Übungen Sequenz- und Expressionsdatenanalyse

Sommersemester 2004

Dr. Ivo Grosse

IPK, Gatersleben
Institut für Informatik,
Universität Halle

Blatt 10

Aufgabe 10.1 (4 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.

Berechnen Sie – für jede der vier Sequenzen seq_1, seq_2, seq_3 und seq_4 – $\tilde{Y}_1(i,j)$ und $\tilde{Y}_2(i,j)$ für alle Positionen $i,j = 1,2,...,W$, und stellen Sie die acht $W \times W$ Matrizen $\tilde{Y}_1$ und $\tilde{Y}_2$ grafisch dar. Unter der Annahme der Nullhypothese, daß es keine statistischen Abhängigkeiten zwischen X_i und X_j gibt, ist sowohl $\tilde{Y}_1$ als auch $\tilde{Y}_2$ χ^2-verteilt mit 3 Freiheitsgraden. Für welche Paare (i,j) finden Sie *statistisch signifikante* Abhängigkeiten, wenn Sie einen P-Wert kleiner als 0.01 als signifikant betrachten? Vergleichen Sie die acht $W \times W$ Matrizen $\tilde{Y}_1$ und $\tilde{Y}_2$ mit den acht $W \times W$ Matrizen Y_1 und Y_2 aus Aufgabe 7.3. Welche Gemeinsamkeiten und welche Unterschiede können Sie beobachten, und wie würden Sie diese interpretieren?

Aufgabe 10.2 (8 Punkte) Benutzen Sie die Datensätze seq_1 und seq_3 zum Trainieren von MDD Modellen. Setzen Sie im MDD Algorithmus die Schwelle T so, dass die Wahrscheinlichkeit für $S_i > T$ gleich 0.01 ist, und setzen Sie die Schwelle $U = 100$. Drucken Sie den Entscheidungsbaum sowie die PWMs an dessen Blättern aus. Plotten Sie die ROC Kurve für die Datensätze seq_2 und seq_4, und vergleichen Sie die ROC Kurve des MDD Modells mit den ROC Kurven der PWM Mischmodelle aus Aufgabe 9.2. Welche Modellkombination liefert die genaueste Klassifizierung von Donorstellen und Nicht-Donorstellen? Sei K die Anzahl der Blätter des Entscheidungsbaums für Datensatz seq_1, und sei L die Anzahl der Blätter des Entscheidungsbaums für Datensatz seq_3. Benutzen Sie nun seq_1 zum Trainieren eines PWM Mischmodells mit K Klassen, benutzen Sie seq_3 zum Trainieren eines PWM Mischmodells mit L Klassen, und plotten Sie die ROC Kurve für seq_2 und seq_4. Vergleichen Sie diese ROC Kurve mit der ROC Kurve für das MDD Modell.

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 ROC Kurven des MDD Modells und des dazugehörigen PWM Mischmodells?

Aufgabe 10.3 (10 Punkte) **Freiwillige Zusatzaufgabe:**

Benutzen Sie den Datensatz seq_1 zum Trainieren eines PWM Mischmodells mit K Klassen und eines Weight Array Mischmodells mit K Klassen, und benutzen Sie seq_3 zum Trainieren eines PWM Mischmodells mit L Klassen und eines Weight Array Mischmodells mit L Klassen. Berechnen Sie für gegebenes K und L die Fläche unter der ROC Kurve, $F(K, L)$, für jede der vier Modellkombinationen (PWM Mischmodell für Donor und PWM Mischmodell für Nicht-Donor, PWM Mischmodell für Donor und Weight Array Mischmodell für Nicht-Donor, Weight Array 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. Schreiben Sie für $K = 1, 2, ..., 5$ und $L = 1, 2, ..., 5$ die Werte $F(K, L)$ in eine 5×5 Matrix. Vergleichen Sie die vier 5×5 Matrizen – eine Matrix füer jede der vier Modellkombinationen – miteinander. Welche Modellkombination liefert die genaueste Klassifizierung von Donorstellen und Nicht-Donorstellen? Was sagt uns das über die "Struktu" der zu Grunde liegenden Daten? D. h. welche statistischen Abhängigkeiten sind modellierungswürdig, welche Modelle neigen zum Übertraining, etc.?

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 vier 5×5 Matrizen?

Abgabetermin: 23. Juni (in der Übung)