Wintersemester 2007/2008 Abgabe am 28. November 2007
Prof. Dr. Stefan Posch, Dipl.-Bioinf. Jan Grau

6. Übung „Angewandte Bioinformatik mit Perl und R“

1. Wir interessieren uns für ein Gen von *Arabidopsis thaliana*, das einen Transkriptionsfaktor kodiert, der an der Blütenbildung beteiligt ist. Das entsprechende Gen heißt *SEPALLATA1*. Für dieses Gen sind bei NCBI (Genbank, `http://www.ncbi.nlm.nih.gov/Genbank/index.html`) mehrere Sequenzen gespeichert, unter anderem auch eines der Chromosomen von Arabidopsis, auf dem dieses Gen liegt, das uns für eine Analyse allerdings zu groß ist.
   Schreiben Sie ein Perl-Skript, das Objekte der Klassen `Bio::DB::Query::GenBank`, `Bio::DB::GenBank` und `Bio::SeqFeatureI` nutzt, um

   a) alle Nukleotidsequenzen mit Ausnahme des Chromosoms und alle Proteinsequenzen für *SEPALLATA1* zu laden,
      Hinweis: Für die Proteinsequenzen darf als Anfrage nur „SEPALLATA1“ angegeben werden.
   b) zu den geladenen Nukleotidsequenzen die entsprechenden gespleißten Sequenzen bestimmt und diese anschließend translatiert,
   c) jede der translatierten Sequenzen mit allen geladenen Proteinsequenzen zeichenweise vergleicht. Dabei sollen Übereinstimmungen mit einem * gekennzeichnet werden. Stimmen die Aminosäuren nicht überein, sollen die entsprechenden Aminosäuren aus beiden Sequenzen gegenübergestellt werden.

   (5 Punkte)

2. Wir betrachten nun das Genom von *Escherichia coli*, das bei NCBI (Genbank) unter der GI 110341805 verfügbar ist. Auf diesem Genom sind Coding-Sequences (CDS) annotiert, die für kodierende Bereiche jeweils eine Start- und eine Endposition auf dem Genom definieren.
   Schreiben Sie ein Perl-Skript, das Objekte der Klassen/Interfaces `Bio::DB::GenBank`, `Bio::SeqFeatureI`, `Bio::AnnotationCollectionI`, `Bio::AnnotationI` und `Bio::LocationI` nutzt, um festzustellen, ob es überlappende CDS gibt (CDS sind als Features der Genomsequenz repräsentiert). Falls eine Überlappung gefunden wird, sollen Start und Ende der beiden überlappenden CDS und ihre Annotationen ausgegeben werden. (4 Punkte)
   Geben Sie mindestens zwei mögliche Begründungen für Ihre Ergebnisse an. (1 Zusatzpunkt)
   Himweis: Schauen Sie sich zunächst an, in welcher Reihenfolge die Features zurückgegeben werden und von welcher Methode man die Information erhält, ob das aktuelle Feature eine CDS ist.

3. Wir betrachten schließlich das Chromosom 13 von *Saccharomyces cerevisiae* (GI 44829554). Über dieses Chromosom wollen wir folgende Informationen sammeln:

   a) Wieviele Coding Sequences sind für dieses Chromosom bekannt?
   b) Wieviele der Coding Sequences werden gespleißt?
   c) Wie ist die Anzahl der Exons pro Coding Sequence verteilt (wieviele CDS mit einem Exon, wieviele mit zwei Exons,...)?

d) Welchen Anteil haben kodierende (Exons) und nichtkodierende (Introns und intergenische Bereiche) Abschnitte am gesamten Chromosom?

e) Wie sind die Coding Sequences über das Chromosom verteilt?

Schreiben Sie ein Perl-Skript, das diese Informationen mit Hilfe der Klassen/Interfaces `Bio::DB::GenBank`, `Bio::SeqFeatureI` und `Bio::LocationI` sammelt. (4 Punkte)
Visualisieren Sie die Verteilung der Exons aus Teilaufgabe c) und die Verteilung der CDS aus Teilaufgabe e) auf eine von Ihnen gewählte Art. (2 Zusatzpunkte)
Hinweis: Subklassen der Interfaces haben eventuell zusätzliche Methoden, die Klasse eines Objektes kann man durch die Funktion `ref()` *feststellen.*