

2. November 2005

Abgabe: 9. November 2005

Aufgabe 1: In einer rein zufällig aus $\{A, C, G, T\}$ zusammengesetzten Sequenz der Länge n werden mit dem Knuth-Morris-Pratt-Algorithmus (einzeln) die Muster AAA und ACA gesucht. Beantworten Sie folgende Fragen durch wahrscheinlichkeitstheoretische Argumentation oder approximativ mittels Computersimulation:

- (a) Geben Sie die jeweilige Funktion π an.
- (b) Wie oft kommen die beiden Muster in der Sequenz vor (in Erwartung)?
- (c) Berechnen Sie für jedes der beiden Muster die erwartete Anzahl an Positionen bis das Muster zum ersten Mal beginnt. (Oder begründen Sie zumindest ob die Erwartungswerte für beide Muster gleich sind.)
- (d) Wieviele Vergleiche werden im Schnitt (Erwartungswert!) durchgeführt bis das Muster zum ersten Mal beginnt?
- (e) Zwei Spieler wetten welches der beiden Muster als erstes in der Sequenz vorkommt. Ist das ein faires Spiel? Begründen Sie!
- (f) Beide Muster sollen simultan mit dem Aho-Corasick-Algorithmus gesucht werden. Geben sie den Schlüsselwort-Baum mit *failure links* an.
- (g) Wieviele Vergleiche führt die AC-Suche für diese Muster im Mittel bei einer unendlich langen zufälligen Sequenz aus bis eines der Muster gefunden ist?

Aufgabe 2: Unter www.informatik.uni-frankfurt.de/~metzler/WS0506/fin den Sie Sequenzdaten, die Sie mit BLAST (www.ncbi.nlm.nih.gov/blast) auswerten sollen.

- (a) Bilden Sie aus den 10 Sequenzen aus `ZehnProteine` 5 Paare durch zufälliges Zuordnen. Welche Score-Schemata und Gap-Penalties sind für die einzelnen Paare besonders gut geeignet? Untersuchen Sie das möglichst genau und Begründen Sie Ihr Ergebnis.
- (b) Führen Sie die zu (a) analogen Untersuchungen mit 3 zufällig zusammengesetzten Paaren von Sequenzen aus `ZehnDNAs` durch.
- (c) Die Sequenzen in `ZehnPseudos` wurden durch einen Zufallsgenerator erzeugt. Einige davon enthalten jedoch mutierte Fragmente aus echten Protein-Sequenzen. Welche der Sequenzen sind das und aus welchen Proteinen stammen die Fragmente?

(d) freiwillige Zusatzaufgabe: Wie und in welchem Sinne sind die Sequenzen aus `ZehnProteine` (bzw. `ZehnDNAs`) miteinander verwandt?

Aufgabe 3: Wir interessieren uns nun für lokale Alignments, die die Eigenschaft haben, dass es kein Alignment mit höherem Score gibt, das mit dem selben Basenpaar beginnt oder mit dem selben Basenpaar endet wie das gegebene Alignment. Schreiben Sie ein möglichst effizientes Programm, welches für ein gegebenes Sequenzenpaar den Score jedes solchen Alignments ausgibt, falls dieser eine eingegebene Schranke u übersteigt.