732A51 Bioinformatics Lab 1

Raymond Sseguya, Martin Smelik, Duc Duong

2018 M11 7

Task 1

Task 1.1

With the help of wikipedia, we can compute easily the probabilities p_1 and q_1 in the first generation. p_1 = frequency(AA) + frequency(Aa)/2 = p^2 + pq = p(p+q) = p*1 = p q_1 = frequency(aa) + frequency(Aa)/2 = q^2 + pq = q(p+q) = q

Now let's have in the nth generation frequency(A) = p and frequency(a) = q. Then we can compute the frequencies for (n+1)st generation:

$[f_(n+1)(AA), f_(n+1)(Aa), f_(n+1)(aa)] = f(AA)^2 + 2f(AA)f(Aa)[1/2, 1/2, 0] + 2f(AA)f(aa)[0, 1, 0] + f(Aa)^2[1/4, 1/2, 1/4] + 2f(Aa)f(aa)[0, 1/2, 1/2] + f(aa)^2[0, 0, 1] = [(f(AA) + f(Aa)/2)^2, (2f(AA) + f(Aa))(f(aa) + f(Aa)/2), f(aa) + f(Aa)^2/2 = [p^2, 2pq, q^2]$

Therefore $f(AA) = p^2$, $f(Aa) = 2pq$, $f(aa) = q^2$

Task 1.2

Total number of people = 357 + 485 + 158 = 1000

Total allele population = 2 times 1000 = 2000

Total number of M is 2 times 357 added to 485 = 1199

Total number of N is 2 times 158 added to 485 = 801

Probability of getting M is 1199 out of 2000 = 0.5995 (assuming diploid)

Probability of getting N is 801 out of 2000 = 0.4005 (assuming diploid)

Creating vector of number of homozygotes and heterozygotes, R

Creating vector of Probabilities of M and N alleles, S

```
R <- c(357, 485, 158)
S <- c(0.5995*0.5995, 2*0.5995*0.4005, 0.4005*0.4005)

chisq.test(R, p=S)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  R
## X-squared = 0.099938, df = 2, p-value = 0.9513
```

The null hypothesis under the chi-square test for goodness of fit, that the population follows the Hardy Weinberg equilibrium, IS ACCEPTED.

Task 2

Task 2.1

According to the information in the FEATURES section, the protein product is **RecQ type DNA helicase**

Task 2.2

The first four amino acids are MVVA, which is **Methionine, Valine, Valine and Alanine**

Task 2.3

Attached "after_backtranseq.FASTA"

Task 2.4

The coding strand sequence obtained from using "backtranseq" exactly matches the **reversed complemented** nucleotide sequence provided. For instance, the last letter **"D"** representing **Aspartic acid** exactly macthes the last three letter sequence **"GAT"**

After reversing and complementing, the nucleotide sequence is in the 3' to 5' prime direction but amino acid contents are exactly the same. For example, the third amino acid in the reversed and complemented nucleotide sequence is **"AAC"** which is the same as the **"GTT"** in the sequence produced by "backtranseq", which is also the same as **"V"**.

Attached is "after_reverse_complement.FASTA"

Task 2.5

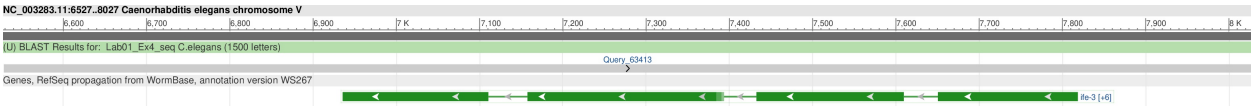
The nucleotide number range is **1 to 5661** and there is no stop codon which means that the gene is incomplete. The genome lies on chromosome *I*.

Task 3

3.1

According to Wikipedia, C. elegans is being extensively used as a model organism.It was the first multicellular organism to have its whole genome sequenced, and as of 2012, is the only organism to have its connectome (neuronal "wiring diagram") completed. The C. elegans genome contains an estimated 20,470 protein-coding genes. About 35% of C. elegans genes have human homologs. Remarkably, human genes have been shown repeatedly to replace their C. elegans homologs when introduced into C. elegans. Conversely, many C. elegans genes can function similarly to mammalian genes.

3.2



3.3

The database genomic sequence progresses from 6529 to 8028, and the query sequence from 1 to 1500, both essentially in the same direction.

When the query sequence is reverse complemented, it runs in the direction which is from 8028 to 6529, opposite to 1 to 1500.

However, the two look exactly the same in the Genome Data Viewer.

3.4

The query sequence is found on chromosome V and positions are **6936 to 7110 for exon** , **7158 to 7393 for exon 2**, **7433 to 7609 for exon 3**, and **7651 to 7818 for exon 4**

3.5

protein code of exon 1: **MSTSVAENKALSASGDVNASDASVPPELLTRHPLQNRWALWYLKA-DRNKEWEDCLK**

protein code of exon 2: **LNTSFIDFFQMVSLFDTVEDFWSLYNHIQSAGGLNWGSDYYLFKEGIKPMWED-VNNVQGG**

protein code of exon 3: **RRTQLLDHYWLELLMAIVGEQFDEYGDYICGAVVNVRQKGD-KVSLWTRDATRDDVNLRIGQVLKQKLSIPDTEILR**

protein code of exon 4: **YEVHKDSSARTSSTVKPRICLPAKDPAPVKEKGPAATTSP-SNPGTEATGTSPATPTP***

When compare these protein codes with the one in question 2 - FASTA version link

We see that there is no relevent between these code.

3.6

The symbol of the gene is ife-3. It is located on chromosome V and has 4 exon as we could see also in the earlier excercises. Interesting information about this gene is that except of some simple organisms, also Eukaryots have it and therefore also humans. This gene mediates the mRNA, that do not content spliced-leader sequence to ribosomes.