technische universität dortmund

Bioinformatik, Fakultät für Informatik

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

Genominformatik, Medizinische Fakultät

Prof. Dr. Sven Rahmann

Wintersemester 2019/20

Algorithmische Bioinformatik
Übungsblatt 4

Ausgabe: 05. November 2019 · **Besprechung:** 12. November

Aufgabe 4.1 Textmodelle und HMMs sind äquivalent.

Beweise die eine Richtung dieser Aussage, indem Du zu einem gegebenen HMM (Q, q_0, A, E, e) ein Textmodell (C, c_0, Σ, ϕ) angibst, so dass die gleichen Beobachtungs-Sequenzen mit gleichen Wahrscheinlichkeiten generiert werden.

Beweise die zweite Richtung dieser Aussage, indem Du zu einem gegebenen Textmodell (C, c_0, Σ, ϕ) ein HMM (Q, q_0, A, E, e) angibst, so dass die gleichen Beobachtungs-Sequenzen mit gleichen Wahrscheinlichkeiten generiert werden.

Aufgabe 4.2 Der proteincodierende Bereich eines Gens in einem Prokaryoten (wie zum Beispiel des Bakteriums *E. coli*) beginnt mit einem Startcodon (oft ATG, steht auch für die Aminosäure Methionin) und endet mit einem Stoppcodon; die Codons dazwischen codieren die Aminosäuren des Proteins. Die Codons kommen nicht gleichverteilt vor, sondern mit einer bestimmten Häufigkeit, die auch von der Häufigkeit der jeweiligen Aminosäure abhängt (vgl. Aufgabe 1.2 von Blatt 1). Entwirf die Topologie eines *CDS finders* (coding sequence finders), also eines HMMs, das codierende Sequenzen in einem Genom erkennen kann. Man braucht vermutlich(!) je drei Zustände für ein Codon, irgend eine Struktur für die nichtkodierenden intergenischen Bereiche und sinnvolle Übergänge dazwischen. Kritisch betrachtet werden sollte die mit dem Modell erzeugte Längenverteilung einer codierenden Sequenz. (Dies ist keine Programmier-, sondern eine Designaufgabe.)

Szenario Wir betrachten ein (kriminelles!) Casino, das ein Würfelspiel veranstaltet. Ein Gewinn wird nur beim Wurf einer Sechs gezahlt. Der Spieler glaubt, dass ein fairer Würfel verwendet wird.

Tatsächlich gibt es aber drei verschiedene Würfel: einen fairen (F), einen unfairen (U), wobei jede Augenzahl zwischen 1 und 5 mit Wahrscheinlichkeit $1/5$ auftritt, sowie einen Würfel zum Anlocken naiver Kunden (A), bei dem die Sechs mit Wahrscheinlichkeit $2/7$ und die restlichen Augenzahlen mit $1/7$ auftreten.

Begonnen wird mit dem fairen Würfel ($q_0 = \mathrm{F}$). (Vor dem ersten Wurf findet aber bereits ein Übergang statt!) Die Übergänge sind wie folgt. Zu 80% wird der aktuelle Würfel beibehalten. Von F wird mit je 10% Wahrscheinlichkeit zu U bzw. A gewechselt. Von U oder A wird mit 20% zu F gewechselt.

Aufgabe 4.3 Erstelle 100 Simulationsreihen (in verschiedenen Dateien) zu je 1000 Würfen. Speichere dabei sowohl die Beobachtungen als auch die Zustände.

Aufgabe 4.4 Berechne die Logarithmen der Wahrscheinlichkeiten der 100 Beobachtungsreihen (Forward-Algorithmus). Stelle die 100 Ergebnisse in einem Histogramm dar.

Aufgabe 4.5 Berechne für jede der 100 Beobachtungsreihen den Viterbi-Pfad. Welchen Anteil hat der Viterbi-Pfad jeweils an der Gesamtwahrscheinlichkeit der Beobachtung?

Aufgabe 4.6 Vergleiche Viterbi-Pfad und wahren Zustandspfad. Wie oft sagt der Viterbi-Pfad den korrekten Zustand vorher?