

# Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2004/05

Utz J. Pape · Ben Rich · Dr. Stefan Röpcke · Prof. Dr. Martin Vingron

**Blatt 8a · Ausgabe am 6.12.2004**

**Abgabe am 13.12.2004 vor Beginn der Vorlesung**

**Aufgabe 31 (Markov-Ketten).** Auf einem vorherigen Übungszettel haben wir gesehen, dass man ein Nullmodell z.B. für die Simulation von Alignment-Scores benötigt. Um möglichst vielsagende p-values zu erhalten, sollte das Nullmodell so realistisch wie möglich sein.

Um Zufallssequenzen zu generieren, hatten wir uns damals eines sehr einfachen Nullmodells bedient: Die einzelnen Buchstaben ( $a \in \Sigma$  mit  $\Sigma = \{A, C, G, T\}$  und  $|\Sigma| = n = 4$ ) waren *iid* verteilt. Zusätzlich haben wir die Gleichverteilung angenommen, d.h. der Buchstabe  $a \in \Sigma$  tritt an der Position  $i$  mit der Wahrscheinlichkeit  $P(X_i = a) = 1/n$  auf. Wir können uns dieses Modell als Markovkette 0ter Ordnung vorstellen. Die Transitionsmatrix  $\Pi \in \mathbb{R}^{n \times n} = (\pi(b, a))_{a \in \Sigma, b \in \Sigma}$  enthält die Übergangswahrscheinlichkeiten  $P(X_i = a | X_{i-1} = b) = \pi(b, a)$ . Dabei muss  $\Pi$  eine stochastische Matrix sein, d.h. alle Einträge sind positiv und es gilt  $\sum_{a \in \Sigma} \pi(b, a) = 1 \quad \forall b \in \Sigma$ . In obigem Modell mit der Gleichverteilung besteht  $\Pi$  dann nur aus den identischen Einträgen  $1/n$ .

Dieses Modell können wir leicht erweitern, indem wir die Hintergrundwahrscheinlichkeit  $\mu(a)$  für alle Buchstaben  $a \in \Sigma$  aus den Daten schätzen. Bezeichnen wir die Anzahl von Vorkommen eines Buchstaben  $a$  in den Daten mit  $N(a)$ , so ist der Maximum-Likelihood-Schätzer gegeben durch  $\hat{\pi}(b, a) = N(a)/n \quad \forall a, b \in \Sigma$ . *Bonus: Beweisen Sie dies!* Dieses Modell wird meist mit M0 bezeichnet. Die 0 deutet dabei an, dass unser Markov Modell 0ter Ordnung ist.

Das Modell können wir nun erweitern, indem wir die Unabhängigkeit der Positionen lockern. Nehmen wir an, dass die Wahrscheinlichkeit eines Buchstabens von seinem Vorgänger abhängt. Wir erhalten ein Markov Modell 1ter Ordnung (also M1). Der Maximum-Likelihood-Schätzer ist gegeben durch

$$\hat{\pi}(b, a) = \frac{N(ba)}{N(b\circ)}$$

*Bonus: Beweisen Sie dies!* Dabei zählt  $N(ba)$  die Anzahl der Vorkommen bei denen  $a$  direkt auf  $b$  folgt.  $N(b\circ)$  ist die Anzahl der Vorkommen von  $b$  bei denen noch mindestens ein Zeichen auf  $b$  folgt (d.h. die letzte Position in der Sequenz wird nicht mitgezählt).

Die Ordnung der Markov-Kette lässt sich prinzipiell beliebig erhöhen. Allerdings braucht man auch immer mehr Daten zum Schätzen der Parameter. In dem Modell Mm hat die Transitionsmatrix die Form  $\Pi \in \mathbb{R}^{n^m \times n}$ . Ein Matrix Eintrag ist dann die Wahrscheinlichkeit eines Buchstabens unter der Bedingung, dass bestimmte Buchstaben an den  $m$  vorhergehenden Positionen aufgetreten sind, d.h.  $\pi(w, a) = P(X_i = a | X_{i-1} = w_m, \dots, X_{i-m} = w_1)$  mit  $a \in \Sigma, w \in \Sigma^m$ . Der Maximum-Likelihood-Schätzer lässt sich folgendermassen notieren:

$$\hat{\pi}(w, a) = \frac{N(wa)}{N(w\circ)}$$

Welches Modell sollte man nun nehmen? Wir wollen natürlich ein so leichtes Modell wie möglich benutzen. Trotzdem möchten wir keine signifikanten Eigenschaften ausser Acht lassen.

Es ist naheliegend zu testen, ob sich die Parameter in dem mächtigeren Modell signifikant von jenen des weniger mächtigen Modells unterscheiden. Vergleichen wir die Modelle Mm und Mm-1, so ist die Nullhypothese  $H_0$  die Frage, ob die Parameter gleich sind:  $H_0 : \pi(bw, a) = \pi(w, a)$  mit  $a, b \in \Sigma, w \in \Sigma^{m-1}$ .

Nun können wir die Pearson  $\chi^2$  Statistik benutzen (Summe der quadrierten Differenzen dividiert durch die geschätzten erwarteten Vorkommen):

$$\chi^2 = \sum_{a,b \in \Sigma, w \in \Sigma^{m-1}} \frac{[N(bwa) - N(bw\circ)\hat{\pi}(w, a)]^2}{N(bw\circ)\hat{\pi}(w, a)}$$

Unter der Nullhypothese folgt diese Statistik asymptotisch einer  $\chi^2$ -Verteilung mit Freiheitsgraden  $n^{m-1}(n-1)^2$ . Beim Vergleich von M1 und M0 ist  $w$  das leere Wort und daher müssen wir  $\hat{\pi}(w, a)$  durch den Quotienten von  $N(\circ a)$  und der Summe aller um eins verkürzten Sequenzlängen ersetzen. *Bonus: Warum?*

In dieser Aufgabe wollen wir nun Sequenzen mit den verschiedenen Modellen simulieren und mit den Simulationen Alignment-Statistiken erstellen.

1. *Bonus:* Wenn Sie M0 und M1 vergleichen, ist die Statistik  $\chi^2$  mit 9 Freiheitsgraden verteilt. Erklären Sie, wie man auf diese Anzahl von Freiheitsgraden kommt, ohne die Formel zu benutzen.
2. Laden Sie die Sequenzdateien von der Vorlesungsseite herunter und schätzen Sie auf Grundlage jeder Datei die Parameter für das Nullmodell. Gehen Sie dabei so vor, dass Sie zunächst das einfache Modell (M0) benutzen, daraufhin testen, ob M1 eine bessere Wahl ist und führen dies fort bis Sie die Nullhypothese nicht mehr auf einem Signifikanzniveau von 1% ablehnen können. Geben Sie in der Lösung für jede Datei die Modelle an, die Sie getestet haben inklusive der Werte der Statistik und (wenn möglich) der p-values. Sollten Sie keine Möglichkeit haben, die p-values direkt zu berechnen, schauen Sie in der Tabelle 1 nach, ob Ihr Statistik-Wert signifikant ist.

Freiheitsgrade	$x$	Freiheitsgrade	$x$
9	21,7	144	202
36	58,7	576	686

Tabelle 1: Werte für  $\chi^2$  ab denen die Nullhypothese mit 1% abgelehnt werden kann.

3. Nehmen Sie nun die Sequenzen der 1. Datei zur Parameterschätzung und simulieren Sie 1000 Sequenzen der Länge 500 mit M0 und anschliessend mit dem Modell ab dem Sie die Nullhypothese nicht mehr ablehnen konnten. Erstellen Sie jeweils eine Smith-Waterman Alignmentstatistik. Bewerten Sie einen Match mit 2,3 und einen Mismatch mit  $-3,1$ . Die Gapkosten sind gegeben durch  $g(k) = 13,1 \cdot k$  wobei  $k$  die Länge des Gaps ist. Nehmen Sie nun die erste Sequenz und modifizieren diese angemessen, so dass Sie im nächsten Schritt sinnvolle p-values erhalten. Berechnen Sie dann den Alignment-Score zwischen der ursprünglichen und der modifizierten Sequenz und die p-values unter den beiden Nullmodellen aus. Zum Schätzen des p-values können Sie direkt die Simulation benutzen, Sie brauchen keine Verteilung zu fitten. Interpretieren Sie das Ergebnis!