Übungen zur Vorlesung Algorithmische Bioinformatik

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

Blatt 6 · Ausgabe am 21.11.2016
Abgabe am 28.11.2016 vor Beginn der Vorlesung

Name: Matrikelnummer: Übungsgruppe:

Aufgabe 1 (*30 Punkte; Theorie*)**.** GATA2 ist Teil der Familie der GATA-Transkriptionsfaktoren, die unter anderem eine wichtige Funktion während der Differenzierung von hämatopoetischen Zellen haben. 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 der 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. Bestimmen Sie $p^\star$ und $q^\star$, für die 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 +20 + 10 Bonuspunkte; Programmieren*)**.** Suchen Sie 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. Schreiben Sie eine Funktion, die aus den Sequenzen in *y_ genes.txt* eine Markovkette 2. Ordnung für die Gene (das Gen-Modell G) schätzt.

   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 den Sequenzen aus *y_ genes.txt* ist. Geben Sie die Transitionsmatrix für das G-Modell an.

2. Schätzen Sie aus den Sequenzen in *y_ ncregions.txt* eine Markovkette 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 für das NC-Modell an.

3. Untersuchen Sie jetzt die Sequenz in der Datei *test.txt*[3]. Schreiben Sie eine Funktion, die über diese Sequenz one-by-one ein Fenster schiebt und für jede Position des Fensters die Log-Likelihood-Ratio berechnet:

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

   Die Funktion übernimmt dabei die Fenstergröße $w = 100bp$ als Parameter. Speichern Sie den Ergebnisvektor S in einer Datei.

4. Erstellen Sie einen Plot, in dem Sie den Ergebnisvektor S der Log-Likelihood-Ratios darstellen. Fügen Sie eine geglättete Version des Ergebnisvektors hinzu (z.B. in `R` mit `smooth.spline()`) und außerdem eine Nulllinie. Wie viele Gene erkennen Sie? Geben Sie den Plot aus.

5. In dieser Aufgabe wurden die Reading Frames nicht berücksichtigt. Erklären Sie, worum es sich dabei handelt und beschreiben Sie, wie der Code angepasst werden müsste, um dieses Problem zu berücksichtigen.

6. *Bonus A: Geben Sie den veränderten Code inklusive Ergebnisse (Transitionsmatrizen, Plot, Genvorhersage) an, wo die Reading Frames berücksichtigt werden. Vergleichen Sie Ihre Ergebnisse mit dem ursprünglichen Code.*

7. *Bonus B: Wiederholen Sie 3 und 4 für unterschiedliche Fenstergrößen. Was beobachten Sie?*

[1] Material 1: `https://www.molgen.mpg.de/3707399/y_genes.txt`
[2] Material 2: `https://www.molgen.mpg.de/3707390/y_ncregions.txt`
[3] Material 3: `https://www.molgen.mpg.de/3707381/test.txt`