Übungen zur Vorlesung Sequenzanalyse I

Universität Bielefeld, WiSe 2009/2010
Prof. Dr. Jens Stoye · Dipl.-Inform. Nils Hoffmann

`http://gi.cebitec.uni-bielefeld.de/teaching/2009winter/sequenzanalyse`

Blatt 1 vom 15.10.2009
Abgabe in einer Woche vor der Vorlesung

Organisatorisches

Wichtige Hinweise: Bitte unbedingt in eine der Übungslisten am TechFak-Brett eintragen, falls Du dies nicht bereits in der ersten Vorlesung am 15.10.2009 getan hast und im eKVV für die Vorlesung und die Übungen registrieren. Auf der Homepage der Veranstaltung steht eine elektronische Version zum Download bereit. Die gedruckte Version des Skripts wird in der ersten Übung verteilt.

Abgabe der Übungszettel: Auch wenn die Bearbeitung der Aufgaben in einer Gruppe (maximal zwei Personen) erfolgt ist, so müssen die Lösungen zu den Aufgaben separat und unter eigenem Namen vor Beginn der Vorlesung abgegeben werden.

Teilnahme an der Klausur: Die Klausur kann am Ende mitschreiben, wer bis dahin A&D bestanden, mindestens 50% der Übungspunkte erreicht und mindestens zweimal in den Übungen eine Aufgabe vorgerechnet hat. Die Klausur findet am 11.02.2010 statt.

Aufgaben

Aufgabe 1 (Textsuche) Beschreibe, welche Algorithmen zur Textsuche Du bereits aus der Vorlesung *Algorithmen und Datenstrukturen* kennst. Welche Problemstellung lösen diese Algorithmen? Welche Problemstellungen gibt es darüber hinaus in der Textsuche?

Aufgabe 2 (FASTAReader) Das FASTA Format ist in der Sequenzanalyse sehr weit verbreitet, um Sequenzdaten, z.B. RNA-Sequenzen, zu speichern und standardisiert auszutauschen. Es besteht aus einer ersten Zeile, dem sogenannten *header*, der mit dem Zeichen `>` eingeleitet wird und der Sequenz in der üblichen Buchstabennotation für die verschiedenen Basen. Je nach Standard folgen in der ersten Zeile nach dem Zeichen `>` beliebig komplexe Metadaten, die die folgende Sequenz näher beschreiben. Wir betrachten die gesamte erste Zeile (ohne `>`) zunächst nur als ID der folgenden Sequenz.
Im folgenden ist ein Beispiel einer FASTA-Datei angegeben.

```
>Name meiner Sequenz
ACGGTAGATATCGACTCGTCGATCGA
ACGATCGCGAGTGCGAGCGGATGATG
```

Eine echte FASTA-Datei mit mehreren Sequenzen kannst Du auf der Seite zur Veranstaltung herunterladen (s.o.). Es ist üblich, dass z.B. bei Suchanfragen die Ergebnisse in einer Datei zurückgegeben werden. Es gibt also mehrere Abschnitte, die jeweils einer Sequenz entsprechen und mit dem Zeichen `>` beginnen. Es können jedoch auch Leerzeichen (Whitespace) zwischen dem letzten Sequenzzeichen und dem nächsten Abschnitt stehen. Diese müssen entfernt werden.
Entwickle eine JAVA-Klasse `FASTAReader`, die das FASTA-Format lesen kann und die Sequenzdaten in einem Objekt vom Typ `java.util.HashMap<String,String>` speichert. Beachte dabei die Prinzipien der Objektorientierung, d.h. deine Klasse sollte wiederverwendbar sein und einfache Erweiterungen der Funktionalität durch Vererbung ermöglichen. Das Lesen der Dateien soll in einer Methode

```
public HashMap<String,String> read(File f)
```

ausgeführt werden. Um das `>` Zeichen zu entfernen, solltest Du einen Blick auf die API zur `java.lang.String`-Klasse werfen[1]. Die dort erwähnten Methoden helfen auch beim Entfernen von Leerzeichen. Weitere Ideen findest Du in der API im Abschnitt `java.io` (`File` und `StringReader`).

[1] http://java.sun.com/javase/6/docs/api/java/lang/String.html