B. Morgenstern / M. Stanke

Übungen zur Vorlesung "Algorithmen der Bioinformatik II" Wintersemester 2005/2006

Blatt 7

1. Führe den Viterbi-Algorithmus im Fall des Spielzeug-GHMMs aus der Vorlesung und der Eingabesequenz $\sigma = 51562612524$ durch. Was ist also die wahrscheinlichste Genstruktur gegeben σ?

   4 Punkte

2. Die Vorwärts-Variablen eines GHMMs und einer Emission σ der Länge t sind wie folgt definiert

   $$\alpha_{q,\ell} := \mathrm{P}(\Phi_\ell \text{ ist ein Parse der Länge } \ell \text{ endend in } q,\ S[1..\ell] = \sigma[1..\ell]) \quad (q \in Q, 1 \leq \ell \leq t)$$

   $$\alpha_{q_{\text{init}},0} = 1 \text{ und } \alpha_{q,0} = 0 \text{ für alle } q \neq q_{\text{init}}.$$

   Beweise folgende Rekursion für $q \in Q$ und $1 \leq \ell \leq t$.

   $$\alpha_{q,\ell} = \sum_{\substack{1 \leq \ell' < \ell, q' \in Q \\ \text{oder } q' = q_{\text{init}}, \ell' = 0}} \alpha_{q',\ell'} \cdot a_{q',q} \cdot e_{q'}(\sigma(\ell', \ell]).$$

   4 Punkte

3. Betrachte das in Abbildung 1.8 skizzierte GHMM für die Genvorhersage bei Eukaryoten. Gebe Pseudocode für die Berechnung der Viterbi-Rekursion

   $$\gamma_{q,\ell} = \max_{\substack{1 \leq \ell' < \ell, q' \in Q \\ \text{oder } q' = q_{\text{init}}, \ell' = 0}} \gamma_{q',\ell'} \cdot a_{q',q} \cdot e_{q',q}(\sigma(\ell'..\ell])$$

   an, wobei $q = \text{TERM}$ der terminale Exonzustand sei und $1 \leq \ell \leq t$ beliebig ($t := |\sigma|$). Die Emissionswahrscheinlichkeit sei dabei von der Form

   $$e_q(\sigma(\ell', \ell])) = r(\ell - \ell') \cdot s(\ell') \cdot s(\ell' + 1) \cdot \cdots \cdot s(\ell)$$

   für irgendwelche bereits gespeicherte Wahrscheinlichkeiten $r(k), (k = 1, 2, \ldots)$ und $s(i), (i = 1, 2, \ldots, t)$. Sei dabei möglichst effizient: Berücksichtige, dass für viele Kombinationen aus q' und ℓ' der zu maximierende Term verschwindet. (Weil der Leserahmen des Vorgängerintrons q' nicht zur Länge des terminalen Exons passt oder weil zwischen ℓ' und ℓ ein Stoppkodon im Leserahmen liegt.)

6 Punkte

4. Implementiere den Viterbi-Algorithmus für das folgende Paar-Hidden-Markow-Modell. $\Sigma = \{a, c, g, t\}$, $Q = \{1, 2, 3\}$,

A	q_{init}	1	2	3	q_{term}
q_{init}	0	0.5	0.25	0.25	0
1	0	0.8	0.04	0.04	0.02
2	0	0.9	0.09	0	0.01
3	0	0.9	0	0.09	0.01
q_{term}	0	0	0	0	1

Für $\sigma, \tau \in \Sigma^*$ ist

$$e_1(\sigma, \tau) := \begin{cases} 0.19 & \text{, falls } \sigma = \tau \text{ und } |\sigma| = 1 \\ 0.02 & \text{, falls } \sigma \neq \tau \text{ und } |\sigma| = |\tau| = 1 \\ 0 & \text{, sonst} \end{cases}$$

$$e_2(\sigma, \tau) := \begin{cases} 0.25 & \text{, falls } \tau = \varepsilon \text{ und } |\sigma| = 1 \\ 0 & \text{, sonst} \end{cases}$$

$$e_3(\sigma, \tau) := \begin{cases} 0.25 & \text{, falls } \sigma = \varepsilon \text{ und } |\tau| = 1 \\ 0 & \text{, sonst} \end{cases}$$

Wende ihn auf die Eingabesequenzen $u = gctttcagggtatcggtatcgagattgctttctgacgtatcgtatgcattg$ und $v = aggctcgcttcagggtatgggtatcgcgatttgacgtatcgtatgcatca$ an. Stelle den Viterbi-Biparse durch eine Tabelle der y_i und z_i (also ein Alignment) dar.

6 Punkte

Abgabe bis Dienstag, den 13. Dezember (Programm in Aufgabe 4 bis zum 3. Januar 2006). Lösungen werden am Dienstag, den 13. Dezember, besprochen.