Bioinformatics Algorithms:

HW V

**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.

Cod for this assignment:   
<https://colab.research.google.com/drive/1NIYglJnZGvV8z3iMsogeDtXy0e__2-vU?usp=sharing>

**Multiple Sequence Alignment**

1. **Given the following sequences, answer the questions based on CLUSTALW algorithm.**

**S1: GTAATGGACGT**

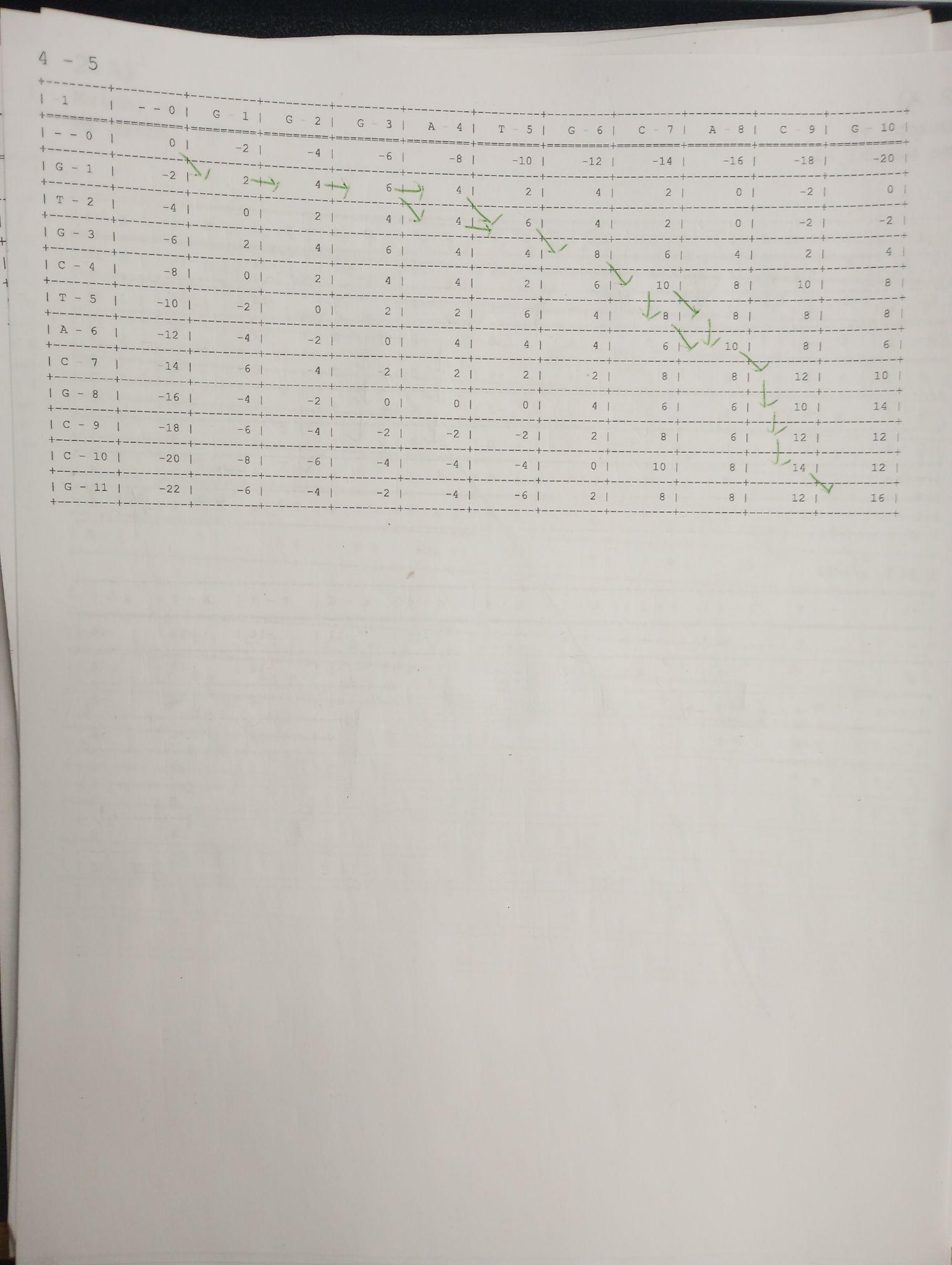
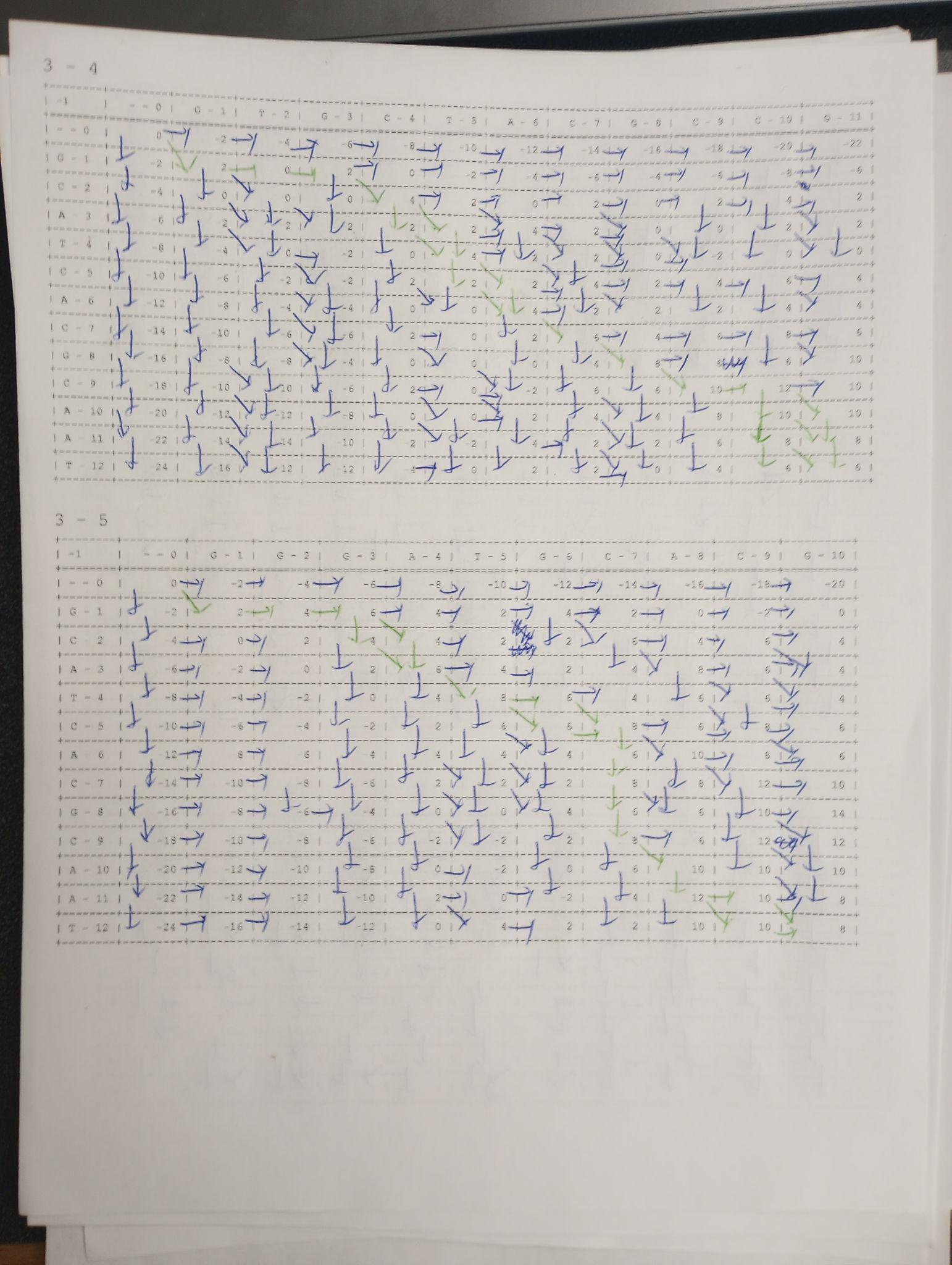
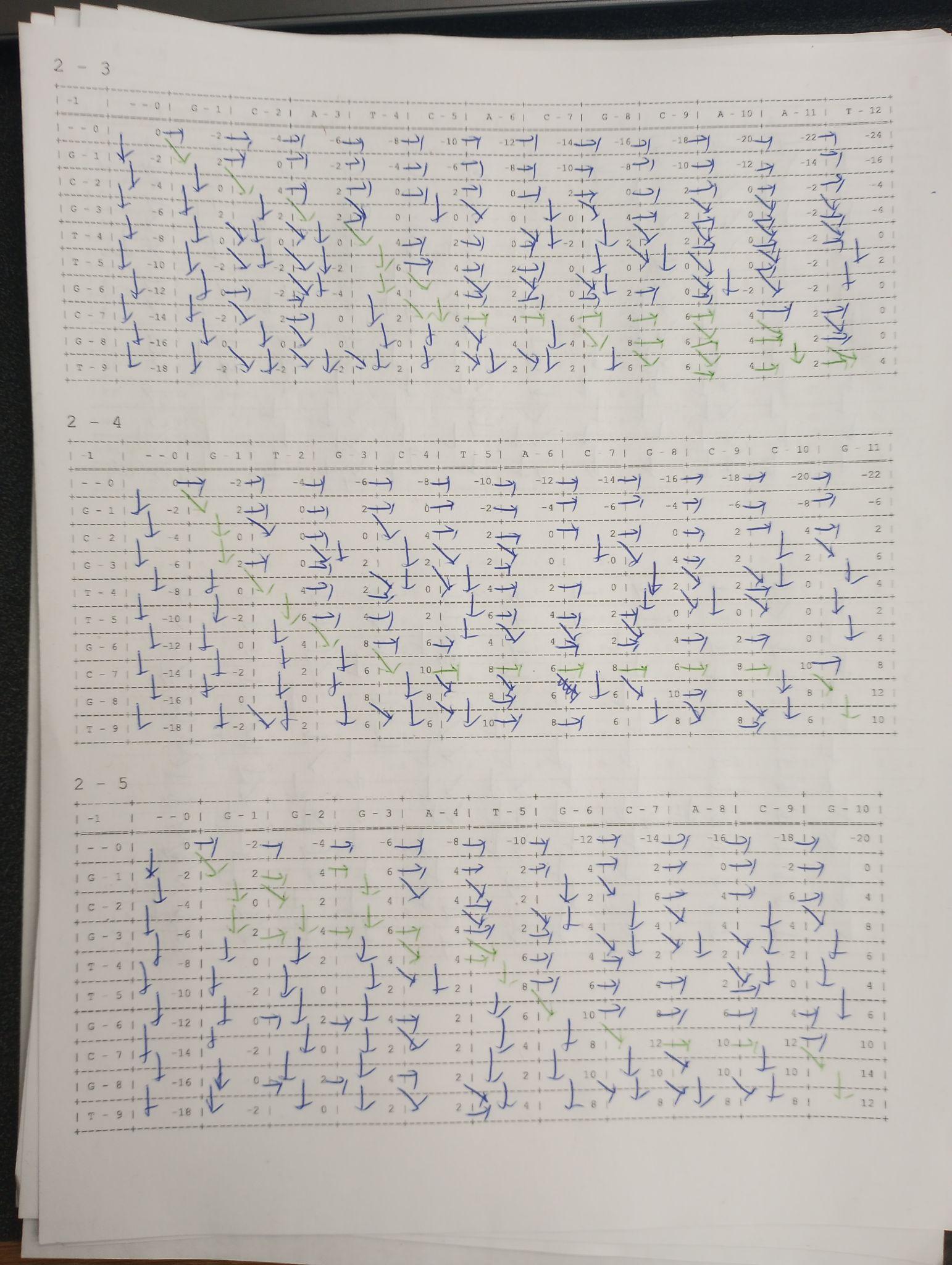
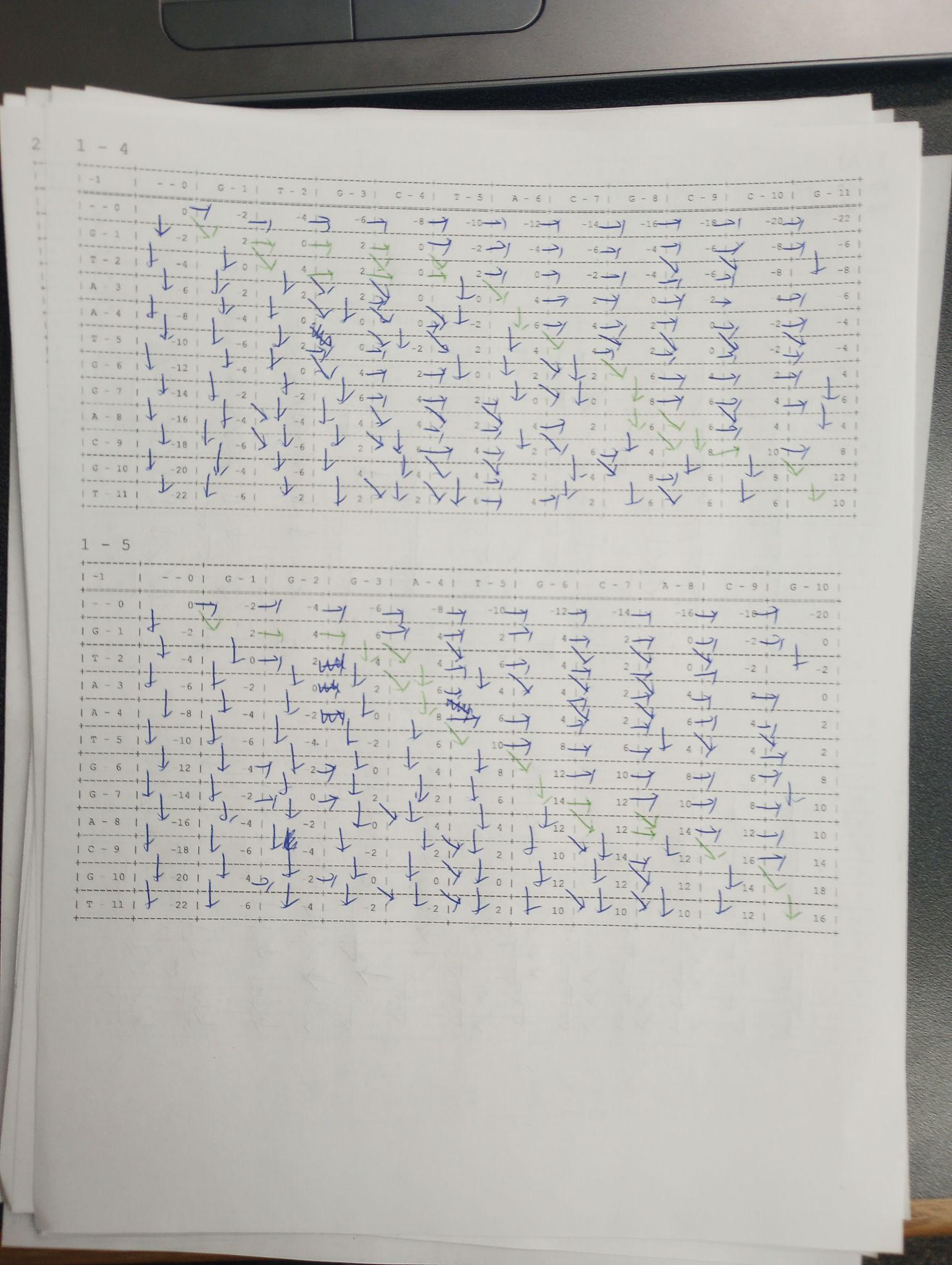
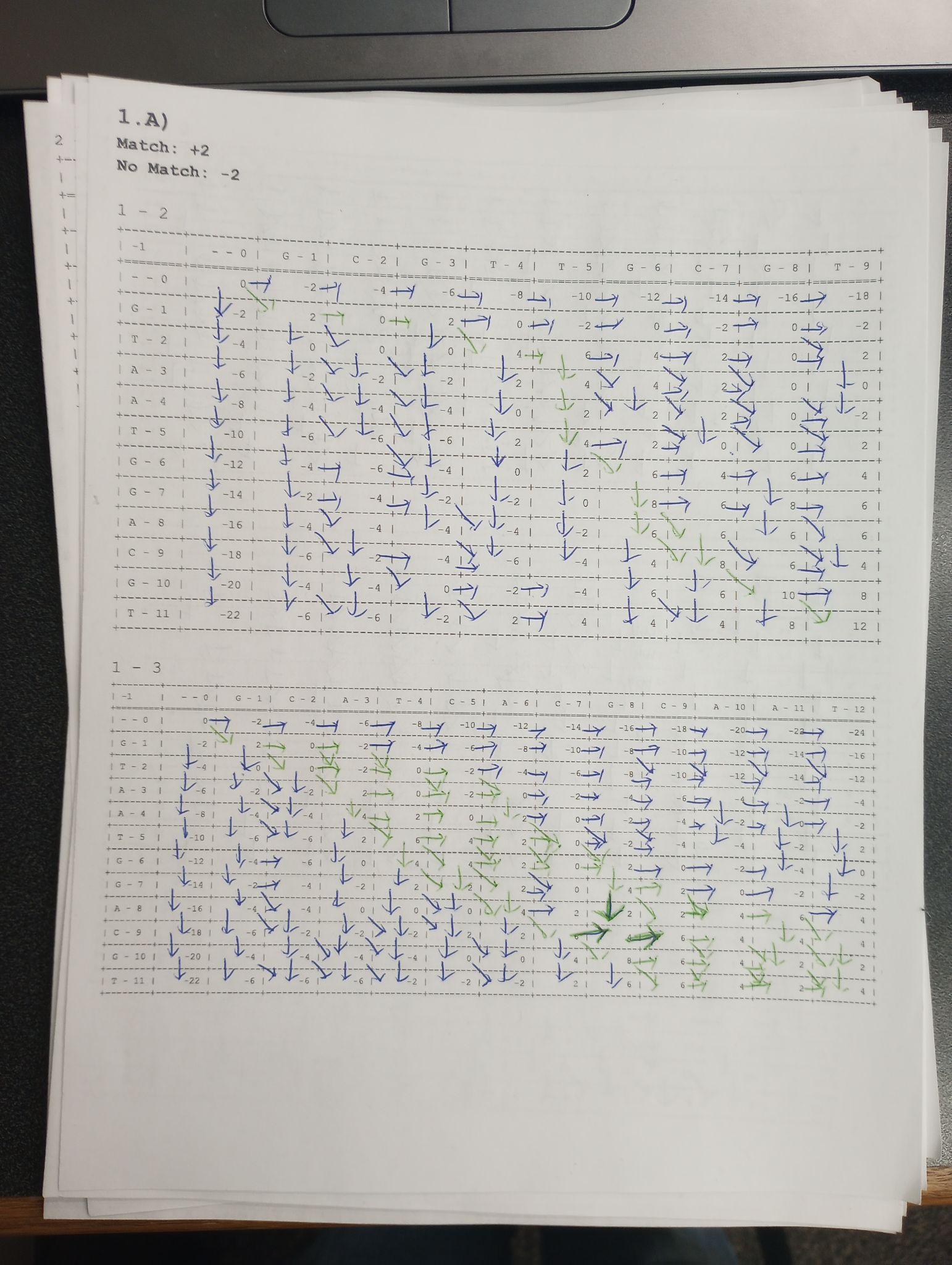
**S2: GCGTTGCGT**

**S3: GCATCACGCAAT**

**S4: GTGCTACGCCG**

**S5: GGGATGCACG**

* 1. **(15 pts.) Compute pairwise alignment to compute similarity matrix. For each pair, perform optimal global alignment and compute the percentage identity (i.e., percentage of matched bases) and the distance will be 1-similarity. You could manually derive the optimal alignment or draw the dynamic programming table.**



S1: G--T-AATGGACGT

S2: GCGT---TG-C-GT

Not Gap: 7

Match: 6

Percent: 0.14286

S1: G--T-AATGG-A-CGT

S3: GCATCA-CG-CAAT–-

Not Gap: 8

Match: 5

Percent: 0.375

S1: G---TAATGGAC-GT

S4: GTGCTA-CG-C-CG-

Not Gap: 7

Match: 5

Percent: 0.28571

S1: G--TAATGG-ACGT

S5: GGGA--TG-CACG-

Not Gap: 7

Match: 6

Percent: 0.14286

S2: GCGTTGC-----GT

S3: GCAT-C-ACGCAAT

Not Gap: 7

Match: 4

Percent: 0.42857

S2: GCGTTGC------GT

S4: G--T-GCTACGCCG-

Not Gap: 5

Match: 5

Percent: 0.0

S2: GCG--TTGC--GT

S5: GG-GAT-GCACG-

Not Gap: 6

Match: 5

Percent: 0.16667

S3: G--CATCACGC-AAT

S4: GTGCT-A-CGCCG--

Not Gap: 8

Match: 5

Percent: 0.375

S3: G--CAT-CACGCAA-T

S5: GGGA-TGC----A-CG

Not Gap: 6

Match: 4

Percent: 0.33333

S4: G---TGCTACGCCG

S5: GGGATGCA-C---G

Not Gap: 7

Match: 6

Percent: 0.14286

|  | **S1** | **S2** | **S3** | **S4** | **S5** |
| --- | --- | --- | --- | --- | --- |
| **S1** | - | 0.14286 | 0.375 | 0.28571 | 0.14286 |
| **S2** |  | - | 0.42857 | 0.0 | 0.16667 |
| **S3** |  |  | - | 0.375 | 0.33333 |
| **S4** |  |  |  | - | 0.14286 |
| **S5** |  |  |  |  | - |

* 1. **(20 pts.) Apply UPGMA algorithm to compute the guide tree. Show your work.**

D1

|  | **S1** | **S2** | **S3** | **S4** | **S5** |
| --- | --- | --- | --- | --- | --- |
| **S1** | - | 0.14286 | 0.375 | 0.28571 | 0.14286 |
| **S2** |  | - | 0.42857 | 0.0 | 0.16667 |
| **S3** |  |  | - | 0.375 | 0.33333 |
| **S4** |  |  |  | - | 0.14286 |
| **S5** |  |  |  |  | - |

len(S2, S4) = 0.0 / 2 = 0.0

D2((S2, S4), S1) = (D1(S2, S1) + D1(S4, S1)) / 2 = (0.14286 + 0.28571) / 2 = 0.214285

D2((S2, S4), S3) = (D1(S2, S3) + D1(S4, S3)) / 2 = (0.42857 + 0.375) / 2 = 0.401785

D2((S2, S4), S5) = (D1(S2, S5) + D1(S4, S5)) / 2 = (0.16667 + 0.14286) / 2 = 0.154765

D2

|  | **(S2, S4)** | **S1** | **S3** | **S5** |
| --- | --- | --- | --- | --- |
| **(S2, S4)** | - | 0.214285 | 0.401785 | 0.154765 |
| **S1** |  | - | 0.375 | 0.14286 |
| **S3** |  |  | - | 0.33333 |
| **S5** |  |  |  | - |

len(S1, S5) = 0.14286 / 2 = 0.07143

D3((S1, S5), S3) = (D2(S1, S3) + D2(S5, S3)) / 2 = (0.375 + 0.33333) / 2 = 0.354165

D3((S1, S5), (S2, S4)) = (D2(S1, (S2, S4)) + D2(S5, (S2, S4))) / 2 = (0.214285 + 0.154765) / 2 = 0.184525

D3

|  | **(S2, S4)** | **(S1, S5)** | **S3** |
| --- | --- | --- | --- |
| **(S2, S4)** | - | 0.184525 | 0.401785 |
| **(S1, S5)** |  | - | 0.354165 |
| **S3** |  |  | - |

len((S2, S4), (S1, S5)) = 0.184525 / 2 = 0.0922625

D4((S2, S4), (S1, S5), S3) = (D3((S2, S4), S3) + D3((S1, S5), S3)) / 2 = (0.401785 + 0.354165) / 2 = 0.377975

D4

|  | **(S2, S4), (S1, S5)** | **S3** |
| --- | --- | --- |
| **(S2, S4), (S1, S5)** | - | 0.377975 |
| **S3** |  | - |

len(S3, ((S2, S4), (S1, S5))) = 0.377975 / 2 = 0.1889875

|  | | | | |  | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | 8 | |  |  |  |  |
|  | |  | | | |  | |  |  |
|  | 6 | |  | | 7 | |  |  |  |
| S2 | | S4 | | S1 | | S5 | | S3 | |

* 1. **(5 pts.) Based on the guide tree, show the order of alignments that need to be done (Note: You don’t need to perform the alignments, just show their order)**

First S2 with S4.

Then S1 with S5.

Then (S2, S4) with (S1, S5).

Then S3 with ((S2, S4), (S1, S5)).

1. **Star alignment.**

**Given the following sequences, answer the questions below.**

**S1: GCCCCTA**

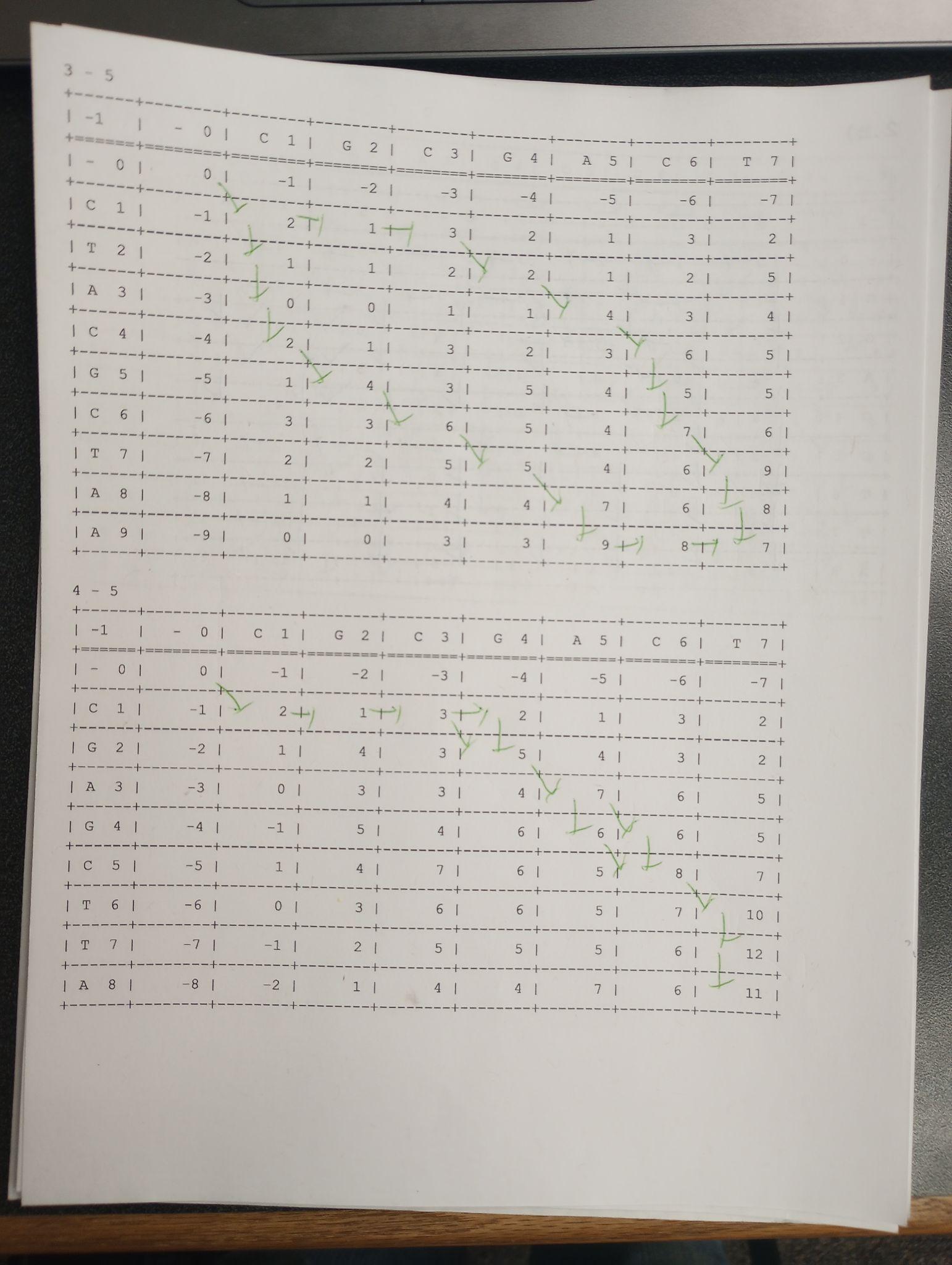
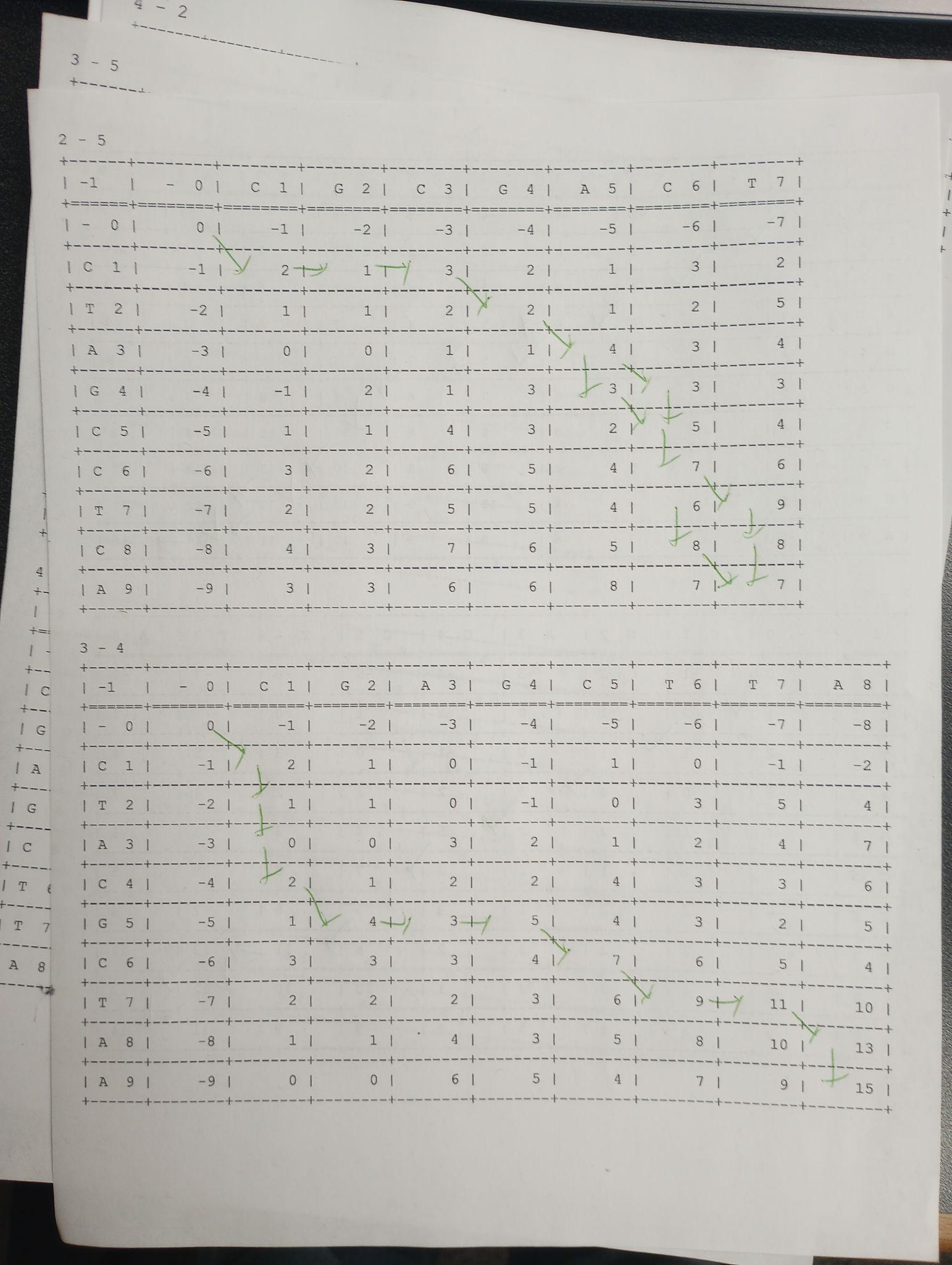
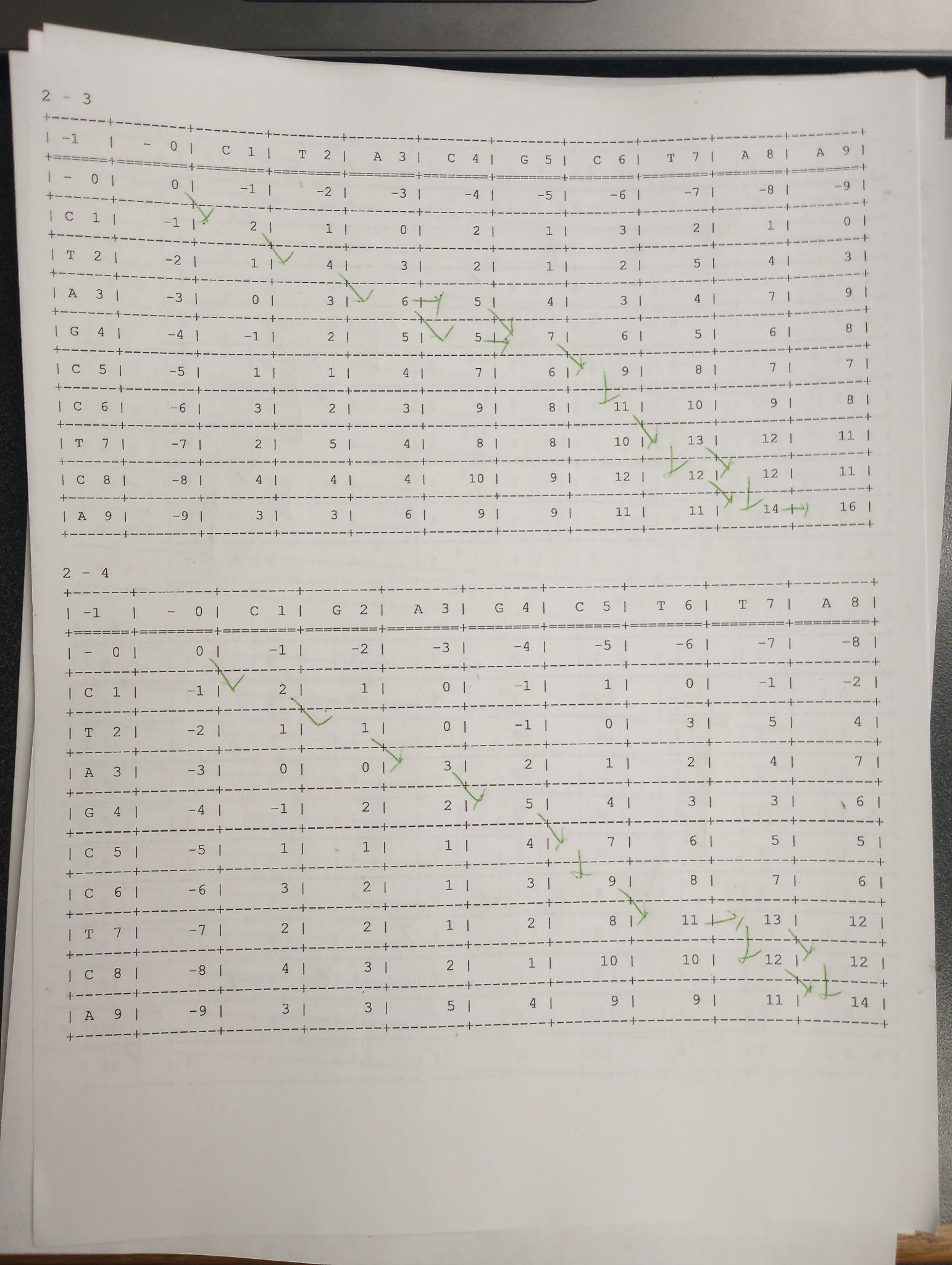
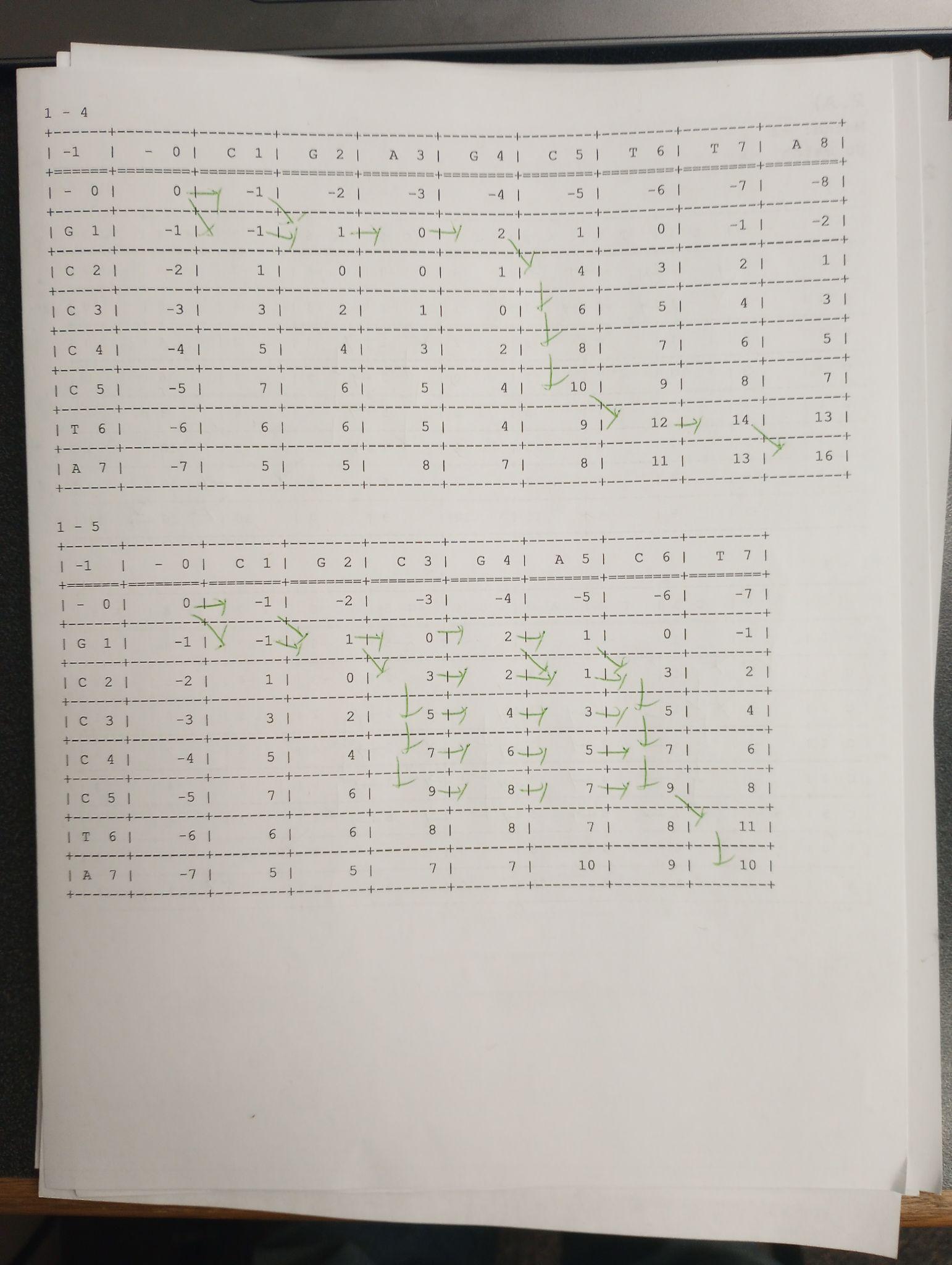
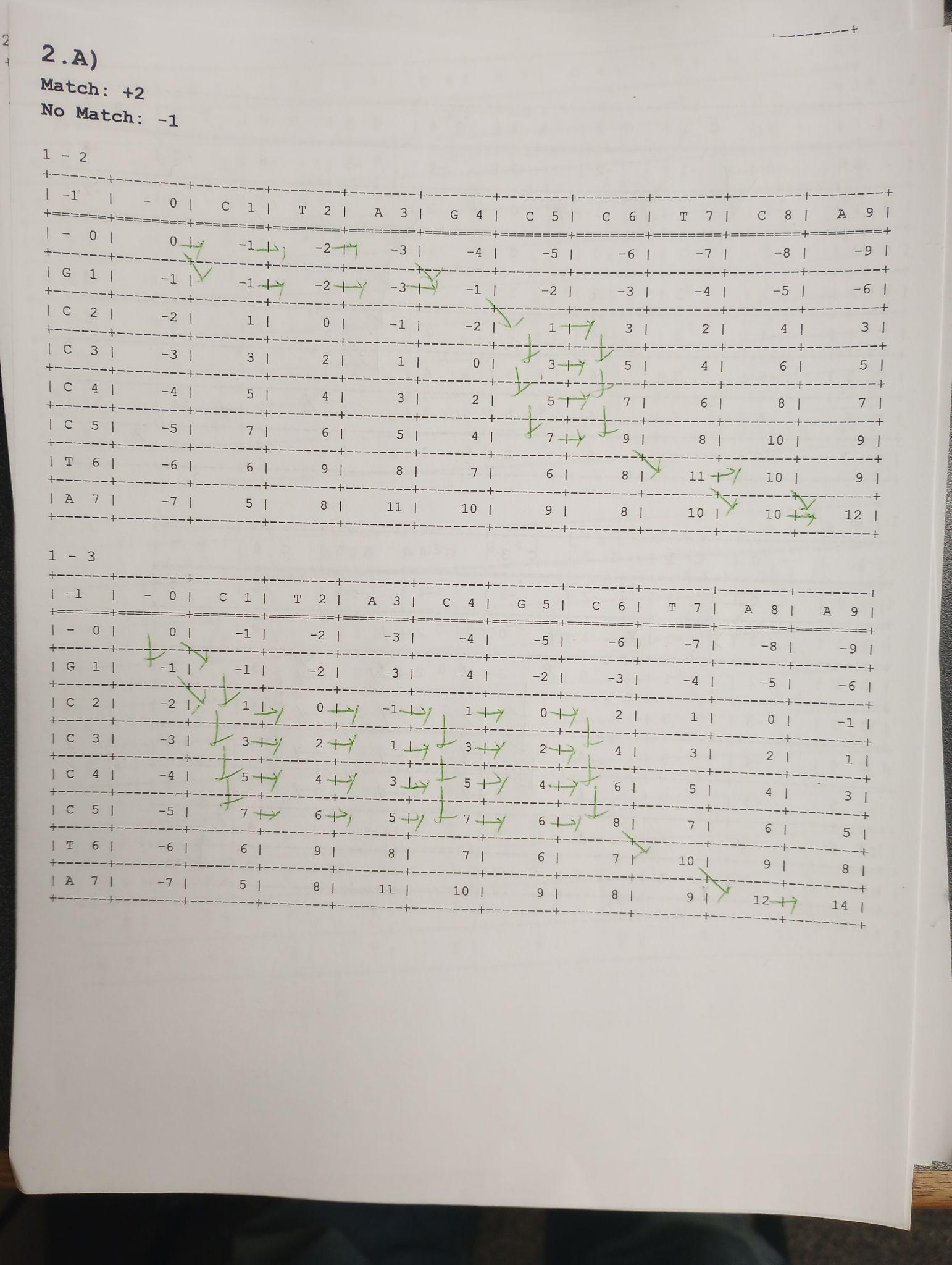
**S2: CTAGCCTCA**

**S3: CTACGCTAA**

**S4: CGAGCTTA**

**S5: CGCGACT**

**a) (20 pts.) Identify the center sequence based on Star alignment algorithm. Use +2 for match, -1 for mismatch or gap.**



+--------+----+----+----+----+----+-------+

| | 1 | 2 | 3 | 4 | 5 | Score |

+--------+----+----+----+----+----+-------+

| 1 | | 12 | 14 | 16 | 10 | 52 |

+--------+----+----+----+----+----+-------+

| 2 | 12 | | 16 | 14 | 7 | 49 |

+--------+----+----+----+----+----+-------+

| 3 | 14 | 16 | | 15 | 7 | 52 |

+--------+----+----+----+----+----+-------+

| 4 | 16 | 14 | 15 | | 11 | 56 |

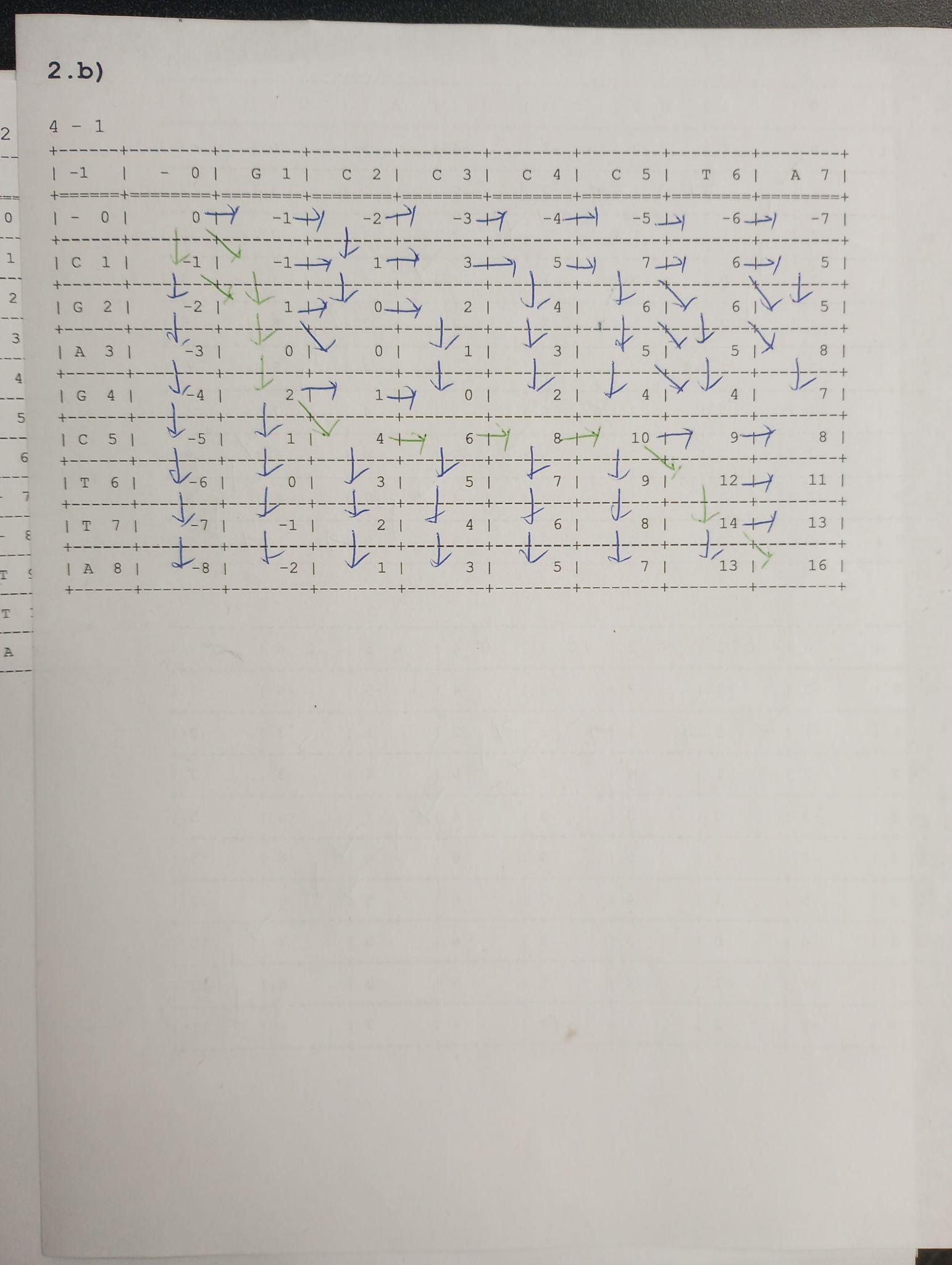
+--------+----+----+----+----+----+-------+

| 5 | 10 | 7 | 7 | 11 | | 35 |

+--------+----+----+----+----+----+-------+

The largest score is 56, so we select S4 to be the center sequence.

**b) (20 pts.) Apply the center star algorithm to perform multiple sequence alignment of these sequences.**



S4: CGAGC---TTA

S1: G---CCCCT-A

S4: CGAGC---TTA

S2: CTAGCC--TCA

S4: CGAGC---TTA

S3: CTA-CGC-TAA

S4: CGAGC---TTA

S5: CGCG---ACT-

**c) (20 pts.) Compute sum-of-pair score for this multiple sequence alignment. Use +2 for match, -1 for mismatch or gap.**

S1: G---CCCCT-A

S2: CTAGCC--TCA

S3: CTA-CGC-TAA

S4: CGAGC---TTA

S5: CGCG---ACT-

Pos1: SP-Score(G, C, C, C, C) = 4 \* 𝛿(G, C) + 6 \* 𝛿(C, C) = 4 \* -1 + 6 \* 2 = 8

Pos2: SP-Score(-, T, T, G, G) = 2 \* 𝛿(-, T) + 2 \* 𝛿(-, G) + 𝛿(T, T) + 2 \* 𝛿(T, G) + 𝛿(G, G) = 2 \* -1 + 2 \* -1 + 2 + 2 \* -1 + 2 = -2

Pos3: SP-Score(-, A, A, A, C) = 3 \* 𝛿(-, A) + 3 \* 𝛿(A, A) + 3 \* 𝛿(A, C) = 3 \* -1 + 3 \* 2 + 3 \* -1 = 0

Pos4: SP-Score(-, G, -, G, G) = 6 \* 𝛿(-, G) + 𝛿(-, -) + 3 \* 𝛿(G, G) = 6 \* -1 + 2 + 3 \* 2 = 2

Pos5: SP-Score(C, C, C, C, -) = 6 \* 𝛿(C, C) + 4 \* 𝛿(C, -) = 6 \* 2 + 4 \* -1 = 8

Pos6: SP-Score(C, C, G, -, -) = 𝛿(C, C) + 2 \* 𝛿(C, G) + 4 \* 𝛿(C -) + 2 \* 𝛿(G, -) + 𝛿(-, -) = 2 + 2 \* -1 + 4 \* -1 + 2 \* -1 + 2 = -4

Pos7: SP-Score(C, -, C, -, -) = 3 \* 𝛿(C, -) + 𝛿(C, C) + 3 \* (-, -) = 3 \* -1 + 2 + 3 \* 2 = 5

Pos8: SP-Score(C, -, -, -, A) = 3 \* 𝛿(C, -) + 𝛿(C, A) + 3 \* 𝛿(-, -) + 3 \* 𝛿(-, A) = 3 \* -1 + -1 + 3 \* 2 + 3 \* -1 = -1

Pos9: SP-Score(T, T, T, T, C) = 6 \* 𝛿(T, T) + 4 \* 𝛿(T, C) = 6 \* 2 + 4 \* -1 = 8

Pos10: SP-Score(-, C, A, T, T) = 𝛿(-, C) + 𝛿(-, A) + 2 \* 𝛿(-, T) + 𝛿(C, A) + 2 \* 𝛿(C, T) + 2 \* 𝛿(A, T) + 𝛿(T, T) = -1 + -1 + 2 \* -1 + -1 + 2 \* -1 + 2 \* -1 + 2 = -7

Pos11: SP-Score(A, A, A, A, -) = 6 \* 𝛿(A, A) + 4 \* 𝛿(A, -) = 6 \* 2 + 4 \* -1 = 8

Total: SP-Score = 8 + -2 + 0 + 2 + 8 + -4 + 5 + -1 + 8 + -7 + 8 = 25

**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.**