Global Alignment with Affine Gap Penalties

The Project:
The aim of this project was to write and implement a code to find the best possible alignment of 2 strings of Nucleotides (DNA or RNA) or amino acids (proteins). Global Alignments are useful in order to compare different DNA or protein sequences, however, the way you score/penalize alignments is important. Using a single constant penalty for all insertions/deletions may not produce the best possible alignment due to the nature of mutations. Mutations tend to rearrange large chunks of DNA in a single event. Therefore, in order to create alignments that abide by this property, initializing gaps should be penalized more than extending gaps. This change improves upon the constant gap alignment in that it scores strings with consecutive matches higher than multiple small-dispersed alignment.

I decided to use this global alignment function to compare the OXTR gene between species. I decided to do this because I read a lot of neuroscience papers that utilize mouse models of Autism Spectrum Disorder (ASD) in order to study the disorder and test possible treatments. The OXTR gene is thought to be relevant to ASD, as an abnormal form is found in many individuals with ASD. Additionally, this gene encodes for the protein receptor for oxytocin, a hormone involved in many social processes such as bonding and maternal behavior. I thought comparing this gene across species would be useful in order to justify using such a model.

In my comparisons I aimed to find out how similar the mouse and human OXTR genes are. Is the sequence largely conserved between the species? What are the types of differences? (i.e. maybe a only a few insertions/deletions). What is there score, and how does this score compare to the scores of humans/mice and other animals?

The Data:
The data used in this project were sequences of DNA and mRNA taken from the National Center for Biotechnology Information (NCBI) online database. Here, I found DNA sequence of this gene in humans, mice (Mus Musculus), and rhesus monkeys (Macaca Mulatta). I additionally found the mRNA sequence for this gene in humans, mice, rhesus monkeys, rats (Rattus Norvegicus), and the boxer breed of dog (Canis Lupus familiaris). The DNA and mRNA sequences did not need to be shortened or altered in any way.
Steps of Program:
First, I wrote a function to read fasta files and output the sequence in a string.

Function to calculate scores of possible Alignments:
• This program required 3 tables containing the scores of possible alignments, and 3 backtrack tables to keep track of moves within each table. Each node of each table refers to index i of string 1, and index j of string 2. The value at each node [i][j] is set to the highest possible score. The corresponding backtrack table recorded what “move” the highest value corresponded with.
• Best Possible Score at indices:
  o In the upper and lower tables – scores corresponded to either initializing a gap or extending a gap
    ▪ Ex: lower table
      • Extending a gap: middle[i-1][j] – initializing gap penalty
      • Initializing a gap: lower [i-1][j]- gap extension penalty
  o For the middle table there are 3 options at each node:
    ▪ Match/mismatch: middle [i-1][j-1]+score
      • The score was calculated using a function from ‘scoring functions’ that used the BLOSUM62 scoring matrix
    ▪ End a gap from either lower or upper table
      • EX: ending gap on lower: middle [i][j]=lower [i][j]
      • No penalties because ending a gap was not penalized
• As the function fills in the scores of various alignments, each backtrack table kept track of what possibility was the highest score – and what “move” it corresponded to
  o Again, each of the 3 tables had its own backtrack table
  o The letters used were ‘L’ ‘M’ and ‘U’
    ▪ These referred to whatever table the score just came from.
      • Ex: initializing a gap would be marked ‘M’
      • Ex: extending a gap via upper table would be marked ‘U’ since it was moving from the upper table to the upper table
• Maximum Global Alignment Score:
  o Since there was no penalty of ending a gap, the maximum score possible can be found in the lower right corner of the middle table
    ▪ Think of this as a “free ride” to the middle table

Next, a function to figure out the best alignment of the strings was calculated
Inputs: lower, upper, and middle backtrack tables, string 1 and string 2
• First 2 empty strings which would represent the alignment strings were created
• This function worked backwards, and therefore started at the lower right corner of the middle table. In order to iterate in this direction, values of i and j were set to the length of their corresponding strings.
• A string called Table was created. This value keeps track of which table the function in in. As the alignment moves between tables, this string is altered. This value is initially set to 'M' as it starts in the middle table

• To iterate backwards, use a while statement: While backtrack does not equal starting value
  o Use if statements to make commands for each possible “move” in each node at the 3 backtrack tables
  o For Upper and Lower Tables:
    ▪ If table value matches table:
      • Add letter at index of string 1/2 to corresponding alignment string and decrement i/j
        o Upper table – j, lower table – j
        o Add dash to other alignment string
      • Look at backtrack value at that index, if it corresponds to a different table – change value of table to letter that corresponds to that table
      • If backtrack value has letter that corresponds to current table, do not change it
  o Middle table:
    ▪ If backtrack value corresponds to another table, change table value to that table
    ▪ If backtrack value is M, then add corresponding letters from strings 1&2 to their corresponding alignment strings
      • Decrement both I and j
      • Do not change value of table
  • Reverse both alignment strings and then Done!

Results:
Validation of constant vs. affine gap penalties:

My first goal was to compare the alignments of DNA sequences to their corresponding mRNA sequences with constant versus affine gap penalties. As expected, using affine gap resulted in alignments that showed a large amount of consecutive matches/mismatches, and consecutive insertions rather than dispersed ones. These gaps show possible introns. The rhesus monkey sequences provide a good example of this property. In both types of global alignment, the DNA alignment string consists of 1 massive string with no insertions. The mRNA alignment string with affine gap penalties is one large block of continuous text (matches/mismatches) followed by one extremely long sequence of insertions, and then a small block of text at the very end. The alignment with constant penalties has many single letter matches, and ends with a long string of insertions.
A. Global alignment with constant penalties

B. Global alignment with affine gap penalties

Figure 1. A shows a small section of the rhesus monkey mRNA string that came out of global alignment with constant gap penalties of mRNA and DNA. B shows a shortened section of the mRNA alignment string from global alignment with affine gap penalties. The length of insertions has been shortened here, but the letters represent all the matches/mismatches seen in this alignment. The full alignments can be seen in Results-DNA-mRNA-comparison-rhesus.docx.

Comparisons of OXTR gene:

<table>
<thead>
<tr>
<th>Organism</th>
<th>mRNA sequence length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rat</td>
<td>5,335</td>
</tr>
<tr>
<td>Mouse</td>
<td>4,568</td>
</tr>
<tr>
<td>Human</td>
<td>4,361</td>
</tr>
<tr>
<td>Dog (boxer)</td>
<td>1,275</td>
</tr>
<tr>
<td>Rhesus Monkey</td>
<td>1,253</td>
</tr>
</tbody>
</table>

Table 1. Length of OXTR mRNA sequences of various organisms in order of decreasing length.

As expected, the OXTR gene was very similar between humans and mice, and mice and rats. It was unexpected, however, to see a large difference between rat and human mRNA, especially given the fact that both were similar to the mouse gene. A summary of the difference between alignment scores of these species can be seen in Figure 2. This figure also shows the difference in length of mRNA sequence for each comparison. Here, it becomes evident that the low alignment score between human and rat mRNA sequences can be explained in part by the different
lengths of these sequences. The length comparison also highlights the weight of the similarities between the rat and mouse mRNAs. Despite the fact that this comparison has a much larger length difference than the human-mouse comparison, it has a much higher score.

Figure 1. Global alignment scores between human, mouse and rat mRNA sequences of the OXTR sequence. Both affine gap penalties, and constant penalties are shown. The difference in length of the mRNA’s for each comparison are also shown.

Overall this data demonstrates that the OXTR mRNA sequence is somewhat conserved between rodents and humans. This sequence is highly conserved, however, between rats and mice. Additional comparisons demonstrate that this sequence is not highly conserved within all mammals. The dog and rhesus monkey mRNA’s had much lower alignment scores with all other sequences. The primary reason for their lower alignment scores is the length of their mRNA sequences; dog and rhesus monkey OXTR mRNA’s were much shorter than the same sequence in rats, mice, and humans. This property can be seen in the extremely low alignment score of human and rhesus monkey mRNAs. In order to better compare these sequences, perhaps a local alignment function should be used.