9. Übung zur Vorlesung
"Bioinformatische Methoden in der Genomforschung"

Sebastian Böcker, Martin Engler

Ausgabe: 14.01.2016
Abgabe: 21.01.2016

Aufgabe 1 (1 Punkte)

Leiten Sie den Maximum-Likelihood-Schätzer der Parameter eines (a) homogenen und (b) inhomogenen Markov-Modells 0. Ordnung her.

Aufgabe 2 (5 Punkte)

Bestimmen Sie den Maximum-Likelihood-Schätzer

1. für ein homogenes Markov-Modell 0. Ordnung basierend auf dem Datensatz `seq_hMM_1` und berechnen Sie die Log-Likelihoods der Datensätze `seq_hMM_1` und `seq_hMM_2`.
2. für ein inhomogenes Markov-Modell 0. Ordnung basierend auf dem Datensatz `seq_iMM_1` und berechnen Sie die Log-Likelihoods der Datensätze `seq_iMM_1` und `seq_iMM_2`.

Aufgabe 3 (5 Punkte)

Versuchen Sie, den Maximum-Likelihood-Schätzer der Parameter eines OOPS-Modells analytisch zu bestimmen. Wo bzw. warum scheitert die analytische Maximierung der Log-Likelihood?

Aufgabe 4 (5 Punkte)

Sei C eine Konstante bzgl. θ und $\delta_{i,j}$ das Kronecker-Delta mit $\delta_{i,j} = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j \end{cases}$

Außerdem sei:

$$H_a^l = \sum_{n=1}^{N} \sum_{u=1}^{L-W+1} \gamma_{u,n}^{(t)} \delta_{X_{n,u+l-1},a}$$

Betrachten Sie den EM-Algorithmus für das OOPS-Modell und zeigen Sie, dass für die Q-Funktion folgendes gilt:

$$Q(\theta, \theta^{(t)}, X) = C + \sum_{l=1}^{W} \sum_{a \in \{A,C,G,T\}} H_a^l \ln \theta_a^l$$

Bonusaufgabe (15 Punkte)

Gegeben sei der Datensatz `H1-hESC`. Dieser enthält 1088 Sequenzen mit variabler Länge und jede Sequenz enthält genau ein Motiv der Länge 20 bp.

Implementieren Sie den EM-Algorithmus für das OOPS-Modell und wenden Sie ihn auf den Datensatz `H1-hESC` an. Initialisieren Sie Ihre γ_{iu_i}, in dem Sie zufällig ein u_i von einer Gleichverteilung ziehen und das zugehörige $\gamma_{iu_i} = 1$ setzen. Plotten Sie $\ln(P(\underline{\underline{x}}|\theta^{(t)}))$ für jeden Iterationsschritt und stoppen Sie Ihren Algorithmus, wenn $\ln(P(\underline{\underline{x}}|\theta^{(t+1)})) - \ln(P(\underline{\underline{x}}|\theta^{(t)})) < 10^{-5}$ ist. Wiederholen Sie den EM-Algorithmus 10 mal.

1. Bestimmen Sie das Maximum der erreichten $\ln(P(\underline{\underline{x}}|\theta^{(t)}))$. In wie vielen der 10 EM-Läufe wurde diese maximale Log-Likelihood erreicht? Wie lautet die maximale $\ln(P(\underline{\underline{x}}|\theta^{(t)}))$ und wie lauten die dazugehörigen Modellparameter $\hat{\theta}_{ML}$?
2. Stellen die Modellparameter $\hat{\theta}_{ML}$ zusätzlich als Sequenzlogo dar.