Übungen Sequenzanalyse II
Wintersemester 2007/08

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

Institut für Informatik
Universität Halle

Blatt 5

Aufgabe 5.1 (10 Punkte)

Benutzen Sie die Datensätze `seq_1` und `seq_3` zum Trainieren eines PWM-Mischmodells mit $K = 2$ Klassen und eines WAM-Mischmodells mit $K = 2$ Klassen. Verwenden Sie hierfür als A-Priori-Dichte ein Produkt aus Dirichlet-Dichten mit einer *equivalent sample size* $\epsilon = 64$ für jedes Mischmodell. 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 4.1.

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

Wiederholen Sie die Analyse für $K = 1$, $K = 3$ und $K = 4$. Vergleichen Sie Ihre Ergebnisse für $K = 1$ mit Ihren Ergebnissen aus Aufgabe 4.1. Welche der acht Modellkombinationen ($\{K = 1, \ldots, 4\} x \{PWM, WAM\}$) 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 (AUC).

Aufgabe 5.2 (5 Punkte)

Variieren Sie nun für $K = 1$ und $K = 2$ ϵ zwischen dem kleinstmöglichen Wert und 256, und plotten Sie die 4 AUC-Werte als Funktionen von ϵ. Plotten Sie zusätzlich die 4 AUC-Werte der entsprechenden Maximum-Likelihood-Klassifikatoren als horizontale Linien, und vergleichen Sie die 4 Kurvenpaare.

Abgabetermin: 28. November