BIO310 Introduction to Bioinformatics
Computer Lab 3 and Homework 2 Spring 2019

March 7, 2019

Instructions:

- We expect you to start working on this assignment in the lab, at the end of the lab you will submit how far you came along. Your overall effort will be graded and it will eventually contribute to your lab grade. This grade will be assigned as a number from 1-5; no effort being 1 and full effort being 5. You are not expected to finish the entire assignment during the lab; you will have a chance to submit the final version till the due date and this will be your homework 2 grade, which is out of 100.
- For the homework submission, submit a PDF document for the answers of the write-up questions, the plots should be appropriately labeled, figures should have captions and should be appropriately cited within the main text. Name your submission as `BIO310-HWx-YourName.pdf` where you substitute in your first and last names into the file name in place of 'YourName' and X with the current homework number. Upload your final report on SuCourse by the due date.
- Upload the code online on SuCourse by the due date. You may code in any programming language you may prefer, but your assistant will only provide help in Python. The code you submit should be in a format that is ready to run and you should include a README file that specifies how to run the code in our machines. In submitting the code on SuCourse, compress it as a ZIP file with the name `BIO310-HWXcode-YourName.zip` where you substitute in your first and last names into the file name in place of 'YourName' and X with the current homework number.
- If you are considering to submit the homework late, please see the late submission policy in the syllabus.
- Please follow the submission instructions, not adhering the submission standards will lead to point deduction.

Part 1: A little puzzle

0	-3	-6	-9	-12	-15	-18	-21	-24	-27
-3	-5	-8	2	-1	-4	-7	-10	-13	-16
-6	-8	3	0	-3	-6	4	1	-2	-5
-9	-11	0	11	8	5	2	-1	-4	-7
-12	-14	-3	8	19	16	13	10	7	4
-15	-17	-6	5	16	14	11	8	5	2
-18	-20	-9	2	13	11	22	19	16	13
-21	-10	-12	-1	10	21	19	17	27	24
-24	-13	-2	-4	7	18	29	26	24	22
-27	-16	-5	-7	4	15	26	37	34	31
-30	-19	-8	-10	1	12	23	34	32	29

Figure 1: The dynamic programming matrix of two sequences.

In Figure 1, you have the scoring matrix of a global pairwise sequence alignment for a pair of sequences generated with a positive match score, m, a mismatch score s and a gap penalty g.

1. Looking at the matrix content discover what m, s and g values should be.
2. Come up with a pair of DNA sequences that could lead to such a scoring matrix.
3. Figure out the trace back path, show it in the scoring matrix clearly.
4. Finally, write down the alignment.

Part 2: Global sequence alignment

1. Implement the global sequence alignment algorithm with linear gap penalty. The user should be able to specify the input filename, and the mismatch, gap penalty and match scores. The two DNA sequences should be in two separate lines of the input file. The first sequence will form the rows of the scoring matrix, the second sequence will be the columns.
2. You should write the output into an output file. The file should include, the alignment of the two sequences. The score achieved by the alignment and the scoring matrix produced.
3. Test your program with several test cases you came up. Especially test edge cases carefully. For example, how would your algorithm run if two very short strings are input, for example 'A' vs 'T' alignment.
4. Input the two sequences you came up in Part 1, the s,m and g values. Did it produce the same scoring matrix in Part 1.
5. We will provide you additional test cases separately and you will submit the output of these test cases.