THE UNIVERSITY *of* EDINBURGH

Bioinformatics 1

Assignment 1

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

October 26, 2016

Contents

1 Question

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

I have selected *Huntington Disease*, also known as *Huntington Chorea*.

HD is a genetic brain disorder which causes jerky movements, emotional problems, and loss of cognition.

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

A DNA segment known as a CAG trinucleotide repeat has been consistently detected in people who have been subjected to *Molecular Genetic Testing* and suffer from *HD*.

It is believed that an increase of the *CAG segment* length causes *huntingtin* proteins to be longer. Furthermore, these abnormal proteins get split into toxic segments which accumulate in neurons. This compromises neurons' normal behaviour, and might lead to their death.

The manifestation of these events damage areas of the brain, thus originating the symptoms of a person with *HD*.

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

As one can see in the NCBI *huntingtin* page, the gene is also known as *HD*, and *IT15*.

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

Yes, it is. The *NCBI* database contains orthologs of numerous species, including Mus musculus and Drosophila melanogaster.

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?

NCBI page: HTT huntingtin [Homo sapiens (human)]

It is located in **Chromosome 4 - *NC_000004.12***, from position *3074510* to *3243960*.

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

It has ***67* exons**, and therefore ***66* introns** *(67 - 1)*.

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.

2.3 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?

I downloaded both the ***Complete Record*** *FASTA* and ***Coding Sequences*** *FASTA nucleotide* files from the NCBI Gene DB.

The sequence unique identifier is *NC_000004.12*, from *GRCh38.p7*.

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

The *Complete Record* file has a length of **169451** nucleotides. The *Coding Sequences* file is **9429** nucleotides long. Therefore, we can say that only **5.56%** of the entire gene correspond to coding sequences:

$$\frac{9429}{169451} = 0.05564$$

3 Question

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

The translation of the *cDNA* contains **3143** amino acids, which makes sense given the *Coding Sequence* is **9429** nucleotides long:

$$9429/3 = 3143$$

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

All of the 64 codons have been used.

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

The most common amino acid in the *huntingtin* protein is **Leucine** *(Leu/L)* - see figure 1.

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

Leucine has 6 coding amino acids. Below is a table with their respective frequencies in the *huntingtin* protein.

Codon	Frequency
UUA	70
UUG	156
CUU	168
CUC	178
CUA	68
CUG	295

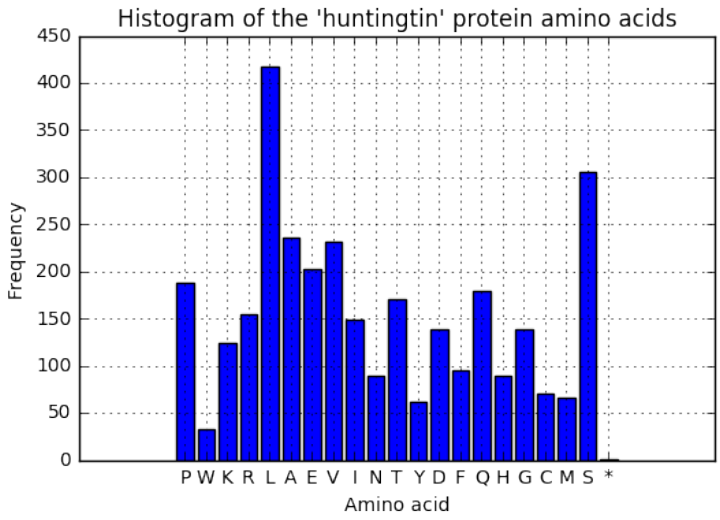

Figure 1: Histogram obtained with the Jupyter Notebook.

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	?	?

6 Sources

Genetics Home Reference
https://ghr.nlm.nih.gov/condition/huntington-disease

Wikipedia
https://en.wikipedia.org/wiki/Huntington%27s_disease