

22. November 2006

Abgabe: 29. November 2006

Aufgabe 1: Implementieren Sie das “Viterbi-Training” zum Schätzen der Modellparameter für das in der Vorlesung beschriebene Modell zum Finden von CpG-islands. Erproben Sie das Programm an den in Aufgabe 1 des vorherigen Übungsblatts simulierten Sequenzen und Vergleichen Sie an Hand der Ergebnisse die Genauigkeit und den Rechzeitaufwand mit dem des Baum-Welch-Algorithmus aus Aufgabe 2 des vorherigen Übungsblatts.

Aufgabe 2: Gegeben sei eine Wahrscheinlichkeitsverteilung $\mathcal{P} = (p_1, \dots, p_n)$ auf einem Alphabet $\mathcal{A} = (a_1, \dots, a_n)$. Die Positionen einer zufälligen Sequenz $S = (S_1, \dots, S_m)$ seien unabhängig voneinander gemäß \mathcal{P} mit Elementen aus \mathcal{A} besetzt. Für $i \leq n$ sei N_i die Anzahl an Positionen in S , an denen ein a_i steht.

- Wie berechnet man für eine gegebene Folge $s = (s_1, \dots, s_m)$ die Wahrscheinlichkeit, dass $S = s$ gilt?
- Wie berechnet man für $v = (v_1, \dots, v_n) \in \mathbb{N}_0^n$ die Wahrscheinlichkeit, dass $(N_1, \dots, N_n) = v$ gilt?
- Wie berechnet man für eine gegebene Sequenz den Maximum-Likelihood-Schätzer $(\hat{p}_1, \dots, \hat{p}_n)$ für (p_1, \dots, p_n) ?

Aufgabe 3: Besorgen Sie sich 5 möglichst verschiedene, lange biologische Sequenzen (Proteine, DNA, . . .) und schätzen Sie für diese sowie für die 5 simulierten Sequenzen in der Datei Markoffordnung.txt die Markoffordnung. Erläutern Sie, wieso die dabei von Ihnen verwendete Methode sinnvoll ist.