732A51 Bioinformatics Lab1

Milda Poceviciute, Fanny Karelius, Rab Nawaz Jan Sher and Saman Zahid

7 november 2018

Question 1

1.1

Initially

$$total \;\; population = 2N$$

Frequency of genome

$$f_1(a) = q, \qquad f_1(A) = p, \qquad f_1(p+q) = 1$$

$$f_1(AA) = p^2, \qquad f_1(aa) = q^2, \qquad f_1(Aa) = 2pq$$

Proportions in offspring population:

$$(p+q)^2 = p^2 + q^2 + 2pq = 1$$

$$P(A) = f_1(AA) + \frac{1}{2} f_1(Aa)$$
$$= p^2 + \frac{1}{2}(2pq) = p^2 + pq$$

$$P(a) = f_1(aa) + \frac{1}{2} f_1(Aa)$$
$$= q^2 + \frac{1}{2}(2pq) = q^2 + pq$$

$$P(Aa \;\; or \;\; aA) = pq + pq = 2pq$$

Second generation:

$$p(AA) = (p^2 + pq)^2 = p^4 + 2p^3q + p^2q^2 = p^2(p^2 + 2pq + q^2) = p^2$$

In the same way:

$$p(aa) = (q^2 + pq)^2 = q^2$$
$$p(Aa \text{ or } aA) = 2(p^2 + pq)(q^2 + pq) = 2(2p^2q^2 + pq^3 + p^3q) = 2pq(p^2 + 2pq + q^2) = 2pq$$

The proportions of the second generation are the same as in the first generation. No, a population in Hardy-Weinberg equilibrium cannot deviate from it with random mating.

1.2

```r
MM <- 357
MN <- 485
NN <- 158

p<-(MM+MN/2)/sum(MM+MN+NN)
q<-(NN+MN/2)/sum(MM+MN+NN)

chisq.test(c(MM,MN, NN), p = c(p^2, 2*p*q,q^2))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(MM, MN, NN)
## X-squared = 0.099938, df = 2, p-value = 0.9513
```

As our p-value is above 0.05 we cannot reject the null hypothesis (at 5% significance level) that the population is in Hardy-Weinberg equilibrium.

Question 2

2.1

According to the GenBank, the protein product is named "RecQ type DNA helicase".

Loading the data

In this question we created Fasta files from the protein sequence from GenBank, corresponding nucleotides sequence. As well as, the best prediction of the nucleotide sequence, and the inverse and complements of the both nucleotide sequences.

```r
#install.packages("seqinr")
library(seqinr)
protein_seq <- read.fasta("Fasta_file.fasta")
EMBOSS_seq <- read.fasta("backtranseq.txt")
True_seq <-  read.fasta("truenucleotideseq.txt")
Reversed_True <- read.fasta("Reversed_compl.txt")
Reversed_Emboss <- read.fasta("Reversed_compl_backtranseq.txt")
emboss <- EMBOSS_seq$A
real <- True_seq$`CU329670.1:1-5662`
inv_real <- Reversed_True$`1-5662`
inv_emboss <- Reversed_Emboss$A
```

2.2

The first four amino acids are:

```r
protein_seq$`protein_id=CAC05745.1,product=RecQtypeDNAhelicase`[1:4]
```

```
## [1] "m" "v" "v" "a"
```

The full names of them are: Methionine, Valine, Valine, Alanine.

2.3, 2.4

We can see that the length of the sequences do not match - the "true" nucleotide sequence from the GenBank has one additional nucleotide as the best predicted sequence by the *backtranseq*.

```
length(emboss)
```

```
## [1] 5661
```

```
length(real)
```

```
## [1] 5662
```

```
length(inv_real)
```

```
## [1] 5662
```

```
length(inv_emboss)
```

```
## [1] 5661
```

Below we print out 10 first nucleotides of each sequence. We deduce that the "true" nucleotide sequence starts with an additional "g", which is omitted in the best prediction sequence:

```
## [1] "Predicted nucleotide seq:"
##  [1] "a" "t" "g" "g" "t" "t" "g" "t" "t" "g"
## [1] "GenBank nucleotide seq:"
##  [1] "g" "a" "t" "c" "a" "c" "g" "t" "a" "c"
## [1] "Inverse and complement of GenBank nucleotide seq:"
##  [1] "a" "t" "g" "g" "t" "c" "g" "t" "c" "g"
## [1] "Inverse and complement of predicted nucleotide seq:"
##  [1] "a" "t" "c" "a" "c" "g" "a" "a" "c" "a"
```

Below we compare the sequences, and find the percentage of the matching nucleotides:

```
## [1] "Comparison of the Predicted and GenBank nucleotide seq:"
## [1] 25.96714
## [1] "Comparison of the Predicted and inverse&complement of GenBank nucleotide seq:"
## [1] 78.97898
## [1] "Comparison of the GenBank and inverse&complement of Predicted nucleotide seq:"
## [1] 78.97898
```

When RNA is done, the way the enzyme copies a string of DNA is that it creates a string that is inverted and made of the complement nucleotides (compared to the original DNA string). The *backtranseq* function imitates the work of the enzyme, hence the resulting sequence is actual the inverse of the original sequence, and with the complemented nucleotides substituted instead the actual ones. Therefore, the inverted and complemented predicted sequence is matching considerably better to the GenBank's sequence than both non-inverted (and complemented) sequences, and vice versa. However, the match is not 100%. This may be due to that the predicted sequence is derived from the amino acids sequence by the provided website. As some of the acids can potentially be composed of the nucleotides in a few different ways, it is reasonable that the resulting nucleotide sequence can potentially differ a little bit from the "real" one provided at the GenBank.

2.5

The nucleotide number range that corresponds to these amino acids is 5661. The reversed and complemented sequence starts with the starting M protein ("ATC" nucleotides), hence we chose this version and looked for the stop codan below. The stop condons are TAA, TAG or TGA. We found several stop codons, but the first one is at location 19:

```
j = 1
stop_co1 <-c("t","a","a")
stop_co2 <-c("t","a","g")
stop_co3 <-c("t","g","a")
result <- c()
i = 1
while (j < 5661){
  acid <- inv_emboss[j:(j+2)]
  #print(acid)
  if (all(acid == stop_co1) || all(acid ==stop_co2) || all(acid == stop_co3)){
    result[i] <- j
    i <- i+1
  }
  j= j+3
}
stop_loc <- result[1]
stop_loc
```

```
## [1] 19
```

```
inv_emboss[stop_loc:(stop_loc+2)]
```

```
## [1] "t" "a" "a"
```

Given the information in GenBank, this protein sequence lies on the Chromosome 1.

Question 3

3.1

C. elegans is a free-living transparent roundworm that lives in temperate soil environments. It is one of the simplest organisms with a nervous system and this makes it important for the scientific community because it is used as a model organism for research on neurological development in animals. It was the first multicellular organism to have its whole genome sequenced. The neurons of C. elegans are very similar to that of humans thats why the developmental and genetic experiments that are not possible to directly implement on human or are very time consuming and costly to implement on humans, C. elegans are used instead.

3.2, 3.3

Numbering of the sequences in the alignment:

- query sequence = 1 - 1500
- subject sequence = 6529 - 8028

The direction of database sequence is opposite to query sequence.

Reverse Numbering of the sequences in the alignment:

- query seq = 1 - 1500
- subject sequence = 8028 - 6529

The direction of database sequence is same as query sequence.

3.4

Chromosome 5. Gene: ife-3

3.5

Extracting exons

```
library(seqinr)
complete_seq <- read.fasta(file = "files/allseq_6936to7818.FASTA")

e1 <- complete_seq$`NC_003283.11:6936-7818`[1:174]
e2 <- complete_seq$`NC_003283.11:6936-7818`[(174+48):(174+48+235)]
e3 <- complete_seq$`NC_003283.11:6936-7818`[(457+40):(457+40+176)]
e4 <- complete_seq$`NC_003283.11:6936-7818`[(673+42):(673+42+168)]
```

Exon translation

```
proteins_from_exons <- read.fasta(file = "files/exons_translation.FASTA")
print(proteins_from_exons)
```

```
## $`exons_6936-7818_1`
##   [1] "l" "r" "s" "w" "g" "g" "w" "r" "s" "s" "c" "s" "l" "r" "a" "g" "i"
##  [18] "r" "w" "r" "s" "r" "c" "g" "w" "s" "l" "f" "l" "h" "w" "c" "w" "i"
##  [35] "l" "g" "w" "k" "t" "y" "a" "w" "l" "d" "s" "r" "*" "g" "a" "s" "r"
##  [52] "r" "v" "l" "v" "n" "f" "v" "p" "q" "n" "l" "s" "i" "r" "n" "a" "q"
##  [69] "f" "l" "l" "q" "n" "l" "s" "d" "a" "k" "i" "d" "i" "i" "a" "s" "s"
##  [86] "i" "t" "s" "p" "q" "g" "n" "l" "v" "t" "l" "l" "t" "n" "i" "h" "d"
## [103] "s" "s" "a" "d" "v" "v" "s" "v" "l" "v" "e" "l" "l" "s" "n" "n" "s"
## [120] "h" "q" "q" "l" "q" "p" "v" "v" "i" "e" "q" "l" "r" "s" "s" "l" "k"
## [137] "l" "l" "l" "i" "d" "n" "n" "q" "p" "t" "s" "t" "l" "n" "v" "v" "d"
## [154] "v" "l" "p" "h" "w" "l" "d" "s" "f" "l" "e" "q" "v" "i" "i" "g" "s"
## [171] "p" "v" "q" "s" "s" "g" "r" "l" "n" "v" "i" "v" "q" "r" "p" "e" "v"
## [188] "l" "d" "s" "v" "e" "k" "*" "n" "h" "l" "q" "t" "i" "l" "p" "f" "l"
## [205] "v" "t" "v" "s" "f" "q" "v" "p" "e" "s" "p" "a" "i" "l" "e" "g" "v"
## [222] "s" "g" "e" "k" "l" "w" "r" "n" "*" "s" "i" "g" "r" "i" "h" "i" "a"
## [239] "g" "s" "*" "q" "c" "f" "v" "f" "r" "y" "g" "c" "a" "h"
## attr(,"name")
## [1] "exons_6936-7818_1"
## attr(,"Annot")
## [1] ">exons_6936-7818_1"
## attr(,"class")
## [1] "SeqFastadna"
```

Complete Sequence translation

```
proteins_from_6936to7818 <- read.fasta(file = "files/complete_translationq3.FASTA")
print(proteins_from_6936to7818)
```

```
## $`6936-7818_1`
```

```
##   [1] "l" "r" "s" "w" "g" "g" "w" "r" "s" "s" "c" "s" "l" "r" "a" "g" "i"
##  [18] "r" "w" "r" "s" "r" "c" "g" "w" "s" "l" "f" "l" "h" "w" "c" "w" "i"
##  [35] "l" "g" "w" "k" "t" "y" "a" "w" "l" "d" "s" "r" "*" "g" "a" "s" "r"
##  [52] "r" "v" "l" "v" "n" "f" "v" "s" "g" "n" "i" "l" "l" "r" "*" "q" "i"
##  [69] "l" "k" "l" "*" "n" "y" "l" "k" "i" "s" "v" "s" "g" "m" "l" "n" "f"
##  [86] "c" "f" "k" "t" "c" "p" "m" "r" "r" "l" "t" "s" "s" "r" "v" "a" "s"
## [103] "r" "v" "h" "k" "e" "t" "l" "s" "p" "f" "*" "r" "t" "f" "t" "t" "a"
## [120] "p" "q" "m" "*" "s" "p" "y" "s" "s" "n" "c" "s" "p" "t" "i" "a" "i"
## [137] "n" "s" "s" "n" "q" "*" "*" "s" "s" "n" "c" "v" "l" "l" "*" "s" "f"
## [154] "y" "d" "s" "l" "n" "k" "i" "y" "f" "s" "k" "r" "t" "c" "l" "s" "t"
## [171] "t" "t" "n" "q" "r" "p" "p" "*" "t" "l" "l" "t" "s" "s" "h" "i" "g"
## [188] "l" "i" "p" "s" "l" "n" "k" "*" "*" "s" "d" "p" "q" "f" "n" "p" "p"
## [205] "a" "d" "*" "m" "*" "l" "y" "s" "d" "q" "k" "s" "s" "t" "v" "s" "k"
## [222] "s" "e" "t" "i" "w" "k" "k" "s" "i" "k" "d" "v" "f" "k" "n" "l" "l"
## [239] "p" "s" "d" "n" "p" "p" "i" "p" "c" "y" "g" "q" "l" "s" "s" "t" "r"
## [256] "e" "p" "s" "d" "s" "g" "g" "g" "v" "w" "*" "e" "a" "l" "e" "e" "l"
## [273] "k" "h" "r" "t" "h" "s" "h" "r" "r" "k" "l" "t" "m" "l" "c" "f" "p"
## [290] "l" "r" "m" "c" "s" "x"
## attr(,"name")
## [1] "6936-7818_1"
## attr(,"Annot")
## [1] ">6936-7818_1 Caenorhabditis elegans chromosome V"
## attr(,"class")
## [1] "SeqFastadna"
```

Some parts of protein translation obtained from exons matches to the protein translation of the entire sequence. The protein sequence obtained from the complete sequence has a number of stop codes, the translation is quite vague as at some places the sequence stops as soon as it starts or start protein occurs again multiple times before stop.

3.6

ife-3 is the Eukaryotic translation initiation factor that is used in early steps of protein synthesis to provide stability. It is most similar to the human gene eIF4E. It is the only isoform required for viability. ife-3 is found in humans, chimpanzees, Rhesus monkeys, dogs, cows, mice, rats, chickens, zebrafish, fruit flies, mosquitoes, rice, and frogs.

There are 3 isoforms of the given sequence and 4 exons are formed.