Computational Biology HS19 – Exercise 1 answers to theory questions

7th October 2019

1. Ideally, we want to incorporate as few gaps into the alignment as possible. The more gaps we are incorporating during alignment, the less it resembles the original aligned sequences. When incorporating gaps, we are essentially creating new sequences by introducing unspecified insertions/deletions and forcefully aligning them together. Mismatches just signify there are substitution differences between the sequences, without any frameshift changes. Hence to find the optimal alignment the scoring system should this effect into account, and penalize gaps more heavily than mismatches.

2. Allowing gaps at the start and end of a local alignment more or less gives the same alignment when gaps are disallowed, except that the score of the overall alignment would be lower due to the gap penalty (-2). So to get the optimal alignment with the highest score it makes sense that gaps are disallowed at the start and end.

3. The short sequence would likely be able to align to multiple different subregions of the much longer one, and we don't have any further information to determine which is the correct position.

4. Needleman-Wunsch requires roughly $3(m \times n) + (m+n) = 3 \times 100 \times 100 + (200) = 30200$ steps = 4 orders of magnitude of steps to align two sequences of this length. This is 10^{71} times less exhaustive that the brute-force approach of finding all possible alignments.

5. Theoretically we could carry out multiple sequence alignment using Needleman-Wunsch by using an array of k-dimensions. But k sequences of length m require m^k steps and requires extensive computational time and memory.