Ali Ramazani

Question 1a:

Algorithm Idea: The algorithm will fill in a table to compute the scores of all possible alignments of substrings of v and w. The key idea is to combine local alignments of v with global alignments of w to find the optimal fitting alignment.

Initialization:

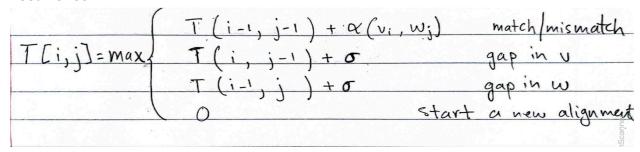
T is a table of size $(m+1) \times (n+1)$ with zeros.

T[i, j] gives us the score of the fitting alignment of the substrings v[1...i] and w[1...j].

T(i, 0) = 0 for $0 \le i \le m$

T(0, j) = 0 for $0 \le j \le n$

Recurrence:



Return:

The optimal score for the fitting alignment is stored in the last row of Table T $max\{T(m, j)\}\ for\ 0 \le j \le n$

Here, m is the length of string v, and n is the length of string w.

Runtime: O(nm)

- **1b)** For a single 250bp read, fitting alignment might be a reasonable approach. Considering the run-time of our fitting algorithm, O(250 * 3 billion) = O(750 billion) base pairs, which would take about 1.83 GB of storage space, assuming 3 billion bp takes about 725 MB.
- **1c)** Considering O(250 million * 3 billion) = O(750 trillion) base pairs, which takes about 1826 GB of storage. Therefore, fitting alignment would not make sense here because aligning a million short reads to a larger genome using dynamic programming requires significant memory, which can be a constraint on a typical desktop computer.

Note: I used this <u>stackoverflow</u> to help me answer parts b and c. I also talked with Carlos and shared ideas to help each other.

Data Analysis:

python3 local_alignment.py TP53-Human.fasta TP53-Cat.fasta scoring1.txt

Alignment Score: 28.0 Optimal Local Alignment:

GCACATCTGCAT-TTTCACCC-CACCCTTCCCCTCCTTCTCCC-TTTTTATATCCCATTTTTATA CG GCACATCTGCGTATTTC-CCCACACCCTTCCCC-C-T-CTCCCCTTTTTATATCCCCTTTTTATATCG

python3 local_alignment.py TP53-Rat.fasta TP53-Cat.fasta scoring1.txt

Alignment Score: 8.0 Optimal Local Alignment:

GACTCTGT GACTCTGT

python3 local_alignment.py TP53-Rat.fasta TP53-Human.fasta scoring1.txt

Alignment Score: 8.0
Optimal Local Alignment:

GGCCAGCC GGCCAGCC

Based on the alignment scores, the Human and Cat sequences appear to be the most similar in the aligned regions.