Project 4: Gene Sequencing

Function: fullAlign()
Input: two DNA sequences, number of bases to compare
Output: edit distance, first 100 characters of dna sequences with modifications to sit side by side
Cut both strings down to the number of bases to compare
Create 2 tables with dimensions n+1*m+1, where n is the length of sequence 1 and m is the length of sequence 2, fill one with 0s and the other with empty strings ***O(nm) and space nm as well***
Set (0,0) as "source" in second table
For i from 1 to n set (i, 0) to (i-1, 0) + 5 in table 1 and "delete" in table 2 ***O(n), space already allocated***
For j from 1 to m set (0, j) to (0, j-1) + 5 in table 1 and "insert" in table 2 ***O(m), space already allocated***
For i from 1 to n+1
    For j from 1 to m+1 ***O(nm)***
        If characters from each string i-1 and j-1 match
            Sub_or_Match = (i – 1, j - 1) - 3
        Else
            Sub_or_Match = (i – 1, j - 1) + 1
        Insert = (i, j-1) + 5
        Delete = (i-1, j) + 5
        Get min of Sub_or_Match, Insert, and Delete
        Insert value to table 1 and string representing type to table 2 at (i, j) ***space already allocated***
Current_Type = (n, m) from table 2
While Current_Type is not "source" ***O(n+m) because the max loops would come from running edges of table***
    If Current_Type is Sub_or_Match
        n = n - 1
        m = m - 1
    If Current_Type is "insert"
        m = m - 1
        insert a dash at position m in sequence 1
    If Current_Type is "delete"
        n = n - 1
        insert a dash at position n in sequence 2
    Current_Type = (n, m) from table 2
Cut both strings down to 100 characters
Edit_Distance = (n, m) from table 1
Return Edit_Distance and both sequence strings
Final O is nm + n + m + nm + n + m = O(nm)
Final space is nm, allocated at the beginning

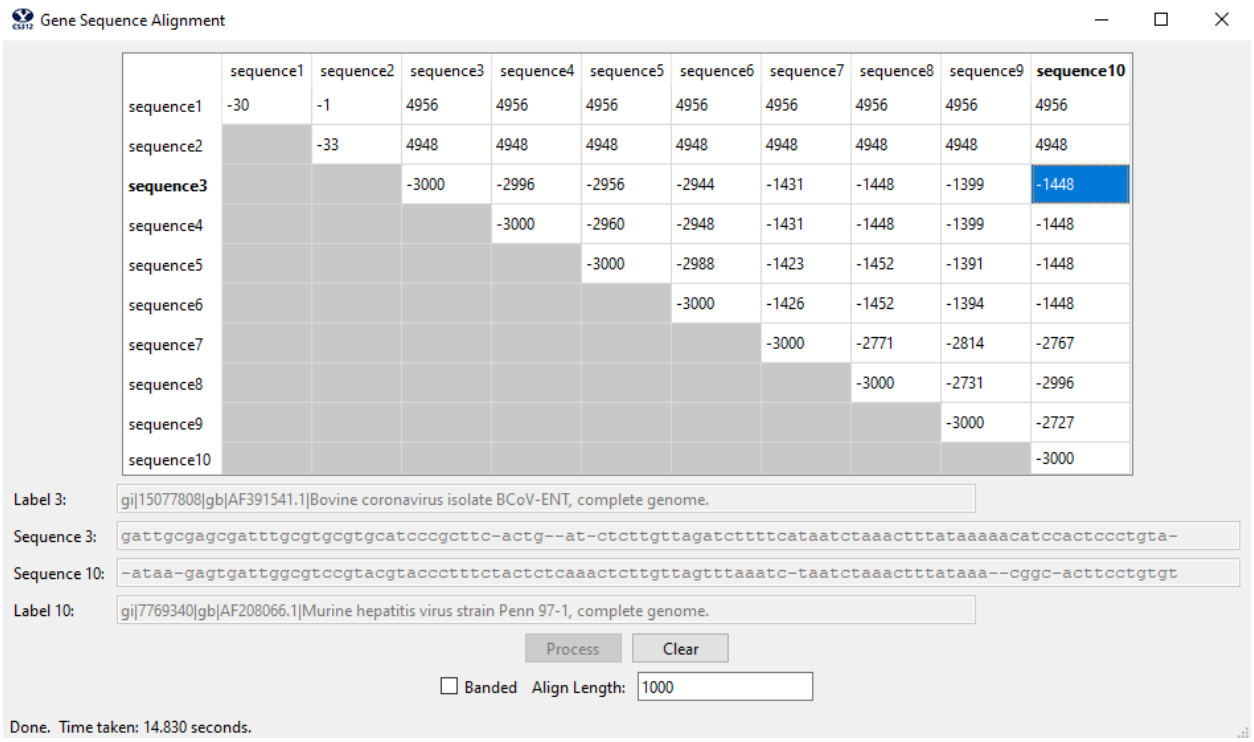
Function: bandedAlign()
Input: two DNA sequences, number of bases to compare
Output: edit distance, first 100 characters of dna sequences with modifications to sit side by side
D = 3 will be the band distance on either side of the center point
k = 2D+1
Cut both strings down to the number of bases to compare
If the difference between the string lengths is greater than D
    Return that they cannot be compared
Swap to put the longer string first if necessary (since |m-n| is less than or equal to D, m≈n)
Create 2 tables with dimensions k*n+1, fill one with infinity and the other with empty strings ***O(kn) and space kn***
Set (D, 0) as 0 in table 1 and "source" in table 2
For j from 1 to n+1
    For i from 0 to k ***O(kn)***
        If a character can be matched (can't match forward from points with infinity)
            If characters from each string at corresponding point match
                Sub_or_Match = (i, j - 1) - 3
            Else
                Sub_or_Match = (i, j - 1) + 1
        Else
            Sub_or_Match = infinity

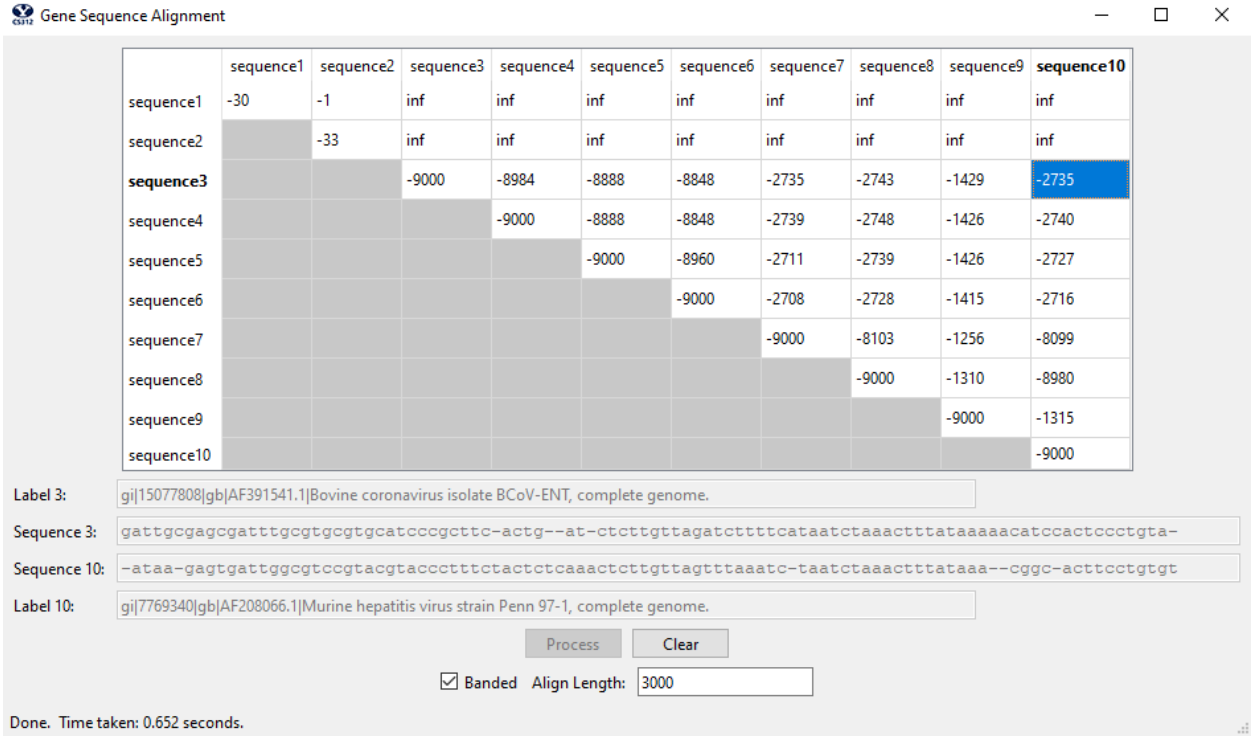
        If a character can be inserted (can't pass table limit)
            Insert = (i + 1, j - 1) + 5
        Else
            Insert = infinity
        If a character can be deleted (can't pass table limit)
            Insert = (i - 1, j) + 5
        Else
            Insert = infinity
        Get min of Sub_or_Match, Insert, and Delete
        Insert value to table 1 and string representing type to table 2 at (i, j) ***space already allocated***
Current_Type = (D, n) from table 2
While Current_Type is not "source" ***O(2n) because the max loops would come from a zigzag***
    If Current_Type is Sub_or_Match
        m = m - 1
    If Current_Type is "insert"
        D = D + 1
        m = m - 1
        insert a dash at position D + m - 3 in sequence 1
    If Current_Type is "delete"
        D = D - 1
        insert a dash at position n in sequence 2
    Current_Type = (n, m) from table 2
Cut both strings down to 100 characters

If strings were swapped before, swap back
Edit_Distance = (n, m) from table 1
Return Edit_Distance and both sequence strings
Final O is kn + kn + 2n = O(kn)
Final space is kn, allocated at the beginning

The way my alignment extraction algorithm works is iterating column by column and filling in alignment values based on the minimum possible from inserting, deleting, and substituting/matching based on the previous values. At the time of inserting that value, it also stores in another table whether it was an insert, delete, or substitute/match that achieved the lowest value. The item at the bottom corner (or what would represent the bottom corner in the case of the banded algorithm) is then the lowest possible edit distance. The item at that location of the second table can be used to trace back the path used to get there, while simultaneously editing the sequence strings to add dashes to the first where there were insertions, and to the second where there were deletions so that in the end the strings line up perfectly.

Results


Gene Sequence Alignment
sequence1 sequence2 sequence3 sequence4 sequence5 sequence6 sequence7 sequence8 sequence9 sequence10
sequence1 -30 -1 4956 4956 4956 4956 4956 4956 4956 4956
sequence2 -33 4948 4948 4948 4948 4948 4948 4948 4948
sequence3 -3000 -2996 -2956 -2944 -1431 -1448 -1399 -1448
sequence4 -3000 -2960 -2948 -1431 -1448 -1399 -1448
sequence5 -3000 -2988 -1423 -1452 -1391 -1448
sequence6 -3000 -1426 -1452 -1394 -1448
sequence7 -3000 -2771 -2814 -2767
sequence8 -3000 -2731 -2996
sequence9 -3000 -2727
sequence10 -3000
Label 3: gi|15077808|gb|AF391541.1|Bovine coronavirus isolate BCoV-ENT, complete genome.
Sequence 3: gattgcgagcgatttgcgtgcgtgcatcccgcttc-actg--at-ctcttgttagatcttttcataatctaaactttataaaaacatccactccctgta-
Sequence 10: -ataa-gagtgattggcgtccgtacgtacccttttctactctcaaactcttgttagtttaaatc-taatctaaactttataaa--cggc-acttcctgtgt
Label 10: gi|7769340|gb|AF208066.1|Murine hepatitis virus strain Penn 97-1, complete genome.
Process
Clear
Banded
Align Length: 1000
Done. Time taken: 14.830 seconds.


Gene Sequence Alignment
sequence1 sequence2 sequence3 sequence4 sequence5 sequence6 sequence7 sequence8 sequence9 sequence10
sequence1 -30 -1 inf inf inf inf inf inf inf inf
sequence2 -33 inf inf inf inf inf inf inf inf
sequence3 -9000 -8984 -8888 -8848 -2735 -2743 -1429 -2735
sequence4 -9000 -8888 -8848 -2739 -2748 -1426 -2740
sequence5 -9000 -8960 -2711 -2739 -1426 -2727
sequence6 -9000 -2708 -2728 -1415 -2716
sequence7 -9000 -8103 -1256 -8099
sequence8 -9000 -1310 -8980
sequence9 -9000 -1315
sequence10 -9000
Label 3: gi|15077808|gb|AF391541.1|Bovine coronavirus isolate BCoV-ENT, complete genome.
Sequence 3: gattgcgagcgatttgcgtgcgtgcatcccgcttc-actg--at-ctcttgttagatcttttcataatctaaactttataaaaacatccactccctgta-
Sequence 10: -ataa-gagtgattggcgtccgtacgtacccttttctactctcaaactcttgttagtttaaatc-taatctaaactttataaa--cggc-acttcctgtgt
Label 10: gi|7769340|gb|AF208066.1|Murine hepatitis virus strain Penn 97-1, complete genome.
Process
Clear
Banded
Align Length: 3000
Done. Time taken: 0.652 seconds.