Übungen Sequenz- und Expressionsdatenanalyse

Sommersemester 2004

Dr. Ivo Grosse

IPK, Gatersleben
Institut für Informatik,
Universität Halle

Blatt 3

Aufgabe 3.0 (6 Punkte – Verlängerung der Abgabefrist 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.

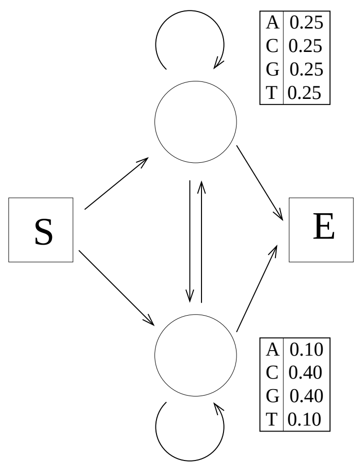
Aufgabe 3.1 (4 Punkte)

(a) erweitern Sie Ihr Progamm um den stochastischen EM Algorithmus
(b) bestimmen Sie das Sequenzmotif (also die PWM) der Länge 12 bp aus dem Datensatz der Aufgabe 2.3 (= Aufgabe 3.0) und
(c) vergleichen Sie die Ergebnisse beider EM-Varianten

Aufgabe 3.2 (3 Punkte) Betrachten Sie das Beispiel-HMM aus der Vorlesung:

Die Übergangswahrscheinlichkeiten $C_{l,j}$ seinen:

		j			
l	S	0	0.5	0.5	0
	1	0	0.6	0.3	0.1
	2	0	0.3	0.6	0.1
	E	0	0	0	1



(a) Geben Sie formal ein äquivalentes HMM **ohne** Start- und Endzustand an.
(b) Wieviele mögliche Pfade können (für beide Varianten der HMM) zur Erzeugung der Sequenz $\vec{x}$ =`AGT` führen.
(c) Berechnen Sie jeweils $P(\pi, \vec{x}|\vec{\theta})$ sowie $P(\vec{x}|\vec{\theta})$.

(d) Wieviele mögliche Pfade führen zur Erzeugung einer (beliebigen) Sequenz $\vec{x}$ der Länge $L'(\vec{x})$.

Aufgabe 3.3 (3 Punkte)

Wie müssen im Forward-Algorithmus die Startwerte $\alpha_j(0)$ für **nicht** emittierende Zustände $j \notin \mathcal{E}$ gewählt werden.

Verwenden Sie die in der Vorlesungen getroffenen Vorraussetzungen:

- jeder Zustand $j \notin \mathcal{E}$ hat höchstens einen nicht-emittierenden Vorgänger ($\notin \mathcal{E}$)
- es gibt keinen (gerichteten) Zyklus mit ausschließlich nicht-emittierenden Zuständen

Abgabetermin: 28. April (in der Übung)