

Übungen zum Bioinformatik-Tutorium

Blatt 7

Termin: Dienstag, 04.12.2018, 11 Uhr

1. Fasta Das FASTA-Format ist ein textbasiertes Format zur Darstellung und Speicherung der Sequenz von Nukleinsäuren (Nukleinsäuresequenz) und Proteinen (Proteinsequenz) in der Bioinformatik. Die Nukleinbasen bzw. Aminosäuren werden durch einen Ein-Buchstaben-Code dargestellt. Das Format erlaubt es, den Sequenzen einen Namen und Kommentare voranzustellen¹.

Schreibe eine Klasse `Proteome` die eine Menge von Proteinen speichern kann. Die Klasse soll folgende Elemente enthalten:

- (a) eine private Instanzvariable `ArrayList<Protein> proteins`; (speichert alle Protein-Objekte aus der Fasta-Datei)
- (b) einen Konstruktor `Proteome(File fastaFile)` (liest Fasta-Datei als Protein-Objekte ein). Die Klasse `StringBuilder` könnte hier hilfreich sein.
- (c) eine Methode `getProteinCount()` (gibt die Anzahl der geladenen Proteine zurück)
- (d) eine Methode `getProtein(int i)` (gibt das i-te Protein-Objekt zurück)

Hinweise: Eine `ArrayList` funktioniert ähnlich wie ein Array, allerdings muss die Länge nicht vorher bekannt sein. Die spitzen Klammern deuten darauf hin, dass `ArrayList` eine sogenannte generische Klasse ist und in diesem Fall nur Protein-Objekte speichern kann (siehe später).

- Man kann bequem Objekte über die `add(Protein p)` Methode ans Ende hinzufügen.
- Die Methode `size()` liefert die Anzahl der enthaltenen Objekte
- Über die `get(int i)` Methode kann man sich einzelne Objekte zurückgeben lassen (beginnt bei 0 zu zählen)
- Ein Blick in das javadoc der Java API kann hilfreich sein:
<https://docs.oracle.com/javase/7/docs/api/java/util/ArrayList.html>

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

2. Proteom-Statistiken

Erweitere die Proteome-Klasse um eine Methode `void writeStatistics(File output)`, die folgende Statistiken der geladenen Proteine in die gegebene Datei in einem selbst gewählten Format schreibt:

- (a) Minimale Sequenzlänge
- (b) Maximale Sequenzlänge
- (c) Mittelwert der Sequenzlänge

3. Aminosäure-Statistiken

Als zusätzliche Statistik soll jetzt auch noch die Anzahl jeder Aminosäure mit ausgegeben werden. Erweitere dazu die Protein-Klasse um die Methode `int countAminoAcids(char aa)`, die die Anzahl der gegebenen Aminosäure in der Sequenz des Proteins zurückgibt. Erweitere dann die Proteome-Klasse um die Methode `int countAminoAcids(char aa)`, die `countAminoAcids` für jedes Protein aufruft und die Gesamtzahl zurückgibt.

Die Methode `void writeStatistics(File output)` aus der Proteome-Klasse soll dann diese Methode für jede Aminosäure (aus `Protein.aminoAcids`) aufrufen und die Ergebnisse mit in die Ausgabe-Datei schreiben.

Hinweis: Dass beide geforderten Methoden `countAminoAcids` heißen, ist natürlich kein Problem, da sie zu verschiedenen Klassen gehören.