        GFIVVAHHMFTVGMVDVTTRAYFTSATMIIAIPITGVKVFSLATLHGNNMK 333
gi|1352141|sp|Q08855|COX1_RHIL      GFVVVAHHMKTIVGMDLDTTAYFVSATMIIAVPTGIKIFSWIATMWGGSIE 357
gi|2114418|gb|AAB58264.1|          YIWAGPHHLHYTALPDWAQTLGMVFSIMLWMPWSWGGMINGLMTLSGAWDK 359
gi|22960507|ref|ZP_00008147.1|      YIWAGPHHLHYTALPDWAQTLGMVFSIMLWMPWSWGGMINGLMTLSGAWDK 392
gi|3850275|gb|AAC72071.1|          YIWAGPHHLHYTALPDWAQTLGMVFSIMLWMPWSWGGMINGLMTLSGAWDK 317
```

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.

```
/seqprg/slib/bin/lalign -N 5000 -s P250 -f -16 -g -4 -w 75 -q @ @
```

LALIGN finds the best local alignments between two sequences
version 2.1u05 August 2003
Please cite:
X. Huang and W. Miller (1991) Adv. Appl. Math. 12:373-381

Dataset 2

```
alignments < E( 0.05):score: 57 (50 max)
Comparison of:
(A) @      gi|13449404|ref|NP_085587.1| cytochrome c oxidase - 527 aa
(B) @      gi|2114418|gb|AAB58264.1| cbb3-type cytochrome oxi - 576 aa
using matrix file: PAM250, gap penalties: -16/-4 E(limit) 0.05

17.3% identity in 231 aa overlap (198-423:260-487); score: 147 E(10000): 5.3e-07

200      210      220      230      240      250      260      270
gi|134 LLSLPVLAGAITMLLTDRNFNTFFDPAGGGDPILYQHLFWFFGHPEVYILILPGF-GIISHIVSTFGSKPVFG
.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:
gi|211 FIVTIAMLHVNNLAVPASFLGSKSYSVSSGVQDALTDQ---WWYGHNAVGFLLTAGFLGMMYYFVPKQANRPVYS
260      270      280      290      300      310      320      330

280      290      300      310      320      330      340
gi|134 YLGMVYAMISIGYLGFLVVAHHMFTVGLDVOTRAYFTAATMIIAVPTGIKIFSWIATMWGGSIQYKTPMLFAVGF
:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:
gi|211 YRLSIIHFVALIFMYIWAGPHHLHYTALPDWAQTLGMVFSIMLWMPWSWGGMINGLMTLSGAWDKIRTDPIIRMMI
340      350      360      370      380      390      400

350      360      370      380      390      400      410
gi|134 --IFLFTIGGLTGIVLANSGLDIALHDTYYVVAHFHYVLSMGAVFALFAGFYVWVGKIFGRT--YPETLGQIHFW
:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:
gi|211 VAIAFYGMSTFEGPMMSVKTVNSLSHYTEWTIGHVHSGALGVWGMITFGAIYYLTPKLWGRERLYSLRMVNVHFW
410      420      430      440      450      460      470      480

420
gi|134 ITFFGV
..:..
gi|211 LATFGI
```

1.5 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.

Choose: (A) Program, (B) Query (sequence/accession), (C) Database and (D) Start Search:

(A) Program: FASTA: protein:protein Compare your own sequences: Compare sequences

(B) Query sequence: FASTA format Subset range: ☐ Use Subset range

>gi|13449404|ref|NP_085587.1| cytochrome c oxidase subunit 1 [Arabidopsis thaliana]
 MKNLVRWLFSTNHKDIGTLYFIFGAIAGVMGTCFSVLIRMELARPGDQILGNGHQLYNVL
 ITAHLAFILMIFMVPAMIGGFGNWFVPILIGAPDMAFPRLNNISFWLLPPSLLLLSSAL
 VEVCSTGCTWTVYPPLSGITSHSGGAVDLAIFSLHLSGVSSILGSINFITTFNMRGPGMT
 MHRPLFLVWSVLVTAFLLLSLPVLAGAITMLLTDRNFNTTFDPAGGGDPILYQHLFWF

☒ Protein ☐ DNA (both-strands) ☐ DNA (forward only) ☐ DNA (rev-comp only)

(C) Database: **Protein** **DNA**
 PIR1 Annotated (rel. 66) GB163.0 Primate Search Database

☒ Exclude low complexity (seg) Reset

Other search options: **Scoring matrix:** open: ext: Ktup: Statistical estimates
 Blosum50 -10 -2 ktup = 2 Default

Output limits: E0: Best E0:

Interpret the results?

```
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
```

Annotation symbols:
 = : active site
 * : phosphorylation
 # : binding site
 ^ : site
 ! : metal binding

The best scores are:

seq	length target seq	Raw score	Bit score	E value	% id	% sim	alen	(length match)
sp P14578 COX1_ORYSJ Cytochrome c oxidase subunit 1; Cy	(524)	3237	898.8	0.000	0.977	0.985	524	align
sp P08742 COX1_MAIZE Cytochrome c oxidase subunit 1; Cy	(528)	3225	895.5	0.000	0.971	0.985	524	align
sp P08743 COX1_OENBE Cytochrome c oxidase subunit 1; Cy	(527)	3212	891.9	0.000	0.939	0.956	527	align
sp P07506 COX1_SOYBN Cytochrome c oxidase subunit 1; Cy	(527)	3209	891.1	0.000	0.947	0.956	527	align
sp P00401 COX1_YEAST Cytochrome c oxidase subunit 1; Cy	(534)	2276	634.1	4.8e-182	0.642	0.872	531	align
sp P03945 COX1_NEUCR Cytochrome c oxidase subunit 1; Cy	(557)	2186	609.3	1.5e-174	0.691	0.868	524	align
sp P00400 COX1_DROYA Cytochrome c oxidase subunit 1; Cy	(511)	2183	608.5	2.3e-174	0.690	0.881	513	align
sp P00402 COX1_EMENI Cytochrome c oxidase subunit 1; Cy	(567)	2152	599.9	1e-171	0.674	0.859	524	align
sp P31833 COX1_BRAJA Cytochrome c oxidase subunit 1; Cy	(541)	2149	599.1	1.7e-171	0.683	0.873	520	align
sp P00395 COX1_HUMAN Cytochrome c oxidase subunit 1; Cy	(513)	2148	598.9	1.9e-171	0.703	0.888	511	align
sp P00396 COX1_BOVIN Cytochrome c oxidase subunit 1; Cy	(514)	2137	595.8	1.5e-170	0.693	0.890	511	align
sp P00399 COX1_DROME Cytochrome c oxidase subunit 1; Cy	(511)	2126	592.8	1.2e-169	0.690	0.879	513	align
sp P00397 COX1_MOUSE Cytochrome c oxidase subunit 1; Cy	(514)	2109	588.1	3.2e-168	0.691	0.874	517	align
ref NP_008136 cytochrome c oxidase subunit I	(518)	2090	582.9	1.2e-166	0.689	0.883	511	align

```

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:
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:*
110     1      1:*
112     2      1:*
114     0      1:*
116     2      0:==
118     3      0:==
120     1      0:==
122     1      0:==
124     0      0:

```

one = represents 3 library sequences

inset = represents 1 library sequences


```

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:
sp|P14578|COX1_ORYSJ Cytochrome c oxidase subunit 1; Cy ( 524) 3237 898.8      0 0.977 0.985 524
align
sp|P08742|COX1_MAIZE Cytochrome c oxidase subunit 1; Cy ( 528) 3225 895.5      0 0.971 0.985 524
align
sp|P08743|COX1_OENBE Cytochrome c oxidase subunit 1; Cy ( 527) 3212 891.9      0 0.939 0.956 527
align
sp|P07506|COX1_SOYBN Cytochrome c oxidase subunit 1; Cy ( 527) 3209 891.1      0 0.947 0.956 527
align
sp|P00401|COX1_YEAST Cytochrome c o

```

Query: TMP.q
 1>>>gi|13449404|ref|NP_085587.1| cytochrome c oxidase subunit 1 [Arabidopsis thaliana]
 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:*=	Observed	Expected
50	6	5:*=	number of	number of
52	18	15:=====*	target	target
54	28	32:=====*	sequences	sequences
56	64	55:=====	with a given	with a given
58	71	79:=====	match score	match
60	48	99:=====		score
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:=====		

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

5122066 residues in 13144 sequences
 Statistics: Expectation_n fit: rho(ln(x))= 7.2005+/-0.00321; mu= 9.5405+/- 0.174

1.6 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

```
ATGACATCAGCGACTCTGACGCCAGGGGCGGCCCTGGGCAGCCAGCGGGTGTGCGAAAATGTGCGTTACTACGAAGACGC
CGTCCGACTCTTCGTTCATCGCTGCAGTGTTCCTGGGGCGTCGTGCGCTTCCTCGCCGGCGTCTTCATCGCGCTGCAGCTGG
CTTTTCCGGCGCTGAATCTCGGCCCTTGAGTGGACGAGCTTCGGGCGCCTGCGGCCGGTCCACACCTCGGCCGTGATCTTC
GCGTTTGGCGGCAACGTCCTGTTTCGCCACCTCGCTCTACTCCGTGCAGCGCACACAGCCGCCAGTTCCTGTTTCGGCGGCGA
GGGCTTCGCGAAGTTCGTCTTCTGGAACACACATCTTCATCGTCTGGCGGCGCTCAGCTACGTGCTCGGCTACACCC
AGGGCAAGGAGTATGCAGAGCCGGAGTGGATCCTCGACCTCTACCTGACGGTCATCTGGGTCTCTACGCCATCCAGTTC
GTCGGCACGGTGATGACCCGCAAGGAGTCGCACATCTACGTGCGCAACTGGTTCTTCATGGCGTTTCATCCTGACCGTCGC
GATCCTCCACATCGGCAACAACGTCAACGTCCCGGTGTCGCTGACCGGGATGAAGTCCTACCCGTTTCGTCTCGGGCGTGC
AGAGCGCCATGGTGCAGTGGTGGTACGGCCACAACGCGGTGCGCTTCTTCCTGACCGCCGGCTTCCTCGGCATCATGTCT
ACTTCGTTCCGAAGCGCGCGGAGCGGCCGGTCTATTTCGTACCGCCTGTCGATCGTGCACCTTCTGGACGCTGATCTTCCTC
TACATCTGGGCGGCGCCGACACCTGCACTACACGGCCCTGCCGATTGGGCGCAGACGCTGGGCATGACCTTCTCGGT
CATGCTGTGGATGCCGTCTGGGGCGGCATGATCAACGGCATCATGACCTGTGCGGTGCCTGGGACAAGCTGCGCACCG
ACCCGGTCTGCGCTTCTTCGTGACGTGCGGTGGCCTTCTACGGCATGTGACCTTCGAGGGCCCGCTGATGTGCGGTGAAG
CCGGTCAACGCCCTGTGCACTACACCGACTGGACGATCGGCCACGTGCACTCCGGTGCCTCGGCTGGGTGGCCTTCAT
CTCCTTCGGCGCGATCTACTATCTGGTCCCGGTCTGTGGAAGCGCTCGCAGCTCTACAGCCTGCGTCTGGTCAGCTACC
ACTTCTGGACCGCCACCATCGGCATCGTGCTCTACATCACCGCCATGTGGGTGTGCGGCATCATGCAGGGCCTGATGTGG
CGCGCTACGACAACCTCGGCTTCCTCCAGTACTCGTTCGTGAGACGGTCGCGGCCATGCATCCCTTCTACGTGATCCG
TGGCTGGGCGGCGTCTGTTCTGGCTGGTGGCCTGATCATGGTCTACAACCTGTGGCGCACGGCCAAGGGTGACGTCC
GCATCGAGAAGCCCTATGCCTCCGCCCGCACAAAGGCGGCGGTGCGTGCAGCCTGA
```

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?

Apps [google](#) [Applications](#) [https://wet.kuleuven...](#) [KU Leuven Stouterpr...](#) [https://www.facebo...](#) [New Tab](#) [https://www.google...](#) [New Tab](#)

[Download](#) [GenPept](#) [Graphics](#) Sort by: [E value](#) [Next](#) [Previous](#) [Descriptions](#)

Cytochrome c oxidase fixN chain [Magnetospira sp. QH-2]
Sequence ID: [emb|CCQ72120.1](#) Length: 489 Number of Matches: 2

Range 1: 233 to 486 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
430 bits (1105)	0.0	Compositional matrix adjust.	210/254(83%)	232/254(91%)	0/254(0%)	+3
Query 720	YFVPKRAERPVSYSRLSIVHFWTLIFLYINAGPHHLHYTALPDWAQTLGHTFSSVLMWIPS	899				
Sbjct 233	YFVPK+A RPVSYSRLSIVHFW LIFLYINAGPHHLHYTALPDW QTLGHTFSS+MLWIPS	292				
Query 900	WGGINGIMTLSGAWDKLRTDPVLRFLVTSVAFYGNSTFEGPLNSVKPNALSHYTDWTI	1079				
Sbjct 293	WGGINGIMTLSGAW+KLRTDPVLRFLVTS+AFYGNSTFEGP MS+K VII+LSHYTDWTI	352				
Query 1080	GHHVSGALGWAFISFGAIYVLPVWLKRSQLYSLRLVSYHFNATIGIVLYITAMVSG	1259				
Sbjct 353	GHHVSGALGWAF+SFGA+YYL P LW R +LVSRLVSYHFW ATIGIVLYIT+MH+SG	412				
Query 1260	IMQGLMRAVDLGFQYSFVETVAAMHPFYVIRalggviflagalIMVYNLIRAKGDV	1439				
Sbjct 413	IMQGLMRAVD LGFLQYSF+E+V AMHP+VIR LGGLF+ G+LIMVYN +TA+GD+	472				
Query 1440	RIEKPYSAPHKAA 1481					
Sbjct 473	RHEEAFVEAPHATS 486					

Range 2: 7 to 231 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
326 bits (835)	0.0	Compositional matrix adjust.	159/225(71%)	185/225(82%)	3/225(1%)	+1
Query 52	SENVRYVEDAVRLFVIAAVFNGVVGFLAGVFIALQLAPPALNLGLEWTSFGRLRPVHTSA	231				
Sbjct 7	+E V Y E+ V+ FVAA FNGV+GF+AGVFIA QLAPP LNLGLEWTS+FGRLRPVHTSA	66				
Query 232	VIFAFGGNVLFATSLVSORTSRQFLFGGGLAKFVFWYVIFIVLAALSIVLGYTQKE	411				
Sbjct 67	VIFAFGGN++ TS V VQRT RQ LFGGGL FVFW + + IVLAA YVLG TQKE	126				
Query 412	YAEPEWLDLTVINVLVIAIQFVGTVMTRKESHIVANMFFHAFILTVAILHIGNWNV	591				
Sbjct 127	YAEPEW +DL LTV+NV Y + F GT+VTR+ESHIVANMF++AFILTV+LH+GNW V	186				
Query 592	PVSLTG---MKSVPFVSGVQSAWVQWVYGHNAVGGFFLTAGFLGIH 717					
Sbjct 187	PV+L G +SY +GV+ AM QWVYGHNAVGGFFLTAGFLGIH 731					