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 acccession number (choose nucleotide blast, and the database RefSeq RNA).

Question 1A Which gene is it, and in which organism?

Solution Gene: Alcohol Dehydrogenase 1A

Organism: *Homo sapiens* (human)

Question 1B Which other organisms does it seem to be highly conserved in?

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 the sequences $A =$ `TAKTQHTSLPL` and $B =$ `TGTCTTCTGGGTTCAGCAAA` which stem from the same patient and are supposedly from the same gene.

Question 2A What kind of sequences are these (DNA, RNA, Protein)?

Solution `A` is a protein, `B` is a DNA sequence.

Question 2B If you BLAST these sequences, for which sequence do you expect a lower *E-value*?

Solution Even though sequence `A` is shorter, as a sequence of amino acids it holds more information and should be more unique than sequence `B`. Hence the E-value for `A` should be smaller.

Question 2C BLAST both these sequences. Which gene do the sequences come from? Which BLAST result gave you a lower *E-value* and why? (use db: nucleotide collection and non-redundant protein sequences)

Solution For sequence `A` one needs to use blastp for sequence `B` one needs to use blastn. Result is Forkhead box protein FBXW10. Sequence `A` gives an E-value of 0.27. `B` gives $E = 0.76$ (these numbers can change due to changes in the database, the important message is that `A` $<$ `B` in this case). The reason for this is stated in the answer to the last question.

Exercise 3

You are given a nucleotide query sequence $q =$ `ATAC`, and a nucleotide database sequence $s =$ `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.

Question 3A Generate all k-length words of the query sequence.

Solution

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

Question 3B List all possible words for the first k-length word (AT) that have a score of at least $T_1 = 1$.

Solution

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

Question 3C Scan the database for exact matches for the words from the question 3B.

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

Question 3D Extend the exact matches that you found in the question 3C to the left/right and report all MSPs with a score greater than 4.

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
                    |||
                    AAA     with score 6
```

AC:

```
Pos: 5              A
                    |||
                    AAA     with score 6
```

MSPs start in the template at index 0 and 3.

Question 3E What happens if we vary the parameters k and T_1$?

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)
