

Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2002/03

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

Blatt 1 · Ausgabe am 24.10.2002

Abgabe am 31.10.2002 vor Beginn der Vorlesung

Aufgabe 1 (TATA-Box). Die sogenannte TATA-Box ist ein häufig vorkommendes Element in eukaryotischen Promoterregionen. Von Ph. Bucher¹ wird sie wie folgt beschrieben.

	-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 Zahlen in der Tabelle geben an, wie oft ein Nukleotid an einer bestimmten Position beobachtet wurde. Position 0 wurde so gewählt, dass dort das am besten konservierte Nukleotid (T) steht (374 von 389 Beobachtungen). Die folgenden Aufgaben beziehen sich nur auf die Teilmatrix der Positionen -2 bis $+4$.²

1. Die obenstehende (Teil-)Matrix mit den Beobachtungen nennen wir auch *count matrix* C . Transformieren Sie diese in ein Profil (*profile*) P , in dem jede Spalte eine Wahrscheinlichkeitsverteilung darstellt. Um Unmöglichkeiten (Wahrscheinlichkeiten von Null) zu vermeiden, soll zuvor noch jeweils ein *pseudo-count* addiert werden.
2. Wie sieht das Sequenzlogo aus, d.h., wie hoch ist der “Stack” in jeder Position, und wie hoch sind die einzelnen Buchstaben (jeweils in bits³)?
3. Berechnen Sie auf hundertstel Bits genau die positionsspezifische log-odds Scorematrix (PSSM) S 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 TATATAT (unter der Teilmatrix von Position -2 bis $+4$)?

Aufgabe 2 (IUPAC-Codes). DNA-Sequenzen können mehrdeutige Symbole enthalten. Zum Beispiel: R (puRin) bedeutet “A oder G”, Y (pYrimidin) “C oder T”, W (Weak) “A oder T”, S (Strong) “C oder G”, und N (aNy) bedeutet “A, C, G oder T”. Die Scorematrix S aus Aufgabe 1 liefert aber nur positionsspezifische Scores für die eindeutigen Nukleotidsymbole ACGT. Erweitern Sie die Scoredefinition sinnvoll auf mehrdeutige Symbole wie R oder Y und begründen Sie Ihre Definition. Welchen Score würden Sie intuitiv einem N zuweisen? Welchen Score weist Ihre Definition einem N zu? Vergleichen Sie den Score der Sequenzen TATAAAA, TATATAT (siehe Aufgabe 1), und TATAWAW.

¹Philipp Bucher. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212**, 563–578 (1990).

²Wenn Sie computerunterstützt arbeiten, dürfen Sie aber auch die gesamte Matrix behandeln.

³Ein *bit* ist die Informationseinheit, die man aus einem Logarithmus zur Basis 2 erhält. Bei \log_e (natürlicher Logarithmus) redet man von *nats* und bei \log_{10} gelegentlich von *dits*.

