Institut für Informatik
Praktische Informatik und Bioinformatik
Prof. Dr. Ralf Zimmer

Wintersemester 2019/20

Übungen zum Propädeutikum Programmierung in der Bioinformatik

Blatt 4

Termin: Dienstag, 12. November 2019

Hinweis: Denkt daran, dass einige der in den Aufgaben verlangten Dinge bereits als vorgefertigte Methoden existieren. Im Zweifel also erst mal in der Java API nachsehen, ob die gewollte Funktion bereits existiert.

Übung 1 *ArrayList*

1. Erstelle eine ArrayList mit Strings und füge ihr fünf Elemente hinzu. Gib danach den Inhalt der ArrayList mit Hilfe eines for-each Loops auf der Konsole aus.
2. Entferne das dritte Elemente aus der ArrayList.
3. Vertausche das zweite mit dem ersten Element der ArrayList.
4. Sortiere die ArrayList.
5. Kehre die Reihenfolge der Elemente in der ArrayList um.

Übung 2 *HashSet*

1. Erstelle ein HashSet mit Integers und füge ihm fünf Elemente hinzu. Gib danach den Inhalt des HashSets mit Hilfe eines for-each Loops auf der Konsole aus.
2. Lass dir die Mächtigkeit des Sets auf der Konsole ausgeben.
3. Wandele das HashSet in einen Array um.
4. Entferne alle Elemente aus dem HashSet.

Übung 3 *HashMap*

1. Erstelle eine HashMap mit Datentyp String für die Schlüssel und Datentyp Integer für die Werte.
2. Füge der HashMap fünf *mappings* hinzu. Die Schlüssel sollen dabei die fünf bevölkerungsreichsten Städte Deutschlands sein und die Werte die dazugehörigen Einwohnerzahlen.
3. Gib die Einwohnerzahl von Berlin auf der Konsole aus.

Übung 4 *Einfaches Einlesen einer Datei & relative Häufigkeit*

In dieser Aufgabe soll ein FASTA-File eingelesen werden. Das FASTA-Format[1] wird genutzt um Nukleotid- oder Aminosäuresequenzen zu speichern. Dabei steht immer ein *Header*, also eine Kopfzeile, vor jeder Sequenz der mit einem `>` beginnt und den Namen und andere Informationen zur nachfolgenden Sequenz enthält. Nach einem Header folgen eine oder mehrere Zeilen der *Sequenz*, die durch den Ein-Buchstaben-Code für Nukleotide oder Aminosäuren dargestellt wird. Hier beispielsweise eine Sequenz des PLS-Gens aus *Arabidopsis thaliana*:

```
>NC_003075.7:c18329699-18329094 Arabidopsis thaliana chromosome 4 sequence
GTATCGCATTTGTTTCAAGTTTTTTTTTCTATAATGTTTCTTCGAAATCCATGATCATATAGTATATAAG
AAGCATGTATTTATAATGTTCCACTTAATATATTAGTATTGGAGACTAAAGCGAACATATAAAACCCAAA
TAAACCTTTCTTTAAGTTTTATTAAAAGTCTAAACACTTGATTTGTGTTTTAGTTTGGGTAGTAGTGAGA
AAAGAAAAATAAATAATCAAAAAGATTAAAGAAGAAAGAATTTGAAAGCAAGGAACACGAAATCCGAAGA
GCGAGGGGAGCGAAGACAGTCCACGTAGCTGCAGAGAGAAAGAGAAGAGCACGTGAGGCACACGTTCCTT
GTGTAAGACTTGTTGTGGTGATGTTGGCGCAGTGTCTCACTGAAACATGAATGAAACCCAGACTTTGTTT
TAATTTCAGGCGAAGGTCCATTTCTCCATGTTATATATCAATCTCTTATTTATTAGTAGCAAAATTGTTT
AAACTTTTTAAAATCCATTGATCACCCTATCATTTTCAATATCTACATACAATCTTATGTCTCGATAAAG
GTTTATCTTTATCTTATTATGCAATACATATCCCTCCCATTTCTAT
```

Erstelle nun eine Klasse `Sequence`. Die Klasse soll folgende Anforderungen erfüllen:

1. Die Klasse soll eine FASTA-Datei einlesen und die Nukleotidsequenz als String speichern. In *dieser* Aufgabe gehen wir der Einfachheit halber davon aus, dass in der Datei nur ein Header und eine Sequenz vorhanden sind. Die FASTA-Datei steht auf der Propädeutikums-Website zum Download bereit.

2. Die Klasse soll die relativen Häufigkeiten der Nukleotide in der Sequenz berechnen und auf der Konsole ausgeben. Die korrekten Werte für die angegebene Datei sind:

```
A=0.3415841584158416; C=0.16831683168316833; G=0.1551155115511551; T=0.334983498349835
```

Übung 5 *Knobelaufgabe: Computing GC-Content*

Bearbeite die Aufgabe Computing GC-Content in Rosalind. Hier ist zu beachten, dass in der FASTA-Datei nicht mehr nur eine, sondern mehrere Header und Sequenzen stehen.

[1] https://en.wikipedia.org/wiki/FASTA_format