Übungen Sequenzanalyse I
Sommersemester 2007

Dr. Ivo Große

Institut für Informatik
Universität Halle

# Blatt 6

**Aufgabe 6.1** (2 Punkte)

Der Datensatz `sigma_70` besteht aus 238 Sigma-70-Bindungsstellen der Länge $L = 12$ bp (eine Bindungsstelle je Zeile).

(a) Bestimmen Sie die Gewichtsmatrix $P$ dieser Bindungsstellen, und nutzen Sie zur Schätzung der Wahrscheinlichkeiten den Bayesschätzer $\hat{p}_{\ell m} = \frac{N_{\ell m}+1}{N_\ell+M}$, wobei $N_{\ell m}$ die Anzahl des Nukleotids $m \in \{A, C, G, T\}$ an Position $\ell \in \{1, 2, ..., L\}$ definiert, $N_\ell = \sum_m N_{\ell m}$ die Anzahl der Bindungsstellen definiert und $M = 4$ die Anzahl der verschiedenen Nukleotide definiert.

(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?

**Aufgabe 6.2** (2 Punkte)

Leiten Sie den M-Schritt des EM-Algorithmus für Sequenzmotive her.

**Aufgabe 6.3** (12 Punkte)

Die Datensätze `seq_5` und `seq_6` enthalten je 238 DNA Sequenzen der Länge 300 bp, die alle genau ein Sequenzmotiv der Länge 12 bp enthalten.

Implementieren Sie den EM-Algorithmus, den stochastischen EM-Algorithmus und den Gibbs-Sampling-Algorithmus, und bestimmen Sie mit jedem Algorithmus das Sequenzmotiv (also die PWM) der Länge 12 bp aus jedem der beiden Datensätze. Nutzen Sie im Fall des stochastischen EM-Algorithmus und des Gibbs-Sampling-Algorithmus zur Schätzung der Wahrscheinlichkeiten der PWM den MP-Schätzer $\hat{p}_{\ell m} = \frac{N_{\ell m}+1}{N_\ell+M}$, und setzen Sie in allen Fällen die Wahrscheinlichkeiten für die Hintergrundnukleotide fest auf $p_{0m} := q_m = 1/M$. Betrachten Sie weiterhin die Parameter $p_{iu_i}$ nicht als im E-Schritt zu schätzende *interne* Parameter, sondern als fest vorgegebene *externe* Parameter, und setzen Sie diese externen Parameter $p_{iu_i} = \frac{1}{L_i-W+1}$.

Plotten Sie in jedem Iterationsschritt $i$ das Log-Likelihood-Verhältnis

$$\sum_{n=1}^{N=238} \sum_{\ell=1}^{L=12} \log \frac{p_{\ell x_{u_n+\ell-1}}}{p_{0x_{u_n+\ell-1}}}$$

als Funktion von $i$, wobei $u_n$ die Startposition des Motivs in Sequenz $n$ und $x_j$ das Nukleotid an Position $j$ in Sequenz $n$ definiert. Plotten Sie weiterhin im Fall der beiden Varianten des EM-Algorithmus die Log-Likelihood als Funktion von $i$. Definieren Sie ein geeignetes Initialisierungs- und Abbruchkriterium für jeden der drei Algorithmen, und führen Sie für jeden der beiden Datensätze 50 Simulationen mit jedem der drei Algorithmen durch.

Vergleichen Sie die drei Algorithmen in Bezug auf die Qualität der gefundenen Motive, das Konvergenzverhalten der Algorithmen und die Konvergenzgeschwindigkeit (also die Laufzeit) der Algorithmen.

**Abgabetermin: 4. Juli**