

Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2008/09

Roland Krause · Hannes Luz · Utz J. Pape · Martin Vingron

Blatt 3 · Ausgabe am 12.11.2008

Abgabe am 19.11.2008 vor Beginn der Vorlesung

Aufgabe 12 ((nicht-parametrischer) Bootstrap in der Phylogenetik). Beim Bootstrapping werden die Spalten $1, \dots, n$ eines multiplen Alignments zufällig mit Zurücklegen gezogen. Wir betrachten eine beliebige feste Spalte i . Die Zufallsvariable N zähle, wie oft Spalte i in einem Bootstrap-sample gezogen wird.

1. Welches ist die exakte Verteilung von N ? Welche Verteilung lässt sich gut zur Approximation bei großem n verwenden?
2. Wie groß ist die Wahrscheinlichkeit $p(b)$, dass Spalte i in *keinem* von b Bootstrap-Samples auftritt für $n \rightarrow \infty$?
3. Wie groß muss man b wählen, damit $p(b) \leq \frac{1}{100n}$ gilt?
4. Betrachten Sie folgenden Baumrekonstruktions-“Algorithmus” für n Sequenzen: Im ersten Schritt werden die erste und zweite Sequenz als Nachbarn angesehen und zu einem Unterbaum zusammengeclustert. Im k -ten Schritt wird der Unterbaum mit den Sequenzen $1, \dots, k$ mit der $(k+1)$ -ten Sequenz zu einem neuen Unterbaum zusammengeclustert. Alle Kantenlängen werden auf 1 gesetzt. Wie sehen qualitativ die Bootstrap-Werte für den rekonstruierten Baum bei entfernt verwandten Eingabesequenzen aus? Was folgt daraus für die Interpretation von Bootstrap-Werten?

Aufgabe 13 (Distanzmatrizen). 1. Es seien folgende Distanzmatrizen gegeben

(a)	<table><tr><td></td><td>a</td><td>b</td><td>c</td><td>d</td></tr><tr><td>a</td><td>0</td><td>9</td><td>9</td><td>9</td></tr><tr><td>b</td><td></td><td>0</td><td>5</td><td>5</td></tr><tr><td>c</td><td></td><td></td><td>0</td><td>2</td></tr><tr><td>d</td><td></td><td></td><td></td><td>0</td></tr></table>		a	b	c	d	a	0	9	9	9	b		0	5	5	c			0	2	d				0	(b)	<table><tr><td></td><td>a</td><td>b</td><td>c</td><td>d</td></tr><tr><td>a</td><td>0</td><td>5</td><td>7</td><td>8</td></tr><tr><td>b</td><td></td><td>0</td><td>8</td><td>9</td></tr><tr><td>c</td><td></td><td></td><td>0</td><td>2</td></tr><tr><td>d</td><td></td><td></td><td></td><td>0</td></tr></table>		a	b	c	d	a	0	5	7	8	b		0	8	9	c			0	2	d				0
	a	b	c	d																																																	
a	0	9	9	9																																																	
b		0	5	5																																																	
c			0	2																																																	
d				0																																																	
	a	b	c	d																																																	
a	0	5	7	8																																																	
b		0	8	9																																																	
c			0	2																																																	
d				0																																																	

Repräsentieren diese Matrizen eine additive Metrik oder eine Ultrametrik? Warum?

2. i) Wie würden Sie auf möglichst einfache Weise eine Distanzmatrix konstruieren, die eine additive Metrik repräsentiert und gleichzeitig nicht ultrametrisch ist?
ii) Geben Sie eine Distanzmatrix auf sechs Objekten an, die additiv aber nicht ultrametrisch ist.

Aufgabe 14 (Maximum Likelihood Parameter-Schätzung). Stellen Sie sich folgendes Zufallsexperiment vor. Sie werfen eine Münze n mal. Dabei beobachten Sie k mal Kopf und $(n - k)$ mal Zahl. Wir beschreiben diesen Ausgang des Experiments (d.h. die beobachteten Daten) durch $D = (n, k)$. Die Wahrscheinlichkeit, D zu beobachten, wird durch die Binomialverteilung $\Pr(X = k) = B(k|p, n)$ beschrieben, wobei der Parameter p ($0 \leq p \leq 1$) die Wahrscheinlichkeit für Kopf bei einem Münzwurf ist.

Maximum Likelihood: Wir betrachten nur einen einzigen Ausgang des Zufallsexperimentes ¹ und wollen mittels Maximum-Likelihood (ML) den Parameter p schätzen. Ein Ausgang D (die Daten) wird also fixiert, während die Likelihood $\mathcal{L}(p)$ als Funktion des Modellparameter p aufgefasst wird.

Die Likelihood $\mathcal{L}(p)$ ist die Wahrscheinlichkeit, die Daten D unter dem Modellparameter p zu beobachten, $\mathcal{L}(p|D) = \Pr(D|p)$.

1. Geben Sie die Likelihood Funktion

$$\mathcal{L}(p) = \Pr(D|p)$$

für den Ausgang $D = (n, k)$ des oben beschriebenen Münzwurf-Experiments in Abhängigkeit des Modellparameters p an.

2. Für die Maximum-Likelihood-Schätzung \hat{p} gilt

$$\hat{p} = \underset{p}{\operatorname{argmax}} \log \mathcal{L}(p|D)$$

Erstellen Sie jeweils einen Plot der Likelihood-Funktion $\log \mathcal{L}(p|D_i)$ für die drei Beobachtungen $D_1 = (100, 45)$, $D_2 = (500, 225)$ und $D_3 = (1000, 450)$. Sie könnten den Plot z.B. mit matlab erstellen. Skalieren Sie die x-Achse mit p ($0 \leq p \leq 1$) und die y-Achse mit $\ln \mathcal{L}(p)$. Welche Werte haben $\hat{p}_1, \hat{p}_2, \hat{p}_3$? Stellen Sie die drei log-Likelihood-Funktionen in einem Diagramm dar und vergleichen Sie diese ².

Haben Sie eine Idee, wie die Varianz einer ML-Schätzung berechnet wird?

Aufgabe 15 (freiwillige Zusatz-Aufgabe über Jukes-Cantor Modell, Simulationen und Distanzen). In dieser Aufgabe sollen Sie Sequenz-Evolution nach dem Jukes-Cantor Modell (Gleichverteilung der Nukleotide, alle Substitutionen sind gleich wahrscheinlich) für DNA-Sequenzen in einem 2-Blatt-Baum simulieren und danach auf den simulierten Alignments die evolutionäre Distanz zurückschätzen.

Nehmen Sie dazu an, zwei Nukleotid-Sequenzen der Länge l divergieren von einer gemeinsamen Vorgängersequenz, indem sich Substitutionen nach dem Jukes-Cantor Modell und dem Baum

$$(\text{human:25, chimpanzee:35})$$

anhäufen. Die Zahlen hinter den Taxa stehen für evolutionäre Abstände des entsprechenden Taxons zum gemeinsamen Vorfahren in PAM-Kalibrierung. Da der Markov-Prozess unter dem Jukes-Cantor Modell zeit-reversibel ist, und wegen $P(s+t) = P(s)P(t)$ (Chapman-Kolmogorov-Gleichung), können wir uns auch denken, dass wir eine Ursprungssequenz aus der Gleichverteilung generieren, und dann auf jeder der l Sequenzpositionen den Markovprozess über die Zeit $t = 60$ laufen lassen.

Es gibt mehrere Programme, mit denen Sie die Simulation durchführen können. Zwei davon:

¹das beschriebene Münzwurfexperiment hat 2^n Ausgänge.

²Der Range des Plots müsste vertikal ($-800 \leq \ln \mathcal{L} \leq 0$) sein, wenn in der Likelihood-Funktion wie in der Vorlesung keine Binomialkoeffizienten als Vorfaktoren vorkommen.

- das Program *evolver* aus dem Programmpaket PAML von Ziheng Yang. Von <http://abacus.gene.ucl.ac.uk/software/paml.html> können Sie die C-Quellen von PAML herunterladen und mit einem 'make' Kommando kompilieren. Ein Beispiel für eine Konfigurationsdatei 'MCbase.dat' zur Simulation von 1000 paarweisen Alignments der Länge 100, die von *evolver* beim Start gelesen wird, finden Sie auf der Vorlesungsseite unter http://lectures.molgen.mpg.de/Algorithmische_Bioinformatik_WS0607/MCbase.dat. In dieser Konfigurationsdatei ist der Baum mit "(human :0.25,chimpanzee :0.35)" angegeben. Die Kantenlängen sind um den Faktor 100 kleiner, da *evolver* von einer Kalibrierung des Prozesses ausgeht, bei der in der Zeit $t = 1$ im Erwartungswert eine Substitution pro Sequenzposition stattfindet.
- REFORM (Random Evolutionary Forests Model) von Sven Rahmann, siehe <http://gi.cebitec.uni-bielefeld.de/people/rahmann/reform/>. *Reform.pl* ist ein Perl-Skript, das das 'CPAN Parse::RecDescent' Modul von Damian Conway verwendet. Ein Beispiel für eine Konfigurationsdatei zur Simulation eines paarweisen Alignments der Länge 1000 nach obigem Baum, die Sie *Reform.pl* beim Start übergeben können, finden Sie auf der Vorlesungsseite unter http://lectures.molgen.mpg.de/Algorithmische_Bioinformatik_WS0809/for-jc-correction.reform.

Ihre (freiwillige) Aufgabe: Erzeugen Sie sich durch oben beschriebenes Evolutionsmodell zwei Datensätze, einen mit 1000 Alignments der Länge $l_1 = 100$ und einen mit 1000 Alignments der Länge $l_2 = 1000$. Bestimmen Sie für jedes Alignment

- i) die PAM kalibrierte Hamming-Distanz

$$D = 100d/l_i$$

(d -Anzahl mismatches),

- ii) die PAM kalibrierte evolutionäre Distanz unter dem Jukes-Cantor Modell:

$$\hat{t} = -\frac{300}{4} \ln\left(1 - \frac{4d}{3l_i}\right)$$

Dabei ist d die Anzahl der mismatches in einem paarweisen gap-losen Nukleotid-Alignment.

Vergleichen Sie für die beiden Datensätze mit $l_1 = 100$ und $l_2 = 1000$ jeweils die zwei Verteilungen der Werte von D und von \hat{t} . Stellen Sie dazu alle vier Verteilungen als Histogramme dar. Wie ist \hat{t} verteilt? Können Sie etwas über die Varianz des \hat{t} -Schätzers sagen?