Assignment 2 : Computational Biology CSEP527

Krishan Subudhi (Student Number: 2040900) ([ksubudhi@uw.edu](mailto:ksubudhi@uw.edu))

1. strings x1="deadly" and x2="ddgearlyk"
   1. Run your code to align and score the pair of strings *x1*="deadly" and *x2*="ddgearlyk".
   2. print the optimal score,
   3. the alignment produced by your trace-back,
   4. and the full alignment matrix. Also calculate the empirical p-value. For consistency, please do exactly 999 permutations of *x2* for this test.

deadly vs ddgearlyk

[[ 0 0 0 0 0 0 0 0 0 0]

[ 0 6 6 2 2 0 0 0 0 0]

[ 0 2 8 4 7 3 0 0 0 1]

[ 0 0 4 8 4 11 7 3 0 0]

[ 0 6 6 4 10 7 9 5 1 0]

[ 0 2 2 2 6 9 5 13 9 5]

[ 0 0 0 0 2 5 7 9 20 16]]

1 D-EADLY

D EA LY

2 DGEARLY

Score = 20

Emperical P-value (999 permutations )= 3.800E-02

# Protein Alignment score matrix

**10 by 10 upper triangular matrix showing the score you found for the alignment of protein *i* to protein *j* (10 choose 2 non-identical pairs)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Proteins** | **P15172** | **P17542** | **P10085** | **P16075** | **P13904** | **Q90477** | **Q8IU24** | **P22816** | **Q10574** | **O95363** |
| P15172 | 0 | 143 | 1500 | 1020 | 978 | 893 | 428 | 368 | 118 | 56 |
| P17542 | 0 | 0 | 128 | 129 | 128 | 112 | 144 | 123 | 156 | 66 |
| P10085 | 0 | 0 | 0 | 1043 | 1002 | 925 | 440 | 367 | 118 | 52 |
| P16075 | 0 | 0 | 0 | 0 | 1147 | 1093 | 448 | 414 | 120 | 61 |
| P13904 | 0 | 0 | 0 | 0 | 0 | 1104 | 450 | 410 | 120 | 72 |
| Q90477 | 0 | 0 | 0 | 0 | 0 | 0 | 449 | 410 | 117 | 62 |
| Q8IU24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 446 | 125 | 45 |
| P22816 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 124 | 74 |
| Q10574 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 67 |
| O95363 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

# **The two p-values from III(c).**

Say how many random trials you did for each

P15172.fasta vs Q10574.fasta

Emperical P-value (999 permutations )= 1.000E-03

Emperical P-value (5000 permutations )= 2.000E-04

P15172.fasta vs O95363.fasta

Emperical P-value (999 permutations )= 5.470E-01

The low p-values suggest that the scores are statistically significant and the sequences are not aligned by chance, but rather have a common ancestor.

# Notes from step I(b)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Species** | **Name** | **Description** | **Accession** |
| **1** | [Homo sapiens (Human)](https://courses.cs.washington.edu/courses/csep527/20au/hw/hw2pics/sapiens.jpg) | MYOD1\_HUMAN | Myoblast determination protein 1 | P15172 |
| **2** | [Homo sapiens (Human)](https://courses.cs.washington.edu/courses/csep527/20au/hw/hw2pics/sapiens.jpg) | TAL1\_HUMAN | T-cell acute lymphocytic leukemia protein 1 (TAL-1) | P17542 |
| **3** | [Mus musculus (Mouse)](https://courses.cs.washington.edu/courses/csep527/20au/hw/hw2pics/Mouse.jpg) | MYOD1\_MOUSE | Myoblast determination protein 1 | P10085 |
| **4** | [Gallus gallus (Chicken)](https://courses.cs.washington.edu/courses/csep527/20au/hw/hw2pics/Chicken.jpg) | MYOD1\_CHICK | Myoblast determination protein 1 homolog (MYOD1 homolog) | P16075 |
| **5** | [Xenopus laevis (African clawed frog)](https://courses.cs.washington.edu/courses/csep527/20au/hw/hw2pics/Xenopus.jpg) | MYODA\_XENLA | Myoblast determination protein 1 homolog A (Myogenic factor 1) | P13904 |
| **6** | [Danio rerio (Zebrafish)](https://courses.cs.washington.edu/courses/csep527/20au/hw/hw2pics/Zebrafish.jpg) | MYOD1\_DANRE | Myoblast determination protein 1 homolog (Myogenic factor 1) | Q90477 |
| **7** | [Branchiostoma belcheri (Amphioxus)](https://courses.cs.washington.edu/courses/csep527/20au/hw/hw2pics/Amphioxus.jpg) | Q8IU24\_BRABE | MyoD-related | Q8IU24 |
| **8** | [Drosophila melanogaster (Fruit fly)](https://courses.cs.washington.edu/courses/csep527/20au/hw/hw2pics/Drosophila.jpg) | MYOD\_DROME | Myogenic-determination protein (Protein nautilus) (dMyd) | P22816 |
| **9** | [Caenorhabditis elegans](https://courses.cs.washington.edu/courses/csep527/20au/hw/hw2pics/Caenorhabditis.jpg) | LIN32\_CAEEL | Protein lin-32 (Abnormal cell lineage protein 32) | Q10574 |
| **10** | [Homo sapiens (Human)](https://courses.cs.washington.edu/courses/csep527/20au/hw/hw2pics/sapiens.jpg) | SYFM\_HUMAN | Phenylalanyl-tRNA synthetase, mitochondrial | O95363 |

1. MYOD1\_HUMAN (P15172)

Acts as a transcriptional activator that promotes transcription of muscle-specific target genes and plays a major role in regulating muscle differentiation.

2. TAL1\_HUMAN (P17542)

hemopoietic : Pertaining to or related to the formation of blood cells.

Implicated in the genesis of hemopoietic malignancies. It may play an important role in hemopoietic differentiation.

blood cells differentiate from stem cells constantly. Stem cells are kind of base class for all blood cells.

3. MYOD1\_MOUSE (P10085)

Acts as a transcriptional activator that promotes transcription of muscle-specific target genes and plays a role in muscle differentiation

-- Looks very similar to Myod1\_human (no 1)

4. MYOD1\_CHICK (P16075)

Acts as a transcriptional activator that promotes transcription of muscle-specific target genes and plays a role in muscle differentiation.

similar to 1

5. MYODA\_XENLA (P13904)

\*\*May\*\* act as a transcriptional activator that promotes transcription of muscle-specific target genes and plays a role in muscle differentiation.

Might be similar to other myod1 protein but looks slightly different.

6. MYOD1\_DANRE (Q90477)

May act as a transcriptional activator that promotes transcription of muscle-specific target genes and plays a role in muscle differentiation

Similar to 1

7. Q8IU24\_BRABE (Q8IU24)

Organism: Amphioxus (Amphioxi are small marine animals found widely in the coastal waters of the warmer parts of the world)

This is a MyoD-related protein.

Not much information could be found. Documentation does not specifically say that it does muscle differentiation.

8. MYOD\_DROME (P22816)

May play an important role in the early development of muscle.

Does not say that it causes muscle differentiation. So might be slightly different than other MYOD proteins above which facilitate in muscle differentiation.

9. LIN32\_CAEEL (Q10574)

Organism: C Elegans

Essential for the specification of the neuroblast cell fate in the development of peripheral sense organs.

neuroblast :a neuroblast or primitive nerve cell is a postmitotic cell that does not divide further, and which will develop into a neuron after a migration phase.

looks like a gene involved in cell differentiation but not muscle cells. This should not have much similarity with the MYOD proteins.

10. SYFM\_HUMAN (O95363)

Is responsible for the charging of tRNA(Phe) with phenylalanine in mitochondrial translation. The gene for this protein is present in chromosome 6 of human DNA

tRNA - any of a class of small, cloverleaf forms of RNA that transfer unattached amino acids in the cell cytoplasm to the ribosomes for protein synthesis.

Mitochondrial translation is responsible for the maintenance of the cellular energetic balance through synthesis of proteins involved in oxidative phosphorylation.

Mitochondrial translation is specifically defined as the process within mitochondria whereby mitochondrial mRNA (mt-mRNA) is translated by mitochondrial ribosomes (mitoribosomes) to generate an amino acid polypeptide.

This is completely separate from other proteins as it helps in attaching amino acid to t-RNA

## **write** whether you expect each protein to be similar to the others or not.

Based on the study, the MYOD related proteins should be very similar to each other. TAL1\_HUMAN(P17542) is related to blood cells and LIN32\_CAEEL (Q10574) is related to nerve cells. Hence they should have less similarities with MYOD proteins. SYFM\_HUMAN (O95363) is a completely different protein which does not do cell differentiation and should have least similarity with other proteins.

## whether your assessments of the similarities of these proteins changed as a result of seeing these alignments and their scores.

A picture containing calendar

Description automatically generated

Based on the score table, all the MYOD proteins have high alignment scores with each other except Q8IU24\_BRABE (Q8IU24) and MYOD\_DROME (P22816). Even though they were MYOD related, the documents did not mention it doing muscle differentiation, hence solidifying my doubt.

TAL1\_HUMAN (P17542) has low alignment score as expected since it’s related to blood cells while others are not. Similarly LIN32\_CAEEL (Q10574) is also related to nerve cells hence has low alignment scores. Since these two proteins help in cell differentiation, they still have a small positive alignment scores. The empirical p-value between P15172 and Q10574 suggests that the alignment is not by chance and the score is statistically significant.

Finally, SYFM\_HUMAN (O95363) is having the least alignment score when aligned with other proteins since this in not at all related to cell differentiation and does an entirely different job.