732A51 Bioinformatics Lab1

rabnawaz and saman

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 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 genetical 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

```r
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

```r
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

```r
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.