Prof. Dr. Knut Reinert
Enrico Siragusa
Sascha Meiers
Christoph Hartmann

Institut für Informatik
AG Algorithmische Bioinformatik

Algorithmen und Datenstrukturen in der Bioinformatik

Dreizehntes Übungsblatt WS 11/12

Abgabe Montag, 30.01.2012, 15:00 Uhr

Name: Übungsgruppe: A☐ B☐ C☐

Matrikelnummer:

Niveau I

Aufgabe 1: Viterbi und *posterior decoding*

Betrachten Sie das HMM mit den Transitionswahrscheinlichkeiten A und den Emissionswahrscheinlichkeiten e:

$A :=$

	0	P	Q
0	0	0.2	0.8
P	0.5	0.3	0.2
Q	0.7	0.3	0

$e :=$

	x	y
P	0.5	0.5
Q	0.2	0.8

a) Berechnen Sie für die Sequenz yxy den Viterbi- und den *posterior decoding*-Pfad.

b) Erklären Sie, warum sich die beiden unterscheiden! Warum kann das Ergebnis des "falschen" Pfades trotzdem von Interesse sein?

Aufgabe 2: Training von HMM's

Gegeben die Sequenz *cdcccd* und die dazugehörige Zustandsfolge $AABAAB$

a) Bestimmen sie mit der *maximum likelihood*-Methode die Parameter für einen HMM, der diese Sequenz erzeugt

b) Welche Probleme können auftreten, wenn man einen HMM mit zu wenigen Sequenzen trainiert? Wie kann man diese beheben?

Niveau II

Aufgabe 3: Das wilde Waschmaschinenrennen

Im Badezimmer von WG Y wird jede Woche ein Rennen zwischen Waschmaschine und Trockner abgehalten: Die Waschmaschine bewegt sich jede Minute *unabhängig* von der Bewegung der Vorminute entweder 1cm nach vorne (v) oder 1cm rückwärts (r), mit jeweils gleichen Wahrscheinlichkeiten.

Der Trockner bewegt sich auch entweder 1cm nach vorne (v) oder 1cm zurück (r), er behält jedoch mit 75% Wahrscheinlichkeit die Richtung bei.

Beide Geräte starten jeweils mit 50% Wahrscheinlichkeit mit einer Vorwärtsbewegung. Die Ziellinie ist in 2 cm Entfernung, beide Geräte haben beliebig viel Platz in beide Richtungen und beliebig viel Zeit.

a) Modellieren sie beide Geräte als *Markov-Ketten* (Keine HMMs!).

b) Mit welcher Wahrscheinlichkeit wird die Kette *rvvv* von beiden Geräten erzeugt?

c) Mit welcher Wahrscheinlichkeit kommt die Waschmaschine ans Ziel? Wie lange braucht sie dafür durchschnittlich?

d) (Intuition) Auf welches der beiden Geräte würden Sie ihr Geld setzen?

Programmieraufgabe (Abgabe Montag, 06.02.2012, 15:00)

P-Aufgabe 7: Implement Horspool algorithm by specializing a given *generic* naive search algorithm.

Program input arguments from the command line are:

a) a file containing the text T

b) a pattern string P

Your program must:

a) read T

b) read P

c) search P into T using the original naive algorithm

d) search P into T using your Horspool algorithm implementation

e) write to standard output the name of each algorithm, elapsed time in milliseconds and the number of occurrences (see below)

Example:

```
user@linux:~$ ./aufgabe7 english.50MB whatever
Naive, 119, 687
Horspool, 37, 687
```

Hints:

Check out the material at `https://svn.imp.fu-berlin.de/agbio/aldabi/ws11/documents/aufgabe7`. You will find a code template *aufgabe7.cpp* containing the naive implementation to specialize.

To test your program, you can download and unpack the input file `http://pizzachili.dcc.uchile.cl/texts/nlang/english.50MB.gz`.