

15. November 2006

**Abgabe: 22. November 2006**

Die Aufgaben 1 und 2 beziehen sich auf das in der Vorlesung beschriebene Modell zum Finden von CpG-islands mit den versteckten Zuständen  $C^+$ ,  $G^+$ ,  $N^+$ ,  $C^-$ ,  $G^-$ ,  $N^-$ . Lassen Sie (insbesondere in Aufgabe 2) für die Emissionswahrscheinlichkeiten von A und T aus  $N^+$  und  $N^-$  auch andere Werte als  $\frac{1}{2}$  zu.

**Aufgabe 1:** Schreiben Sie ein Programm zum Simulieren von DNA-Sequenzen gemäß dem CpG-islands-HMM.

**Aufgabe 2:** Programmieren Sie den Baum-Welch-Algorithmus für das CpG-islands-HMM und erproben Sie Ihr Programm an simulierten Sequenzen aus Aufgabe 1 sowie einigen menschlichen DNA-Sequenzen aus den einschlägigen Datenbanken.

**Aufgabe 3:** Es sei  $X_1, \dots, X_n$  die versteckte Kette eines HMM mit Emissionen  $S_1, \dots, S_n$ . Geben Sie einen Algorithmus an, der für  $i \leq n$ ,  $y \in \mathcal{Z}$  und Beobachtungen  $(s_1, \dots, s_n) \in \mathcal{A}^n$  folgendes berechnet:

$$\max_{(x_1, \dots, x_n) \in \mathcal{Z}^n} \text{Ws}(X_1 = x_1, \dots, X_n = x_n, | X_i = y, S_1 = s_1, \dots, S_n = s_n)$$