Computer Lab 2
Bioinformatics

Linköpings Universitet, IDA, Statistik

2024 XI 22

Kurskod och namn:	732A51 Bioinformatics
Datum:	2024 XI 14—2024 XI 25 (lab session 22 XI 2024 SU25)
Delmomentsansvarig:	Krzysztof Bartoszek, Ying Luo
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 ying.luo@liu.se), by **23:59 25 November 2024** at latest. Notice there is a deadline for corrections **23:59 2 February 2025** and a final deadline of **23:59 2 March 2025** 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. Some questions are marked with an asterisk. From these **at least one** must be done, according to your choice.

Question 1: DNA sequence acquisition and simulation

In this exercise you will perform statistical analysis of three nucleotide data sets. First download the sequences from GenBank and save them in a fasta file. For this use the provided `R` script, `732A51_BioinformaticsHT2023_Lab02_GenBankGetCode.R`. This is a dataset of the RAG1 gene sequences from 33 lizard species. You are encouraged to read in detail the references in the script as they indicate many useful tools. Explore the dataset using the tools provided by the `ape` and `seqinr` packages. Take note of the lengths of all the sequences and the base composition.

Question 1.1

Simulate an artificial DNA sequence dataset. It should contain 33 sequence. The lengths of the sequences should be the same as in the lizard dataset, i.e. for each real sequence simulate an artificial one. The simulation rule is as follows, each nucleotide is to be independently and randomly drawn from the distribution given by the base composition (frequencies) in the true lizard sequences. Save your dataset in a fasta format file. Remember to give unique names to your sequences. Report on the base composition in your simulated data.

Question 1.2*

Simulate a second artificial DNA sequence dataset. It should contain 33 sequence. The lengths of the sequences should be the same as in the lizard dataset, i.e. for each real sequence simulate an artificial one. First simulate a phylogenetic tree with 33 tips in `phylo` format (i.e. `ape`). Plot your resulting tree. For simulating the tree explore the functions of the `ape`, `TreeSim` or other `R` packages. Choose a simulation function and model yourself.

Now simulate sequences on this using e.g. `phangorn::simSeq()`. Choose the sequence length yourself, but try to make it so that it will be comparable with the original lizards dataset. You need to also specify the `Q` matrix—the transition rate matrix. Choose one yourself, however try to make the stationary distribution equal to the base composition (frequencies) of the lizard sequences (look at EG Ch. 14.3.3). If you cannot obtain such a transition matrix, choose some another one. Save your dataset in a fasta format file. Remember to give unique names to your sequences. Report on the base composition in your simulated data. Comment on if it is what you expect.

HINT: It could be useful to read `https://sites.google.com/site/eeob563/computer-labs/old-labs/lab-2` and `https://en.wikipedia.org/wiki/Models_of_DNA_evolution#F81_model_(Felsenstein_1981)`

Question 2: Sequence analysis

Question 2.1

Report some basic statistics on each sequence dataset: individual base composition, GC content, CG, AT content. Also translate your sequences into protein sequences (see Lab 1) and report on the amino acid composition. In your simulated sequences, how many times did you observe a stop codon inside your sequence? Does this occur in your true sequences? Comment.

Question 2.2*

Try to fit a Markov chain to your three datasets from Question 1. What Markov chain order would you expect to obtain for your two simulated datasets? What order do you obtain for the true lizard sequences? Comment

Explore `R` packages (or other software) to find a package to fit the Markov chain. The `R` package `markovchain` can be your starting point. If your chosen software does not support multiple samples from the chain, think how to deal with this, e.g. concatenate the sequences or estimate separately for each sequence and propose a way to pool the results.

Think of the assumptions behind the analyses and how your data could be or is violating them. You do not need to correct for the violations but should report them.

Question 2.3

Align your sequences using software of your choice (a starter for `R`: `https://stackoverflow.com/questions/4497747/how-to-perform-basic-multiple-sequence-alignments-in-r`, you can also look what `Biopython`, `BioPerl` offer, use the `Clustal` family of programs or something else of your choice).

Choose a distance measure between sequences, calculate for each alignment the distances between all pairs of sequences. Then, plot heatmaps visualizing the distances. Comment on what you can observe.

Question 3: Phylogeny reconstruction

Question 3.1

Construct (using algorithm and software of your choice) phylogenetic trees from the three multiple alignments (or distance matrices) done in Question 2.3. You might want to look at the functions offered by `ape`, `phangorn` (`https://cran.r-project.org/web/packages/phangorn/vignettes/Trees.pdf`) or go for some completely different software. Plot the inferred trees. Are the two based on the simulated data similar to expected?

Perform a phylogenetic bootstrap analysis and report the bootstrap support for the individual clades, you can look at `ape::boot.phylo()`.

Question 3.2*

Compare your inferred trees and also your simulated one. Apart from visualizing the trees one may calculate various indices related to them and distances between the trees. Explore what indices and metrics the `ape`, `distory`, `phangorn`, `phyloTop`, `TotalCopheneticIndex` or `treespace` `R` packages offer, choose some and report the results in a meaningful way. You might have to save your tree to drive and then read in it using e.g. `ape`'s tree reading functionality.