Exercise sheet 6: BLAST

Exercise 1

You are given accession number NM_000667.3. Use the BLAST web server to find out about the gene that belongs to this accession number (choose `nucleotide blast`, and the database `reference RNA sequences (refseq_rna)`).

1a)

Which gene is it, and in which organism?

Hide

Solution Gene: Alcohol Dehydrogenase 1A

Organism: *Homo sapiens* (human)

1b)

Which other organisms does it seem to be highly conserved in?

Hide

Solution

- *Gorilla gorilla*: gorilla
- *Pan troglodytes*: common chimpanzee
- *Pan paniscus*: bonobo
- *Nomascus leucogenys*: northern white-cheeked gibbon
- *Cebus capucinus*: white-headed capuchin

Many more…

Exercise 2

You are given a nucleotide query sequence $q = \texttt{ATAC}$, and a nucleotide database sequence $s = \texttt{ATAAAACGGGGGG}$. The word-size $k = 2$. Use a simple scoring scheme that assigns a score of 2 for a match and a score of -1 for a mismatch.

2a)

Generate all k-length words of the query sequence.

Hide

Solution

- $w_1 = AT$
- $w_2 = TA$
- $w_3 = AC$

2b)

List all possible words for the first k-length word (AT) that have a score of at least $T_1 = 1$.

Hide

Solution

- $s(AA) = 1$
- $s(AC) = 1$
- $s(AG) = 1$
- $s(AT) = 4$
- $s(CT) = 1$
- $s(GT) = 1$
- $s(TT) = 1$

2c)

Scan the database for exact matches for the words from the question 2B.

Hide

Solution AA at position 2,3,4. AC at position 5, AT at position 0.

2d)

Extend the exact matches that you found in the question 2C to the left/right and report all MSPs with a score greater than 4.

Hide

Solution AA:

```
Pos: 2              ATA
                    |||
                    AAA     with score 3
```

```
Pos: 3              ATAC
                    ||||
                    AAAC    with score 5
```

```
Pos: 4              AT
                    ||
                    AA    with score 1
```

AT:

```
Pos: 0              ATA
                    |||
                    ATA     with score 6
```

AC:

```
Pos: 5              AT
                    ||
                    AC     with score 1
```

MSPs start in the template at index 0 and 3.

2e)

What happens if we vary the parameters k and T_1?

Hide

Solution

- Higher T_1, k: - faster (less seeds), - less sensitive (some hits will be missed)
- Lower T_1, k: - slower (more seeds), - more sensitive (less hits will be missed)

Exercise 3 - Programming assignment

For the programming tasks, please follow the instructions given in GitHub Classroom under the following link.

https://classroom.github.com/a/nxAqfoYx