Übungen Sequenz- und Expressionsdatenanalyse

Sommersemester 2004

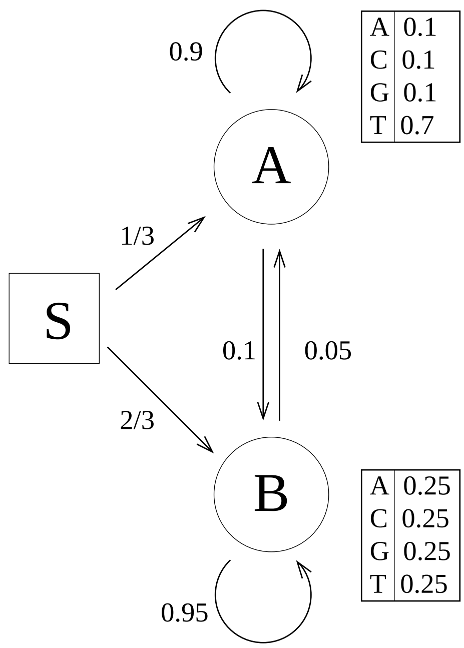
Dr. Ivo Grosse

IPK, Gatersleben
Institut für Informatik,
Universität Halle

Blatt 5

Aufgabe 5.1

Gegeben ist nebenstehendes HMM und die ersten 500 Nukleotide einer Realisierung dieses HMMs in `pgrc-16.ipk-gatersleben.de/~grosse/Lehre/SEDA/SS04/Uebungen/Serie_5/seq_1.txt`.



(a) Berechnen Sie den wahrscheinlichsten Pfad des HMMs für diese Sequenz.

(b) Berechnen Sie für jede Position $l = 1, \ldots, 501$ die Wahrscheinlichkeit $P(s_l = i|\vec{x}, \vec{\theta})$ für das Auftreten des Zustandes $s_l = i, i \in \mathcal{S} = \{A, B, S\}$.

Ordnen Sie nun jeder Position $l > 1$ den Zustand A zu, wenn $P(s_l = A|\vec{x}, \vec{\theta}) \geq 0.5$, anderenfalls den Zustand B.

Vergleichen Sie den entstandenen Pfad mit dem wahrscheinlichsten Pfad aus 1(a). Überlegen Sie sich hierzu ein geeignetes Vergleichskriterium oder Vergleichsmaß.

(c) Vergleichen Sie diese beiden Pfade mit dem durch die Realisierung erzeugten "tatsächlichen" Pfad in `pgrc-16.ipk-gatersleben.de/~grosse/Lehre/SEDA/SS04/Uebungen/Serie_5/seq_2.txt`.

Aufgabe 5.2

Gegeben seien folgende fünf Sequenzen und die wahrscheinlichsten Pfade eines HMMs mit der Standardarchitektur (Plan 9).

`CAEFTPAVH`	$m_1 m_2 m_3 m_4 m_5 m_6 m_7 m_8 m_9$
`CKETTPADH`	$m_1 m_2 m_3 m_4 m_5 m_6 m_7 m_8 m_9$
`CAETPDDH`	$m_1 m_2 m_3 d_4 m_5 m_6 m_7 m_8 m_9$
`CAEFDDH`	$m_1 m_2 m_3 m_4 d_5 d_6 m_7 m_8 m_9$
`CDAEFPDDH`	$m_1 i_1 m_2 m_3 m_4 d_5 m_6 m_7 m_8 m_9$

Geben Sie das induzierte multiple Alignment an.

Abgabetermin: 12. Mai (in der Übung)