Übungen Sequenz- und Expressionsdatenanalyse

Sommersemester 2004

Dr. Ivo Grosse

IPK, Gatersleben
Institut für Informatik,
Universität Halle

Blatt 2

Aufgabe 2.1 (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.

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

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

Aufgabe 2.2 (4 Punkte)

Lesen Sie den Artikel [Bai95] über MEME (`www.informatik.uni-halle.de/~posch/AG/Lehre/SEDA_SS04/PS/bailey95.pdf`)

(a) Vergleichen Sie die Notation aus der Vorlesung und im Artikel und erstellen Sie eine Tabelle mit den Äquivalenzen.

(b) Korrigieren Sie den Pseudocode von MEME auf S. 56.

(c) Leiten Sie die Formel für die Log-Likelihood auf S. 54 des Artikels her.

(d) Diskutieren Sie die Algorithmen `Squash` und `Smooth` im Anhang des MEME-Artikels zur Realisierung der zwei Einschränkungen bei der Schätzen der Wahrscheinlichkeiten $\gamma_{iu}^{(t)}$ im *n-per*-Modell.

Aufgabe 2.3 (6 Punkte Punkte)

Die Datensätze

- `pgrc-16.ipk-gatersleben.de/~grosse/Lehre/SEDA/SS04/Uebungen/Serie_2/seq_1.txt`
- `pgrc-16.ipk-gatersleben.de/~grosse/Lehre/SEDA/SS04/Uebungen/Serie_2/seq_2.txt`

enthalten je 238 DNA Sequenzen der Länge 300 bp, die alle genau ein Sequenzmotif der Länge 12 bp enthalten.

Implementieren Sie den EM Algorithmus und bestimmen Sie das Sequenzmotif (also die PWM) der Länge 12 bp aus jedem Datensatz.

Literatur

[Bai95] T. L. Bailey, C. Elkan: *Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization*, *Mach. Learn.*, Bd. 21, Nr. 1-2, 1995, S. 51–80.

Abgabetermin: 21. April, in der Übung