Computer Lab 1
Bioinformatics

Linköpings Universitet, IDA, Statistik

2018/11/07

Kurskod och namn:	732A51 Bioinformatics
Datum:	2018/11/06—2018/11/14 (lab session 07 November 2018)
Delmomentsansvarig:	Krzysztof Bartoszek
Instruktioner:	This computer laboratory is part of the examination for the Bioinformatics course Create a group report, on the solutions to the lab as a **.PDF** file. Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments. **All R code should be included as an appendix into your report.** In the report reference **ALL** consulted sources and disclose **ALL** collaborations. The report should be handed in via LISAM (or alternatively in case of problems e–mailed to krzysztof.bartoszek@liu.se), by **23:59 14 November 2018** at latest. Notice there is a final deadline of **23:59 3 February 2019** after which no submissions nor corrections will be considered and you will have to redo the missing labs at the next course opportunity. The report has to be written in English.

Question 1: Hardy–Weinberg equilibrium

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 $2pq$.

Question 1.1

Show that with random mating (i.e. both alleles of the offspring 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?

Question 1.2

We look at the MN blood group (`https://en.wikipedia.org/wiki/MNS_antigen_system`), it has two possible co–dominating (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 chi–square goodness of fit test to test if the population is in Hardy–Weinberg equilibrium.

Question 2: Exploring a genomic sequence

This is exercise 2.1 from J. Momand, A. McCurdy. Concepts in Bioinformatics and Genomics, Oxford, 2017. Oxford University Press.

For this exercise, you will need to access GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`) by going to the NCBI website and using the dropdown menu to search "Nucleotide". Note that the definition of the coding strand (`https://en.wikipedia.org/wiki/Coding_strand`) is the strand of DNA within the gene that is identical to the transcript (see e.g. the lecture slides for the genetic code, i.e. translation of DNA triples to amino acids). On the other hand, the template strand is the strand that is complimentary to the coding strand.

Use the following accession number to access the nucleotide sequence in GenBank: CU329670. For the following questions you need to go to the FEATURES section of the record (scroll down) and there link to the CDS (protein coding sequence, from CoDing Sequence) to gain access to the first 5662 nucleotides of the sequence.

Question 2.1

Name the protein products of the CDS.

Question 2.2

Write the first four amino acids.

Question 2.3

Save (and submit) the nucleotide sequence of the coding strand that corresponds to these amino acids as a FASTA format file.. Use `backtranseq` (`https://www.ebi.ac.uk/Tools/st/emboss_backtranseq/`, note species used) to obtain the sequence from the protein sequence.

Question 2.4

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. 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.

Question 2.4

Using the sequence shown in the record, give the nucleotide number range that corresponds to these amino acids (protein sequence). Find and report the stop codon in the nucleotide sequence. On which chromosome does the genomic sequence lie?

Question 3: Exploring a genomic sequence

This is exercise 2.2 from J. Momand, A. McCurdy. Concepts in Bioinformatics and Genomics, Oxford, 2017. Oxford University Press.

Genes in eukaryotes are often organized into exons and introns, which require splicing to produce an mRNA that can be translated. The gene organisation is the order of the DNA segments that comprise the gene starting 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 first exon, the first intron, the second exon and so on. The interspersed introns can make gene identification difficult in eukaryotes—particularly in higher eukaryotes with many introns and alternative spliced (a single gene codes for different proteins, through different exons used, see `https://en.wikipedia.org/wiki/Alternative_splicing`) mRNAs. Prediction of many genes and their organisation has been based on similarity searches between genomic sequence and known protein amino acid sequences and genomic sequence and the corresponding full–length cDNAs (complementary DNA—DNA synthesized from a single stranded RNA, see `https://en.wikipedia.org/wiki/Complementary_DNA`). cDNAs are reverse–transcribed mRNAs and therefore do not contain intron sequences, hence cDNAs can be considered mRNAs. A comparison of genomic sequence (with introns) to its corresponding cDNAs will reveal where introns begin and end. GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`) will contain the genomic sequence and the cDNA sequence. To find out the structure of the gene (i.e. the arrangement of the exons and introns), we simply need to perform a sequence comparison between the genomic sequence and the cDNA sequence. In the file `732A51_BioinformaticsHT2018_Lab01Ex03.fasta` you can find a genomic sequence from the species *C. elegans*, you may also find the DNA sequence at `http://global.oup.com/us/companion.websites/9780199936991/stu_res/eoc/tex/` . The Basic Local Alignment Sequence Tool (BLAST), can be used to elucidate part of the gene organisation (arrangement of exons and introns) of a genomic sequence. BLAST can be used to compare genomic DNA with all RNA sequences (i.e., cDNA sequences) in GenBank. The top hit of the output will be a sequence comparison between your sequence (the query sequence) and the most similar sequence in the database (subject sequence). Subsequent hits will display sequence comparisons between the query sequence and subject sequences that are increasingly less similar. If all hits have 100% identity, use the hit with the most extensive percent coverage to report on.

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.

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.

Question 3.3

Note the numbering of the sequences in the alignment (i.e. pairing of query and database sequences). Does the database genomic sequence progress in the same direction as the query sequence? 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?

Question 3.4

On what chromosome and what position is the query sequence found?

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&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome`) to obtain protein sequences. How do they compare to your translation?

Question 3.6

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.