

# Übungen zur Vorlesung Bioinformatik I

## Wintersemester 2003/2004

### Übungsblatt 4

Abgabe: bis 14.11.03, 16.30 Uhr in INF 580

1. Die sogenannte TATA-Box ist ein häufig vorkommendes Element in eukaryotischen Promoterbereichen. Bucher (1990) hat das Vorkommen der Nukleotide an einzelnen Positionen in 389 nicht verwandten Promoterbereichen gezählt:

	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
A	61	16	352	3	354	268	360	222	155	56	83	82	82	68	77
C	145	46	0	10	0	0	3	2	44	135	147	127	118	107	101
G	152	18	2	2	5	0	20	44	157	150	128	128	128	139	140
T	31	309	35	374	30	121	6	121	33	48	31	52	61	75	71
		T	A	T	A	A	A	A							

Die Position 0 wurde so gewählt, daß dort das am besten konservierte Nukleotid steht. Betrachten Sie im folgenden nur die Teilmatrix der Positionen -2 bis +4.

- (a) Transformieren Sie diese Matrix in eine Wahrscheinlichkeitsmatrix für das Vorkommen der Nukleotide (*Profil*). Um Wahrscheinlichkeiten von 0 zu vermeiden, addieren Sie jeweils 1 pseudocount.
- (b) Wie sieht das Sequenzlogo aus, d.h. wie hoch ist der "Stack" in jeder Position, und wie hoch sind jeweils die einzelnen Buchstaben (in bits)?
- (c) Berechnen Sie auf hundertstel bits genau die positionsspezifische log-odds-Scorematrix unter der Annahme der Gleichverteilung  $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$  auf {A,C,G,T} als Hintergrundverteilung. Welchen Score erzielt die Sequenz TATAAAT (unter der Teilmatrix der Positionen -2 bis +4)?

(5 Punkte)

2. Die Jukes-Cantor-Formel für den Erwartungswert der Austausch je Position,

$$K = -\frac{3}{4} \ln \left( 1 - \frac{4p}{3} \right)$$

ist für  $p > \frac{3}{4}$  nicht definiert. Geben Sie eine plausible Erklärung, warum dieses Verhalten sogar gewünscht sein kann.

Hinweis: Überlegen Sie, an wie vielen Positionen zwei zufällige, nicht verwandte Sequenzen, übereinstimmen könnten.

(2 Punkte)