Übungen Sequenzanalyse II
Wintersemester 2007/08

Prof. Dr. Ivo Große, Dipl.-Bioinf. Jan Grau

Institut für Informatik
Universität Halle

Blatt 4

Aufgabe 4.1 (5 Punkte)

Die Datensätze `seq_1` und `seq_2` enthalten Alignments von Donorstellen der Länge $L = 9$ bp, und 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`. Verwenden Sie hierfür als A-Priori-Dichte ein Produkt aus Dirichlet-Dichten mit einer *equivalent sample size* $\epsilon = 16$ für jedes Modell. 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?

Aufgabe 4.2 (5 Punkte)

Variieren Sie nun ϵ zwischen dem kleinstmöglichen Wert und 100, und plotten Sie die vier AUC-Werte als Funktionen von ϵ. Plotten Sie zusätzlich die vier AUC-Werte der Maximum-Likelihood-Klassifikatoren als horizontale Linien, und vergleichen Sie die vier Kurvenpaare.

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 Ergebnisse?

Abgabetermin: 21. November