Übungen Algorithmen der Bioinformatik II
Wintersemester 2007/08

Prof. Dr. Stefan Posch, Dr. Birgit Möller

Institut für Informatik
Universität Halle

# Blatt 11

**Aufgabe 11.1** Der Datensatz "sigma70_fg.txt" auf der Homepage zur Vorlesung besteht aus 238 Sigma-70 Bindungsstellen der Länge 12. Der Datensatz "sigma70_bg.txt" ist ein zugehöriger Hintergrunddatensatz von ebenfalls 238 Sequenzen der Länge 12. Teilen Sie beide Datensätze jeweils in zwei Hälften, so dass Sie zwei TFBS-Datensätze $F_1$ und $F_2$ und zwei Hintergrunddatensätze $B_1$ und $B_2$ erhalten. Für diese Datensätze sollen nun verschiedene Bayes-Klassifikatoren trainiert und getestet werden.

(a) 1. Konstruieren Sie zunächst zwei Bayes-Klassifikatoren $K_1^{ML}$ und $K_2^{ML}$, indem Sie jeweils auf $B_i$ und $F_i, i \in \{1, 2\}$, ein inhomogenes bzw. homogenes MM(0) für die TFBS bzw. den Hintergrund trainieren. Schätzen Sie die Parameter der Modelle dabei mittels ML-Ansatz.

2. Konstruieren Sie dann zwei Bayes-Klassifikatoren $K_1^{MAP}$ und $K_2^{MAP}$, indem wieder jeweils auf $B_i$ und $F_i, i \in \{1, 2\}$, ein inhomogenes bzw. homogenes MM(0) für die TFBS bzw. den Hintergrund trainiert wird. Diesmal sollen die Parameter der Modelle jedoch mittels MAP-Ansatz geschätzt werden.
Nehmen Sie für die MAP-Schätzung einen Dirichlet-Prior $D(\vec{p}|\vec{a})$ an, für dessen Parametervektor $\vec{a} \in R^D$ gelten soll: $a_i = \lambda, \forall i = 1 \ldots D$.
Setzen Sie für diesen ersten Trainingsdurchgang $\lambda = 2$.

(b) Wenden Sie die Klassifikatoren zum Testen jeweils auf beide Datensätze an und bestimmen Sie die Fehlerraten.

(c) Variieren Sie nun für $K_1^{MAP}$ und $K_2^{MAP}$ den Parameter $\lambda$ des Priors und schätzen Sie die Modelle neu. Welchen Einfluss hat $\lambda$ auf die Klassifikationsergebnisse? Vergleichen Sie insbesondere die Fehlerraten eines Klassifikators auf seiner Trainingsmenge mit der Rate auf dem jeweils unbekannten Datensatz.

(d) Berechnen Sie für alle Markov-Modelle aus $K_1^{ML}$, $K_2^{ML}$, $K_1^{MAP}$ und $K_2^{MAP}$ jeweils das Sequenz-Logo. Für $K_1^{MAP}$ und $K_2^{MAP}$ sollen dabei die Modelle mit $\lambda = 2$ zu Grunde gelegt werden.

**Aufgabe 11.2** Gegeben seien $N$ DNA-Sequenzen der Länge $L$ in einem Trainingsdatensatz. Auf diesen Sequenzen wollen wir homogene und inhomogene Markov-Modelle der Ordnung $d$ trainieren. Geben Sie allgemein für ein homogenes bzw. inhomogenes Markov-Modell die Anzahl der zu schätzenden Parameter in Abhängigkeit von $N$, $L$ und $d$ an.

---

Abgabe: 11.01.08