Noah M. Jorgenson

Bioinformatics

Lab 5

A.

The subst\_score takes two arguments and returns whether they are the same or not, 1 if they are and -1 if they are not. The while is traversing over the length of the first sequence in the argument from the end of the sequence and compares the value of that nucleotide with the value of the nucleotide in the second argument sequence (since they are required the be the same length we can do it in this way). It returns the result of the subst\_score function with the two nucleotides at the specific index in the while loop (from length of first argument sequence, -1 until it hits -1).

B.

Below are the results after running the function for the sequences mentioned in part B. The results appear to demonstrate that the funciton is working, for example WHAT and WHY produced a traceback matrix where the match was something like WHY-, but for WHEY and WHY the traceback produced WH-Y instead.

Score matrix for WHAT and WHY with gap penalty of -1:

[[ 0. -1. -2. -3. -4.]

[-1. 1. 0. -1. -2.]

[-2. 0. 2. 1. 0.]

[-3. -1. 1. 1. 0.]]

Traceback matrix:

[[ 0. 1. 1. 1. 1.]

[ 0. 2. 1. 1. 1.]

[ 0. 0. 2. 1. 1.]

[ 0. 0. 0. 2. 1.]]

Score matrix for WHEY and WHY with gap penalty of -1:

[[ 0. -1. -2. -3. -4.]

[-1. 1. 0. -1. -2.]

[-2. 0. 2. 1. 0.]

[-3. -1. 1. 1. 2.]]

Traceback matrix:

[[ 0. 1. 1. 1. 1.]

[ 0. 2. 1. 1. 1.]

[ 0. 0. 2. 1. 1.]

[ 0. 0. 0. 2. 2.]]

Score matrix for AAAC and AGC with gap penalty of -1:

[[ 0. -1. -2. -3. -4.]

[-1. 1. 0. -1. -2.]

[-2. 0. 0. -1. -2.]

[-3. -1. -1. -1. 0.]]

Traceback matrix:

[[ 0. 1. 1. 1. 1.]

[ 0. 2. 1. 1. 1.]

[ 0. 0. 2. 1. 1.]

[ 0. 0. 0. 2. 2.]]

C.

The resulting matrices appear to be what I imagined the alignment looking like in my head when using the traceback matrix. When adjusting the gap penalty however, when it’s made smaller the traceback is more likely to make a dash since the argmax function will tend ot pick the dashes over the match even if it doesn’t match since the gap penalty is now small and thus the result of score+gap penalty is larger than score + subst\_score. We needed the extra while loops following the end of the first while loop in order that we might empty the rest of the values out of the traceback matrix. It prematurely stops once we’ve reached the end of one sequence but not necessarily both of them.

Score matrix for WHAT and WHY with gap penalty of -0.5:

[[ 0. -0.5 -1. -1.5 -2. ]

[-0.5 1. 0.5 0. -0.5]

[-1. 0.5 2. 1.5 1. ]

[-1.5 0. 1.5 1. 0.5]]

Traceback matrix:

[[ 0. 1. 1. 1. 1.]

[ 0. 2. 1. 1. 1.]

[ 0. 0. 2. 1. 1.]

[ 0. 0. 0. 0. 0.]]

['W', 'H', 'A', 'T', '-']

['W', 'H', '-', '-', 'Y']

Score matrix for WHEY and WHY with gap penalty of -0.5:

[[ 0. -0.5 -1. -1.5 -2. ]

[-0.5 1. 0.5 0. -0.5]

[-1. 0.5 2. 1.5 1. ]

[-1.5 0. 1.5 1. 2.5]]

Traceback matrix:

[[ 0. 1. 1. 1. 1.]

[ 0. 2. 1. 1. 1.]

[ 0. 0. 2. 1. 1.]

[ 0. 0. 0. 0. 2.]]

['W', 'H', 'E', 'Y']

['W', 'H', '-', 'Y']

Score matrix for AAAC and AGC with gap penalty of -0.5:

[[ 0. -0.5 -1. -1.5 -2. ]

[-0.5 1. 0.5 0. -0.5]

[-1. 0.5 0. -0.5 -1. ]

[-1.5 0. -0.5 -1. 0.5]]

Traceback matrix:

[[ 0. 1. 1. 1. 1.]

[ 0. 2. 1. 1. 1.]

[ 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 2.]]

['A', 'A', 'A', '-', 'C']

['A', '-', '-', 'G', 'C']

The resulting matrices appear to be what I imagined the alignment looking like in my head when using the traceback matrix. When adjusting the gap penalty however, when it’s made smaller the traceback is more likely to make a dash since the argmax function will tend ot pick the dashes over the match even if it doesn’t match since the gap penalty is now small and thus the result of score+gap penalty is larger than score + subst\_score. I think the gap penalty of -2 is better for the AAAC & AGC pair.

Score matrix for IQIFSFIFRQEWNDA and QIFFIFRMSVEWND with gap penalty of -0.5:

[[ 0. -0.5 -1. -1.5 -2. -2.5 -3. -3.5 -4. -4.5 -5. -5.5 -6. -6.5 -7.

-7.5]

[-0.5 -1. 0.5 0. -0.5 -1. -1.5 -2. -2.5 -3. -3.5 -4. -4.5 -5. -5.5

-6. ]

[-1. 0.5 0. 1.5 1. 0.5 0. -0.5 -1. -1.5 -2. -2.5 -3. -3.5 -4.

-4.5]

[-1.5 0. -0.5 1. 2.5 2. 1.5 1. 0.5 0. -0.5 -1. -1.5 -2. -2.5

-3. ]

[-2. -0.5 -1. 0.5 2. 1.5 3. 2.5 2. 1.5 1. 0.5 0. -0.5 -1.

-1.5]

[-2.5 -1. -1.5 0. 1.5 1. 2.5 4. 3.5 3. 2.5 2. 1.5 1. 0.5

0. ]

[-3. -1.5 -2. -0.5 1. 0.5 2. 3.5 5. 4.5 4. 3.5 3. 2.5 2.

1.5]

[-3.5 -2. -2.5 -1. 0.5 0. 1.5 3. 4.5 6. 5.5 5. 4.5 4. 3.5

3. ]

[-4. -2.5 -3. -1.5 0. -0.5 1. 2.5 4. 5.5 5. 4.5 4. 3.5 3.

2.5]

[-4.5 -3. -3.5 -2. -0.5 1. 0.5 2. 3.5 5. 4.5 4. 3.5 3. 2.5

2. ]

[-5. -3.5 -4. -2.5 -1. 0.5 0. 1.5 3. 4.5 4. 3.5 3. 2.5 2.

1.5]

[-5.5 -4. -4.5 -3. -1.5 0. -0.5 1. 2.5 4. 3.5 5. 4.5 4. 3.5

3. ]

[-6. -4.5 -5. -3.5 -2. -0.5 -1. 0.5 2. 3.5 3. 4.5 6. 5.5 5.

4.5]

[-6.5 -5. -5.5 -4. -2.5 -1. -1.5 0. 1.5 3. 2.5 4. 5.5 7. 6.5

6. ]

[-7. -5.5 -6. -4.5 -3. -1.5 -2. -0.5 1. 2.5 2. 3.5 5. 6.5 8.

7.5]]

Traceback matrix:

[[ 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 2. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 2. 0. 2. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 2. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 2. 1. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 0. 2. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 2. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 0. 2. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0. 2. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 2. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 2. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 2. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 2. 1.]]

['-', 'Q', 'I', 'F', 'S', 'F', 'I', 'F', 'R', 'Q', '-', '-', '-', 'E', 'W', 'N', 'D', 'A']

['Q', 'Q', 'I', 'F', '-', 'F', 'I', 'F', 'R', '-', 'M', 'S', 'V', 'E', 'W', 'N', 'D', '-']

Alignment matches: 11/18 [61.1111%]

As you can see, 11 our of the 18 values in the alignments matched when using a gap penalty of -0.5, which is about 61.1% correct matches.

Score matrix for IQIFSFIFRQEWNDA and QIFFIFRMSVEWND with gap penalty of -3.5:

[[ 0. -3.5 -7. -10.5 -14. -17.5 -21. -24.5 -28. -31.5 -35. -38.5

-42. -45.5 -49. -52.5]

[ -3.5 -1. -2.5 -6. -9.5 -13. -16.5 -20. -23.5 -27. -30.5 -34.

-37.5 -41. -44.5 -48. ]

[ -7. -2.5 -2. -1.5 -5. -8.5 -12. -15.5 -19. -22.5 -26. -29.5

-33. -36.5 -40. -43.5]

[-10.5 -6. -3.5 -3. -0.5 -4. -7.5 -11. -14.5 -18. -21.5 -25.

-28.5 -32. -35.5 -39. ]

[-14. -9.5 -7. -4.5 -2. -1.5 -3. -6.5 -10. -13.5 -17. -20.5

-24. -27.5 -31. -34.5]

[-17.5 -13. -10.5 -6. -5.5 -3. -2.5 -2. -5.5 -9. -12.5 -16.

-19.5 -23. -26.5 -30. ]

[-21. -16.5 -14. -9.5 -5. -6.5 -2. -3.5 -1. -4.5 -8. -11.5

-15. -18.5 -22. -25.5]

[-24.5 -20. -17.5 -13. -8.5 -6. -5.5 -3. -4.5 0. -3.5 -7.

-10.5 -14. -17.5 -21. ]

[-28. -23.5 -21. -16.5 -12. -9.5 -7. -6.5 -4. -3.5 -1. -4.5

-8. -11.5 -15. -18.5]

[-31.5 -27. -24.5 -20. -15.5 -11. -10.5 -8. -7.5 -5. -4.5 -2.

-5.5 -9. -12.5 -16. ]

[-35. -30.5 -28. -23.5 -19. -14.5 -12. -11.5 -9. -8.5 -6. -5.5

-3. -6.5 -10. -13.5]

[-38.5 -34. -31.5 -27. -22.5 -18. -15.5 -13. -12.5 -10. -9.5 -5.

-6.5 -4. -7.5 -11. ]

[-42. -37.5 -35. -30.5 -26. -21.5 -19. -16.5 -14. -13.5 -11. -8.5

-4. -7.5 -5. -8.5]

[-45.5 -41. -38.5 -34. -29.5 -25. -22.5 -20. -17.5 -15. -14.5 -12.

-7.5 -3. -6.5 -6. ]

[-49. -44.5 -42. -37.5 -33. -28.5 -26. -23.5 -21. -18.5 -16. -15.5

-11. -6.5 -2. -5.5]]

Traceback matrix:

[[ 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 2. 2. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 2. 2. 2. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 2. 2. 2. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 2. 2. 2. 2. 1. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 2. 0. 2. 2. 2. 1. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 2. 0. 2. 2. 2. 1. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 2. 0. 2. 0. 2. 1. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 2. 0. 2. 0. 2. 1. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 2. 0. 2. 0. 2. 0. 2. 1. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 2. 0. 2. 0. 2. 0. 2. 1. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 0. 2. 0. 2. 0. 2. 0. 2. 1. 1.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 2. 0. 2. 0. 2. 0. 2. 1.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 0. 2. 0. 0. 0. 2. 1. 2.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 2. 0. 0. 0. 2. 1.]]

['I', 'Q', 'I', 'F', 'S', 'F', 'I', 'F', 'R', 'Q', 'E', 'W', 'N', 'D', 'A']

['Q', 'I', 'F', 'F', 'I', 'F', 'R', 'M', 'S', 'V', 'E', 'W', 'N', 'D', '-']

Alignment matches: 6/15 [40.0%]

When utilizing a larger gap penalty, it was less likely to take gaps over simply accepting a mismatch and rsulted in a match in the alignments of only 6/15, or 40%.

We could potentially improve the scoring by adjusting the values of our gap penalty and zeroing in on the best possible penalty for this sequence.