Genomics

Freie Universität Berlin, Institut für Informatik
Peter Robinson
Wintersemester 2014/2015

3. Übungsblatt
Diskussion am 19. November 2014

Aufgabe 1.

Offenes Leseraster: Eine Folge von Tripletts, die für Aminosäuren kodieren und nicht von einem Stoppsignal unterbrochen werden. Ein "langes" Leseraster gilt als Hinweis auf eine proteinkodierende Sequenz, aber wie "lang" ist lang"? Berechnen Sie die erwartete Länge eines zufälligen offenen Lesenrasters, d.h., die Entfernung zwischen dem ersten nicht-Stopp-Kodon und dem nächsten Stoppcodon

Aufgabe 2.

Die DNA ist also prinzipiell in 6 Leserastern ablesbar, dreimal in die eine und dreimal in die andere Richtung. Gehen Sie nun zur Nukleotiddatenbank von NCBI:
`http://www.ncbi.nlm.nih.gov/nuccore/`.
Suchen Sie die Genomsequenz des HI-Virus unter der Accession-Nummer AF033819.3. Sie können in der GEnbank-Datei die Lokalisationen der Gene nachlesen:

Genname	Nukleotidposition
GAG	336..1838
POL	<1631..4642
VIF	4587..5165
VPR	5105..5341
VPU	5608..5856
ENV	5771..8341
NEF	8343..8714

Nun laden Sie die FASTA-Version dieser Seqeunz herunter. Untersuchen SIe die Sequenz mit dem Program `plotorf`, welches Sie hier finden können:

http://emboss.bioinformatics.nl/cgi-bin/emboss/plotorf

Welche ORFs in der Tabelle können Sie in dem Ergebnis von plotorf identifizieren? Auf welchem Strang (+,-) sind sie gelegen? Welche sind die entsprechenden Leseraster? (F1, F2, F3 ...?)

Aufgabe 3.

Schreiben Sie ein Computerprogramm um das Doppelverdau-Problem zu lösen. Sie dürfen den Code in den Vorlesungsfolien als Ausgangspunkt nehmen.

- Verdau mit A: 9,8,4
- Verdau mit B: 11,7,3

- Doppelverdau: 9,6,3,2,1

Welche Reihenfolge von Schnittstellen ist mit diesen Beobachtungen vereinbar?
Nun wenden Sie ihr Programm auf folgende Daten an:

- Verdau mit A: 3;6;8;9
- Verdau mit B: 4;5;7;11
- Doppelverdau: 1;2;3;3;5;6;7

Aufgabe 4.

Beschreiben Sie soviele Unterschiede zwischen dem Mitochondriongenom und dem Kerngenom wie Ihnen einfallen

Aufgabe 5.

Wähle einen Bereich des humanen Genoms aus und visualisieren Sie ihn im USCS Genombrowser. Schätzen Sie welchen Anteil der Sequenz repetetive Elemente aufweist. (zum Beispiel chr5:9,954,995-10,055,004). Nun untersuchen Sie auf analoge Art das Hefegenom (S. cerevisiae, unter "Other"). Sehen Sie einen Unterschied in Hinblick auf die DIchte von repetitiven Sequenzen? S pekulieren Sie über den Grund

Aufgabe 6.

Erklären Sie den Unterschied zwischen einem prozessierten und unprozessierten Pseudogen.

Aufgabe 7.

Wieviele menschliche Pseudogene gibt es? Konsultieren Sie die Webseiten:
http://pseudogene.org/Human/
oder
http://www.gencodegenes.org/

Aufgabe 8.

Welches menschliche Gen hat die meisten Pseudogene? Wie viele sind es?

Um diese Frage zu beantworten, können Sie Daten über alle menschlichen Psuedogene von **Pseudopipe** (http://pseudogene.org/pseudopipe/) herunterladen. Klicken Sie von dort auf den Link *Pipeline Predictions on human genome* und laden Sie den Datensatz "Brent Pseudogenes" herunter. Sie werden eine TSV-Datei mit Angaben über die Pseudogene sowie über das jeweilige Urspungsgen ("Parent protein") erhalten. Sie können dann mit einem Skript die Anzahl von Pseudogenen für jedes Parentprotein bestimmen und sortieren. Konsultieren Sie www.ensembl.org um die entsprechenden Proteinnamen zu finden.