

Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2004/05

Utz J. Pape · Ben Rich · Dr. Stefan Röpcke · Prof. Dr. Martin Vingron

Blatt 9 · Ausgabe am 13.12.2004

Abgabe am 03.01.2005 vor Beginn der Vorlesung

Aufgabe 32 (Forward-Algorithmus). Auf dem Notationsblatt ist in der Formel (4) angegeben, wie die Wahrscheinlichkeit einer Sequenz in einem HMM $P(O|\lambda)$ mit Hilfe des Forward-Algorithmus berechnet werden kann. Beweisen Sie diese.

Aufgabe 33 (Modellierung der Längenverteilung). In der Programmieraufgabe des letzten Blattes haben Sie bereits mit den *E. coli* ORFs (offene Leseraster) gearbeitet. Wie sieht nun deren Längenverteilung aus? Die Daten wie auch ein Hinweisblatt können Sie auf der Veranstaltungsseite finden.

1. Stellen Sie die Längenverteilung aller für Proteine kodierenden Gene graphisch dar.
2. Fitten Sie eine geometrische Verteilung an die Daten. Welche Methode haben Sie verwendet und wie lauten die Parameter? Stellen Sie das Ergebniss graphisch dar.
3. Modellieren Sie die Verteilung als Markovkette wie in der Vorlesung vorgestellt. k Zustände mit Selbstübergangswahrscheinlichkeit p werden hintereinandergeschaltet. Die Übergangswahrscheinlichkeit von Zustand i nach $i + 1$ beträgt jeweils $1 - p$. Die sich ergebende Verteilung wird auch als *Negativ Binomial* bezeichnet (Siehe Hinweisblatt). Die Frage lautet hier: Welches ist das optimale k für die *E. coli* ORFs?
4. Stellen Sie die Anpassung im Histogramm dar.
5. Verwenden Sie den Chi-Quadrat-Anpassungstest um die Güte der Anpassung bewerten?

Aufgabe 34 (Markovketten - CpG Islands). CpG Islands treten häufig in der upstream Region von Genen auf. Daher spielen sie bei der Genvorhersage eine wichtige Rolle. In dieser Aufgabe beschäftigen wir uns zunächst mit dem Klassifizieren von Sequenzen in die beiden Klassen CpG+ und CpG-. Anschliessend versuchen wir auf Sequenzen die Positionen von CpG Islands vorherzusagen.

1. Durch welchen Mechanismus entstehen CpG Islands?
2. Auf der Vorlesungsseite finden Sie zwei Dateien mit Sequenzen. Die CpG+ Datei enthält Sequenzen, die aus CpG Islands bestehen. CpG- hingegen 'normale' Sequenzen.
 - (a) Konstruieren Sie zwei first order Markovketten CpG+ und CpG- mit jeweils vier Zuständen (A,C,G,T) und schätzen Sie die Parameter aus den entsprechenden Sequenzen. Geben Sie die Transitionsmatrizen an und kommentieren Sie diese.
 - (b) Um nun zwischen beiden Klassen zu unterscheiden, berechnen wir den log-odd Score von CpG+ zu CpG-. Berechnen Sie die log-odd-Score Matrix.

- (c) Berechnen Sie die Scores aller Sequenzen aus beiden Dateien. Muss der Score für jede Sequenz normalisiert werden? Schlagen Sie eine geeignete Methode vor und berechnen Sie die normalisierten Scores. Berechnen Sie für die CpG+ und die CpG- Sequenzen jeweils das Histogramm der normalisierten Scores und stellen Sie diese gemeinsam (!) dar.
3. Nun wollen wir nicht mehr Sequenzen klassifizieren, sondern innerhalb von Sequenzen CpG Islands finden. Auf der Vorlesungsseite finden Sie eine Datei mit annotierten Sequenzen.
- (a) Eine Möglichkeit zur Annotation von CpG Islands besteht darin, über die Sequenzen ein Fenster laufen zu lassen und den Score jedes Fensters zu berechnen. Wählen Sie eine geeignete Fenstergrösse. Führen Sie dies nun mit der ersten annotierten Sequenzen durch und plotten Sie den Score gegenüber der Position. Markieren Sie auch die schon vorher annotierten CpG Islands. Was beobachten Sie? Welche Probleme treten bei dieser Methode auf?
- (b) Um Abhilfe zu schaffen, verwenden wir nun ein labelled HMMs (auch CHMMs wegen Class HMMs). Zeichnen Sie das CHMMs mit den acht Zuständen und notieren Sie zu jedem Zustand auch das zu emitierende Label.
- (c) Berechnen Sie die Transitionsmatrix mit Hilfe des Maximum-Likelihood (ML) Schätzers auf Basis der annotierten Sequenzen. Da jeder Zustand nur ein Symbol emittiert, brauchen wir nicht Baum-Welch zu benutzen.
- (d) Annotieren Sie nun die Sequenzen neu mit Hilfe Ihres CHMMs. Wie sehen die Ergebnisse aus?
- (e) Wir haben ML zum Schätzen der Parameter benutzt. In der Vorlesung haben Sie auch die conditional ML (CML) kennengelernt. Was wird dabei maximiert? Wie sieht die Formel aus? Warum könnte es sinnvoller sein, CML anstatt ML zu benutzen?

Aufgabe 35 (Silent States in HMMs). In HMMs kann es sinnvoll sein, silent states zu benutzen, die kein Symbol emittieren.

1. Geben Sie ein Beispiel dafür.
2. Gehen wir zunächst davon aus, dass es keine Übergänge zwischen zwei silent states gibt. Was muss in diesem Fall trotzdem bei der Berechnung des Forward-Algorithmus beachtet werden?
3. Nun erlauben wir auch Übergänge zwischen silent states allerdings mit der Einschränkung, dass die durch silent states induzierten Graphen azyklisch sind. Warum ist diese Einschränkung wichtig? Wie muss der Forward Algorithmus berechnet werden? Welchen zusätzlichen Graph-Algorithmus benötigen Sie und wie funktioniert er?