Practical Computing for Bioinformatics - HW1

Adnan Kaan Ekiz(r0776549) - Olympia Gennadi(r0829391)

October 2020

Contents

1 Question 1

1.1 Part A

To search the IDs, we have checked several databases and we found out that the information is kept in **GenBank** database which is composed by NCBI. GenBank provides several sources of information regarding these sequences which are actually six isoforms of serine/threonine-protein kinase MRCK alpha.

1.2 Part B

Through related information in GenBank, we can detect the corresponding gene and the organism. The official name of the gene is **CDC42 binding protein kinase alpha** and the official symbol is **CDC42BPA**. Furthermore, it exists in **Homo Sapiens**, widely known as Humans.

We have noticed that given codes are corresponding to the protein accession number and we have obtained the gene code by looking at the following isoform proteins:

- NP_001374479 - serine/threonine-protein kinase MRCK alpha isoform F
- NP_003598 - serine/threonine-protein kinase MRCK alpha isoform B
- NP_055641 - serine/threonine-protein kinase MRCK alpha isoform A
- NP_001352939 - serine/threonine-protein kinase MRCK alpha isoform D
- NP_001352940 - serine/threonine-protein kinase MRCK alpha isoform E
- NP_001352948 - serine/threonine-protein kinase MRCK alpha isoform C

1.3 Part C

NCBI gives us the opportunity to identify homologuous genes of a specific gene in different organisms by choosing HomoloGene Database. Homologuous genes can also be detected by using other databases such as FlyBase. By utilizing these ways, we concluded that the homologuous gene in C.elegans (Caenorhabditis elegans) is called **mrck-1** which is described as Myotonic dystrophy-Related, Cdc42-binding Kinase homolog;Non-specific serine/threonine protein kinase;Protein kinase C;Serine/threonine-protein kinase mrck-1.

It has been shown that mrck-1 gene is related to the activation of myosin 2 that is highly related to the aktin filaments in the cell. In the paper "Myosin II regulation during C. elegans embryonic elongation: LET-502/ROCK, MRCK-1 and PAK-1, three kinases with different roles" it is also mentioned that depletion of let-502 with either pak-1 or mrck-1 leads to arrest at the 1.2-fold stage, which we consider as an absence of elongation suggesting that MRCK is a kinase responsible for myosin II activation. Without this gene there won't or there would be limited amount of cell migration in addition to the some problems occurring in cytokinesis of the cell which would cause problems in terms of morphogenetic processes.

1.4 Part D

First possible solution

In the last part of the exercise we retrieved the GO terms assigned to the gene using **UniProt** Database.

Through GenBank (from UniProt source) we observed that there are specific GO terms assigned by all sequences and we found their corresponding function in UniProt database. We verified the results in Gene Ontology Resource where we searched for the corresponding gene of each GO term. The IDs and functions of three GO terms shared by all sequences are listed below :

- GO:0005524 **ATP binding** : molecular function
- GO:0006468 **Protein phosphorylation** : biological function
- GO:0042802 **Identical protein binding** : molecular function

We used the following website to retrieve the GO annotations :

- **Uniprot :** https://www.ebi.ac.uk/QuickGO/annotations?geneProductId=Q5VT2

Second possible solution

When protein sequences of mrck-1 (C.elegans) and CDC42BPA (Homo Sapiens) are considered, we found the shared GOs in three different categories as follows:

- GO (Molecular Function) : nucleotide binding, protein kinase activity, protein serine/threonine kinase activity, ATP binding, kinase activity, transferase activity, metal ion binding
- GO (Biological Process) : protein phosphorylation, phosphorylation, peptidyl-threonine phosphorylation, actomyosin structure organization, intracellular signal transduction
- GO (Cellular Component) : cytoplasm, cytoskeleton

After determining the genes, we used the following website to check for the GO of the genes in 2 different organisms (homo sapiens, c.elegans).

- **For C. Elegans:** https://www.ensembl.org/Caenorhabditis_elegans/Gene/Ontologies/cellular_component ?db=core;g=WBGene00006437;r=V:6257392-6267287
- **For Homo Sapiens:** https://www.ensembl.org/Homo_sapiens/Gene/Ontologies/cellular_component?db=core;g=ENSG00000143776;r=1:226989865-227318474

2 Question 2

2.1 Part A

When the sequence file is analyzed with **Nucleotide BLAST**, we can see that 4 different genes from C. Elegans matched 100% with our file (**unknown_sequence.fsa**). After further analysis, it has been decided that only one of the genes was producing a protein that is required to answer the question. We want the match that is the gene and not just a part of the full genome that contains the sequence. Details of the gene that we have found are given below:

- Name of the gene is **clk-1 5-demethoxyubiquinone hydroxylase, mitochondrial**
- Name of the protein produced by the gene is **5-demethoxyubiquinone hydroxylase, mitochondrial**
- This gene sequence belongs to **C. Elegans** organism

2.2 Part B

HomoloGene database in NCBI helped us to find the homologuous genes of **clk-1** in zebrafish and human:

- Zebrafish (D.rario) : **coq7 coenzyme Q7 homolog, ubiquinone**
- Human (Homo Sapiens) : **COQ7 coenzyme Q7, hydroxylase**

We retrieved the protein sequences produced by those genes from **UniProt** database. Combining the three protein sequences of the genes we created a multiple sequence alignment. The alignment is illustrated below :

```
CLUSTAL

sp|Q99807|COQ7_HUMAN/1-217   MSCAGAAAAPRLWRLRPGARRSLSAYGRRTSVRFRSSGMTLDNISRAAVDRIIRVDHAGE
tr|F1QW05|F1QW05_DANRE/1-223 MQTAGKCAVRLDLSRVFCPSSVWNCRAVNRVNGLVSCRRYSVIPPPRDEQEKAMLDRMLR
sp|P48376|COQ7_CAEEL/1-187   MFRVITRGAHTAASRQALIEKIIRVDHAGELGADRIYAGQLAVLQGSSVGSVIKKMWDEE

sp|Q99807|COQ7_HUMAN/1-217   YGANRIYAGQMAVLGRTSVGPVIQKMWDQEKDHLKKFNELMVTFRVRPTVLMPLWNVLGF
tr|F1QW05|F1QW05_DANRE/1-223 VDHAGEYGANRIYAGQMAVLGRTQTGPLIQHMWDQEKIHLEKFNEILGEHRVRPTLLLPL
sp|P48376|COQ7_CAEEL/1-187   KEHLDTMERLAAKHNVPHTVFSPVFSVAAYALGVGSALLGKEGAMACTIAVEELIGQHYN

sp|Q99807|COQ7_HUMAN/1-217   ALGAGTALLGKEGAMACTVAVEESIAHHYNNQIRTLMEEDPEKYEELLQLIKKFRDEELE
tr|F1QW05|F1QW05_DANRE/1-223 WNIAGFALGACTALLGKEGAMACTVAVEESISEHYNSQIRTLMEADPDRYTELLQLIKEF
sp|P48376|COQ7_CAEEL/1-187   DQLKELLADDPETHKELLKILTRLRDEELHHHDTGVEHDGMKAPAYSALKWIIQTGCKGA

sp|Q99807|COQ7_HUMAN/1-217   HHDIGLDHDAELAPAYAVLKSIIQAGCRVAIYLSERL------
tr|F1QW05|F1QW05_DANRE/1-223 RDDEIEHHDTGLEHDAESVPGYMLLKTAIQAGCTAAIYISQRI
sp|P48376|COQ7_CAEEL/1-187   IAIAEKI-----------------------------------
```

Figure 1: Multiple sequence alignment

To determine the degree of homology between the protein sequences we used **JalView** program. After inserting the fasta file in JalView we calculated pairwise alignments to receive the percent identity between the sequences for each of the three pairs. The results are listed below :

Homology between Homo Sapiens and C.elegans

The percentage identity for Human and C.elegans sequences is 52.94%.

```
---------------------
Score = 5210.0
Length of alignment = 187
Sequence  sp|P48376|COQ7_CAEEL/1-187 (Sequence length = 187)
Sequence sp|Q99807|COQ7_HUMAN/31-217 (Sequence length = 217)

 sp|P48376|COQ7_CAEEL/1-187 MFRVITRGAHTAASRQALIEKIIRVDHAGELGADRIYAGQLAVL
                              |  . |       .| ...|||||||| ||.||||||.|||
sp|Q99807|COQ7_HUMAN/31-217 SVRFRSSGMTLDNISRAAVDRIIRVDHAGEYGANRIYAGQMAVL

 sp|P48376|COQ7_CAEEL/1-187 QGSSVGSVIKKMWDEEKEHLDTMERLAAKHNVPHTVFSPVFSVA
                              .|||.||.||||.||.||   . |     |  ||. |. .|
sp|Q99807|COQ7_HUMAN/31-217 GRTSVGPVIQKMWDQEKDHLKKFNELMVTFRVRPTVLMPLWNVL

 sp|P48376|COQ7_CAEEL/1-187 AYALGVGSALLGKEGAMACTIAVEELIGQHYNDQLKELLADDPE
                             ..||| |.||||||||||||.|||| |..|||.|.. |. .|||
sp|Q99807|COQ7_HUMAN/31-217 GFALGAGTALLGKEGAMACTVAVEESIAHHYNNQIRTLMEEDPE

 sp|P48376|COQ7_CAEEL/1-187 THKELLKILTRLRDEELHHHDTGVEHDGMKAPAYSALKWIIQTG
                               |||... ..|||||.||| |..||.  ||||. || |||.|
sp|Q99807|COQ7_HUMAN/31-217 KYEELLQLIKKFRDEELEHHDIGLDHDAELAPAYAVLKSIIQAG

 sp|P48376|COQ7_CAEEL/1-187 CKGAIAIAEKI
                             |. || ..|..
sp|Q99807|COQ7_HUMAN/31-217 CRVAIYLSERL

Percentage ID = 52.94
```

Figure 2: Pairwise alignment 1

Homology between Homo Sapiens and D. rerio

The percentage identity for Human and Zebrafish sequences is 62.21%.

```
Score = 6850.0
Length of alignment = 217
Sequence tr|F1QW05|F1QW05_DANRE/11-223 (Sequence length = 223)
Sequence    sp|Q99807|COQ7_HUMAN/1-217 (Sequence length = 217)

tr|F1QW05|F1QW05_DANRE/11-223 LDLSRVFCPSSVWNCRAVNRVNGLVSCRRYSV----IPPPRD
                              .  .    .. .|  |.  | .     || ||         |
   sp|Q99807|COQ7_HUMAN/1-217 MSCAGAAAAPRLWRLRPGARRSLSAYGRRTSVRFRSSGMTLD

tr|F1QW05|F1QW05_DANRE/11-223 EQEKAMLDRMLRVDHAGEYGANRIYAGQMAVLGRTQTGPLIQ
                              .  .| .||..||||||||||||||||||||||  ||.||
   sp|Q99807|COQ7_HUMAN/1-217 NISRAAVDRIIRVDHAGEYGANRIYAGQMAVLGRTSVGPVIQ

tr|F1QW05|F1QW05_DANRE/11-223 HMWDQEKIHLEKFNEILGEHRVRPTLLLPLWNIAGFALGACT
                               |||||| || ||||..   |||||.|.||||. |||||| |
   sp|Q99807|COQ7_HUMAN/1-217 KMWDQEKDHLKKFNELMVTFRVRPTVLMPLWNVLGFALGAGT

tr|F1QW05|F1QW05_DANRE/11-223 ALLGKEGAMACTVAVEESISEHYNSQIRTLMEADPDRYTELL
                              ||||||||||||||||||..|||.||||||| ||..| |||
   sp|Q99807|COQ7_HUMAN/1-217 ALLGKEGAMACTVAVEESIAHHYNNQIRTLMEEDPEKYEELL

tr|F1QW05|F1QW05_DANRE/11-223 QLIKEFRDDEIEHHDTGLEHDAESVPGYMLLKTAIQAGCTAA
                              |||| |||.|.|||| ||.||||  |.| .||. |||||  |
   sp|Q99807|COQ7_HUMAN/1-217 QLIKKFRDEELEHHDIGLDHDAELAPAYAVLKSIIQAGCRVA

tr|F1QW05|F1QW05_DANRE/11-223 IYISQRI
                              ||.|.|.
   sp|Q99807|COQ7_HUMAN/1-217 IYLSERL

Percentage ID = 62.21
```

Figure 3: Pairwise alignment 2

Homology between D.rerio and C.elegans

The percentage identity for Zebrafish and C.elegans sequences is 47.06%.

```
---------------------
Score = 4740.0
Length of alignment = 187
Sequence    sp|P48376|COQ7_CAEEL/1-187 (Sequence length = 187)
Sequence tr|F1QW05|F1QW05_DANRE/39-223 (Sequence length = 223)

   sp|P48376|COQ7_CAEEL/1-187 MFRVITRGAHTAASRQALIEKIIRVDHAGELGADRIYAGQLA
                               . ||   .     .|......||||||| ||.||||||.|
tr|F1QW05|F1QW05_DANRE/39-223 RYSVIP--PPRDEQEKAMLDRMLRVDHAGEYGANRIYAGQMA

   sp|P48376|COQ7_CAEEL/1-187 VLQGSSVGSVIKKMWDEEKEHLDTMERLAAKHNVPHITVFSPV
                               ||  .  |..|. |||.|| ||.  . . . | |  |.. |.
tr|F1QW05|F1QW05_DANRE/39-223 VLGRTQTGPLIQHMWDQEKIHLEKFNEILGEHRVRPTLLLPL

   sp|P48376|COQ7_CAEEL/1-187 FSVAAYALGVGSALLGKEGAMACTIAVEELIGQHYNDQLKEL
                               ..|..|||  .||||||||||||.|||| |..||| |.. |
tr|F1QW05|F1QW05_DANRE/39-223 WNIAGFALGACTALLGKEGAMACTVAVEESISEHYNSQIRTL

   sp|P48376|COQ7_CAEEL/1-187 LADDPETHKELLKILTRLRDEELHHHDTGVEHDGMKAPAYSA
                               .  ||.   |||...  .||.|..|||||.|||.   |.|
tr|F1QW05|F1QW05_DANRE/39-223 MEADPDRYTELLQLIKEFRDDEIEHHDTGLEHDAESVPGYML

   sp|P48376|COQ7_CAEEL/1-187 LKWIIQTGCKGAIAIAEKI
                               ||  ||.|| .|| |...|
tr|F1QW05|F1QW05_DANRE/39-223 LKTAIQAGCTAAIYISQRI

Percentage ID = 47.06
```

Figure 4: Pairwise alignment 3

Due to the high identity percentages and the low number of gaps between the sequences we can assume that all pairs are likely to share functional similarity. As we can see, the two pairs in which Human is involved present higher percentage sequence identity than the third pair (C.elegans and Zebrafish). For that reason, those two pairs should present higher similarity in sequences than the last pair.

Afterwards, we analyzed the multiple sequence alignment and isolated a part of our graphical results as it can be shown in the following picture.

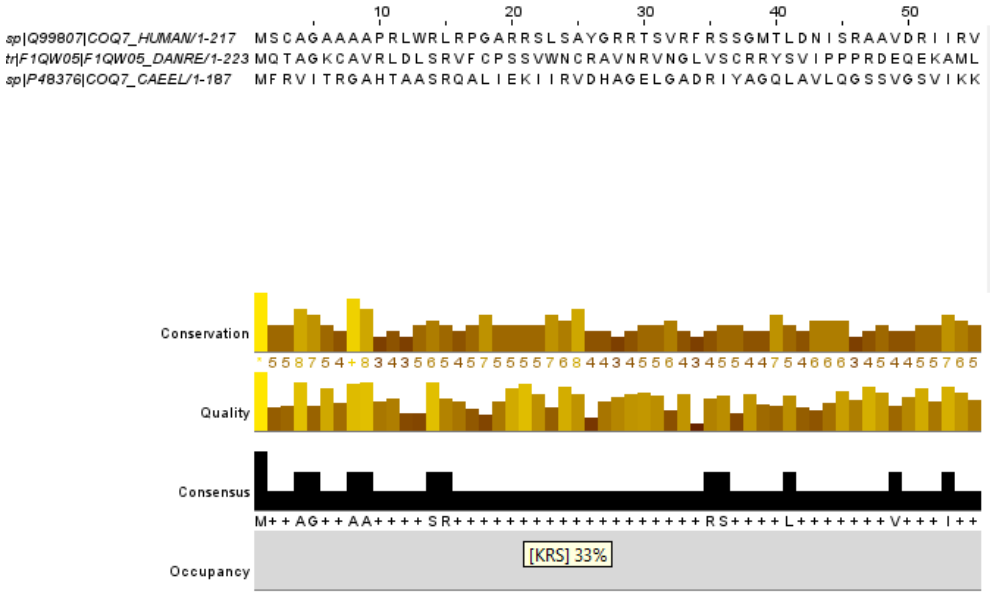

Figure 5: Graphical Evaluation of Multiple Sequence Alignment

- Conservation : Alignment conservation annotation is a quantitative numerical index reflecting the conservation of the physico-chemical properties for each column of the alignment. The histogram of conservation indicates that generally there is a moderate group conservation in each column. In some columns we can recognize a strong conservation of physico-chemical properties -the highest bars- between the amino acids in the multiple sequence alignment but the majority of them present mild conservation. Therefore the physico-chemical properties between protein sequences do not show high maintenance.

- Quality : Alignment quality annotation is an ad-hoc measure of the likelihood of observing the mutations in a particular column of the alignment. In general, the average quality score of the alignment is closer to 1 than to 0 meaning that there are not a lot of mutations in most of the columns or most mutations are observed to be favourable. In addition, the majority of histogram bars are sufficiently high indicating that we have a high quality annotation without a great number of mutations.

- Consensus : Alignment consensus annotation reflects the percentage of the different residue per column. As we can see, most columns include three different amino acids with a 33.3% ($\frac{1}{3}$) percentage each one. In some columns we can see that the same amino acid appears twice and is presented as the prevalent amino acid with a percentage 67% ($\frac{2}{3}$), while in other columns we have 100% match between the three amino acids (for example in the first column). Moreover, if we hide the gaps, we will observe that the consensus percentage between amino acids increases, mostly at the last part of the multiple sequence alignment.

Phylogenetic Tree

With the method BLOSUM62 (in JalView) we designed the phylogenetic tree of the organisms based on the given sequences.

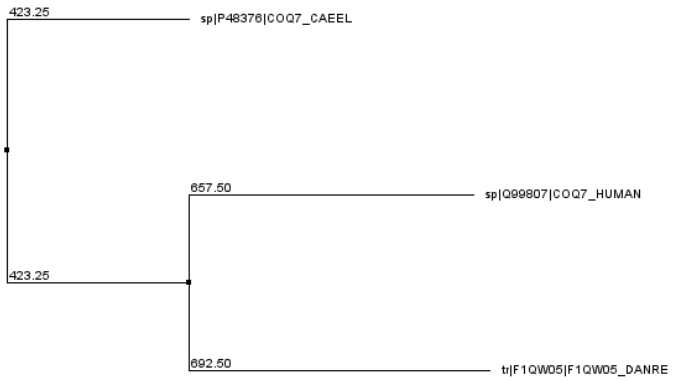

Figure 6: Phylogenetic Tree : Result of the method BLOSUM62

The above tree indicates the relationship between the three species. As we can see, human and zebrafish are the most closely related species. Therefore, it is more possible to share the same ancestor. On the other hand, C.elegans is placed on a different cluster and the possibility to share a recent common ancestor with the other two organisms is not quite high.

3 Question 3

3.1 Part A

To solve the problem and retrieve the job IDs, we have generated a script called "**script1.sh**" that is submitted alongside the report. To be precise, we have decided to explain each line of the script as follows:

1. **> error_jobID.txt** : This part is used to generate an empty text file "error_jobID.txt". It is used to make sure that the text file is empty when starting the script.
2. **for FILE in 'find . -name 'stderr*'; do** : For every file in the directory that starts with the string "stderr"
3. **echo $FILE;** : Print the name of the file
4. **if (grep -q 'Error' $FILE) || (grep -q 'error' $FILE); then** : If file contains the string "error" or "Error"
5. **echo 'Found!'** : Then print a message to show that the string is present in the file
6. **echo $FILE | cut -c10-17 >> error_jobID.txt** : Append the job ID to our newly created text file
7. **ct=($(wc -l error_jobID.txt))** : Count the number of lines to determine the total number of job IDs in the text file
8. **lineN=${ct[0]}** : Only retrieve the total line number
9. **echo "Words(Error, error) found in $lineN files!"** : Print the total occurrences of the strings "error" and "Error" for the "stderr" files

– Full script is listed below –

```
> error_jobID.txt

for FILE in 'find . -name 'stderr*'';
do

echo $FILE;

if ( grep -q 'Error' $FILE ) || ( grep -q 'error' $FILE )
then
echo 'Found!'
echo $FILE | cut -c10-17 >> error_jobID.txt
fi

done;

ct=($(wc -l error_jobID.txt))
lineN=${ct[0]}

echo "Words(Error, error) found in $lineN files!"
```

– End of the script –

3.2 Part B

Based on the output of the script, **86** job IDs have been found.

4 Question 4

We have generated another script called "**script2.sh**" to determine the highest CPU time in two different conditions as mentioned in the assignment pdf. To show the different commands and the explanations clearly, we will again present every line of the code with its explanation:

1. **longest="00:00:00"; longestjname="";** : First we set the longest job name as empty string and longest job time as "00:00:00" to further compare these variables with the variables that we will obtain in the "stdout" files
2. **for FILE in 'find . -name 'stdout*'; do** : For every file in the directory that starts with the string "stdout"
3. **if [$ -eq 0] ; then** : If no parameter is given when running the script
4. **var=' grep 'cput' $FILE '** : Receive the line in the file that includes the string "cput"
5. **btime=$(echo $var | cut -c22-29)** : Obtain the CPU time by cutting the line obtained by the earlier command
6. **var2=' grep 'Job Name' $FILE '** : Receive the line in the file that includes the string "Job Name"
7. **bjname=$(echo $var2 | cut -c11-)** : Obtain the job name by cutting the line obtained by the earlier command
8. **if [["$btime" > "$longest"]]; then** : If obtained time from the current file is more than our current longest time
9. **longest=$btime; longestjname=$bjname;** : Then change the longest time and job name to our current time and job name
10. **if ["$bjname" = "$1"]; then** : If current job name is the same as parameter that is given
11. **echo "Maximum CPU time of $longest for job $longestjname "** : Printing the results if no parameter is given
12. **echo "CPU time of $longest for job $longestjname "** : Printing the results if a parameter is given

– Full script is listed below –

```
longest="00:00:00"
longestjname=""
for FILE in `find . -name 'stdout*'`; do

        if [ $# -eq 0 ] ; then

                var=` grep 'cput' $FILE `
                btime=$(echo $var | cut -c22-29)
                echo $btime

                var2=` grep 'Job Name' $FILE `
                bjname=$(echo $var2 | cut -c11-)
                echo $bjname

                if [[ "$btime" > "$longest" ]]; then

                        longest=$btime
                        longestjname=$bjname
                fi

        else
                var=` grep 'cput' $FILE `
                btime=$(echo $var | cut -c22-29)
                echo $btime

                var2=` grep 'Job Name' $FILE `
                bjname=$(echo $var2 | cut -c11-)
                echo $bjname

                if [ "$bjname" = "$1" ]; then

                        if [[ "$btime" > "$longest" ]]; then

                                longest="$btime"
                                longestjname="$bjname"
                        fi
                fi
        fi
done

if [ "$#" -eq "0" ]; then
echo "Maximum CPU time of $longest for job $longestjname "

else
echo "CPU time of $longest for job $longestjname "
fi
```

– End of the script –