Übungen Sequenz- und Expressionsdatenanalyse
Sommersemester 2004
Dr. Ivo Grosse

IPK, Gatersleben
Institut für Informatik,
Universität Halle

Blatt 9

Aufgabe 9.1 (4 Punkte) Ein Datensatz X computergenerierter DNA Sequenzen bestehe aus 5000 Sequenzen der Länge $W = 9$ 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 $W = 9$ 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 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 $W \times W$ Matrizen $Y_1(i, j)$ und $Y_2(i, j)$ 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 würde 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) Weight Array 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) Weight Array 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) Weight Array Mischmodell mit 3 Klassen

(xi) Mischung dreier inhomogener Markov Modelle zweiter Ordnung

(xii) Bayes Netz Mischmodell mit 3 Klassen

Aufgabe 9.2 (8 Punkte) Die Datensätze `pgrc-16.ipk-gatersleben.de/~grosse/Lehre/SEDA/SS04/Uebungen/Serie_6/seq_1.txt` und `pgrc-16.ipk-gatersleben.de/~grosse/Lehre/SEDA/SS04/Uebungen/Serie_6/seq_2.txt` enthalten Alignments von Donorstellen der Länge $W = 9$ bp, und die Datensätze `pgrc-16.ipk-gatersleben.de/~grosse/Lehre/SEDA/SS04/Uebungen/Serie_6/seq_3.txt` und `pgrc-16.ipk-gatersleben.de/~grosse/Lehre/SEDA/SS04/Uebungen/Serie_6/seq_4.txt` enthalten Alignments von Nicht-Donorstellen der Länge $W = 9$ bp.

Benutzen Sie die Datensätze seq_1 und seq_3 zum Trainieren eines PWM Mischmodells mit $K = 2$ Klassen und eines Weight Array Mischmodells mit $K = 2$ Klassen. Wie lauten die beiden PWMs bzw. die beiden Dinukleotid-PWMs? Vergleichen Sie diese PWMs bzw. Dinukleotid-PWMs mit der PWM bzw. Dinukleotid-PWM aus Aufgabe 7.1.

Plotten Sie die ROC Kurven für die beiden Modellkombinationen (PWM Mischmodell für Donor und PWM Mischmodell für Nicht-Donor, Weight Array Mischmodell für Donor und Weight Array Mischmodell für Nicht-Donor) für die Datensätze seq_2 und seq_4. Welche Modellkombination liefert die genauere Klassifizierung von Donorstellen und Nicht-Donorstellen?

Vertauschen Sie die Datensätze seq_1 und seq_3 mit den Datensätzen seq_2 und seq_4, und wiederholen Sie die Analyse. Wie robust sind die beiden ROC Kurven?

Wiederholen Sie die Analyse für $K = 1$, $K = 3$, $K = 4$ und $K = 5$. Testen Sie, dass Ihre Ergebnisse für $K = 1$ identisch sind zu Ihren Ergebnissen aus Aufgabe 7.2. Welche der 10 Modellkombinationen liefert die genaueste Klassifizierung von Donorstellen und Nicht-Donorstellen?

Hinweis: als Mass für die Genauigkeit der Klassifikation eignet sich die Fläche unter der ROC Kurve.

Abgabetermin: 16. Juni (in der Übung)