Prof. Dr. Knut Reinert
Dr. Roland Krause
Matthias Winkelmann
Patrick Pett

Institut für Informatik
AG Algorithmische Bioinformatik

Algorithmische Bioinformatik WS 11/12
Übungsblatt 2

Abgabe am 31.10.2011, 12 Uhr

Aufgabe 1: Distanzmatrizen

Es seien folgende Distanzmatrizen gegeben

(a)

	a	b	c	d
a	0	9	9	9
b		0	5	5
c			0	2
d				0

(b)

	a	b	c	d
a	0	5	7	8
b		0	8	9
c			0	2
d				0

Repräsentieren diese Matrizen eine Metrik, eine additive Metrik und/ oder eine Ultrametrik?

Aufgabe 2: Markov-Ketten

In dieser Aufgabe wollen wir die Essensabfolge in der Mensa modellieren. Zur Vereinfachung unterscheiden wir nur zwischen drei Gerichte, von denen täglich genau eines angeboten wird, nämlich:

- M: Fleisch (Meat)
- F: Fisch
- V: Gemüse (Vegetable)

Seit Monaten haben wir die Essensausgabe beobachtet und jeden Tag notiert, welches Gericht es gab, also beispielsweise $MFMVMSSVM...$[1].

[1] Die vollständige Sequenz finden sie unter `http://lectures.molgen.mpg.de/Algorithmische_Bioinformatik_WS1011/dokumente/mensa.txt`

a) Modellieren Sie die Essensausgabe als Markovkette nullter Ordnung. Stellen Sie die Markovkette grafisch dar und schätzen Sie die Parameter. Mit welcher Wahrscheinlichkeit gibt es nach einander $VVVMF$?

b) Nun benutzen wir eine Markovkette erster Ordnung zur Modellierung. Was für eine Wahrscheinlichkeit hat dann $VVVMF$? Geben Sie bitte auch die geschätzten Parameter an.

c) Berechnen Sie die Wahrscheinlichkeit der ursprünglichen Sequenz in beiden Modellen. Was stellen Sie fest? Warum?

d) Wenn es an einem Tag Fleisch gab, mit welcher Wahrscheinlichkeit gibt es dann welches Essen vier Tage später?

[Anwendung und Mathe 120]

Aufgabe 2: *Maximum-Likelihood*-Parameter-Schätzung Bei einem Zufallsexperiment werfen Sie eine Münze n mal und beobachten k mal Kopf und $(n-k)$ mal Zahl. Der Ausgang des Experiments–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.

Wir wollen aus einem Ausgang des Zufallsexperimentes [2] 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, $\mathcal{L}(p|D) = \Pr(D|p)$.

a) 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.

b) 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` oder `R` 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

[2] das beschriebene Münzwurfexperiment hat 2^n Ausgänge.

drei log-Likelihood-Funktionen in einem Diagramm dar und vergleichen Sie diese [3].

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

Aufgabe 4: Gen-Baum einer Proteinfamilie Was können Sie über die Phylogenie von Organismen anhand einer Proteinfamilie aussagen? Was können Sie anhand Ihres Wissen des Speziesbaumes über eine Proteinfamilie aussagen?

Erstellen Sie zunächst einen Stammbaum der Elongationsfaktoren 1-alpha und 2. Laden Sie sich die Sequenzen aus dem Wiki der Vorlesung herunter.

a) Welche Spezies finden Sie in den Sequenzen? Verwenden Sie die Uniprot-Datenbank als Quelle.

b) Wir wollen die Bäume vergleichen, die mit Maximum Likelihood-Methode und Neighbor Joining und UPGMA erstellt werden. Sie können dazu *PhyML*[4] und *Ninja*[5] . oder andere Implementierungen zur Bestimmung der Bäume verwenden.

c) Alignieren Sie die Sequenzen mit *muscle*.

d) Inspizieren Sie das Alignment. Wie viele Spalten in dem Alignment sind informativ? Legen Sie einfache Kriterien fest, und klassifizieren Sie sie. Vorzugsweise schreiben Sie dazu ein Programm.

e) Erstellen Sie die Bäume mit *PhyML*, NJ und UPGMA.

f) Stellen Sie die Bäume mit einem Programm wie *Phylogwidget*[6] oder `Archaeopteryx`[7] dar.

g) Vergleichen Sie die Ergebnisse untereinander. Stellen Sie die Ergebnisse übersichtlich dar und bewerten Sie sie ausführlich. Gehen Sie insbesondere darauf ein, welche Aussagen Sie über die Position der Wurzel ziehen können.

[3] Der 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.

[4] `http://www.atgc-montpellier.fr/phyml/binaries.php`

[5] `http://nimbletwist.com/software/ninja/`. Funktioniert vermutlich nicht unter Windows

[6] `www.phylowidget.org/`

[7] `http://www.phylosoft.org/archaeopteryx/`