Einführung in die Bioinformatik

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

WS 15/16

Übungsblatt 3

Abgabe bis 16.12.2015, 10:00. Lösungen bitte elektronisch (an tobias.badura@uni-ulm.de) mit Namen in Text und Quelltextdateien abgeben.

2. 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" (LEP Human Genome) 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).

3. Aufgabe (7): Parsen von Datenbankeinträgen

Schreiben Sie die Funktionen „parseGenBank" und „parseEMBL" in R. Diese sollen als Eingabe den Pfad zu „obesity.gb" bzw. „leptin.embl" erhalten und die darin enthaltenen Nukleotidsequenzen als String zurückgeben.

Hinweis: Verwenden Sie „readLines" zum Einlesen.

2. Aufgabe (12): Multiples Sequenz Alignment

Implementieren Sie den Center Star Alignment Algorithmus wie in der Vorlesung vorgestellt. Verwenden Sie als scoring Schema für die paarweisen Alignments die folgenden Einstellungen:

- match = 1
- mismatch = -1
- gap = -2

Ermitteln Sie hiermit das multiple Alignment der folgenden Sequenzen:

S1	A	T	T	G	C	C	A	T	T	
S2	A	T	G	G	C	C	A	T	T	
S3	A	T	C	C	A	A	T	T	T	T
S4	A	T	C	T	T	C	T	T		
S5	A	T	T	G	C	C	G	A	T	T