Übungen Sequenzanalyse I
Sommersemester 2007

Dr. Ivo Große

Institut für Informatik
Universität Halle

# Blatt 1

### Aufgabe 1.1

Beweisen Sie, dass das links-rechts Markovmodell erster Ordnung äquivalent zum rechts-links Markovmodell erster Ordnung ist.

Zusatzaufgabe: Beweisen Sie diese Äquivalenz für ein Markov Modell zweiter Ordnung.

Zusatzaufgabe: Beweisen Sie diese Äquivalenz für ein Markov Modell $M$-ter Ordnung.

### Aufgabe 1.2

Beweisen Sie die vier Identitäten aus der Vorlesung für $Y_1$ und $Y_2$.

### Aufgabe 1.3

Betrachten Sie die beiden Vektoren $\vec{q}_1^{\,\mathrm{egal}} = (1/4, 1/4, 1/4, 1/4)$ und $\vec{N}_1 = (N_A, N_C, N_G, N_T) = (2500, 2500, 2500, 2500)$, und generieren Sie einen Datensatz aus 104 Dinukleotiden so, dass genau $N_a$ Dinukleotide mit Nukleotid $a \in \{A, C, G, T\}$ beginnen, und dass das zweite Nukleotid $b \in \{A, C, G, T\}$ statistisch unabhängig vom ersten mit Wahrscheinlichkeit $q_b^{\mathrm{egal}}$ generiert wird. Berechnen Sie für diesen Datensatz die vier Teststatistiken $Y_E, Y_1, Y_2, Y_3$, wiederholen Sie alles 109 mal, und plotten Sie die vier Histogramme.

Wiederholen Sie diese Prozedur für $\vec{N}_2 = (1000, 1000, 1000, 7000)$, $\vec{N}_3 = 100 \times \vec{N}_1$ und $\vec{N}_4 = 100 \times \vec{N}_2$ sowie für $\vec{q}_2^{\,\mathrm{egal}} = (1/10, 1/10, 1/10, 7/10)$. (Für $\vec{N}_3$ und $\vec{N}_4$ enthält jeder Datensatz natürlich 106 Dinukleotide.) Vergleichen Sie nun für jede der vier Teststatistiken die acht erhaltenen Histogramme. Für welche der Teststatistiken hängen die Histogramme nur schwach von $\vec{N}$ und/oder von $\vec{q}_1^{\,\mathrm{egal}}$ ab?