Theresa Steele

Reframing the situation:

Implementing penalties for frameshifting mutations in a sequence alignment function

In translation, mRNA is 'read' three nucleotides at a time, with these three letters coding for a specific protein. Consequently, mutations in mRNA that do not shift the 'frame' of translation – substitutions, or indels that occur in multiples of 3 – should be observed more often in conserved sequences than frameshifting mutations.

To account for this, I adapted the Global Sequence Aligner from homework 6.2 to only penalize indels of length 1 or 2, and to reward diagonals that are 3 nucleotides long or more. My function, FrameAlign, does so by adding two "directions" to the recurrence relation of the GlobalAlign function – an option to move south by three for no penalty (score of zero), and the option to move east by three, also for no cost (as summarized in figure 1)

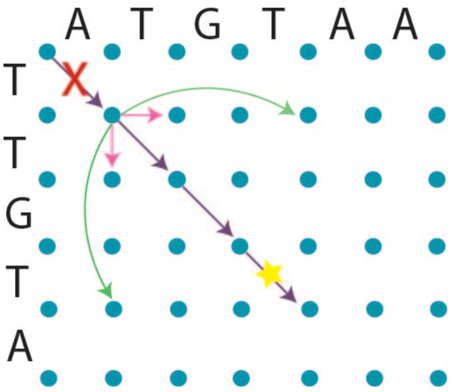


Table 1. Scores used in all sequence aligners

Movement	Score
Indel (1)	-1
Indel (3)	0
Mismatch	0
Match	1
Match(3)	+1

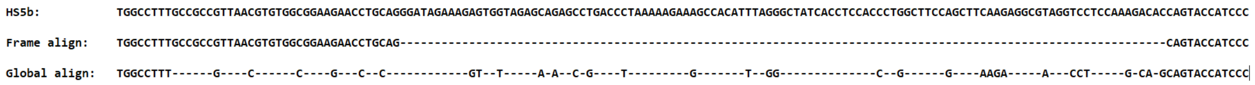
Figure 1. FrameAlign uses five scores: Indel(1) (Pink), representing a move south/east by 1; Indel(3) (green), representing a move south/east by three; mismatch (red X), a diagonal movement where string 1 and string 2 nucleotides do not match; match (purple), a diagonal movement where string1 and string 2 nucleotides do match; and match(3), a bonus given to matches that occur after two previous matches.

To test my function, I first generated toy motifs: a motif of length k, and a second motif identical to the first, but with three nucleotides removed from a random location in the sequence. I ran these toy sequences with the functions Global Align (GA), Local Align (LA), and Frame Align (FA). Table 2 illustrates an example of a toy motif for which all alignment methods return different alignments, and illustrates the advantage of using Frame Align.

Table 2. Alignment of toy motifs.

Function	Alignment	Score
Global Alignment	GCCTGTGACTTA G -C - -TGACTTA	6
Local Alignment	TGTGACTTA GCTGACTTA	7
Frame Alignment	GCCTGTGACTTA GC - - -TGACTTA	14

For the experimental portion of my project, I compared alignments of mRNA sequences made with Global, Local, and Frame Align functions. I chose to align mRNA of the NMDA-type glutamate receptor (subunit 1; NMDAR1) from *Homo sapiens*, *Xenopus tropicalis*, and *Apis mellifera*. NMDAR subunit1 is a useful sequence to compare these alignment functions because splice variants of this receptor have been sequenced in humans. I compared alignments of the *Homo sapiens* GluN1-3b splice variant to the *Homo sapiens* GluN1-5b splice variant as a proof of concept. Frame Align was able to successfully identify the location of alternative splicing between GluN1-3b and -5b, where the Global Align and Local Align (not shown) functions were not.



```
HS5b:          TGGCCTTTGCCGCCGTTAACGTGTGGCGGAAGAACCTGCAGGGATAGAAAGAGTGGTAGAGCAGAGCCTGACCCTAAAAAGAAAGCCACATTTAGGGCTATCACCTCCACCCTGGCTTCCAGCTTCAAGAGGCGTAGGTCCTCCAAAGACACCAGTACCATCCC

Frame align:   TGGCCTTTGCCGCCGTTAACGTGTGGCGGAAGAACCTGCAG---------------------------------------------------------------------------------------------------------------CAGTACCATCCC

Global align:  TGGCCTTT------G----C------C----G---C--C------------GT--T-----A-A--C-G----T---------G-------T--GG--------------C--G------G----AAGA-----A---CCT-----G-CA-GCAGTACCATCCC
```

Figure 2. Section of alignments of GluN1-3b (bottom two rows) to GluN1-5b (top) produced with Frame and Global Align functions.

I then aligned NMDAR 1 mRNA from Xenopus tropicalis, a vertebrate, and Apis mellifera, the western honey bee, with GluN1-5b using the three alignment functions. These sequences were obtained in FASTA format from Nucleotide (NCBI). Because Frame Align adds bonuses for continuous diagonal movement, the raw scores of the three alignment functions are not directly comparable (figure 3, left); however, looking at the score of the *A. mellifera* and *X. tropicalis* sequences as a percentage of the score of the alignment of GluN1-5b with itself using each function (figure 3, right).

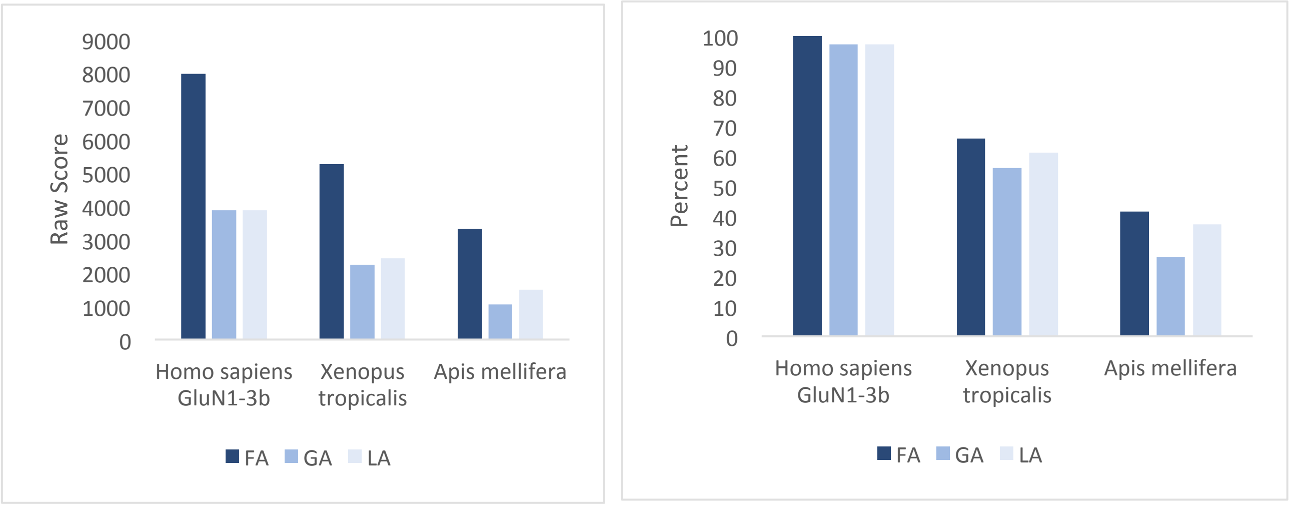
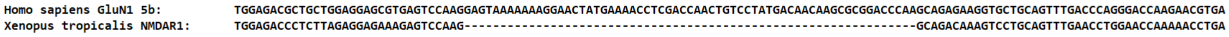


Figure 3. (Left) Alignment scores of *Homo sapiens* GluN1-5b to *Homo sapiens* GluN1-3b, *Xenopus tropicalis* NMDAR 1, and *Apis mellifera* NMDAR 1 using Frame Align (FA), Global Align (GA) and Local Align (LA) functions. (Right) Alignment scores represented as percentages of the GluN1-5b: GluN1-3b score.

I found that Frame Align could successfully identify splice sites within otherwise identical DNA sequences, where the Global and Local Alignment functions could not. The Frame Align function did not substantially improve alignments of NMDA receptor subunit 1 from *X. tropicalis* or *A. mellifera*; However, Frame Align was able to identify what appears to be a large indel or alternative splicing site that in the *X. tropicalis* NMDAR1 mRNA that did not appear in the global or local alignments (Figure 4).



```
Homo sapiens GluN1 5b:         TGGAGACGCTGCTGGAGGAGCGTGAGTCCAAGGAGTAAAAAAAGGAACTATGAAAACCTCGACCAACTGTCCTATGACAACAAGCGCGGACCCAAGCAGAGAAGGTGCTGCAGTTTGACCCAGGGACCAAGAACGTGA
Xenopus tropicalis NMDAR1:     TGGAGACCCTCTTAGAGGAGAAAGAGTCCAAG---------------------------------------------------------------GCAGACAAAGTCCTGCAGTTTGAACCTGGAACCAAAAACCTGA
```

Figure 4. Frame alignment of *X. tropicalis* NMDAR1 with GluN1-5b.

Surprisingly, frame align works slightly better than our previously built Global Align function at detecting large indels; however, it remains to be seen how this function would compare to a function built specifically to account for large indels (such as the affine gap function). On the other hand, accounting for frame in the alignment of mRNA sequences did not substantially improve the scores of these alignments, and actually produced an alignment of GluN1-5b to *A. mellifera* NMDAR1 that contained more gaps than the alignments produced by global or local alignment functions. Frame Align is a useful function when attempting to detect sites of alternative splicing, but is less useful with aligning distantly related sequences.