Übungen Algorithmen der Bioinformatik II
Wintersemester 2004/05
Stefan Posch, Birgit Möller, Denis Williams

Institut für Informatik
Universität Halle

Blatt 10

Aufgabe 10.1 Wir modellieren Sequenzen der Länge L über einem festen Alphabet der Größe D mit einem inhomogenen Markov Modell 1-ter Ordnung. Als Daten stehen uns N Sequenzen der Länge L zur Verfügung, die – wie üblich – als paarweise unabhängige Realisierung diese iMM(1) angenommen werden. Als a priori Verteilung der Parameter des iMM(1) nehmen wir statistisch unabhängige Dirichlet-Verteilungen an.

Geben Sie die explizit die ML-, MAP- und MP-Schätzwerte der Parameter des iMM(1) an.

Aufgabe 10.2 Wir betrachten N Sequenzen über einem festen Alphabet, die nun aber unterschiedliche Längen L_n haben.

Geben Sie für folgende Modelle für Sequenzen der Länge $L, L > 5$, jeweils die ML-Schätzer an:

(a) ein homogenes Markov Modell 5-ter Ordnung

(b) ein "periodisches" inhomogenes Markov Modell 5-ter Ordnung

(also eine iMM(5), für das gilt:
$P_l(X_l|X_{l-M}, \cdots, X_{l-1}) = P_k(X_k|X_{k-M}, \cdots, X_{k-1})$, falls $k = l\%3$)

Aufgabe 10.3 Der Datensatz
`http://www.informatik.uni-halle.de/~posch/AG/Lehre/AdB-II_WS04/uebungen/sigma70.txt`
besteht aus 238 Sigma-70 Bindungsstellen der Länge 12.

a) Modellieren Sie eine Sigma-70 Bindungsstellen als iMM(0) und bestimmen Sie die ML-Schätzwerte der Parameter $p_{il} := P(X_l = a_i)$. Diese p_{il} bilden eine Gewichtsmatrix (PWM), wobei die l-te Spalte die Verteilung von X_l beschreibt.

b) Bestimmen Sie die zu dieser Gewichtsmatrix (und zu diesem Datensatz) gehörende Konsensussequenz.

c) Wie häufig taucht die Konsensussequenz im gegeben Datensatz auf?

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

Abgabe: 22.12.04