June 07, 2021

Assignment 8
Algorithms for Sequence Analysis, Summer 2021

Algorithmic Bioinformatics · Prof. Dr. Sven Rahmann

Hand in date: Monday, June 14, before 20:00

Exercise 1: Metric? (4 Theory)
Let Σ be an alphabet.

(a) Show that the Hamming distance is a metric on Σ^n for each n.

(b) Show that the edit distance is a metric on Σ^*.

(c) Show that the k-mer distance is not a metric in general: Give two distinct strings with k-mer distance zero (you can choose $k \geq 2$ and your alphabet).

Note: Subtask (b) is the real work; (a) and (c) are easy.

Exercise 2: Examples (4 Theory)
Compute the edit distance of the following pairs of strings:

(a) pflanze, panzer

(b) atlantique, cyborg

(c) Generalize the previous example: How large is the edit distance between two sequences of lengths m, n if their character sets are disjoint?

Exercise 3: Faster edit distance computation (4 Theory)
Assume that you know a correct upper bound K for the edit distance of two given strings s, t of the same length n. How can the running time of the DP algorithm for computing the edit distance be improved from $O(n^2)$ to $O(nK)$?
In practice, we usually don't know a correct upper bound before we have computed the edit distance. Nevertheless, can you construct an algorithm that computes the correct (unknown) edit distance k in $O(nk)$ time for any k?
Hint: Start by guessing a small upper bound $K \geq k$. Run the $O(nK)$ algorithm. Check whether your guess was correct (how?). If not, increase K (how?).

Exercise 4: Alignments or edit paths as cigar strings (4 Programming)
How many alignments exist between two strings of length 5?

Write a function that lists all edit paths (alignments) as so-called *cigar strings*.
In a cigar string, `M` means match or mismatch (two aligned characters), `I` means insertion and `D` means deletion. Examples for two length-5 strings are `MMMMM` or `IIIIIDDDDD`.
(The name *cigar string* stems from the fact the these strings look like long cigars.)

Be sure to submit the output of your program in addition to the code.
(The code should be general for lengths m, n; the output for $m = n = 5$.)