Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2016/17
Martin Vingron · Annalisa Marsico · Alena van Bömmel · Edgar Steiger · Thimo Wellner

Blatt 5 · Ausgabe am 14.11.2016
Abgabe am 21.11.2016 vor Beginn der Vorlesung

Name: Matrikelnummer: Übungsgruppe:

Aufgabe 1 (*30 Punkte; Praxis*)**.** Gehen Sie auf die Internetseite der Pfam-Datenbank[1] und finden Sie die folgenden Informationen über die *Bromodomain*:

1. Schauen Sie sich das multiple Alignment vom Kern ('Seed') der Domain an. Welche Positionen sind hoch-konserviert?

2. Laden Sie das multiple Alignment herunter. Wählen Sie ein geeignetes Format, damit Sie das Alignment mit HMMER[2] verwenden können.

3. Suchen Sie auf der HMMER-Website nach weiteren Proteinen in der UniProtKB Datenbank, welche den Kern der Bromodomain ebenfalls beinhalten. Wie viele signifikante Proteine finden Sie?

4. Entnehmen Sie dem multiplen Alignment die letzte Sequenz und lernen Sie das Profile-HMM neu. Alignieren Sie anschließend die Sequenz wieder zum Profile. Dazu müssen Sie HMMER (*hmmbuild* und *hmmalign*) installieren. Hat sich das Alignment verändert? Warum?

Aufgabe 2 (*20 Punkte; Praxis*)**.** In der Vorlesung wurde die Methode *PhyloHMM* vorgestellt. In einer Anwendung von PhyloHMM werden HMMs mit phylogenetischen Modellen kombiniert, so dass man z.B. anhand eines multiplen Alignments ein Genom in hoch-konservierte und nicht konservierte Regionen segmentieren kann. Die *hidden states* im HMM repräsentieren dabei jeweils ein phyolgenetisches Model mit unterschiedlichen Substitutionsraten.
Auf der Vorlesungsseite finden Sie zwei multiple Alignments[3][4], von denen eins gut konserviert ist und eins stark mutiert ist.

1. Berechnen Sie mit PHYLIP (*dnaml*) den Maximum-Likelihood für die Alignments. Verwenden Sie dabei den auf der Vorlesungsseite verlinkten phylogenetischen Baum[5].

2. Verzehnfachen Sie nun alle Kantenlängen in diesem Baum und berechnen Sie den Maximum-Likelihood erneut. Erklären Sie kurz, was Sie durch diese Veränderung modellieren.

3. Erklären Sie anhand der Likelihoods, welches der Alignments Sie für konservierter halten.

[1] `http://pfam.xfam.org/`
[2] `http://hmmer.org`
[3] Material 1: `https://www.molgen.mpg.de/3704635/dist_100.txt`
[4] Material 2: `https://www.molgen.mpg.de/3704644/dist_5000.txt`
[5] Material 3: `https://www.molgen.mpg.de/3704754/outtree_100.txt`

Aufgabe 3 (*40 Punkte; Programmieren*)**.** In folgender Aufgabe sollen Sie eine Alignment-Statistik für simulierte Sequenzen aufstellen, graphisch darstellen und mit einer Normalverteilung vergleichen.

1. Erzeugen Sie 20.000 zufällige DNA-Sequenzen der Länge 2000*bp*. Nehmen Sie an, dass die Nukleotide unabhängig voneinander mit den folgenden Wahrscheinlichkeiten auftreten: $P(A) = P(T) = 0.3$ und $P(C) = P(G) = 0.2$. Implementieren Sie dazu eine Funktion, die einen zufälligen Wert zwischen 0 und 1 als Input bekommt, und dann ein Nukleotid gemäß der obigen Verteilung ausgibt. Schreiben Sie die simulierten Sequenzen in eine fasta-Datei mit Namen *library.fasta*.

2. Erzeugen Sie außerdem eine zweite Fasta-Datei mit Namen *query.fasta*, die eine Sequenz der Länge 200*bp* mit derselben Nukleotidverteilung enthält. Wie oft kommt diese Sequenz in der Library vor?

3. Berechnen Sie die optimalen paarweisen Alignments (Query-Sequenz vs. jede der Library-Sequenzen) mit Hilfe des Smith-Waterman-Algorithmus (*ssearch36*) aus dem FASTA-Paket[6] und einer sehr hohen Gap-Penalty (z.B. *-g -10*) und lassen Sie sich die maximalen Scores pro Library-Sequenz mit Hilfe der Option *-R scores.txt* ausgeben.

4. Stellen Sie die empirische Verteilung der Scores (*scores.txt*) als Histogram graphisch dar (z.B. mit R).

5. Berechnen Sie den Mittelwert und die Standardabweichung der maximalen Scores. Stellen Sie graphisch eine Normalverteilung mit den berechneten Parametern im selben Plot wie das Histogram dar. Achten Sie auf gleiche Maßeinheiten auf der vertikalen Achse. Was fällt auf? Geben Sie auch eine Erklärung zu Ihren Beobachtungen.

Aufgabe 4 (*10 Punkte; Theorie*)**.** Eine Person wurde auf HIV getestet. Mit einer Wahrscheinlichkeit von 99.5% liefert der Test korrekterweise bei einer infizierten Person ein positives Ergebnis (Sensitivität). Die Wahrscheinlichkeit, mit der der Test ebenfalls korrekterweise kein positives Ergebnis bei einer nicht infizierten Person liefert (Spezifität), beträgt 99.3%. Berechnen Sie, wie groß die Wahrscheinlichkeit ist, dass eine Person tatsächlich infiziert ist, wenn der Test ein positives Ergebnis ausgibt und diese Person entweder einer niedrigen oder einen hohen Risikogruppe (mit einer HIV-Prävalenz von entweder 0.01% oder 0.18%) angehört. Nutzen Sie dazu das Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1}$$

[6] `http://fasta.bioch.virginia.edu/fasta_www2/fasta_down.shtml`