Übungen Sequenz- und Expressionsdatenanalyse

Sommersemester 2004

Dr. Ivo Grosse

IPK, Gatersleben
Institut für Informatik,
Universität Halle

Blatt 1

Aufgabe 1.1 (4 Punkte)

Der Datensatz

`http://pgrc-16.ipk-gatersleben.de/~grosse/Lehre/SEDA/SS04/Uebungen/Serie_1/sigma_70.txt`

besteht aus 238 Sigma-70 Bindungsstellen der Länge $L = 12$ (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 1.2 (4 Punkte)

Führen Sie den Beweis zur folgender effizienten Auswertung von $Q(\vec{\theta}; \vec{\theta}^{(t)})$ im EM-Algorithmus für unabhängige Beobachtungen aus der Vorlesung aus:

$$\boxed{\begin{aligned}
Q(\vec{\theta}; \vec{\theta}^{(t)}) &= \sum_{i=1}^{N} \sum_{\vec{u} \in \mathcal{U}} \left(\alpha_{iu_i} \left(\prod_{j=1}^{N} \gamma_{ju_j}^{(t)} \right) \right) \\
&= \sum_{i=1}^{N} \sum_{u_i=1}^{K_i} \alpha_{iu_i} \gamma_{ju_j}^{(t)} \\
&= \sum_{i=1}^{N} \sum_{u=1}^{K_i} \log P_i(U_i = u, \vec{X}_i = \vec{x}_i | \vec{\theta}) P_i(U_i = u | \vec{X}_i = \vec{x}_i, \vec{\theta}^{(t)})
\end{aligned}}$$

Aufgabe 1.3 *

Bestimmen Sie den ML-Schätzwert für die complete data likelihood für die Modellierung von motifs der Vorlesung und N statistisch unabhägige Beobachtungen $\vec{x} = (\vec{x}_1, \ldots, \vec{x}_N)$ und den (eigentlich unbeochateten) Startpositionen des motifs $\vec{u} = (u_1, \ldots, u_N)$:

$$\vec{\theta}^* = \underset{\vec{\theta}}{\operatorname{argmax}}\, P(\vec{X} = \vec{x}, \vec{U} = \vec{u} | \vec{\theta})$$

Aufgabe 1.4 (2 Punkte)

Beweisen Sie die in der Vorlesung angegebene Formel zur iterativen Berechung der Parameter $\vec{q}_j$ und $\vec{p}_i$ für motifs mit dem EM Algorithmus.

$$\boxed{\begin{aligned} p_{iu}^{(t+1)} &= \gamma_{iu}^{(t)} \\ q_{ja}^{(t+1)} &= \frac{H_{ja}^{(t)}}{\sum_{\tilde{a}}^{A} H_{j\tilde{a}}^{(t)}} \end{aligned}}$$

Hinweis: Stellen Sie die notwendige Bedingung für ein Maxium von Q auf, die mittels Lagrange-Multiplikator die Nebenbedingungen der Parameter einbringt.

Abgabetermin: 15. April, in der Übung