TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

Department of Computer Science
Algorithms & Complexity Group

Issue Date: 09.06.2015
Version: 2015-06-09 16:05

Exercise Sheet 4 for Computational Biology (Part 1), SS 15

Hand In: *Until Tuesday, 23.06.2015, 10:00 am, email to* `r_muelle@cs...`*, hand-in box in stairwell 48-6 or in lecture.*

Problem 10

4 points

Consider the set A used in the BLAST heuristic as introduced in the lecture (cf. page 77 of the German lecture notes), i. e., the set of all words $\alpha' \in \Sigma^w$, for which there is a substring $\alpha \in \Sigma^w$ of P, such that the *gap-free* alignment of α and α' has a score better than a given threshold s.

Design an efficient algorithm for computing A, where we assume that w and $c := |\Sigma|$ are small constants.

Problem 11

4 points

In the lecture, we only considered the main principle of the BLAST heuristic. In practice, there are numerous variants, adapted and optimized for the needs of certain use cases, for example:

- Blastn: Search a nucleotide database using a nucleotide query.
- Blastp: Search protein database using a protein query.
- Blastx: Search protein database using a translated nucleotide query.
- Tblastn: Search translated nucleotide database using a protein query.
- Tblastx: Search translated nucleotide database using a translated nucleotide query.

The main application of BLAST is to annotate new sequences based on the search for known homogolous sequences. In this hands-on exercise, you will use the BLAST web service available at `http://www.ncbi.nlm.nih.gov/BLAST` to identify sample sequences (two nucleotide sequences and two proteins). The sequences are available in FASTA format at

`http://wwwagak.cs.uni-kl.de/Veroffentlichungen/AdB-I/AdB-I-SS-15-Exercise-Sheets/AdB-I-SS-15-Sheet-04-Data.html`

Figure out which proteins/genes those sequences represent. Play around with the various parameters of the heuristics and justify your answers based on the results.

Note that there are several databases for searching sequences. Among others there are nr, swissprot, pat, pdb and month for proteins and nr, EST, gss, Chromosome and month for DNA. You should have a look at the documentation of BLAST.

Problem 12

5 points

Consider the asymmetric random walk with probability $p < \frac{1}{2}$ for going one step upwards and $1 - p$ for going one step downwards. Determine the expected length of an excursion of this walk.

Formally, the walk is defined as follows: Let $X_1, X_2, \ldots$ be i. i. d.[1] random variables over $\{-1, 1\}$, with $\Pr[X_i = 1] = p$. Then, the sequence of random variables $S_0, S_1, \ldots$ with $S_n = \sum_{j=1}^{n} X_j$ is called the (asymmetric) random walk with parameter p. The length L of an excursion of this walk is the random variable defined as

$$L = \begin{cases} 0, & \text{if } S_1 = -1; \\ \min\{i : S_i = 0 \wedge i > 0\}, & \text{otherwise.} \end{cases}$$

Your task is thus to compute $\mathbb{E}[L]$ as a function of $p \in [0, \frac{1}{2})$.

[1] **i**ndependent and **i**dentically **d**istributed