Biocomputing:

HW IV

**Important Notes**

1. What to submit:
2. To Canvas: This document with your answers to each question. You can type the answers or include a picture of your hand note.

1. Do not delete the questions in the document. Just add your answers below each question.
2. **For groups**: Submit group member and self-evaluation form: **Each group member must submit this form** to give score for themselves and their fellow group members about their contribution to the project. The form is available at https://forms.office.com/r/kKTUx62gm7
3. Remember to write your name(s) to the header above.
4. Note academic integrity policy reminder below.

Code for this homework: <https://colab.research.google.com/drive/1MM7ma71P2NH3rxS9fSW8-4T76SQwoQBj?usp=sharing>

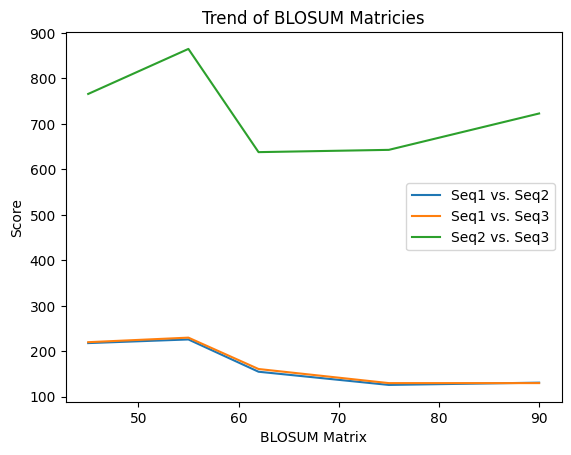
**Pairwise Sequence Alignment**

1. **Three protein sequences with ID are provided in seq.fasta file. Perform global sequence alignment using EMBOSS-Needle in Online tool (**[**https://www.ebi.ac.uk/Tools/psa/emboss\_needle/**](https://www.ebi.ac.uk/Tools/psa/emboss_needle/)**) between each pair of these sequences. Click “More options” button, adjust the substitution matrix while using default parameters for others.**

Chart, line chart, histogram

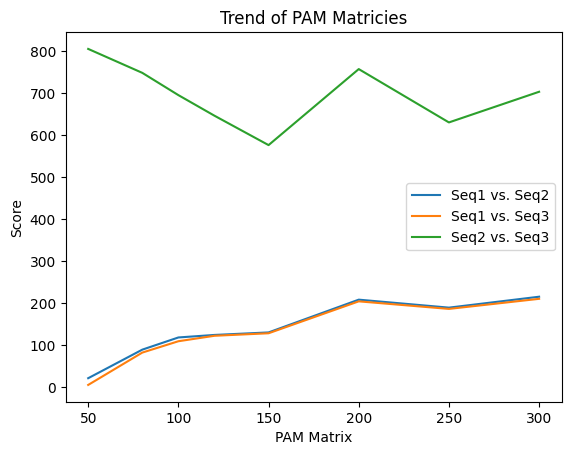
Description automatically generated

1. **(10 pts.) Perform sequence alignment for all sequence pairs using BLOSUM45, BLOSUM55, BLOSUM62, BLOSUM75, BLOSUM90 and plot the alignment score for each substitution matrix choice (see sample figure below). Your plot needs to have three curves, one for each sequence pair. Use a large gap penalty (i.e., gap open 100, gap extend 10). What is the trend of each curve (Does it stay same, increase, or decrease)? Explain why this trend occurs.**



The trend of the curves for Seq1 vs. Seq2 and Seq3 is a shallow downward trend, in comparison to the trend for Seq2 vs. Seq3. The trend peaks at a score of about 240 at matrix 55, and drops in a curve after that. The curve for Seq2 vs. Seq3 is highly erratic in comparison, peaking at matrix 55 with a score of about 875, dropping immediately to the low of about 625 at matrix 62, then having a comparatively shallow upward trend. The overall trend of all three sequence pairs is downward. Since lower numbers BLOSUM matrices are for distantly related proteins and higher numbers are for more closely related proteins, the overall trends indicated that sequence 1 is unrelated to the two others while sequences 2 and 3 are fairly related.

1. **(10 pts.) Repeat the alignment using PAM50, PAM80, PAM100, PAM120, PAM150, PAM200, PAM250, PAM300 substitution matrices. Once again, use a large gap penalty (i.e., gap open 100, gap extend 10). Plot the alignment scores and interpret the results as above.**

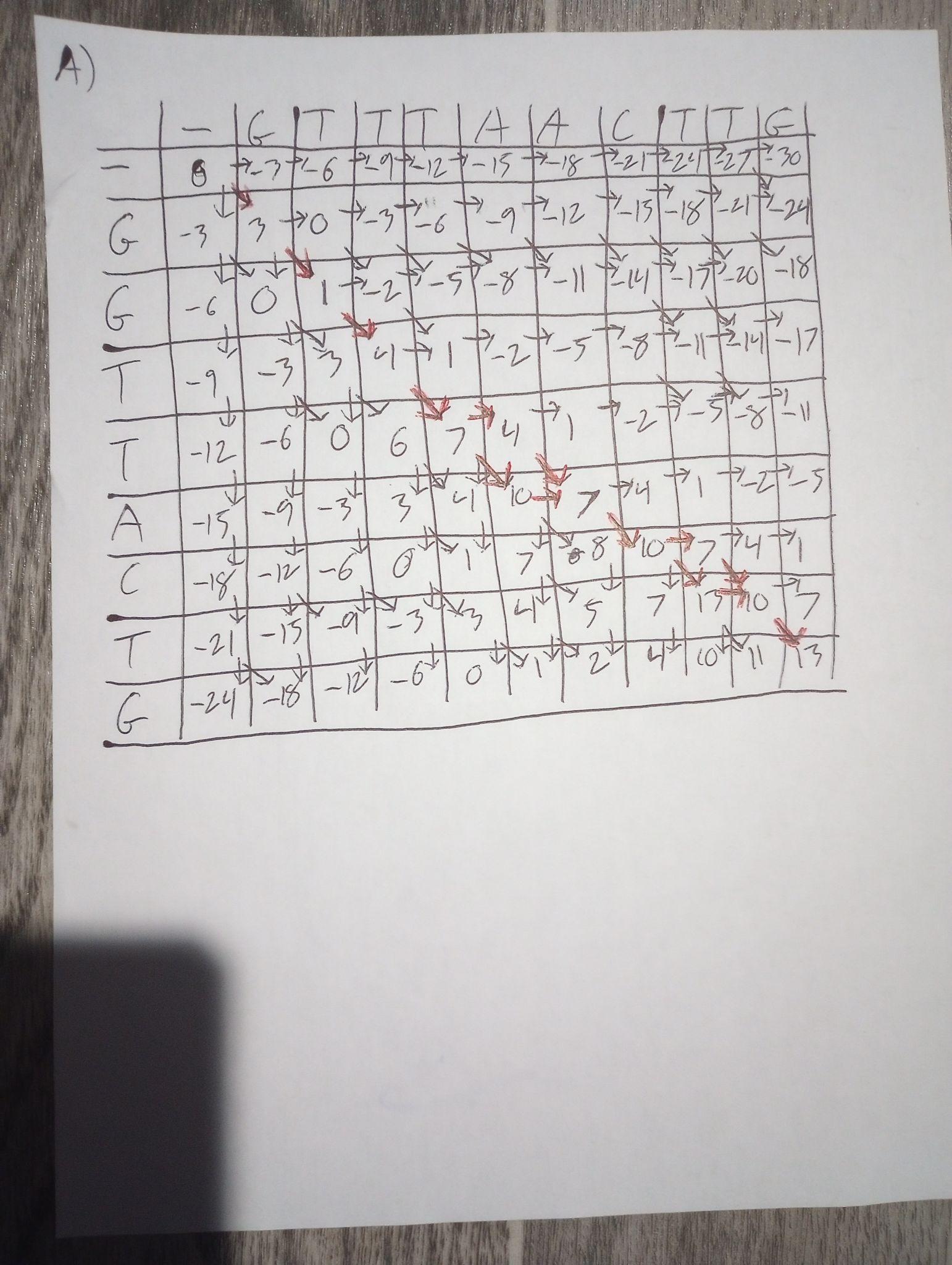


The trend of Seq1 vs. Seq2 and Seq3 is a general upward trend, while the plot between Seq2 and Seq3 is more erratic with a general trend that is either shallowly downwards or flat. since matrices with low numbers are for closely related sequences and high numbers are for more distantly related sequences, it is clear that sequence 1 is highly unrelated to sequences 2 and 3 while sequence 2 and 3 are possibly related to each other.

1. **(10 pts.) Based on your analysis in (a) and (b), which two sequences could be very related to each other? Justify your answer.**

Both plots indicate that sequence 1 is probably not related to either sequence 2 or sequence 3. The plots indicate that sequences 2 and 3 might be related to each other, with high scores in the matrix numbers for closely related sequences. The main concern is that there are also high scores in matrix numbers for distantly related proteins, and the trend on the BLOSUM plot is not in a favorable direction. However, the overall high scores and favorable PAM trend indicate that, of the three sequences, sequence 2 and sequence 3 are the most likely to be related.

1. **Apply dynamic programming (Needleman-Wunsch algorithm) to find the optimal global pairwise alignment between s1= GGTTACTG, s2=GTTTAACTTG. Use match +3, mismatch -2, gap -3. You need to fill in the dynamic programming table with scores and arrows then apply trace back to show the optimal alignment. An empty table is given for you. You may solve this question on paper and provide the picture here, too.** 
   1. **(15 pts.) Complete dynamic programming table. Remember to show arrows and highlight the traceback path.**



* 1. **(5 pts.) Write the final alignment**

| G | G | T | T | - | A | C | - | T | G |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| G | T | T | T | A | A | C | T | T | G |

or

| G | G | T | T | A | - | C | - | T | G |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| G | T | T | T | A | A | C | T | T | G |

or

| G | G | T | T | - | A | C | T | - | G |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| G | T | T | T | A | A | C | T | T | G |

or

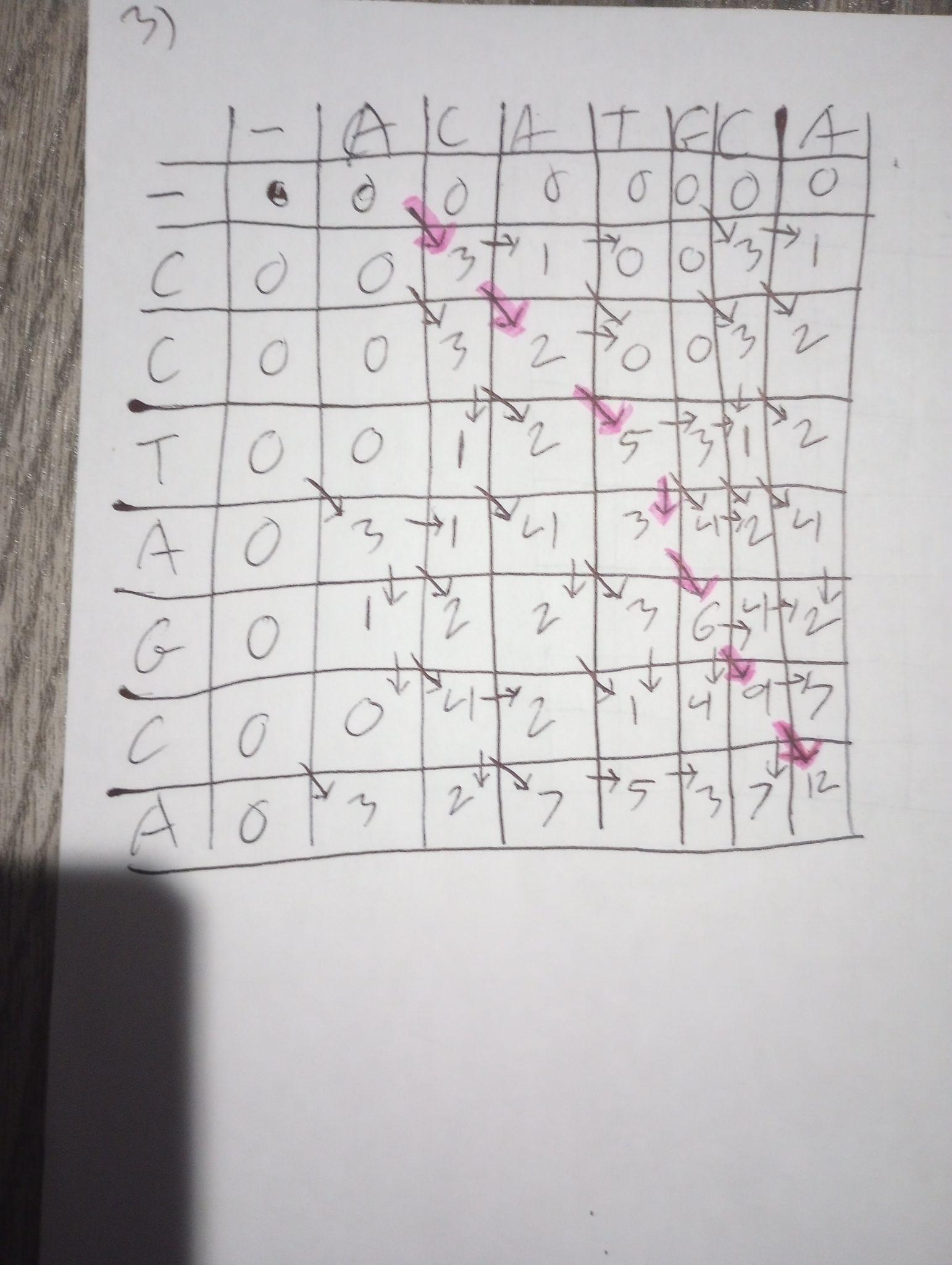
| G | G | T | T | A | - | C | T | - | G |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| G | T | T | T | A | A | C | T | T | G |

* 1. **(5 pts). Write the optimal alignment score.**

13

1. **Suppose that you are given the two sequences below to apply Smith-Waterman local alignment algorithm. The scores are +3 for a match, -1 for a mismatch, and -2 for a gap.**

* 1. **(15 pts.) Fill in the dynamic programming table below. Remember to draw arrows and show the traceback path of the optimal local alignment (you may handwrite on the matrix paper and the put image below).**



* 1. **(5 pts.) What is the optimal alignment score?**

12

* 1. **(5 pts.) Write the optimal local alignment of these sequences.**

| - | C | C | T | A | G | C | A |
| --- | --- | --- | --- | --- | --- | --- | --- |
| A | C | A | T | - | G | C | A |

1. **(20 pts.) Given two DNA sequences ATGATTCCTACATG and GTTAGTACTGATG, draw the dot plot for window size = 5, stringency = 3.**

|  | **G** | **T** | **T** | **A** | **G** | **T** | **A** | **C** | **T** | **G** | **A** | **T** | **G** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **A** |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **T** |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **G** |  |  |  |  |  |  |  |  |  | ⚫ |  |  |  |
| **A** |  |  |  | ⚫ |  |  |  |  |  |  | ⚫ |  |  |
| **T** |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **T** |  |  |  |  |  | ⚫ |  |  |  |  |  |  |  |
| **C** |  |  |  | ⚫ |  |  | ⚫ |  |  |  |  |  |  |
| **C** |  |  |  |  | ⚫ |  |  | ⚫ |  |  |  |  |  |
| **T** |  |  |  |  |  | ⚫ |  |  |  |  |  |  |  |
| **A** |  |  |  |  |  |  | ⚫ |  |  |  |  |  |  |
| **C** |  |  |  |  |  |  |  | ⚫ |  |  |  |  |  |
| **A** |  |  |  |  |  |  |  |  |  |  | ⚫ |  |  |
| **T** |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **G** |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Important Reminder on Academic Integrity Policy**

**This assignment must be done in a group up to 2 people. All solutions/writing must be in your own work.**

Academic honesty policy is strictly enforced in this class. Violators of this policy will be reported to the Office of the Provost and the Office for Student Success. There are no exceptions. Please refer to the course syllabus for the academic honesty policy for more details. If you have any questions about this policy, please contact the instructor. **All submissions will be scanned by Turnitin and other code plagiarism software.**