Übungen zur Vorlesung Algorithmische Bioinformatik

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

Blatt 7 · Ausgabe am 28.11.2016
Abgabe am 5.12.2016 vor Beginn der Vorlesung

Name: Matrikelnummer: Übungsgruppe:

Aufgabe 1 (*10 Punkte; Theorie*)**.** Welche Verteilung würden Sie zur Modellierung folgender Messungen verwenden? Begründen Sie Ihre Aussage.

1. Die Durchmesser von Hefezellen in einer Population.
2. Die Anzahl an Kopf-Würfen bei 100 Münzwürfen.
3. Blitzeinschläge pro km^2.
4. Die jeweils besten Sprintzeiten der Sportlerinnen bei einem Wettkampf in fünf Durchgängen.
5. Die Zeit zwischen dem Kauf eines Mobiltelefons und seinem ersten Sturz auf den Boden.

Aufgabe 2 (*10 Punkte; Theorie*)**.** Berechnen Sie die folgenden Aufgaben:

1. Wie groß ist die Wahrscheinlichkeit, dass in einer Klasse mit 60 Studierenden mindestens eine Person am heutigen Tag Geburtstag hat? Nehmen Sie eine Gleichverteilung der Geburtstage übers Jahr an.
2. Die Zufallsvariable X sei normalverteilt mit dem Mittelwert $\mu = 10$ und einer Standardabweichung σ. Wie groß darf σ höchstens sein, damit die Wahrscheinlichkeit, dass X einen Wert $x \geq 15$ annimmt, nicht größer als 5% ist?

Aufgabe 3 (*40 Punkte; Theorie*)**.** In der Vorlesung wurde die Burrows-Wheeler-Transformation vorgestellt.

1. Beschreiben Sie kurz in Ihren eigenen Worten das Prinzip der Burrows-Wheeler-Transformation. Was ist die Anwendung der Transformation?
2. Führen Sie die Burrows-Wheeler-Transformation auf folgendem String aus:

   ACCCGTGAA$

3. Bestimmen Sie das dazugehörige Suffixarray und schreiben Sie die Offsets der Suffixes neben der Transformationsmatrix. Wofür kann man das Suffixarray verwenden?
4. Berechnen Sie den Originalstring der folgenden Transformation:

   Letzte Spalte: T$TGAACCTAG

   Die erste Spalte erhalten Sie durch lexikographische Sortierung. Wie wird in der Praxis dieser Schritt der Sortierung umgangen und warum?

Aufgabe 4 (*40 Punkte; Praxis*)**.** Wir möchten die auf der Vorlesungsseite verlinkte FastQ-Datei[1] analysieren. Sie enhält die bei der Sequenzierung eines menschlichen Chromosom 21 detektierten Reads.
Benutzen Sie für Ihre Analyse den GALAXY-Server[2], der verschiedene Programme zur einfachen Verarbeitung von Sequenzierdaten bereitstellt (*NGS toolbox*).
Beachten Sie folgende Hinweise: Je nach Auslastung des Servers kann die Analyse ggf. länger dauern. Außerdem müssen Sie sich, um die Aufgabe vollständig bearbeiten zu können, auf der Webseite registrieren. Im Fall eines Ausfalls des Hauptservers können Sie einen der zur Verfügung stehenden öffentlichen Server nutzen. (In diesem Fall geben Sie bitte auch den Namen des benutzten Servers an und achten Sie darauf, dass alle benötigten Tools zur Verfügung stehen.)

1. Laden Sie die FastQ-Datei (Typ *fastqillumina* und Genom *hg19*) auf den Server und schauen Sie sich zuerst die Qualität der sequenzierten Reads an. Nutzen Sie dazu „FASTQC: Read Quality Reports“ (unter *QC and manipulation*). Fassen Sie kurz die Statistiken zu den Reads in ihrer Library (z.B. Länge, Qualität, Sequenzkomposition, etc.) zusammen. Was fällt Ihnen an den positionsspezifischen Quality-Scores auf?

2. Mappen Sie die Reads mit *Bowtie for Illumina* auf das humane Referenzgenom (*hg19*). Nutzen Sie dazu den 'build-in'-Genomindex, erlauben Sie maximal zwei Alignmentfehler und unterdrücken Sie Reads, falls diese nicht eindeutig gemapped werden können. Geben Sie die Optionen, die Sie verändert haben, an.

3. Analysieren Sie die Mapping-Statistiken mit *SAM tools: flagstat*. Werden Reads während des Mappings gefiltert? Wenn ja, wodurch?

4. Berechnen Sie mit Hilfe von *Bedtools: Create a Bedgraph* die Genom-weite Coverage. Laden Sie diese als Custom-Track in den UCSC Genome Browser. Was fällt Ihnen anhand der Coverage auf, an welchen Elementen im Genom ist die Coverage besonders hoch?

5. Wie Sie im Genome Browser sehen können, gibt es einen bekannten SNP an der Position *chr21:30255074*. Ist diese Region durch unsere Library abgedeckt?

6. Schauen Sie, ob der SNP auch in unserer Library detektiert wurde. Generieren Sie dazu eine Zusammenfassung mit *SAM tools: pileup*, bei der für jede Position das Nukleotid im Referenzgenom und das Konsensus-Nukleotid in den Reads angegeben wird.

[1] Material 1: `https://ws.molgen.mpg.de/ws/393235/test_21.fastq`
[2] https://usegalaxy.org/