Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2014/15
Martin Vingron · Juliane Perner · Annkatrin Bressin

Blatt 6 · Ausgabe am 17.11.2014
Abgabe am 24.11.2014 vor Beginn der Vorlesung

Name: Matrikelnummer: Übungsgruppe:

Aufgabe 1 (*30 Punkte; Theorie*)**.** GATA2 ist Teil der Familie der GATA-Transcriptionsfaktoren, die unter anderem eine wichtige Funktion während der Differenzierung von hämatopoetischen Zellen hat. Sein DNA-Bindemotiv wird mit folgender Count Matrix C beschrieben:

	1	2	3	4	5	6	7	8	9	10	11
A	1715	544	3155	0	4380	0	4329	4188	442	2526	2377
C	224	1967	0	0	0	0	0	0	914	765	427
G	1185	1765	0	4380	0	0	0	192	3015	1057	525
T	1256	104	1225	0	0	4380	51	0	9	32	1051

1. Schreiben Sie das Konsensusmotiv, also die Sequenz mit den häufigsten Nukleotiden, für diese Matrix auf.

2. Transformieren Sie die Count Matrix C in eine Häufigkeitsmatrix P, in dem jede Spalte eine Wahrscheinlichkeitsverteilung darstellt. Um Wahrscheinlichkeiten von Null zu vermeiden, muss zuvor ein *pseudo-count* addiert werden.

3. Berechnen Sie die positionsspezifische log-odds Scorematrix (PSSM) S unter der Annahme, dass die Hintergrundverteilung auf (A,C,G,T) folgender Verteilung entspricht:

$$\pi = (0.3, 0.2, 0.2, 0.3)$$

4. Welchen Score erzielt die Sequenz CTAGATAATGA unter dem Motiv?

Aufgabe 2 (*20 Punkte; Theorie*)**.** Motive werden häufig als Motiv-logos dargestellt. Dazu wird die Entropie jeder Position berechnet.

1. Wir betrachten zuerst eine Verteilung auf zwei Symbolen p und q. Zeigen Sie wo die Entropie $H = -\sum_{x\in\{p,q\}} x log_2 x$ ihr Maximum annimmt?

2. Wir wollen nun die Entropie des GATA-motivs berechnen. Berechnen Sie die Entropie H_k jeder Position k als $-\sum_{a\in A,C,G,T} p_{ka} log_2 p_{ka}$.

3. Erstellen Sie eine vereinfachte Version eines Motiv-logos indem Sie ein Balkendiagramm mit jeweils einem Balken pro Position erstellen. Dabei soll jeder Balken so unterteilt sein, dass die Höhe eines Segments den Beitrags eines Buchstaben zur Entropie darstellt (am einfachsten z.B. in R mit *barplot(matrix,beside=FALSE)*).

Aufgabe 3 (*50 Punkte + 10 Bonuspunkte; Programmieren*)**.** Suchen Sie die Gene im Genom von *S.cervisiae* mittels einer Markovkette 2. Ordnung.

1. Laden Sie die Fasta-Dateien mit 1000 proteinkodierenden Gensequenzen[1] und mit nichtkodierenden DNA-Sequenzen[2] herunter. Schätzen Sie aus den Sequenzen in *y_genes.txt* eine Markov Kette 2. Ordnung für die Gene (das Gen-Modell G).

   Die Transitionswahrscheinlichkeiten $a^G_{rs,t}$ können Sie wie folgt berechnen:

   $$a^G_{rs,t} = \frac{c^G_{rs,t}}{\sum_l c^G_{rs,l}}$$

   Wobei $c^G_{rs,t}$ die Zahl der rst-Trinukleotide in der Sequenzen aus *y_genes.txt* ist. Überlegen Sie, ob Sie ein Modell für alle 3 Reading Frames schätzen oder 3 Modelle für jeden Reading Frame separat. Geben Sie die Transitionsmatrix(-matrizen) für das Gen-Modell an.

2. Schätzen Sie aus den Sequenzen in *y_ncregions.txt* eine Markov Kette für ein Hintergrundmodell (das Noncoding-Modell NC). Die Berechnung von $a^{NC}_{rs,t}$ erfolgt analog zu den $a^G_{rs,t}$ mit den Sequenzen aus *y_genes.txt*. Geben Sie die Transitionsmatrix fur das NC-Modell an.

3. Untersuchen Sie jetzt die Sequenz in der Datei *test.txt*[3]. Schieben Sie ein Fenster der Größe $w = 100bp$ über diese Sequenz und berechnen Sie für jede Position des Fensters den Log-Likelihood-Ratio:

   $$S(x_k, ..., x_{k+w-1}) = \log \frac{\Pr(x_k, ..., x_{k+w-2} \text{ |model G})}{\Pr(x_k, ..., x_{k+w-2} \text{ |model NC})} = \sum_{i=k}^{k+w-2} \log \frac{a^G_{x_{i-2}x_{i-1},x_i}}{a^{NC}_{x_{i-2}x_{i-1},x_i}}$$

   Wiederholen Sie die Berechnung des Log-Likelihood-Ratios für jeden Reading Frame. Überlegen Sie sich, wie Sie das Fenster über die Sequenz schieben.

4. Erstellen Sie einen Plot, in dem Sie $S(x_k, ..., x_{k+w-1})$ jeweils für die drei Reading Frames darstellen. Wie viele Gene finden Sie? Geben Sie die Start- und Endpositionen der vorhergesagten Gene in einer Textdatei *Prediction.dat* an.

5. (Optional) Wiederholen Sie (iii)-(iv) für unterschiedliche Fenstergrößen. Was beobachten Sie?

[1] Material 1: `http://www.molgen.mpg.de/Algorithmische-Bioinformatik-WS1415/u6/y_genes`
[2] Material 2: `http://www.molgen.mpg.de/Algorithmische-Bioinformatik-WS1415/u6/y_ncregions`
[3] Material 3: `http://www.molgen.mpg.de/Algorithmische-Bioinformatik-WS1415/u6/test`