THE UNIVERSITY *of* EDINBURGH

Bioinformatics 1

Assignment 1

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

October 27, 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?

Out of the 64 possible codons available, **62** codons are used. **UAA** and **UAG** are not 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

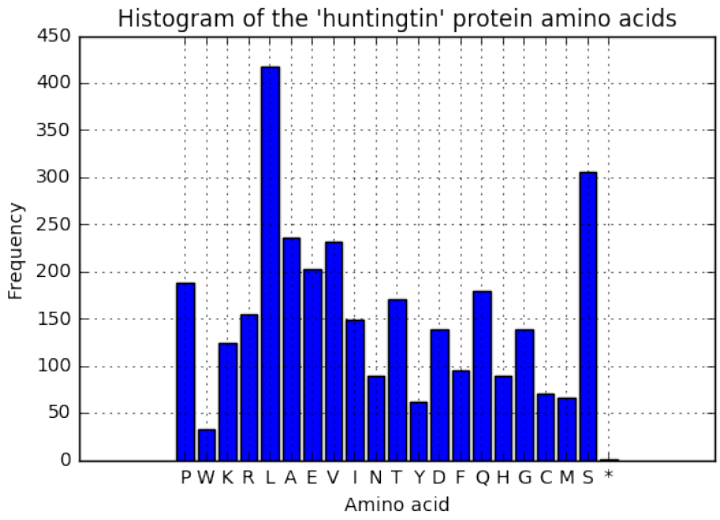

Figure 1: Histogram obtained with the Jupyter Notebook.

4 Question

Now look at the following database:

Kazusa - Codon Usage Database

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.

According to the *Saccharomyces cerevisiae* DB page, yeast have **6534504** codons. Below is a table with the frequency of each codon in the format *[triplet] [frequency:* ***percentage****] ([number]).*

UUU 2.61 (170666) UUC 1.84 (120510) UUA 2.62 (170884) UUG 2.72 (177573)	UCU 2.35 (153557) UCC 1.42 (92923) UCA 1.87 (122028) UCG 0.86 (55951)	UAU 1.88 (122728) UAC 1.48 (96596) UAA 0.11 (6913) UAG 0.05 (3312)	UGU 0.81 (52903) UGC 0.48 (31095) UGA 0.07 (4447) UGG 1.04 (67789)
CUU 1.23 (80076) CUC 0.54 (35545) CUA 1.34 (87619) CUG 1.05 (68494)	CCU 1.35 (88263) CCC 0.68 (44309) CCA 1.83 (119641) CCG 0.53 (34597)	CAU 1.36 (89007) CAC 0.78 (50785) CAA 2.73 (178251) CAG 1.21 (79121)	CGU 0.64 (41791) CGC 0.26 (16993) CGA 0.30 (19562) CGG 0.17 (11351)
AUU 3.01 (196893) AUC 1.72 (112176) AUA 1.78 (116254) AUG 2.09 (136805)	ACU 2.03 (132522) ACC 1.27 (83207) ACA 1.78 (116084) ACG 0.80 (52045)	AAU 3.57 (233124) AAC 2.48 (162199) AAA 4.19 (273618) AAG 3.08 (201361)	AGU 1.42 (92466) AGC 0.98 (63726) AGA 2.13 (139081) AGG 0.92 (60289)
GUU 2.21 (144243) GUC 1.18 (76947) GUA 1.18 (76927) GUG 1.08 (70337)	GCU 2.12 (138358) GCC 1.26 (82357) GCA 1.62 (105910) GCG 0.62 (40358)	GAU 3.76 (245641) GAC 2.02 (132048) GAA 4.56 (297944) GAG 1.92 (125717)	GGU 2.39 (156109) GGC 0.98 (63903) GGA 1.09 (71216) GGG 0.60 (39359)

By examining the table, one can see that codons with a frequency lower than 1% are: `UGU`, `UGC`, `UAA`, `UGA`, `UCG`, `UAG`, `CGU`, `CUC`, `CCC`, `CAC`, `CGC`, `CGA`, `CCG`, `CGG`, `AGC`, `ACG`, `AGG`, `GGC`, `GCG`, `GGG`. Thus, these codons might introduce errors during transcription.

We can even see that `UAA` (0.11%), `UAG` (0.05%), and `UGA` (0.07%) are very rare triplets (for yeast). All these three codons correspond to the **termination** codon - see RNA codon table.

This is not a problem however, because despite the low percentage of these codons, there are thousands of each available, which by far cover the needs of the protein expression.

Another codon which might be considered *rare* is **CGG**, but again that should not be a problem because the expression does not demand a number of that triplet greater than its frequency in yeast.

Having said that, suppose the frequency of **CGG** was much lower - so low that it could introduce transcription errors. One possible work around is to reverse engineer the *mRNA*. **CGG** codes *Arginine*, which is also coded by **CGU**, **CGC**, and **CGA** - which have a much higher frequency (check table above). One could replace **CGG** occurences in the *mRNA* with one of those other three *Arginine* encoding codons, and that should solve the transcription errors.

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. 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?

The resulting protein does not differ, it is the same: `LKREAETLREREG`. Since the encoded amino acids do not change, the protein has exactly the same functionality.

One might theorise that this amino acid encoding redundancy is a biological safeguard for small mutations not to change the functionality of a protein.

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.

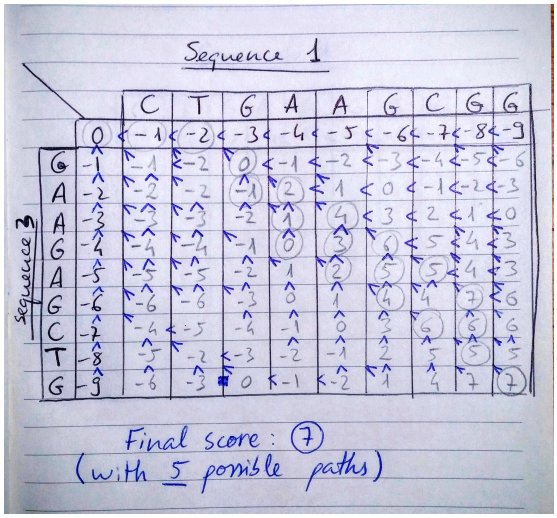

Figure 2: Needleman-Wunsch algorithm result on sequence 1 and 3.

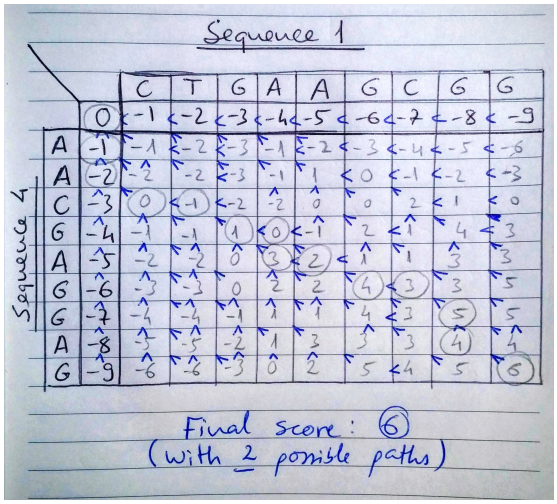

Figure 3: Needleman-Wunsch algorithm result on sequence 1 and 4.

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

By comparing *sequence 1* and *sequence 2* one can see they are very similar. In fact, they encode the same amino acids.

Sequence 3 and *sequence 4* are also pretty similar. The amino acids they encode are `EELKREADNLKDR` and `NEELKREADTLKD`, respectively - they share a common sequence `EELKREAD`.

Furthermore, with the information from the species (see next answer), we can confirm that *sequences 3* and *4* are closely related (they are both fish).

5.4 Extra mark for the gene name and the most likely species for each sequence.

After running a Standard Nucleotide BLAST query for each of the given sequences, we find that it corresponds to the **KCNB1** gene - *potassium voltage-gated channel, member 1*.

Sequence	Species
1	*Mus musculus*
2	*Homo sapiens*
3	*Fundulus heteroclitus*
4	*Danio rerio*

6 Sources

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

National Center for Biotechnology Information
https://www.ncbi.nlm.nih.gov/

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