732A96 Advanced Machine Learning

LAB 1 Bioinformatics

Hugo Morvan
William Wiik

STIMA
Department of Computer and Information Science
Linköpings universitet

2024-11-13

Contents

1 Question 1

We consider a gene locus with two possible alleles (say A and a) and a diploid population with N individuals. Hence, there are $2N$ alleles in the population. Let p be the proportion of As in the allele population and q the population of as (of course $p + q = 1$). A population is said to be in Hardy-Weinberg equilibrium if the proportion of AA homozygotes is p^2, aa homozygotes is q^2 and the proportion of heterozygotes (Aa) is $2pq$.

1.1 Question 1.1

Show that with random mating (i.e. both alleles of the o spring are just randomly, with proportions p and q, drawn from the parental allele population) Hardy-Weinberg equilibrium is attained in the first generation. What is the proportion of A and a alleles in the offspring population? Hence, with random mating, can a population in Hardy-Weinberg equilibrium ever deviate from it?

Answer:

If the mating is random then all individuals inherits alleles independently with probabilities p and q from each parent. Then:

$$P(AA) = p \cdot p = p^2$$
$$P(aa) = q \cdot q = q^2$$
$$P(Aa) = (p \cdot q) + (p \cdot q) = 2pq$$

The population can not deviate from Hardy-Weinberg equilibrium unless factors such as mutation changes the alleles.

1.2 Question 1.2

We look at the MN blood group, it has two possible codominating (both contribute to heterozygotes) alleles L^M (denoted M) and L^N (denoted N). In a population of 1000 Americans of Caucasian descent the following genotype counts were observed, 357 individuals were MM, 485 were MN and 158 were NN. Use a chisquare goodness of t test to test if the population is in Hardy-Weinberg equilibrium.

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

population <- 1000

# p = the proportion of Ms in the allele population
p <- (2 * MM + MN) / (2 * population)
q <- 1 - p

# Hardy-Weinberg equilibrium
HW_MM <- p**2  * population
HW_NN <- q**2  * population
HW_MN <- q * p  * population * 2


# chisq = sum[(Observed - Expected)^2 / (Expected)]
```

```r
chisq <- sum( (c(MM, MN, NN) - c(HW_MM, HW_MN, HW_NN))^2 / (c(HW_MM, HW_MN, HW_NN)) )
# df = 1 because we are testing against one constrain
p_value <- pchisq(chisq, df = 1, lower.tail = FALSE)
p_value
```

```
## [1] 0.7519044
```

There is no significant evidence to reject the null hypothesis, which means that we can not assume the population is NOT in Hardy-Weinberg equilibrium. Hence, we can conclude that the population is likely in Hardy-Weinberg equilibrium.

2 Question 2

In this exercise, you will use GenBank (https://www.ncbi.nlm.nih.gov/genbank/) with the default "Nucleotide" database. We will be interested in the sequence with accession number in GenBank: MK465080. You will find the relevant information in the FEATURES section of the returned record and access the nucleotides of the sequence under CDS (protein coding sequence, from CoDing Sequence). Remember that the coding strand (https://en.wikipedia.org/wiki/ Coding_strand) is the strand of the gene that is identical to the transcript (see the lecture slides for the genetic code—translation of DNA triples to amino acids). The complimentary to it strand is called the template strand.

2.1 Question 2.1

From what species does the sequence come from? Name the protein product of the CDS.

Answer: The sequence comes from the species *Branchipus schaefferi* and the protein product is named *cytochrome c oxidase subunit I*

2.2 Question 2.2

Save (and submit) the nucleotide sequence of the coding strand that corresponds to these amino acids as a FASTA format file.. Use transeq (https://www.ebi.ac.uk/Tools/st/emboss_ transeq/) to translate the nucleotides to a protein. Do you obtain the same protein sequence? Check what is the ORF and codon table (these are provided by GenBank in the FEATURES section). Use backtranseq (https://www.ebi.ac.uk/Tools/st/emboss_backtranseq/) to obtain the sequence from the protein sequence.

2.3 Question 2.3

Compare your obtained coding strand sequence with the nucleotide sequence provided (when following the CDS link). Are they the same or do they differ? Try reversing and taking the complement (e.g., http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp. html or http://www.bioinformatics.nl/cgi-bin/emboss/revseq or write your own code) of the your coding strand DNA. Consider also backtranseqambiq (https://www.ebi.ac.uk/Tools/st/emboss_backtranseqambig/ and check if the resulting nucleotide sequence is compatible with the true one. Do not forget to check the codon table. Explain what happened and why. Save (and submit) the nucleotide sequence of the template strand that corresponds to these amino acids as a FASTA format file.

3 Question 3

Eukaryotic genes are commonly divided into exons and introns and these needed to be spliced in order to produce an mRNA that can be translated into a protein. The gene starts with the promoter ("region of DNA that initiates transcription, i.e. DNA→RNA, of a particular gene", see https://en.wikipedia.org/wiki/Promoter_(genetics)), the the first exon, first intron, second exon, e.t.c. Multiple introns an alternative splicing (a single gene codes for different proteins, through different exons used, see https://en.wikipedia.org/wiki/Alternative_splicing) of mRNAs can make it difficult to identify genes. Finding genes and how they are organized is often due to searching for similar nucleotide sequences within already know protein amino acids, or rather their nucleotide sequences and the corresponding full–length cDNAs (complementary DNA—DNA synthesized from a single stranded RNA, see https://en.wikipedia. org/wiki/Complementary_DNA). cDNAs come from back transcription (reverse–transcription) of mRNAs and hence are without introns—and can be considered as equivalent to mRNA sequences. Comparing a part of the genome (that contains introns) with its cDNA will show the introns' start and end points. GenBank (https://www.ncbi.nlm.nih.gov/genbank/) contains both cDNA sequences and corresponding genomic sequences (if available). If order to discover the structure of the gene we need to compare the cDNA with the genomic sequence. In the file 732A51 BioinformaticsHT2024 Lab01Ex03.fasta you can find a genomic sequence from the species C. elegans, that contains a particular gene. You will use the Basic Local Alignment Sequence Tool (BLAST), to find the gene's organization. BLAST is used to compare a query sequence all sequences (i.e., cDNA sequences) in GenBank. Usually the top scoring hit is the one you want. The next ones will be less and less similar. It can happen that all hits have 100% identity—then consider the percent coverage.

3.1 Question 3.1

Read up on **C. elegans** and in a few sentences describe it and why it is such an important organism for the scientific community.

3.2 Question 3.2

Use the nucleotide BLAST tool to construct a schematic diagram that shows the arrangement of introns and exons in the genomic sequence. In the BLAST tool, https://blast.ncbi.nlm. nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch, choose database RefSeq Genome Database and remember that the species source of the genomic sequence is *Caenorhabditis elegans*. Use the Genome Data Viewer button. Alternatively you may use https://wormbase.org/tools/blast_ blat.

3.3 Question 3.3

How are the sequences numbered in the alignment (i.e., pairing of query and database sequences)? Are the directions the same? What would happen if you reverse complement your query sequence (e.g., http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp. html or http://www.bioinformatics.nl/cgi-bin/emboss/revseq or write your own code) and do the search with such a reverse complemented sequence?

3.4 Question 3.4

On what chromosome and what position is the query sequence found? At which position does the gene begin and end in your query sequence?

3.5 Question 3.5

Extract the DNA code of each exon and using transeq (https://www.ebi.ac.uk/Tools/st/ emboss_transeq/) find the protein code of the gene. You can also use blastx (https://blast. ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&PAG or https://wormbase.org/tools/blast_blat) to obtain protein sequences. How do they compare to your translation?

3.6 Question 3.6

What gene is in the query sequence? Hovering over an exon you should see links to View GeneID and View WormBase. These point to pages with more information on the gene. Follow them and write a few sentences about the gene.