June 28, 2021

Assignment 11
Algorithms for Sequence Analysis, Summer 2021

Algorithmic Bioinformatics · Prof. Dr. Sven Rahmann

Hand in date: Monday, July 05, before 20:00

Exercise 1: Alignment statistics (4 Theory)
Consider the following setting after a local alignment search. An observed score of 60 has a p-value of 0.001 (10^{-3}). An observed score of 80 has a p-value of 10^{-7}. What is approximately the p-value of an observed score of 90 ? Explain your calculations.

Exercise 2: Linear-space local alignment (4 Theory)
The linear-space alignment method was described for global alignments, where the matrix size (m, n) is known in advance. Describe a practical way to obtain an optimal *local* alignment in linear space.

Exercise 3: Limits for the Hamming distance (4 Theory)
Let $\Sigma_k := \{1, \ldots, k\}$ be a generic alphabet of size k. Consider the *normalized Hamming distance* of two strings of length n, given by

$$d_{\mathrm{NH}(n)}(x, y) := |\{i : x_i \neq y_i\}|/n$$

(i.e., the standard Hamming distance divided by the string length).
Argue that for i.i.d. uniform random strings $X, Y \in \Sigma_k^n$, independently of n,

$$\mathbb{E}[d_{\mathrm{NH}(n)}(X, Y)] = \frac{k-1}{k},$$

and therefore also

$$\lim_{n\to\infty} \mathbb{E}[d_{\mathrm{NH}(n)}(X, Y)] = \frac{k-1}{k}\,.$$

Exercise 4: Limits for the Edit distance (4 Programming)
Exercise 3 was just a warm-up. We are really interested in the *normalized edit distance*, not in the Hamming distance. Unfortunately, this is much harder to solve mathematically, so we resort to simulation.
So, consider two i.i.d. uniform random strings $X, Y \in \Sigma_k^n$ (same length n) and their normalized edit distance (i.e., standard edit distance divided by n).
Let $D_{n,k} := \mathbb{E}[d_{\mathrm{NE}}(X, Y)]$ be the expected normalized edit distance of two random strings of length n over Σ_k, and let $D_k := \lim_{n\to\infty} D_{n,k}$.
Using simulation of many (possibly many millions of) strings and fast edit distance computation, determine approximate values for D_k for some small $k = 2, 3, \ldots, 10$.
Compare D_k with the corresponding Hamming distance result $(k-1)/k$.