Project 3

Rune, Astrid og Ditte

Results of brca1 MSA:

The sum of pairs scores for the *brca1* test sequences are:

seqs 1-3: 790 (*sp_exact*)

seqs 1-4: 2,340 (*sp_approx*)

seqs 1-5: 3,313 (*sp_approx*)

seqs 1-6: 5,964 (*sp_approx*)

The sum of pairs scores for the *brca1* full sequences are:

seqs 1-8: 263,904 (*sp_approx*)

FASTA files: https://github.com/RuneWind/AiB/tree/master/Project%203/results

Implementations

sp_exact

- Our exact MSA implementation calculates the sum-of-pairs scores for all pairs of sequences

sp_appox

- Our approximate implementation first defines a center string as the sequence with the lowest summed sum-of-pairs scores with all the other sequences
- MSA is created by aligning the center string with all other sequences

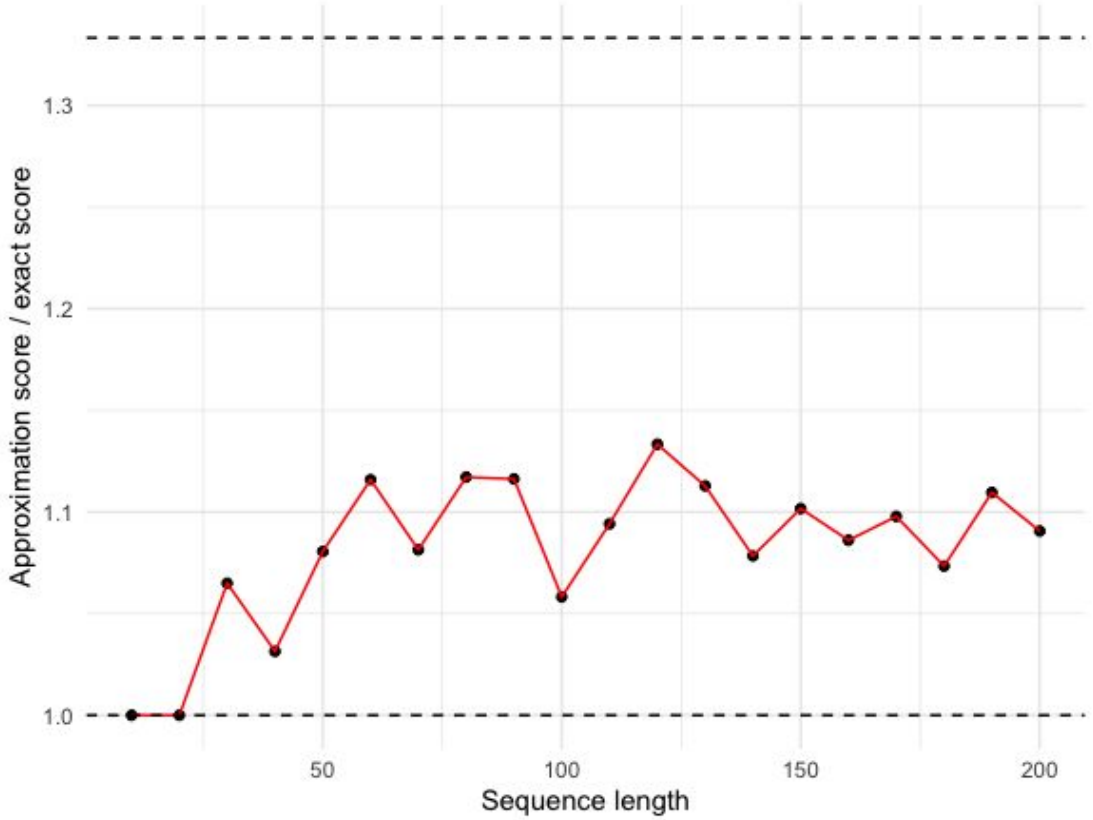
Dealing with non-base characters

- In both implementations characters different from A, T, C, or G are simply replaced by A
- We expect to see very few non-base characters and thus that this will have little effect on the MSA

sp_approx vs. sp_exact

The score of a MSA between 3 sequences obtained from *sp_approx* is expected to at most 4/3 of the score obtained using *sp_exact*.

The plot to the right shows, that this is true for our implementation.



Space consumption and running time

The space consumption in our implementation of both *sp_exact* and *sp_approx* is dominated by our alignment matrix

Our space consumption is $O(n \cdot m)$ or (n^2) if $n = m$