Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2016/17
Martin Vingron · Annalisa Marsico · Alena van Bömmel · Edgar Steiger · Thimo Wellner

Blatt 4 · Ausgabe am 7.11.2016
Abgabe am 14.11.2016 vor Beginn der Vorlesung

Name: Matrikelnummer: Übungsgruppe:

Aufgabe 1 (*20+10 Punkte; Theorie/Rechnen*)**.** Wir haben eine Münze zehnmal geworfen und dabei siebenmal Kopf und dreimal Zahl beobachtet. Wir möchten nun wissen, ob die Münze fair ist. Berechnen Sie dazu den Maximum-Likelihood-Schätzer für einen Münzwurf unter der Annahme, dass unsere Zufallsvariable X (in unserem Fall ist $X = 7$) binomialverteilt ist:

$$f(x; p) = P_p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

1. Bestimmen Sie die Likelihood-Funktion $L(p) = P(\text{Data}|p)$ für $p \in [0, 1]$.
2. Stellen Sie L für $p \in [0, 1]$ grafisch dar. Vergleichen Sie ihren Plot mit der Likelihood-Funktion für den Fall, dass wir 30-mal Zahl und 70-mal Kopf beobachtet hätten. Was fällt auf? Erstellen Sie die Grafik in einem Programm ihrer Wahl (empfohlen R), geben Sie die Grafik ausgedruckt oder als Datei ab.
3. Berechnen Sie auch die Wahrscheinlichkeit, dass die Daten beobachtet werden, wenn $p = 0.5$ ist. Erklären Sie anhand dieser, ob die Münze fair ist.
4. *Bonus: Finden Sie das Maximum der Likelihood-Funktion $L(p)$, indem Sie L nach p ableiten. Bestimmen Sie auch das Maximum der log-Likelihood-Funktion $l(p) = \log[L(p)]$. Was fällt auf?*

Aufgabe 2 (*30 Punkte; Praxis; Multiples Alignment*)**.**

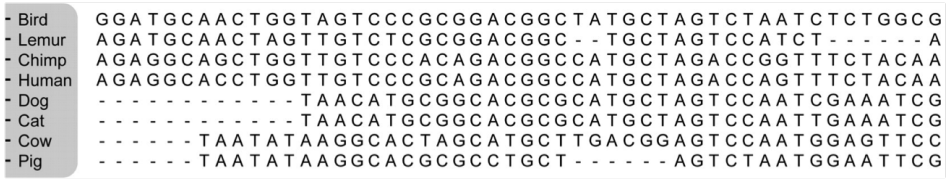
1. Auf der Vorlesungsseite finden Sie zwei Guide-Trees[1] [2] und die dazugehörigen DNA-Sequenzen[3]. Installieren Sie ClustalW[4] und rekonstruieren Sie multiple Alignments dieser Sequenzen unter Verwendung von jeweils einem der Guide-Trees als Input. Sind die Ergebnisse identisch? Geben Sie die beiden Alignments an.
2. Gegeben sind folgende 5 Sequenzen:

$$S_1 = \texttt{CTACGGAGAG}$$
$$S_2 = \texttt{CTCGTTGACA}$$
$$S_3 = \texttt{CCGTTCACG}$$
$$S_4 = \texttt{CACTGAGAT}$$
$$S_5 = \texttt{CCATTGAACG}$$

[1] Material 1: https://www.molgen.mpg.de/3700237/guidetree1.txt
[2] Material 2: https://www.molgen.mpg.de/3700245/guidetree2.txt
[3] Material 3: https://www.molgen.mpg.de/3700285/randomdna1.txt
[4] http://www.clustal.org/clustal2/

Berechnen Sie die optimalen paarweisen Alignments (S_1, S_2), (S_2, S_3), (S_3, S_4) und (S_4, S_5) mit Hilfe vom EMBOSS[5]. Suchen Sie dabei das passende Tool (im Bezug auf die Sequenz und den sinnvollsten Algorithmus) und notieren Sie dieses. Erstellen Sie *per Hand* ein vernünftiges multiples Alignment von allen 5 Sequenzen mit Hilfe der paarweisen Alignments. Was müssen Sie betrachten, wo treten Probleme auf?

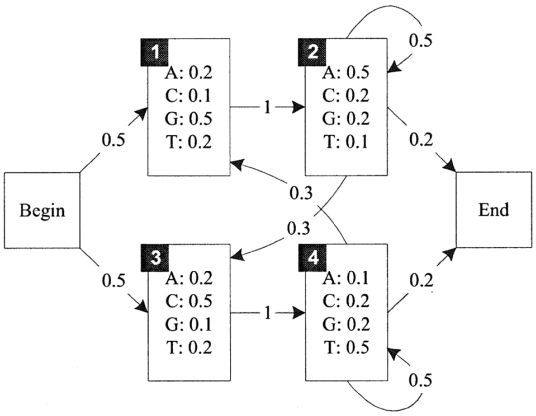
3. Betrachten Sie folgendes multiples Alignment:



Kumar and Filipski, Genome Research, 2007

Erstellen Sie wieder mit Hilfe von EMBOSS ein globales optimales Alignment der Sequenzen von Lemur und Schwein[6]. Was beobachten Sie? Was hat sich im Vergleich zu dem multiplen Alignment geändert? Geben Sie das paarweise Alignment an.

Aufgabe 3 (*35 Punkte; Theorie*)**.** Das folgende HMM generiert DNA-Sequenzen:



1. Geben Sie zu diesem HMM das Alphabet Σ der emitierten Symbole, die Zustandsmenge S der versteckten Zustände (hidden states) und die Matrix der Übergangswahrscheinlichkeiten A an.

2. Bestimmen Sie per Hand den wahrscheinlichsten Pfad für die Beobachtung *TATA* mittels Viterbi-Algorithmus.

3. Berechnen Sie per Hand die Forward- und Backward-Matrix für *TATA*. Welche Wahrscheinlichkeiten lassen sich damit berechnen?

[5] `http://www.ebi.ac.uk/Tools/psa/`

[6] Das Alignment ist auf der Vorlesungsseite als Material4: `https://www.molgen.mpg.de/3700325/multiples_alignment_lemur_pig.txt` aufrufbar.

Aufgabe 4 (*15+10 Punkte; Praxis*)**.** In der Vorlesung wurde beispielhaft das HMM des *occasionally dishonest casino* besprochen. Im Kasino werden zwei Würfel eingesetzt: ein fairer Würfel (F), bei dem alle Ergebnisse gleichverteilt sind, und ein "loaded" Würfel (L), bei dem alle Ergebnisse außer der 6, die mit der Wahrscheinlichkeit von 0.5 fällt, gleichverteilt sind. Die Übergangswahrscheinlichkeiten sind $FF = 0.95$ und $LL = 0.9$.

Sie sollen nun folgende Analysen mit den Beobachtungen, die in der Textdatei[7] auf der Vorlesungseite abgelegt wurden, durchführen. Diese Datei enthält die Augenzahlen von 300 Würfen und den dabei verwendeten Würfel. Nutzen Sie für Ihre Bearbeitung den ausführbaren Python-Code für den Viterbi-Algorithmus von Wikipedia[8].

1. Geben Sie die stationäre Verteilung für die Zustände (L und F) an und vergleichen Sie diese mit der beobachteten Häufigkeiten der Zustände. Warum sind diese ungefähr identisch?

2. Bestimmen Sie für die beobachtete Sequenz des Würfelexperiments den Viterbi-Pfad.

3. *Bonus: Drehen Sie die beobachtete Sequenz um, und bestimmen Sie dann erneut den Viterbi-Pfad. Vergleichen Sie nun den Viterbi-Pfad der ursprünglichen Sequenz mit dem umgedrehten resultierenden Viterbi-Pfad. Begründen Sie ihre Beobachtung.*

[7] Material 5: `https://www.molgen.mpg.de/3700376/casino_new.txt`
[8] `http://en.wikipedia.org/wiki/Viterbi_algorithm`