technische universität dortmund

Bioinformatik / Lehrstuhl 11
Fakultät für Informatik

Dominik Kopczynski
Sven Rahmann

Sommersemester 2011

Algorithmen auf Sequenzen
Übung – Blatt 14

Ausgabe: 07. Juli, **Besprechung:** 12.07. 9:00 Uhr; 14.07. 14:00 Uhr

Aufgabe 14.1

Berechne mit gegebenem Alphabet $\Sigma = \{A, C, G, T\}$, mit folgender Übergangsmatrix und der Startverteilung $p^0 = (0.2, 0.4, 0.1, 0.3)$ die Wahrscheinlichkeiten in einem M1-Modell für folgende Strings:

$$P = \begin{pmatrix} 0.2 & 0.2 & 0.3 & 0.3 \\ 0.4 & 0.1 & 0.15 & 0.35 \\ 0.15 & 0.25 & 0.4 & 0.2 \\ 0.7 & 0.05 & 0.15 & 0.1 \end{pmatrix}$$

- `GATTACA`
- `AGTCAGA`
- `GCCACT`
- `ATAC`

Präsentationsaufgabe 14.2 :

Schreibe ein Programm, das Ziehen von Zahlen mit beliebiger Wahrscheinlichkeitsverteilung realisiert. Gegeben sei das Alphabet mit $\Sigma = \{A, \ldots, Z\}$ und der Wahrscheinlichkeitsverteilung $p_i = \frac{f(i)}{\sum_{j=0}^{|\Sigma|} f(j)}$ mit $i = \{0, \ldots, |\Sigma|\}$ und $f(i) = \sin(0.1i) \cdot \cos(0.3i) + 1$. Implementiere das Ziehen sowohl mit linearer, als auch mit binärer Suche. Lass das Programm 100000x zufällig einen Buchstaben aus dem Alphabet ziehen. Gib anschließend die Wahrscheinlichkeiten und die Anzahl der Ziehungen für alle Buchstaben aus.

Aufgabe 14.3

Zeige, dass beim Schätzen von M0-Modellen die eindeutige Lösung $p_a^* = n_a / \sum_b n_b$ tatsächlich die Maximum-Likelihood-Schätzung für p ist um $\prod_{j=1}^{M} \mathbb{P}(s^j)$ zu maximieren.
Tipp: Extrempunkte einer Funktion können durch das Null-setzen der ersten Ableitung berechnet werden. Optimierungsprobleme mit Nebenbedingungen können mit Hilfe des Lagrange'schen Multiplikators gelöst werden.

Aufgabe 14.4

Als lustiger Schluss für die Übung:

- Suche einen langen englischen Text, z.B. "All's well that ends well"[1] von Shakespeare, im .txt-Format.
- Räume den Text auf, so dass nur Kleinbuchstaben vorkommen, alle Satzzeichen geeignet durch eins aus Punkt, Komma, Fragezeichen (oder nur Punkt) ersetzt werden und beliebige Whitespaces durch ein einzelnes Space ersetzt werden (und ggf. andere Aufräumarbeiten nach Bedarf; insgesamt soll das Alphabet möglichst klein sein und keine Sonderzeichen enthalten).
- Schätze daraus nun ein Markov-Modell möglichst hoher Ordnung (z.B. M4 oder M5, wenn sinnvoll).
- Benutze dieses Modell, um zufällige Shakespeare-ähnliche Text zu erzeugen.
- Lies die so gewonnenen Texte laut vor, klingt das wie Shakespeare?
- Besonders lustig: Bringe noch an prominenter Stelle die Worte "cheap" und "viagra" unter und verschicke eine Million dieser Zufallstexte an eine Million Leute per E-Mail, aber nicht an den Dozenten oder Übungsgruppenleiter ☺

[1] `http://www.gutenberg.org/ebooks/1791`