Einführung in die Bioinformatik

Prof. Dr. Enno Ohlebusch, Prof. Dr. Hans Kestler, Dr. Karlheinz Holzmann,
Tobias Badura

WS 16/17

Übungsblatt 2

Abgabe bis 14.12.2014, 12:00. Lösungen bitte elektronisch an tobias.badura@uni-ulm.de.

1. Aufgabe (3): Gendatenbanken

In Gendatenbanken finden Sie gesammelte Informationen (z.B. Nukleotidsequenzen) zu Genen und Proteinen. GenBank und EMBL („Ensemble") sind zwei der bekanntesten Gendatenbanken.

- Was ist das Fasta Format? Wie ist es aufgebaut?
- Speichern Sie den GenBank Eintrag zum „obesity protein" (U22421) unter „obesity.gb" ab. Erklären Sie daran den generellen Aufbau eines GenBank Eintrages.
- Speichern Sie den EMBL Eintrag zu „leptin" (ENSG00000174697) unter „leptin.embl" ab. Erklären Sie daran den generellen Aufbau eines EMBL Eintrages.

Tipps:
Suchen Sie auf den jeweiligen Seiten nach einer Export Funktion.
Stellen Sie sicher, dass das Format, das Sie exportieren, auch das richtige ist (Fasta $\neq$ GenBank $\neq$ EMBL).

2. Aufgabe (4): Parsen von Fasta-Dateien

In dem EMBL und GenBank Eintrag befindet sich jeweils die Translation zum Protein. Speichern Sie diese jeweils in einer Fasta-Datei. Schreiben Sie die Funktion „parseFasta". Diese soll als Eingabe den Pfad zu der Fasta-Datei erhalten und die darin enthaltene Proteinsequenz als String zurückgeben.

3. Aufgabe (14): Needleman & Wunsch

Schreiben Sie ein Programm, dass den Needleman & Wunsch Algorithmus nutzt, um ein globales Alignment zwischen zwei Proteinsequenzen zu berechnen. Nutzen Sie die PAM250 Matrix als Kostenmatrix. Führen Sie ein Alignment zwischen den beiden Proteinsequenzen aus Aufgabe 2 aus.

Hinweis: Sie können die PAM250.csv verwenden oder die Matrix in R selber definieren, die gap-Kosten sollen hier g = 8 betragen.