

Quiz zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2002/03

Dipl.-Math. Sven Rahmann · Prof. Dr. Martin Vingron

Dieses Blatt dient Ihrer Selbstkontrolle und wird nicht abgegeben.

Pro Aufgabe haben Sie ca. 1 Minute Zeit.

Die Klausuraufgaben werden komplexer sein als diese Quizfragen.

1. Zeichnen Sie qualitativ (ohne genaue Zahlenangaben) das Sequenzlogo der folgenden Count-Matrix, wenn als Hintergrundverteilung die Gleichverteilung angenommen wird.

Pos	1	2	3	4
A	1	25	45	20
C	1	25	5	20
G	97	25	45	30
T	1	25	5	30

2. Welche Zielfunktion (in Worten, nicht in Formeln) wird (a) beim Nussinov-Algorithmus, (b) beim Zuker-Algorithmus optimiert?
3. Wir beobachten: Zwei homologe Sequenzen der Länge n unterscheiden sich in d Positionen. Die Jukes-Cantor-Korrekturformel sagt aus, dass (most likely) die Anzahl der stattgefundenen Mutationen pro site nicht (wie man naiverweise vermuten könnte) $x := d/n$, sondern

$$f(x) := 3 \cdot \ln(1 - 4x/3)/4$$

beträgt. Skizzieren Sie den Verlauf von $f(x)$ im Vergleich zu x und begründen Sie kurz die Notwendigkeit einer Korrekturformel.

4. Welche Verfahren zur Baumrekonstruktion gibt es?
5. Definieren Sie ein ORF (open reading frame).
6. Nennen Sie eine positiv geladene Aminosäure.
7. Betrachten Sie den folgenden Ausschnitt aus der Matrix optimaler Alignment-Scores, die beim Smith-Waterman (SW), bzw. Needleman-Wunsch (NW) Algorithmus gefüllt wird. Ein Match gibt 2 Punkte, ein Mismatch -4 und ein Gap -5 (lineare Gapkosten).

		t_j			
		\vdots			
		<table border="1"><tr><td>2</td><td>4</td></tr></table>	2	4	
2	4				
s_i	\dots	<table border="1"><tr><td>0</td><td></td></tr></table>	0		
0					

Welche Punktzahl erhält das leere Feld in den folgenden vier Fällen? (a) $s_i \neq t_j$, NW; (b) $s_i = t_j$, NW; (c) $s_i \neq t_j$, SW; (d) $s_i = t_j$, SW.

8. Gegeben ist folgendes Alignment (mit derselben Scorefunktion wie bei der vorigen Aufgabe):

```
A T G G C C T A T T
T T G - - G A A T -
```

Welchen Score erhält dieses Alignment? Geben Sie ein besseres Alignment an.

9. Sortieren Sie die folgenden Datenbanksuchprogramme (a) nach Geschwindigkeit, beginnend mit dem schnellsten, (b) nach Sensitivität, beginnend mit dem sensitivsten: BLAST, BLAT, HMMSEARCH (letzteres ist eine Datenbanksuchmethode mit HMMs)
10. Wir beobachten den Zustand eines Nukleotids, das sich im Laufe der Zeit gemäß einer Markovkette entwickelt: AAATTAAGGCATTAACCGGCGCTC. Schätzen Sie aus dieser Beobachtung die 1-Schritt Übergangswahrscheinlichkeit P_{CG} .
11. Sei P eine *symmetrische* Übergangsmatrix einer Markovkette. Welches ist die zugehörige stationäre Verteilung? Wie sieht in etwa $P^{1000000000}$ aus?
12. Definieren Sie *Contig*, *Unitig*, *unique Unitig*.
13. Was versteht man unter der “consecutive-ones-property” einer Matrix?
14. Welche Art von Verteilung (G=Geometrisch, P=Poisson, N=Normal) beschreibt jeweils folgende Zufallsgrößen am besten?
 - Anzahl der Mutationsereignisse in einem kurzen Zeitintervall
 - Smith-Waterman-Score zweier zufälliger Sequenzen
 - Verweildauer in einem HMM-Zustand, der einen Übergang zu sich selbst gestattet.
 - Anzahl der Vorkommnisse des 5-grams ACGTC in einer zufälligen Sequenz der Länge 4096
 - Anzahl der Vorkommnisse des 5-grams ACGTC in einer zufälligen Sequenz der Länge 1000000000
 - Abstand zwischen zwei Read-Startpunkten in der genomischen Sequenz beim whole genome shotgun sequencing
 - Wartezeit bis zum ersten Sechser im Lotto (in Wochen)
 - Anzahl der Read-Startpunkte in einem festen Sequenzfenster