

Übungen zur Vorlesung  
“Algorithmen der Bioinformatik II”  
Wintersemester 2005/2006

---

**Blatt 1**

1. Für einen String  $S$  und  $i \in \{2, \dots, |S|\}$  ist, wie in der Vorlesung erklärt, die *Z-Box* an Position  $i$  der längste Substring, der bei  $i$  anfängt und ein Präfix von  $S$  matcht (d.h. mit einem Präfix von  $S$  identisch ist); die Zahl  $Z_i$  ist die Länge der Z-Box bei  $i$ .
  - (a) Wie kann man in linearer Zeit – d.h. in Zeit  $O(|S|)$  – alle Werte  $Z_i$  berechnen?
  - (b) Gegeben ist ein String  $S = P\$T$ . Wieviele Rechenschritte braucht man, wenn die Werte  $Z_i$  nur für den Teilstring  $T$  von  $S$  ausgerechnet werden sollen?
2. Programmieraufgabe (für Teilnehmer ohne Vorkenntnisse geeignet): Implementiere den “naiven” Algorithmus für das Pattern-Matching.
3. Programmieraufgabe (für Teilnehmer, die den ersten Teil der Vorlesung gehört haben): Implementiere den Z-Box-Algorithmus und seine Anwendung für das Pattern-Matching.
4. Vergleiche die Laufzeiten der Methoden, die in Aufgaben 2 und 3 implementiert wurden, auf realen und simulierten Daten. Die Strings  $T$  und  $P$  sollen dabei folgendermaßen erzeugt werden:
  - (a) Schreibe ein Programm, das Zufallssequenzen über dem Alphabet  $\mathcal{A} = \{A, T, C, G\}$  erzeugt (Wahrscheinlichkeit für jedes Zeichen an jeder Position = 0,25, Positionen unabhängig voneinander).
  - (b) Verwende beliebige genomische Sequenzen aus *GenBank*:  
<http://www.ncbi.nlm.nih.gov/>

Lösungen werden am Donnerstag, 3. November, besprochen.