Lab 1 - Gr. 14 - Bioinformatics (732A93)

tbd

Assignment 1

Question 1: Hardy–Weinberg equilibrium

Question 1.1

We consider a diploid population of N individuals and we focus our attention on a gene locus with two possible alleles, A and a. The resulting population will be something like the following:

 AA   Aa   AA   aa   aA   aA   Aa   …..

We then define the following quantities:

- n_{AA}: the population of AA homozygotes subjects in the parental generation;

- n_{aa}: the population of aa homozygotes subjects in the parental generation;

- n_{Aa}: the population of heterozygotes subjects in the parental generation;

- p: the proportion of As in the allele population, given by $\frac{n_{AA}}{N} + \frac{1}{2} \cdot \frac{n_{Aa}}{N}$;

- q: the proportion of as in the allele population, given by $\frac{n_{aa}}{N} + \frac{1}{2} \cdot \frac{n_{Aa}}{N}$;

- $p + q = 1$

Our goal is to demonstrate how the defined population will be able to come into a state called *Hardy–Weinberg equilibrium*, in only one generation obtained through random mating (both alleles of the offspring are just randomly, with proportions p and q, drawn from the parental allele population). A population in this state present the proportion of AA homozygotes equal to p^2, the proportion of aa homozygotes equal to q^2 and the proportion of Aa/aA heterozygotes equal to $2pq$.

Question 1.2

We now look at the MN blood group, that has two possible co–dominating alleles, L^M (denoted M) and L^N (denoted N). In a population of $n = 1000$ Americans of Caucasian descent, the following genotype counts were observed:

- $n_{MM} = 357$ individuals were MM homozygotes;

- $n_{MN} = 485$ individuals were MN heterozygotes;

- $n_{NN} = 158$ individuals were NN homozygotes;

The relatives proportion, obtained by dividing the genotype counts by the total population, are 0.357, 0.485 and 0.158 respectively. According to the Hardy–Weinberg principle, we expect these quantities to follow the proportions p^2, $2pq$ and q^2, where p and q are the proportions of N and M alleles. In formulas:

$$p = \frac{n_{MM}}{n} + \frac{1}{2} \cdot \frac{n_{MN}}{n} = 0.357 + \frac{1}{2} \cdot 0.485 = 0.5995$$

$$q = \frac{n_{NN}}{n} + \frac{1}{2} \cdot \frac{n_{MN}}{n} = 0.158 + \frac{1}{2} \cdot 0.485 = 0.4005$$

(1)

Performing a chi–square goodness of fit test (`chisq.test()` function in R) result in *p-value* $= 1$, meaning that the hypothesis of the existance of a Hardy–Weinberg equilibrium is accepted.

```
## Warning in chisq.test(c(0.357, 0.485, 0.158), p = c(p_2, pq, q_2)): Chi-
## squared approximation may be incorrect

##
##  Chi-squared test for given probabilities
##
## data:  c(0.357, 0.485, 0.158)
## X-squared = 9.9938e-05, df = 2, p-value = 1
```

Assignment 2

Note that, by convention, a *coding strand* is used when displaying a DNA sequence. "A coding strand, is the segment within double-stranded DNA that runs from 5' to 3', and which is complementary to the antisense strand of DNA, or template strand, which runs from 3' to 5"' (https://en.wikipedia.org/wiki/Sense_strand).

2.1

We go to the following link https://www.ncbi.nlm.nih.gov/nuccore/CU329670, scroll down to "Features", click on "CDS" to see the first 5662 nucleotides of the sequence.

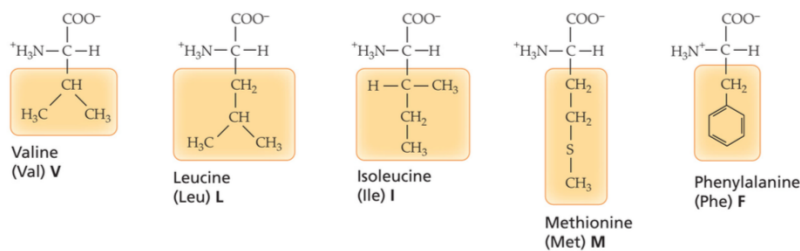The protein product is: **RecQ type DNA helicase**

2.2

Note that proteins are made up of 20 different amino acids (linked together). Amino acids are molecules composed of an alpha carbon, a carboxyl group, an amino group and a side chain. The side chain is what makes an amino acid unique (see p. 5, Concepts in Bioinformatics and Genomics). The 20 different amino acids can be found in the picture below.
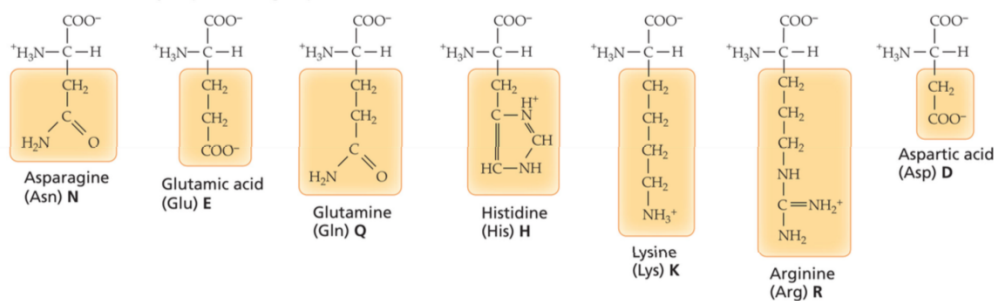
In our case, the first four amino acids are:

1. M (= **Methionine**, coded by ATG, representing the starting sequence)
2. V (= **Valine**)
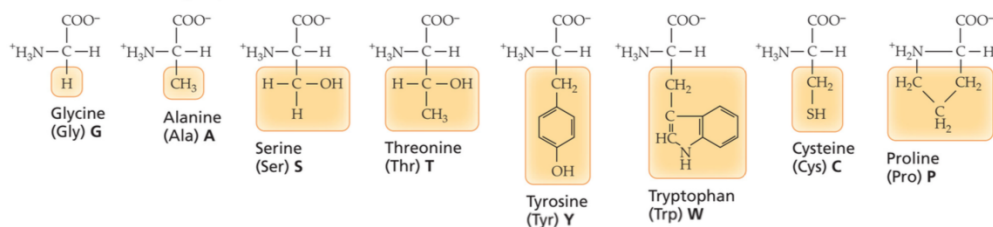3. V (= **Valine**)
4. A (= **Alanine**)

Amino acids with hydrophobic side groups

Valine (Val) V · Leucine (Leu) L · Isoleucine (Ile) I · Methionine (Met) M · Phenylalanine (Phe) F

Amino acids with hydrophilic side groups

Asparagine (Asn) N · Glutamic acid (Glu) E · Glutamine (Gln) Q · Histidine (His) H · Lysine (Lys) K · Arginine (Arg) R · Aspartic acid (Asp) D

Amino acids with side groups that are in between

Glycine (Gly) G · Alanine (Ala) A · Serine (Ser) S · Threonine (Thr) T · Tyrosine (Tyr) Y · Tryptophan (Trp) W · Cysteine (Cys) C · Proline (Pro) P

2.3

Note that nucleotides code for amino acids. They are molecules composed of a sugar, a phosphate and a base. The part of nucleotides that distinguishes them is the base.

There are four possible bases: A (Adenine), C (Cytosine), G (Guanine), T (Thymine)

TABLE 1-1. SINGLE-LETTER ABBREVIATIONS USED FOR DNA NUCLEOTIDE SEQUENCES

ONE-LETTER ABBREVIATION	NUCLEOTIDE NAME	BASE NAME	CATEGORY
A	Adenosine monophosphate	Adenine	Purine
C	Cytidine monophosphate	Cytosine	Pyrimidine
G	Guanosine monophosphate	Guanine	Purine
T	Thymidine monophosphate	Thymine	Pyrimidine
N	Any nucleotide	Any base	NA
R	A or G	A or G	Purine
Y	C or T	C or T	Pyrimidine
– or *	—	—	Gap

3

Specific combinations of these nucleotides code for specific amino acids. In genetic code, three nucleotides (called codon) always code for one amino acid. E.g. ATG codes for met (Methionine) which is a starting point of a protein. Which codos code for which amino acids is shown in the overview below:

1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA[B] Stop (Ochre)		TGA[B] Stop (Opal)		A
	TTG		TCG		TAG[B] Stop (Amber)		TGG (Trp/W) Tryptophan		G
C	CTT	(Leu/L) Leucine	CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A] (Met/M) Methionine		ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

Saving the nucleotide sequence from GenBank:

The complete nucleotide sequence of the coding strand from GenBank (https://www.ncbi.nlm.nih.gov/nuccore/CU329670.1?from=1&to=5662) was saved in FASTA format as **2.3_nucleotid-sequence.FASTA**. Note that only the last 12 characters AGCGACGACCAT actually correspond to the amino acids MVVA if the reverse compliment is taken (see 2.4).

Using backtrack to obtain the amino acid sequence from the protein sequence:

We used backtrack to obtain the amino acid sequence corresponding to the protein sequence `MVVA` (https://www.ebi.ac.uk/Tools/st/emboss_backtranseq/). After pasting `MVVA` and selecting "Schizosaccharomyces pombe (CAI equivalent)", the following sequence is returned: **ATGGTCGTCGCT**

2.4

The above mentioned obtained coding strand sequence ATGGTCGTCGCT does not exist in the provided nucleotide sequence (that was saved in FASTA format as 2.3_nucleotid-sequence.FASTA) since this sequence is not found when searching the file.

Option 1

However, we can modify the displayed sequence to get what we are looking for. On GenBank, under "Display options", one has to select both "Show sequence" and **"Show reverse complement"** (https://www.ncbi.nlm.nih.gov/nuccore/CU329670.1?from=1&to=5662). Afterwards, the displayed nucleotid sequence under "ORIGIN" starts with ATGGTCGTCGCT which codes for our amino acids MVVA.

Option 2

Alternatively, one can take the last 12 characters from the original nucleotid sequence (the one in the FASTA file, i.e. the sequence before selecting "Show reverse complement" on GenBank). These characters are: AGCGACGACCAT. Copy this string and paste it here: http://arep.med.harvard.edu/labgc/adnan/projects/

Utilities/revcomp.html Then click **"reverse complement"**. The result is again the sequence that we are looking for: ATGGTCGTCGCT which codes for our amino acids MVVA.

Conclusion

ATGGTCGTCGCT is exactly what we got in 2.3 using backtranseq to obtain the amino acid sequence from the protein sequence `MVVA`. The only tricky thing here was that we needed to take the reverse complement of the correct characters from the nucleotid sequence (namely the last 12 characters).

2.5

Number range that corresponds to these amino acids (protein sequence):

In the saved FASTA file, the *last 12 characters* AGCGACGACCAT correspond to the amino acids MVVA (after taking the reverse complement). The number range corresponding to them is 5651 to 5662.

If we had selected "Show reverse complement" on GenBank before, then the *first 12 characters* ATGGTCGTCGCT would correspond to the amino acids MVVA. The number range would then be 1 to 12.

Stop codon in the nucleotide sequence:

Note that there are three possible stop codons: TAA, TAG, TGA. Still, it is not easy to identify the stop codon manually. However, one can use: https://www.ncbi.nlm.nih.gov/orffinder and paste the complete (reverse complement) nucleotide sequence in FASTA format. With the default parameters, this is automatically translated to the correct amino acid sequence. Also, after scolling down a bit, one can see that the last stop happens at 5661. We find TC from 5661 to 5662 (and at 5662, we have the last nucleotid). From 5658 to 5660, we have TGA which is one of the above mentioned stop codons. We can conclude that the stop codon in the nucleotid sequence is hence TGA from 5658 to 5660 (looking at the reverse complement nucleotid sequence).

Chromosome on which the genomic sequence lies:

We can see on GenBank (in the header under the definition as well as in the features under source), that this genomic sequence lies on the chromosome 1 of schizosaccharomyces pombe.

Assignment 3

3.1

3.2

Appendix

```
knitr::opts_chunk$set(fig.width = 7, fig.height = 3, echo = FALSE)

library(dplyr)
library(tidyr)
library(magrittr)
library(kableExtra)


# --------------------------------------------------------------------------------
# Question 1.2
# --------------------------------------------------------------------------------
```

```r
# The prob of one allele is the proportion of the homozigote + half the proportion of the heterozygote
p <- 0.357 + 0.485/2
q <- 0.158 + 0.485/2

# Expected values (Hardy-Weinberg equilibrium)
p_2 <- p^2
q_2 <- q^2
pq <- 2*p*q

Xsq <- chisq.test(c(0.357, 0.485, 0.158), p = c(p_2, pq, q_2))
Xsq
# Xsq$observed
# Xsq$expected

knitr::include_graphics("images/2.2_amino-acids.png")
knitr::include_graphics("images/2.3_nucleotides.png")
knitr::include_graphics("images/2.3_codons-aminoacids")
# -------------------------------------------------------------------------
# Question 3.1
# -------------------------------------------------------------------------


# -------------------------------------------------------------------------
# Question 3.2
# -------------------------------------------------------------------------
```