8CM00 Systems Medicine

Assignment 2 Sequence alignment, Bacteria and Graphs, October 26, 2022

Peter Hilbers

The assignment of module 2 consists of part A and part B.

Part A: Sequence alignment

In this exercise a global and a semi-global alignment are to be considered. In all cases an algorithm should be designed to solve the problem. In the first item your own algorithm should be designed, so calls to methods of other programming packages are not allowed. A discussion of the complexity of the algorithm should be included, and the efficiency of your solution is an important aspect in the grading.

1. Design an algorithm to find the optimal global alignments between ACCAATTACCAATTAAG and AATGA using a score of +1 for a match, −2.5 for a mismatch, and −5 for opening and −2 for extending a gap. Include in your solution the total score matrix that your dynamic programming algorithm is generating. When equal maximal scores are produced, all solutions should be generated. The algorithm(s) used to produce the score and the alignment should also be shown and discussed.
2. Here and in the next item BioPython should be used. During the lectures the reference Covid-19 virus (Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)), has been discussed. Design a program by which the reference sequence is obtained both in 'genbank' and 'fasta' format.
3. Each student will have received its own second Covid sequence by email (If not please send me an email). The task is to do an as complete as possible analysis of the differences between that sequence and the reference sequence based upon their semi-global alignment using the same score function as above. The assumption is that matrices with a maximum size of 5000×5000 fit into memory, so a semi-global alignment of both sequences as a whole is not possible. Usage of `globalxx` of BioPython on shorter sequences is allowed. When multiple alignments have a maximum score, then no more than 5 solutions need to be discussed.

In your submission you should include at least answers to the following questions:

(a) Is there more than one solution? What are the differences between the solutions?

(b) What is the number of matches?

(c) What are the mutations? Do they occur in non-coding parts or in the coding parts? When in the coding parts, do they occur in a spike-protein or in other proteins? To which variant does your sequence belong? (On `https://covariants.org` and on NCBI Virus SARS-CoV-2 – Variants Overview fine overviews are given of the variants and mutations).

(d) How dependent is your solution on the score function?

(e) Is the solution approach also applicable to other long sequences? What are the weak points? What are the strong points?

Especially in the semi-global alignment exercise the approach followed should clearly be described. Include why the solution is correct and why it is optimal etcetera. Extra credits will be given to new, creative ideas.

Part B: Bacterial growth, symbiosis

In this assignment the metabolic interactions between pairs of prevalent bacteria of the human colon microbiota are to be investigated. By email you will receive 3 names of bacteria, hence the analysis has to be done for 3 pairs of bacteria.

As discussed during the lectures there are several types of interactions between species. In particular cross-feeding may occur and in the article The Classification and Evolution of Bacterial Cross-Feeding" by Smith et al, Frontiers in Ecology and Evolution, May 2019, Volume 7, Article 153 several types of cross-feeding are discussed.

In the second part of the assignment the question should be addresssed whether the bacteria experience cross-feeding and if so, to classify the type of cross-feeding and type of symbiosis.

In the first part the metabolism graphs of the bacteria are to be analysed and their differences discussed. To that end construct for each of the bacteria the metabolite(substrate) graph, and the metabolite-reaction graph. Assume that these graphs are undirected and unweighted. For each of these graphs analyse its static properties: number of nodes, edges, degree distribution, hubs, connectedness, distance density function, scale-freeness, assortativity.

Next flux balance analyses(FBA) are to be performed for different media.

- For each of the bacteria run an FBA in isolation with as objective function the bacteria's biomass reaction and as medium the one based upon its sbml-file. Define the length of a path as the sum of the fluxes on the path. What are the lengths of the shortest paths from each component of the medium to the biomass? How different are the 3 bacteria?
- Next assume that the medium is DMEM(`https://www.thermofisher.com/nl/en/home/life-science/cell-culture/mammalian-cell-culture/classical-media/dmem.html`) and perform again an FBA for each of the bacteria in isolation with objective function the bacteria's biomass reaction. When the bacteria cannot grow in this medium then add in a systematic way supplements such that the bacteria can grow. Clearly describe how and why these supplements have been added. Detrmine the lengths of the shortest paths from each component of the medium to the biomass? How different are the 3 bacteria? What are the differences as a result of the change in medium?
- Similarly as the previous exercise, but now start with an "empty medium" and with as objective function the bacteria's biomass reaction. Add in a systematic way nutrients to the medium such that the bacteria can grow. Clearly describe how and why these supplements have been added.
- Next the symbiosis between the 3 pairs of bacteria is to be studied. For each of the pairs the following procedure should be followed: First both bacteria are grown in a medium, but there is no interaction possible between them. Next, interaction in the medium is possible where this interaction has to be described clearly, e.g. which of the outputs of one bacteria are inputs to the other. The deficits or profits of each of the bacteria should adequately be shown and where possible explained. Clearly describe the media used and motivate your choice.

Submit your solution (including the codes used) on Canvas in a zip-file.
Deadline for submission of part A and part B: November 13, 2022, 23:59:59.

Success!!!!