Übungen Sequenzanalyse I
Sommersemester 2007

Dr. Ivo Große

Institut für Informatik
Universität Halle

Blatt 3

Aufgabe 3.1 (6 Punkte)

Ein Datensatz X computergenerierter DNA-Sequenzen bestehe aus 5000 Sequenzen der Länge $L = 9$ bp, die durch ein homogenes Markov Modell nullter Ordnung mit den Wahrscheinlichkeiten $p(A) = p(T) = 0.45$ und $p(C) = p(G) = 0.05$ generiert wurden, und aus weiteren 5000 Sequenzen der Länge $L = 9$ bp, die durch ein homogenes Markov Modell nullter Ordnung mit den Wahrscheinlichkeiten $p(A) = p(T) = 0.05$ und $p(C) = p(G) = 0.45$ generiert wurden.

(a) Wie sieht die PWM des Datensatzes X aus? Welche falschen Schlussfolgerungen könnten Sie aus dieser PWM über den Datensatz ziehen, wenn Sie nicht wüssten, wie der Datensatz tatsächlich generiert wurde?

(b) Wie sieht die Dinukleotid-PWM (also die WAM) des Datensatzes X aus? Welche falschen Schlussfolgerungen könnten Sie aus dieser Dinukleotid-PWM über den Datensatz ziehen, wenn Sie nicht wüssten, wie der Datensatz X tatsächlich generiert wurde?

(c) Wie sehen die $L \times L$ Matrizen $Y_1(i, j)$, $Y_2(i, j)$ und $Y_3(i, j)$ des Datensatzes X aus? Welche falschen Schlussfolgerungen könnten Sie aus diesen Matrizen über den Datensatz ziehen, wenn Sie nicht wüssten, wie der Datensatz X tatsächlich generiert wurde?

(d) Welche der folgenden Modelle würden sich gut zur Modellierung der Sequenzen des Datensatzes X eignen? Geben Sie für jedes Modell stichpunktartig Gründe an, warum es sich gut bzw. schlecht zur Modellierung der Sequenzen des Datensatzes X eignen würde.

   (i) PWM-Modell - inhomogenes Markov Modell nullter Ordnung

   (ii) WAM-Modell - inhomogenes Markov Modell erster Ordnung

   (iii) inhomogenes Markov Modell zweiter Ordnung

   (iv) Bayes Netz, welches auch statistische Abhängigkeiten zwischen nicht-nächsten Nachbarn modelliert

   (v) PWM-Mischmodell mit 2 Klassen

   (vi) WAM-Mischmodell mit 2 Klassen

(vii) Mischung zweier inhomogener Markov Modelle zweiter Ordnung

(viii) Bayes Netz Mischmodell mit 2 Klassen

(ix) PWM-Mischmodell mit 3 Klassen

(x) WAM-Mischmodell mit 3 Klassen

(xi) Mischung dreier inhomogener Markov Modelle zweiter Ordnung

(xii) Bayes Netz Mischmodell mit 3 Klassen

Aufgabe 3.2 (6 Punkte)

Berechnen Sie – für jede der vier Sequenzen `seq_1`, `seq_2`, `seq_3` und `seq_4` – $Y_1(i,j)$, $Y_2(i,j)$ und $Y_3(i,j)$ für alle Positionen $i,j = 1,2,...,L$, und stellen Sie die zwölf $L \times L$ Matrizen Y_1, Y_2 und Y_3 grafisch dar. Unter der Annahme der Nullhypothese, daß es keine statistischen Abhängigkeiten zwischen X_{j-1} und X_j gibt, sind Y_1, Y_2 und Y_3 χ^2-verteilt mit 9 Freiheitsgraden. Beantworten Sie die folgenden Fragen für jede der zwölf Matrizen: Für welche Paare (i,j) finden Sie *statistisch signifikante* Abhängigkeiten, wenn Sie einen P-Wert kleiner als 0.01 als signifikant betrachten? Gibt es statistisch signifikante Abhängigkeiten auch zwischen nicht-nächsten Nachbarn? Beschreiben Sie Ihre Beobachtungen. Welche Schlußfolgerungen ergeben sich daraus für die Modellierung von Donorstellen und Nicht-Donorstellen? Sind WAM-Modelle tatsächlich ideal für die Modellierung von Donorstellen und Nicht-Donorstellen geeignet?

Aufgabe 3.3 (4 Punkte)

Der Datensatz `coin` enthält 100 Binärsequenzen der Länge $L = 10$, die durch ein Mischmodell zweier homogener Markov Modelle nullter Ordnung generiert wurden. Hierbei stehen Z für Zahl und W für Wappen. Die beiden Klassenwahrscheinlichkeiten $\pi_1 = \pi_2 = 0.5$ sind extern vorgegeben. Die einzigen zu schätzenden Parameter dieses Modells sind die Wahrscheinlichkeiten q_1 und q_2 der beiden Münzen 1 und 2, Zahl zu werfen.

(a) Schätzen Sie die Parameter q_1 und q_1 mittels Maximum Likelihood Prinzip unter Nutzung der gegebenen Klassenzugehörigkeiten.

(b) Ignorieren Sie für die folgenden drei Teilaufgaben die Klassenzugehörigkeiten, d. h. betrachten Sie die Klassenzugehörigkeiten im folgenden als nicht gegeben. Plotten Sie die Log-Likelihood als Funktion von q_1 und q_2.

(c) Bestimmen Sie die Maxima und Maximalstellen dieser Funktion mit geringem Aufwand durch ein Verfahren Ihrer Wahl (gitterbasierte Abrasterung, Maximumssuche per Auge, Gradientenanstieg, etc.). Vergleichen Sie die Maximalstellen mit den in Aufgabe 3.3 (a) geschätzten Werten, und diskutieren Sie die Unterschiede.

(d) Versuchen Sie, die Maximalstellen analytisch zu bestimmen, indem Sie die Log-Likelihood nach q_1 und q_2 ableiten und beide Ableitungen Null setzen. Worin liegt das Problem, dieses Gleichungssystem (mit lediglich zwei Gleichungen und zwei Unbekannten) analytisch zu lösen?

Abgabetermin: 23. Mai