Übungen Sequenz- und Expressionsdatenanalyse

Sommersemester 2004

Dr. Ivo Grosse

IPK, Gatersleben
Institut für Informatik,
Universität Halle

Blatt 7

Aufgabe 7.1 (2 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.

Schätzen Sie die beiden $4 \times W$ Gewichtsmatrizen (weight matrices) und die beiden verallgemeinerten $16 \times (W-1)$ Dinukleotid Gewichtsmatrizen (weight arrays) dieser Alignments. Benutzen Sie dafür den MP-Schätzer mit einem Pseudo-Zähler, d.h. benutzen Sie $\frac{\hat{p}_{ij}=N_{ij}+1}{N_{.j}+M}$, wobei $M = 4$ im Falle der Gewichtsmatrizen und $M = 16$ im Falle der verallgemeinerten Dinukleotid Gewichtsmatrizen.

Freiwillig: Schätzen Sie die $4 \times W$ Gewichtsmatrix (weight matrix) und die verallgemeinerte $16 \times (W-1)$ Dinukleotid Gewichtsmatrix (weight array) des Alignments von Sigma-70 Faktor Bindungsstellen aus dem Datensatz `http://pgrc-16.ipk-gatersleben.de/~grosse/Lehre/SEDA/SS04/Uebungen/Serie_1/sigma_70.txt` aus Aufgabe 1.1.

Aufgabe 7.2 (6 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 Modells und eines WA Modells, und plotten Sie die ROC Kurven für alle vier Modellkombinationen (PWMM für Donor und PWMM für Nicht-Donor, PWMM für Donor und WAM für Nicht-Donor, WAM für Donor und PWMM für Nicht-Donor, WAM für Donor und WAM für Nicht-Donor) für die Datensätze seq_2 und seq_4. Welche Modellkombination ist optimal für die 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 vier ROC Kurven?

Aufgabe 7.3 (4 Punkte)

Berechnen Sie – für jede der vier Sequenzen seq_1, seq_2, seq_3 und seq_4 – $Y_1(i,j)$ und $Y_2(i,j)$ für alle Positionen $i,j = 1,2,...,W$, und stellen Sie die acht $W \times W$ Matrizen Y_1 und Y_2 grafisch dar. Unter der Annahme der Nullhypothese, daß es keine statistischen Abhängigkeiten zwischen X_{j-1} und X_j gibt, ist sowohl Y_1 als auch Y_2 χ^2-verteilt mit 9 Freiheitsgraden. 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/Nicht-Donorstellen? Sind WA Modelle tatsächlich ideal für die Modellierung von Donorstellen/Nicht-Donorstellen geeignet?

Abgabetermin: 2. Juni (in der Übung)