Übungen zur Vorlesung Algorithmische Bioinformatik

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

Blatt 8 · Ausgabe am 5.12.2016
Abgabe am 12.12.2016 vor Beginn der Vorlesung

Name: Matrikelnummer: Übungsgruppe:

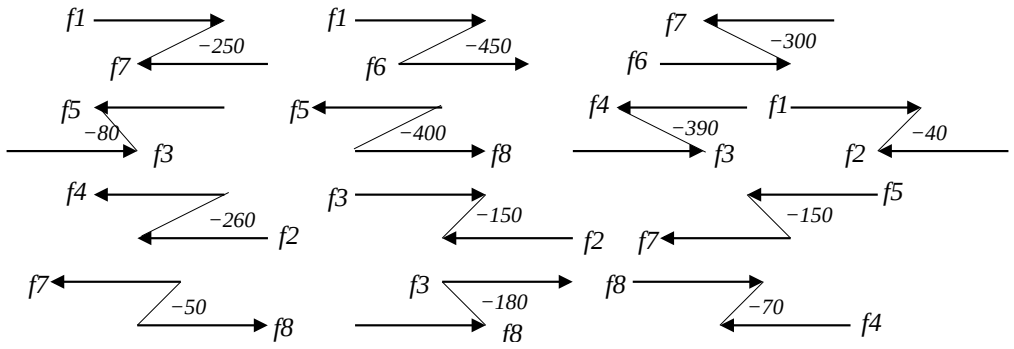
Aufgabe 1 (*10 + 10 Punkte; Theorie*)**.** Auf dem letzten Aufgabenblatt war folgende Burrows-Wheeler-Transformation gegeben:

1.Spalte:	$AAACCGGTTT
Letzte Spalte:	T$TGAACCTAG

1. Wir möchten nun den Teilstring ACG in der Burrows-Wheeler-Transformation suchen. Führen Sie die Suche mit einem möglichst effizienten Algorithmus auf der Transformation aus. Wie oft kommt der Teilstring vor?

2. *Bonus: Welche Informationen würden Sie zusätzlich benötigen, um auch die Position des Teilstrings mit diesem Verfahren effizient bestimmen zu können? Warum?*

Aufgabe 2 (*35 Punkte; Theorie*)**.** Gegeben seien die Reads $F = \{f_1, f_2, ..., f_8\}$ der Länge 500. Sie überlappen wie folgt:



1. Zeichnen Sie den Overlap-Graphen mit den entsprechenden Kantenbeschriftungen.

2. Bestimmen Sie in diesem Overlap-Graphen den minimalen Spannbaum (MST), der alle Read-Kanten enhält.

3. Zeichnen Sie das durch den MST bestimmte Layout der Reads. Geben Sie die globalen Koordinaten der einzelnen Reads an.

4. Sind alle gegebenen Overlaps mit dem Layout konsistent? Falls nicht, wo und warum gibt es Inkonsistenzen?

Aufgabe 3 (*30 + 10 Punkte; Theorie*)**.** In der Vorlesung haben Sie die Lander-Waterman-Formel wie folgt kennengelernt:

$$E = Ne^{-(1-\Theta)L\lambda}$$

Sie beschreibt die erwartete Anzahl an Contigs in Abhängigkeit der Anzahl an Reads N, der Readlänge L, der Genomgröße G ($\lambda = N/G$) und des Parameters Θ, der den Anteil des minimalen geforderten Überlapps zwischen zwei Reads angibt.
Sie wollen nun abschätzen, wieviele Reads einer bestimmten Länge Sie benötigen, um das Hefegenom (ca. $2 \cdot 10^7 bp$) bzw. das Humangenom (ca. $3.27 \cdot 10^9 bp$) vollständig zu sequenzieren.

1. Berechnen Sie zuerst den Anteil der Basen, die durch die Sequenzierung getroffen wurden. Dies entspricht der Wahrscheinlichkeit dafür, dass eine einzelne Base im jeweiligen Genom von mindestens einem Read abgedeckt wird. Betrachten Sie dazu die folgenden Kombinationen an Parametern:

   - $N = \{10.000.000, 50.000.000\}$
   - $L = \{50, 100\}$

2. Erstellen Sie nun jeweils einen Plot für jedes der beiden Genome, in dem der Erwartungswert E in Abhängigkeit der Anzahl an Reads N dargestellt wird. Nehmen Sie an, dass die Readlänge $L = 50$ ist und $\Theta = 0.2$. Wählen Sie der Genomgröße entsprechend realistische Library-Größen.

3. Wie würden Sie N wählen, um ein möglichst komplettes Bild der Genomsequenz zu erhalten? Begründen Sie Ihre Antwort.

4. *Bonus: Verändern Sie nun einmal* $\Theta \in \{0.1, 0.2, ..., 0.5\}$ *und einmal die Readlänge* $L \in \{50, 100, 150\}$. *Was beobachten Sie?*

Aufgabe 4 (*25 Punkte; Theorie*)**.** Gegeben sei die folgende RNA-Sequenz mit ihrer Sekundärstruktur im Vienna-Format.

```
CAUAGGGUAGUGGCUAAGAACCUAAACUCUAAAUUUAGAUGUCCUGAGUUCUAAUCCCAGCUGUAUGC
((((.(((..(((...(((((.((.((((((....)))).)).)).)))))....)))))).)))).
```

1. Zeichnen Sie die reguläre graphische Darstellung der Sekundärstruktur.

2. Zeichnen Sie die zirkuläre Darstellung der Sekundärstruktur.

3. Benennen Sie die unterschiedlichen Sekundärstrukturelemente, die Sie in der RNA-Struktur finden.