ZBI ZENTRUM FÜR BIOINFORMATIK

SIC Saarland Informatics Campus

UNIVERSITÄT DES SAARLANDES

June 21, 2021

Assignment 10
Algorithms for Sequence Analysis, Summer 2021

Algorithmic Bioinformatics · Prof. Dr. Sven Rahmann

Hand in date: Monday, June 28, before 20:00

Exercise 1: Strange scores (4 Theory)

Consider the following scoring scheme: $+1$ for matches, -99 for mismatches, 0 for gaps. For concreteness, consider strings over the DNA alphabet (it does not matter). Under this scoring scheme, what is the difference between the optimal global and the optimal local alignment score? What does the global alignment score represent? (You may want to experiment with some examples using the provided code.)

Exercise 2: General Gap Costs (4 Theory)

Consider optimal global alignment, but with a general gap cost function, where a consecutive run of g gap characters receives a score of $-\gamma(g)$, and $\gamma(g)$ is an *arbitrary* gap cost function with $\gamma(0) = 0$, $\gamma(g) > 0$ for $g \geq 1$. Explain how to find the optimal global alignment in this case by adapting the dynamic programming algorithm (or graph topology). (Hint: It is not sufficient to look at the direct left/upper neighbors.) Analyze the running time of the resulting algorithm.

In practice, this generalization is not used frequently (because of the running time). But it is useful to have gap cost functions like $\gamma(1) = 3$, $\gamma(2) = 4$, $\gamma(3) = 1$, $\gamma(4) = 4, \dots$, where gap lengths that are multiples of 3 (codons) are penalized less than other gap lengths. Such gap cost functions play a role in alignments of protein-coding DNA.

Exercise 3: Jukes-Cantor Model of DNA evolution (4 Theory)

Consider the following rate matrix of a Markov process on DNA (`ACGT`):

$$Q = \begin{pmatrix} ? & \alpha & \alpha & \alpha \\ \alpha & ? & \alpha & \alpha \\ \alpha & \alpha & ? & \alpha \\ \alpha & \alpha & \alpha & ? \end{pmatrix}$$

All off-diagonal rates $\alpha > 0$ are equal. What is the stationary distribution π? Choose α such that Q is calibrated to 1 PEM and determine the diagonal. Compute P, P^{120} and the limit P^{∞}. What is the observed amount of change after 1 PEM, 120 PEM, and infinite time, respectively? (Note: `scipy.linalg.expm()` provides matrix exponentials.)

Exercise 4: Implementation of two alignment variants (4 Programming)

Extend the provided alignment code (or transfer it to a language of your choice and then extend it) to optimal overlap computation (free end gap alignment) and optimal pattern matching (semiglobal alignment). Report optimal alignments (incl. the score), using $+1/-1/-2$ for match / mismatch / gap, for all four alignment variants of

```
x = 'TGTGACAGATTTCATGGAACCGGAAGTGGCTGGAACATAAATCG',
y = 'ATAAATGGCTCTGCCGGCAGTGGCGTTGGACTCTGCAGTGACAG'.
```