Übungen zum Sequenzanalyse-Praktikum

Universität Bielefeld, SoSe 2013
Prof. Dr. Jens Stoye · M.Sc. Stefan Janssen · B.Sc. Linda Sundermann
`http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2013summer/SequaPrak`
`praktikum-seqan@CeBiTec.Uni-Bielefeld.DE`

Übungsblatt 6 vom 17.05.2013
Abgabe 22.05.2013

Exercise 6.1:

- Use the NCBI server to get the mouse leptin nucleotide sequence with accession `U18812`. Make sure you get the sequence from the original paper.
- What is the basic structure of a GenBank entry?
- Check out which other formats are also available.
- The leptin entry was replaced on March 30, 1995, why? Can you find the sequence that was published together with the Nature article from Dec 1, 1994?

Exercise 6.2:

- Use accession number `16846` and retrieve the EMBL entry for the same sequence as in the previous exercise.
- EMBL uses SRS to query its databases. SRS is the Sequence Retrieval System which was originally developed by Thure Etzold at the EBI. Click on Text Entry on the top of the page to see the original EMBL entry in text format. What is the basic structure of an EMBL entry?
- Is the information different from what you have seen at the NCBI?

Exercise 6.3:

- Browse to the Taxonomy database. Let's see if we find some information about our ancestors. Search for *neanderthalensis* as token set. Isn't the Neanderthal man extinct? How could they get a protein sequence?
- Which proteins did they sequence (tell us five) and what do they say about Osteocalcin compared to modern humans?
- Are there more sequences for extinct organisms?
- The Taxonomy database is a very useful database to download all available sequences for a given species. How many sequences are available for human?

Exercise 6.4:

- Browse to the PDB website. Has the structure of leptin been solved? If so, by whom and where was it published?
- Which method was used to solve the structure?
- Download the PDB entry and use *rasmol* to visualize the structure.

Exercise 6.5:

- Browse to the Taxonomy Database at NCBI and download a FASTA formatted file of all nucleotide sequences from Rat-kangaroos (Potoroidae).
- How many sequences are available? Did the download succeed? (Sometimes huge batch downloads break, so it is always a good idea to check if all sequences were retrieved!)
- How many partial coding sequences are included in the file? Extract the sequences to a new FASTA file.
- Extract the GI numbers for all sequences.

Exercise 6.6:

- Browse to the Taxonomy Database at NCBI and download a GenBank formatted file of all nucleotide sequences from Rat-kangaroos (Potoroidae).
- How many sequences are available? Did the download succeed? (Sometimes huge batch downloads break, so it is always a good idea to check if all sequences were retrieved!)
- How many different geni and species are represented in the file?
- Write a small shell script that counts the number of sequences for each genus available in a GenBank file. Apply your script to the file above and count the geni.
- How many coding sequences and corresponding GenBank entries are in the file you downloaded? Extract the protein IDs to a file and use NCBI's Batch Entrez to download the protein sequences corresponding to the entries in your file.

Exercise 6.7:

- Browse to the EBI website and download an EMBL formatted file of all nucleotide sequences from Rat-kangaroos (Potoroidae).
- How many sequences are available? Did the download succeed? (Sometimes huge batch downloads break, so it is always a good idea to check if all sequences were retrieved!)
- How many of the entries have the feature key 5'UTR?
- How many and which tRNA-genes are annotated?
- A very useful tool to convert files between different formats is readseq. Try the web version at http://www.ebi.ac.uk/cgi-bin/readseq.cgi to convert the formats and extract features.