

1. November 2006

**Abgabe: 8. November 2006**

**Aufgabe 1:** In einer rein zufällig aus  $\{A, C, G, T\}$  zusammengesetzten Sequenz der Länge  $n$  werden mit dem Aho-Corasick-Algorithmus simultan die Muster **AAAAA** und **ACACC** gesucht. Beantworten Sie folgende Fragen durch wahrscheinlichkeitstheoretische Argumentation oder approximativ mittels Computersimulation:

- (a) Geben sie den Schlüsselwort-Baum mit *failure links* an.
- (b) 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/WS0607/](http://www.informatik.uni-frankfurt.de/~metzler/WS0607/) finden Sie Sequenzdaten, die Sie mit BLAST ([www.ncbi.nlm.nih.gov/blast](http://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:** Implementieren Sie ein Programm, das nach Eingabe von "Aminosäurehäufigkeiten"  $(p_1, \dots, p_{20})$  und einer Scorematrix  $(s_{ij})_{i,j \leq 20}$  mit  $\sum_{ij} p_i p_j s_{ij} < 0$  die Lösung  $\lambda^* > 0$  der Gleichung  $\sum_{ij} p_i p_j e^{\lambda s_{ij}} = 1$  numerisch berechnet. Erproben Sie Ihr Programm mit verschiedenen Eingabewerten und stellen Sie den Zusammenhang zwischen der Rechengenauigkeit und der benötigten Anzahl an Iterationen der von Ihnen verwendeten numerischen Methode dar.