
Übungen zur Vorlesung
 “Algorithmen der Bioinformatik II”
 Wintersemester 2005/2006

Blatt 6

1. Konstruiere mit dem Applet unter <http://www.gobics.de/mario/genomanalyse/HMMApplet.html> ein zeitweise unehrliches Kasino, also ein HMM mit dem Emissionsalphabet $\Sigma = \{1, 2, 3, 4, 5, 6\}$, mit zwei Zuständen $f(\text{air})$ und $u(\text{nfair})$, mit den Emissionswahrscheinlichkeiten $e_f(1) = \dots = e_f(6) = 1/6$, $e_u(1) = \dots = e_u(5) = 1/10$, $e_u(6) = 1/2$, den Übergangswahrscheinlichkeiten $a_{f,f} = 0.95$, $a_{f,u} = 0.05$, $a_{u,u} = 0.9$, $a_{u,f} = 0.1$. Das unehrliche Kasino soll immer mit dem fairen Würfel starten.

Erzeuge eine zufällige Emission der Länge 300, und ermittle den Viterbi-Parce zu dieser Emission. Bringe einen Ausdruck der Modellparameter und des Viterbi-Parses mit in die Übung.

4 Punkte

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

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

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

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

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

6 Punkte

3. Skizziere ein GHMM für eukaryotische Gene, das nur biologisch konsistente Gene zulässt, aber bei dem die Länge der Exons wie im Skript beschriebenen allgemeinen Modell nur vom aktuellen Zustand abhängt und nicht auch vom Vorgängerzustand. Tipp: Führe interne Exonzustände ein, die den Leserahmen am Ende *und am Anfang* des Exons spezifizieren.

6 Punkte

4. Konstruiere ein HMM, das Donor-Spleißstellen (DSS) findet und wende es auf folgende 5 Sequenzen an, die auch unter <http://gobics.de/mario/genomanalyse> liegen. Gehe davon aus, dass jede der Sequenzen im Exon beginnt und im darauf folgenden Intron endet, also genau eine Donor Splice Site enthält. Die Aufgabe besteht darin, die eigentlich unbekannte (aber hier angegebene) DSS genau zu lokalisieren.

>HS04636

```
cgccccgcacctgctgctgtgctgctgctcctggcgctcagccatacag|gtgagtacctggcgccgcaccggggactccggttccacgca
cccgggcagagtttccgctctgacctcctgggtctatcccagttactccgacttctctccgaatagagaag
```

>HS08198

```
ggctgtgggctccgcttgccgtttcctgcaggagcagggccaccgggcccaggccaccacactgcatgtggctcccagggcacagccat
ggctgtcagtaccttccgaaagct|gtgagtcccagagcagccctgcaccctaaccaccaacctcctctcagccccggacttcagccctg
ctctggcccctgacccccacccggctgtggcctggactag
```

>HS207581

```
agatgctgtggccacatggctaaacctgacctatctcagaagcagaatctcctagccccacag|gtatTTTTaaacttctcataattaaa
ctacagtgatgaaagatagccacactcaggccatttgggctgctcagatgaatcctgccctgctgctggcaaacatgtgcttaggacat
tgactgatctgccatgttggcttctctctg
```

>HS307871

```
atggaagcgaatgggttggg|gtgagttctccagagcacgcggtgtggctagccgggcttctaatttgagtcttccaactcaggactctat
ccctctactcccccttccccacctggagaacctccaacctgaactccgttagctggatcctgaatcctaaaacctgatttttgaga
tgttcatcccagggccttaattcaagggatgcctcaggatttccaaccaggatcttcattctgggaccatcaactctgatccct
```

>HS3112010

```
ttttgttgaccacggtcattgctctcacttgcttggcgctttgcctcccaggccctgtgcctccctctacagccctcaggagctca
ttgaggagctggtcaacatcaccagaaccagaag|gtgagtgtcggctagccagggtcctagctatgagggtccagggtgggtgattcc
caagatgaggtcatgagcag
```

Verwende folgende Positions-Gewichts-Matrix (PWM) für die 6 Positionen um die DSS beginnend mit der letzten Position im Exon.

pos	a	c	g	t
-1	0.11	0.04	0.76	0.09
1	0.0	0.0	1.0	0.0
2	0.0	0.0	0.0	1.0
3	0.5	0.03	0.45	0.02
4	0.72	0.09	0.10	0.09
5	0.06	0.06	0.83	0.05

Diese Tabelle ist beispielsweise so zu lesen: Die letzte Stelle im Exon ist mit Wkeit 0.76 ein g, die ersten beiden Stellen im Intron sind immer ein g bzw. ein t und an der dritten Stelle im Intron kommt mit Wkeit 0.5 ein a, usw.. Gehe davon aus, dass im Exon die Wkeiten der Basen a, c, g, t 0.23, 0.29, 0.28 bzw. 0.2 sind und im Intron 0.26, 0.24, 0.24 bzw. 0.26.

Tipp: Verwende das HMMAppllet unter <http://www.gobics.de/mario/genomanalyse/>. Führe einen jeweils einen Exon und Intronzustand mit einer Selbstschleife ein um die variable (unbekannte) Länge der Exon- und Intronsequenz zu ermöglichen. Wähle dabei die Übergangswkeiten der Selbstschleife für Exon und Intronzustand gleich. Führe für jede der 6 Positionen der PWM einen Zustand mit den entsprechenden Emissionswkeiten ein.

Stelle das HMMApplet auf “allgemeines Modell” und gebe im Eingabefenster die Namen der Zustände, die Startwahrscheinlichkeiten, die Übergangsmatrix, das Emissionsalphabet und die Emissionswahrscheinlichkeiten im angegebenen Format (Dezimalstelle mit Punkt) ein. Lasse den Viterbipfad berechnen und vergleiche die vorhergesagte DSS-Position mit der oben angegebenen. Das HMMApplet erlaubt nicht, einen Endzustand anzugeben. Um zu erzwingen, dass jeder Pfad im Intronzustand endet, führe etwa einen zusätzlichen Endzustand am Ende der Kette deiner Zustände ein, der nur das Zeichen 'x' emittiert, das du am Ende der Eingabesequenzen hinzufügst. Drucke für den späteren Vergleich die Ausgabe der Funktion “zeige Modellparameter” aus und skizziere deinen Zustandsgraphen.

6 Punkte

Abgabe bis Dienstag, den 6. Dezember, bei Mario Stanke. Lösungen werden am Dienstag, den 6. Dezember, besprochen.