THE UNIVERSITY *of* EDINBURGH

Bioinformatics 1

Assignment 1

Henrique Manuel Martins Ferrolho
s1683857 - henrique.ferrolho@gmail.com

October 25, 2016

Contents

1 Question

1.1 What is the name of the disease you have selected?

I have selected Huntington's disease, also known as Huntington's chorea.

1.2 Explain why it is thought there is a genetic basis for this disease.

TODO

1.3 What is the human name for the gene that is thought to be involved?

The official symbol is *HTT*. It's official full name is *huntingtin*. Furthermore, it's gene ID in the NCBI Gene Database is *3064*.

1.4 Is this gene known by any other names?

Yes: *HD*, and *IT15*.

1.5 Whether yes or no, explain how you investigated this.

TODO

1.6 Is this gene present in a model organism such as the mouse or the fruit fly?

TODO

2 Question

2.1 Now investigate the structure of the gene. Find the gene in a database. Which chromosome is it located on, and at which position along the chromosome?

TODO

2.2 What is the structure of the gene, how many exons and introns does it have?

TODO

2.3 Download the gene transcript and the coding sequences for your gene - if multiple are listed, choose the main one, at the top of the list. Where did you get this sequence from and what was the unique identifier used so that someone else could be sure they were looking at the same sequence?

TODO

2.4 How long is the transcript, and what proportion is coding?

TODO

3 Question

3.1 Translate your cDNA sequence into protein/amino acid sequence. How many amino acids does your protein contain?

TODO

3.2 Of the 64 possible codons available, how many are used?

TODO

3.3 What is the most common amino acid in the protein?

TODO

3.4 How many codons for this amino acid exist and how often is each used?

TODO

4 Question

Now look at the following database: http://www.kazusa.or.jp/codon/ The codon usage database lists the frequency which each codon is used in a species (different species prefer different codons). Sequences which have too many rarer codons result in slowing down transcription and inhibition of protein expression - in extreme cases, rare codons are thought to introduce transcription errors when the rare tRNA is not available. If you were to try and express your human cDNA sequence in yeast (Saccharomyces cerevisiae), which codons in your sequence might cause problems for expression. Note there is no hard threshold, but generally codons with 1% usage or less are considered rare.

TODO

5 Question

We now turn to sequence comparison and alignment. You are given the following coding sequence fragments. They encode a homologous proteins in different species, sequence 2 is human (1 extra mark if you can give the gene name and the most likely species for the other sequences). The sequences are aligned to the correct reading frame:

1. `CTGAAGCGGGAGGCTGAGACGCTGCGGGAGCGGGAGGGC`
2. `CTCAAGCGTGAGGCCGAGACCCTACGGGAGCGGGAAGGC`
3. `GAAGAGCTGAAGAGAGAGGCTGACAATTTAAAGGACAGA`
4. `AACGAGGAGCTCAAGCGAGAAGCTGATACGCTGAAGGAC`

5.1 Sequences 1 and 2 differ slightly. How does the resulting protein differ? Could this have functional implications?

TODO

5.2 Now use the Needleman Wunsch algorithm to compare sequence 1 to sequences 3 and 4. Use the scoring: match +2, mismatch -1, indel -1. Perform at least one of these on paper (or both if you wish). On paper, use the first three codons only.

TODO

5.3 Comparing the bare sequences, what can you conclude about the relatedness of the species?

TODO

5.4 Extra mark

TODO

Sequence	Gene name	Species
1	?	?
2	?	human
3	?	?
4	?	?