Bioinformatics - Computer Lab 1

Group 7: Phillip H?lscher (phiho267), Lennart Schilling (lensc874), Thijs Quast (thiqu264), Mariano Maquieira Mariani (marma330)

13 11 2018

Question 1

Hardy-Weinberg equilibrium

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

Two alleles A and a p = A q = a

Question 1.1

The genotype frequencies must sum to one sum = $p^2 + 2pq + q^2 = 1$

What is the proportion of A and a alleles in the offspring population?

f1(A) = f(AA) + 1/2 f(Aa)

p(A) = p^2 + 1/2 2pq = p^2 + pq = p (p + q)

remember: p + q = 1

p(A) = p

f1(a) = f(aa) + 1/2 f(Aa)

p(a) = q^2 + 1/2 2*pq = q^2 + pq = q (p + q)

remember: p + q = 1

p(a) = q

Hence, with random mating, can a population in Hardy-Weinberg equilibrium ever deviate from it?

Yes, but just if the seven assumptions of deviations from Hardy-Weinberg equilibrium are not fulfilled:[1]

- organisms are diploid
- only sexual reproduction occurs
- generations are nonoverlapping
- mating is random

[1] https://en.wikipedia.org/wiki/Hardy%E2%80%93Weinberg_principle

	A (*p*)	**a (*q*)**
A (*p*)	AA (p^2)	Aa (pq)
a (*q*)	Aa (qp)	aa (q^2)

Figure 1: Hardy-Weinberg

$$p = \frac{2 \times \text{obs}(\text{AA}) + \text{obs}(\text{Aa})}{2 \times (\text{obs}(\text{AA}) + \text{obs}(\text{Aa}) + \text{obs}(\text{aa}))}$$

Figure 2: formula to calculate p

- population size is infinitely large
- allele frequencies are equal in the sexes
- there is no migration, gene flow, admixture, mutation or selection

Question 1.2

Alleles:

L^M denoted with M

L^N denoted with N

MM = 357

MN = 485

NN = 158

First we have to calulate the value of p and q.

```
MM = 357
MN = 485
NN = 158
n = MM+MN+NN
p = (2 * MM + MN) / (2*(n))
q = 1 - p
```

Now we can calculate p^2, 2pq and q^2

```
p_power2 <- p^2
q_power2 <- q^2
pq2 <- 2*p*q
```

Now we use the chisq.test function

```
chisq.test(c(MM,MN,NN), p = c(p_power2, pq2, q_power2))
```

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

Question 2

Question 2.1

Name of the protein product of the CDS: *RecQ type DNA helicase*

Question 2.2

By looking at the FEATURES/CDS/translation - section which shows the amino acid sequence we can identify the first four amino acids: `Methionine (M)`, `Valine (V)`, `Valine (V)`, `Alanine (A)`

Question 2.3

For the back-translation of the protein sequence to a nucleotide sequence *backtranseq* was used.[2] The input was taken from the FEATURES/CDS/translation - section which shows the amino acid sequence. As a result, the nucleotide sequence of the coding strand that corresponds to these amino acids can be found in the file *lab1_Q_2_3.fasta* which was submitted in addition to this report.

Question 2.4

The comparison between the obtained coding strand sequence (file *lab1_Q_2_3.fasta*) and the nucleotide sequence provided (accessible by following the CDS link at the ORIGIN sector) shows that they differ.

Provided nucleotide sequence: `GATCACGTAC[...]CGACGACCAT`

Obtained coding strand sequence: `ATGGTTGTTG[...]TGTTCGTGAT`

Reversing the coding strand creates a new strand where every base is reversed - The last base will be the first base, the second last base will be the second base and so on. Complementing the coding strand creates a new strand where every base is complementary to the base of the origin coding strand. Overview about the complementaries for each base:

- A > T
- T > A
- C > G
- G > C

Neither reverse (`TAGTGCTTGT[...]GTTGTTGGTA`) nor complement (`TACCAACAAC[...]ACAAGCACTA`) sequences[3] equal the provided nucletoide sequence.

The nucleotide sequence of the template strand that corresponds to the amino acids (which is the result of complementing the coding strang) can be found in the file *lab1_Q_2_4.fasta* which was submitted in addition to this report.

Question 2.5

Based on the information "complement(<1..5662)" in the CDS section, the nucleotide number range that corresponds to these amino acids is 1 to 5662.

Because the stop codon is not included in this sequence, it is not possible to identify it.

Since the sequence is defined as 'Schizosaccharomyces pombe chromosome I', the genomic sequence lies on chromosome 1.

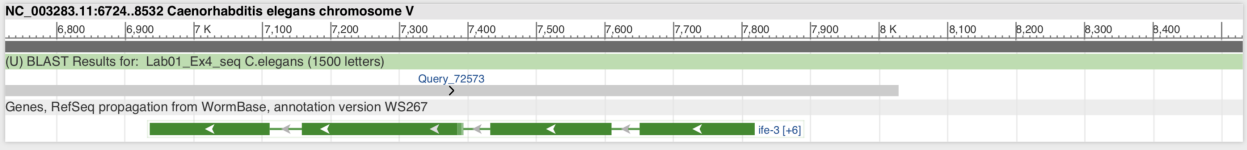
Question 3

Question 3.1

[2] https://www.ebi.ac.uk/Tools/st/emboss_backtranseq
[3] http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp.html

The Caenorhabditis elegans (C. Elegans) is a worm-like specicies of 1mm in length, living in the soil. C. Elegans have neurons, skin, muscles and other features that are very similar to human beings[4]. This organism is important in the scientific community because it is one of the most simple organisms with a nervous system. This makes it very appropriate for scientific research. In addition, one can easily apply genetic manipulation on a c. elegans[5].

Question 3.2



Question 3.3

```
Query  1     ATTTTTAAAAATGTACAAAATCAAACGCCCTACAAATCATGTGTGTGAAGAAGAATAATA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  6529  ATTTTTAAAAATGTACAAAATCAAACGCCCTACAAATCATGTGTGTGAAGAAGAATAATA  6588

Query  61    ACTAACATATCTATTTATATTTACCGAATAAATATATATTCATCAATTAACCTGAAGAAC  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  6589  ACTAACATATCTATTTATATTTACCGAATAAATATATATTCATCAATTAACCTGAAGAAC  6648

Query  121   AAACGAATTCGGCTACAGGCGTCGATCAGTCTCGAATCTAGTAACAACAAGAGAGCAATA  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  6649  AAACGAATTCGGCTACAGGCGTCGATCAGTCTCGAATCTAGTAACAACAAGAGAGCAATA  6708
```

The file contains 1500 characters. The BLAST tool starts matching from 1 to 60, 61 to 120 etc. In the same way, the database genomic sequence starts from 6529 to 6588, 6589 to 6648. Therefore, both the query sequence and database sequence progress in the same direction, namely increasing.

```
Query  1     TTATTGTTTTCCAAGCTTTAATATCAATTTATTGTGCCCGATGTTACCAATTACACTTGA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  8028  TTATTGTTTTCCAAGCTTTAATATCAATTTATTGTGCCCGATGTTACCAATTACACTTGA  7969

Query  61    AAAATCTAAAAAGCTTGGAAACTAGCCGAAAATGTGCAGTAAAACAAAATTTCCTATAAA  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  7968  AAAATCTAAAAAGCTTGGAAACTAGCCGAAAATGTGCAGTAAAACAAAATTTCCTATAAA  7909

Query  121   ATCCGAGTTATTTGAACCAAATTCATACTCTTCTCTATTTTATCGTTTTCCGAGCTCTAA  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  7908  ATCCGAGTTATTTGAACCAAATTCATACTCTTCTCTATTTTATCGTTTTCCGAGCTCTAA  7849
```

If one reverse complements the query sequence, the query sequence progresses increasingly, whereas the database sequence progresses decreasingly.

Question 3.4

The chromosome query sequence is found on chromosome "V". The query sequence notation is found on the range: 6,529 >> 8,028.

Question 3.5

[4]http://www.people.ku.edu/~erikl/Lundquist_Lab/Why_study_C._elegans.html
[5]https://www.quora.com/Why-are-Caenorhabditis-elegans-important-to-biology)

```r
#install.packages("stringi")
library(stringi)
sequence <- "ATTTTTAAAAATGTACAAAATCAAACGCCCTACAAATCATGTGTGTGAA
GAAGAATAATAACTAACATATCTATTTATATTTACCGAATAAATATATATTCATCAATTAAC
CTGAAGAACAAACGAATTCGGCTACAGGCGTCGATCAGTCTCGAATCTAGTAACAACAAGAG
AGCAATACGAAAACCGGTAAATCAATAGGGGGAAGCGAAACAGTAGGTACAAATTGGAGGGG
AAGCACCAATACATTAGGTGGGGGGTACGACTTGAAAAATGAGCTGATTTTCGAATAGTTAA
AGCGATGATCGTGTCCGAAAAACAGTTCATTTTTCAAGACAACATTGAGACTGGGAGTACGG
GGAAGCTCATTTACGGTGAGAGGAATTGGTGAGATCTTTAGAATATGCTTAAGGAGTTGGGG
TGGCTGGAGAAGTTCCTGTAGCCTCCGTGCCGGGATTCGATGGAGAAGTCGTTGCGGCTGGT
CCCTTTTCCTTCACTGGTGCTGGATCCTTGGCTGGAAGACATATGCGTGGCTTGACAGTCGA
TGAGGTGCGAGCCGACGAGTCCTTGTGAACTTCGTATCTGGAAATATTTTACTTAGATAGCA
AATACTAAAATTGTAAAATTACCTCAAAATCTCAGTATCCGGAATGCTCAATTTCTGCTTCA
AAACCTGTCCGATGCGAAGATTGACATCATCGCGAGTAGCATCACGAGTCCACAAGGAAACC
TTGTCACCCTTTTGACGAACATTCACGACAGCTCCGCAGATGTAGTCTCCGTACTCGTCGAA
TTGCTCTCCAACAATAGCCATCAACAGCTCCAACCAGTAGTGATCGAGCAATTGCGTTCTTC
TCTGAAGCTTCTATGATTCATTGAATAAAATATATTTCTCAAAACGTACTTGCTTATCGACA
ACAACCAACCAACGTCCACCTTGAACGTTGTTGACGTCCTCCCACATTGGCTTGATTCCTTC
CTTGAACAAGTAATAATCGGATCCCCAGTTCAATCCTCCGGCAGACTGAATGTGATTGTACA
GCGACCAGAAGTCCTCGACAGTGTCGAAAAGTGAAACCATCTGGAAAAAATCGATAAAAGAC
GTATTTAAAAATCTTCTACCTTCAGACAATCCTCCCATTCCTTGTTACGGTCAGCTTTCAAG
TACCAGAGAGCCCAGCGATTCTGGAGGGGGTGTCTGGTGAGAAGCTCTGGAGGAACTGAAGC
ATCGGACGCATTCACATCGCCGGAAGCTGACAATGCTTTGTTTTCCGCTACGGATGTGCTCA
TTTAGCTGAAAATAGGTAATATTATATACGATTAGAGCTCGGAAAACGATAAAATAGAGAAG
AGTATGAATTTGGTTCAAATAACTCGGATTTTATAGGAAATTTTGTTTTACTGCACATTTTC
GGCTAGTTTCCAAGCTTTTTAGATTTTTCAAGTGTAATTGGTAACATCGGGCACAATAAATT
GATATTAAAGCTTGGAAAACAATAA"
exon_1 <- substring(sequence, 1123, 1290)
exon_2 <- substring(sequence, 905, 1081)
exon_3 <- substring(sequence, 630, 865)
exon_4 <- substring(sequence, 408, 582)

#Reversed complemented using http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp.html
exon_1_reverse_comp<- 'ATGAGCACATCCGTAGCGGAAAACAAAGCATTGTCA
GCTTCCGGCGATGTGAATGCGTCCGATGCTTCAGTTCCTCCAGAGCTTCTCACCAGACA
CCCCCTCCAGAATCGCTGGGCTCTCTGGTACTTGAAAGCTGACCGTAACAAGGAATGGG
AGGATTGTCTGAAG'

exon_2_reverse_comp<- 'ATGGTTTCACTTTTCGACACTGTCGAGGACTTCTGG
TCGCTGTACAATCACATTCAGTCTGCCGGAGGATTGAACTGGGGATCCGATTATTACTT
GTTCAAGGAAGGAATCAAGCCAATGTGGGAGGACGTCAACAACGTTCAAGGTGGACGTT
GGTTGGTTGTTGTCGATAAGCAA'

exon_3_reverse_comp<- 'AAGCTTCAGAGAAGAACGCAATTGCTCGATCACTAC
TGGTTGGAGCTGTTGATGGCTATTGTTGGAGAGCAATTCGACGAGTACGGAGACTACAT
CTGCGGAGCTGTCGTGAATGTTCGTCAAAAGGGTGACAAGGTTTCCTTGTGGACTCGTG
ATGCTACTCGCGATGATGTCAATCTTCGCATCGGACAGGTTTTGAAGCAGAAATTGAGC
ATTCCGGATACTGAGATTTTGAG'

exon_4_reverse_comp<- 'ATACGAAGTTCACAAGGACTCGTCGGCTCGCACCTC
ATCGACTGTCAAGCCACGCATATGTCTTCCAGCCAAGGATCCAGCACCAGTGAAGGAAA
AGGGACCAGCCGCAACGACTTCTCCATCGAATCCCGGCACGGAGGCTACAGGAACTTCT
CCAGCCACCCCAACTCCTTAA'
```

```
#We concatenate them
paste0(exon_1_reverse_comp,exon_2_reverse_comp,exon_3_reverse_comp,exon_4_reverse_comp)
```

```
## [1] "ATGAGCACATCCGTAGCGGAAAACAAAGCATTGTCA\nGCTTCCGGCGATGTGAATGCGTCCGATGCTTCAGTTCCTCCAGAGCTTCTCACCAGAG
```

```
#Using transeq https://www.ebi.ac.uk/Tools/st/emboss_transeq/ we get the below protein
obtained_protein<-'MSTSVAENKALSASGDVNASDASVPPELLTRHPLQNRWALW
YLKADRNKEWEDCLKMVSLFDTVEDFWSLYNHIQSAGGLNWGSDYYLFKEGIKPMWEDVN
NVQGGRWLVVVDKQKLQRRTQLLDHYWLELLMAIVGEQFDEYGDYICGAVVNVRQKGDKV
SLWTRDATRDDVNLRIGQVLKQKLSIPDTEILRYEVHKDSSARTSSTVKPRICLPAKDPA
PVKEKGPAATTSPSNPGTEATGTSPATPTP*'
```

We compare it to the one in BLAST and they are the same.

Results from transeq are different than those from Genomen bank -> Flip strands -> Sequence text viewer.

Question 3.6

The Caenorhabditis elegans (strain: Bristol N2) has a gene symbol ife-3, which is a protein coding gene type. The gene contains 4 exons, As mentioned earlier, the sequence is found at chromosome V (NC_003283.11)[6]. The sequence is also known as B0348.6, which goes under the other name of CELE_BO3048.6, the status of the species is "Live", gene name evidence: Eric Aamodt. Ife-3 is enables one to encode one of five C. elegans homologs[7].

[6]GeneID, https://www.ncbi.nlm.nih.gov/gene/178536
[7]WormBase, https://www.wormbase.org/species/c_elegans/gene/WBGene00002061#0-9g-3