Algorithmen auf Sequenzen

SoSe 2022

Projekt 4

Hidden Markov Modelle

Abgabetermin: 14.07.2022

Ansprechpersonen:

- Georg Grünert: georg.gruenert@mni.thm.de
- Sujaya Shrestha: sujaya.shrestha@mni.thm.de
- Cornelia Meckbach: cornelia.meckbach@mni.thm.de

Abgabe: Beschreiben Sie Ihr Vorgehen und Ihre wichtigsten Erkenntnisse in einem Dokument. Fügen Sie dem Dokument auch Ihren kommentierten Quellcode hinzu, sowie wichtige Ausgaben und speichern Sie es als PDF. Vergessen Sie nicht Ihren Namen dazuzuschreiben. Speichern Sie den Quellcode inkl. ReadMe-file (für Ausführungsdetails), sowie das oben beschriebene PDF-Dokument in einem Ordner und geben Sie diesen entweder als .zip oder als .tar.gz ab.

Wichtig: Beschreiben Sie wichtige Erkenntnisse oder Überlegungen zu der jeweiligen Aufgabenstellung in einem kurzen Text (separiert vom Quellcode).

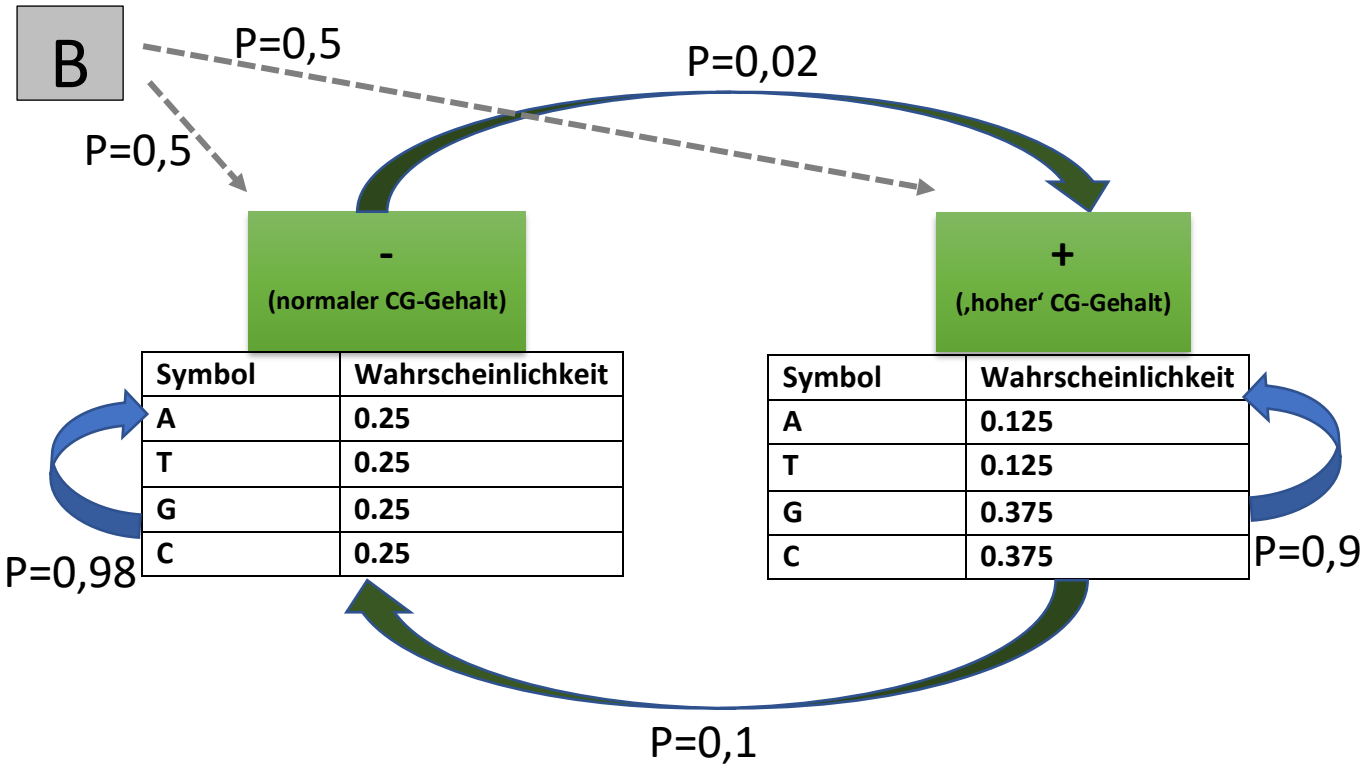
Projekt 4: Hidden Markov Modelle

Aufgabe 1: Implementation

Implementieren Sie den Viterbi-Algorithmus. Ihr Algorithmus sollte eine Sequenz von Emissionen, Übergangswahrscheinlichkeiten von Zuständen (inkl. Startzustand), sowie für jeden Zustand die Emissionswahrscheinlichkeiten übergeben bekommen und den wahrscheinlichsten (Zustands-) Pfad ausgeben.

Aufgabe 2: Sequenzen generieren

Gegeben ist Ihnen das folgende Hidden Markov Modell, welches Sequenzbereiche mit einem erhöhten GC-Gehalt sowie „normale“ Sequenzbereiche modelliert. (Es geht hier wirklich nur um den Gehalt der Nukleotide Guanin und Cytosin im Allgemeinen und keine CpG-Inseln.)



Symbol	Wahrscheinlichkeit
A	0.25
T	0.25
G	0.25
C	0.25

Symbol	Wahrscheinlichkeit
A	0.125
T	0.125
G	0.375
C	0.375

Erstellen Sie eine Funktion, die das oben gezeigte Modell übergeben bekommt und daraus dann Sequenzen (Pfad und Symbolsequenz) einer übergebenen Länge l erzeugt.

Aufgabe 3: Viterbi-Algorithmus Testen

Um eine Einschätzung darüber zu bekommen, wie gut der Viterbi-Algorithmus ist, testen Sie den Viterbi-Algorithmus an den Sequenzen, die Sie mit der Funktion aus Aufgabe 2 erstellen. Ermitteln Sie die Sensitivität und die Spezifität des Viterbi-Algorithmus indem Sie

a) 1000 Sequenzen (Symbol- & Pfadsequenzen) der Länge 1000 mit dem Modell & der Funktion aus Aufgabe 2 generieren.
b) Für jede dieser Symbol-Sequenzen den Viterbi-Algorithmus mit dem Modell aus Aufgabe 2 laufen lassen und anschließend den Viterbi-Pfad mit dem eigentlichen Pfad vergleichen. Berechnen Sie für jede Sequenz Sensitivität und Spezifität und stellen Sie die Verteilungen der Werte in einem Boxplot gegenüber.
c) Vergleichen Sie Sensitivität und Spezifität und erklären Sie das Ergebnis im Kontext des Modells.

Aufgabe 4: Modell anpassen

Ändern Sie das Modell aus Aufgabe 2 für Ihren Viterbi-Algorithmus so ab, dass die Spezifität im Mittel erhöht wird. Erläutern Sie Ihr Vorgehen. Stellen Sie Ihre neuen Modelle im Text (z.B. in Tabellenform) vor und erklären Sie, warum die Spezifität erhöht wurde.