Übungen Sequenzanalyse I
Sommersemester 2007

Dr. Ivo Große

Institut für Informatik
Universität Halle

Blatt 2

Aufgabe 2.1 (4 Punkte)

Die Datensätze `seq_1` und `seq_2` enthalten Alignments von Donorstellen der Länge $L = 9$ bp.

(a) Schätzen Sie die beiden $L \times 4$ Gewichtsmatrizen (position weight matrices) und die beiden verallgemeinerten $(L-1) \times 16$ Dinukleotid-Gewichtsmatrizen (weight array matrices) dieser Alignments. Benutzen Sie dafür den ML-Schätzer.

(b) Ein PWM-Modell ordnet jeder Sequenz der Länge L eine Wahrscheinlichkeit zu. Beweisen Sie allgemein, dass die Sequenz, für die diese Wahrscheinlichkeit maximal ist, gleich der Konsensussequenz ist.

(c) Bestimmen Sie die Konsensussequenz des o. g. Datensatzes, also die Sequenz, für die die PWM-Wahrscheinlichkeit maximal ist.

(d) Ein WAM-Modell ordnet jeder Sequenz der Länge L eine Wahrscheinlichkeit zu. Bestimmen Sie die Sequenz, für die die WAM-Wahrscheinlichkeit maximal ist, und vergleichen Sie diese mit der Konsensussequenz.

(e) Wie häufig tauchen die Konsensussequenz und die Sequenz mit maximaler WAM-Wahrscheinlichkeit im gegeben Datensatz auf?

(f) Welche Sequenz taucht am häufigsten im gegebenen Datensatz auf, und wie häufig?

Aufgabe 2.2 (12 Punkte)

Die Datensätze `seq_3` und `seq_4` enthalten Alignments von Nicht-Donorstellen der Länge $L = 9$ bp.

Benutzen Sie die Datensätze `seq_1` und `seq_3` zum Trainieren eines PWM-Modells und eines WAM-Modells, und plotten Sie die ROC-Kurven für alle vier Modellkombinationen (PWMM für Donor und PWMM für Nicht-Donor, PWMM für Donor und WAMM für Nicht-Donor, WAMM für Donor und PWMM für Nicht-Donor, WAMM für Donor und WAMM für Nicht-Donor) für die Datensätze `seq_2` und `seq_4`. Welche Modellkombination ist optimal für die Klassifizierung von Donorstellen und Nicht-Donorstellen, wenn Sie die Fläche unter der ROC-Kurve (AUC) als Gütemaß verwenden?

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 und die vier AUC-Werte?

Abgabetermin: 9. Mai