582685 Project in Biodatabases (Spring 2012), 2 cr

Contact Veli Mäkinen (vmakinen@cs.helsinki.fi) if you plan to take this project.[1] Next deadline for returning your answers is 31.1.2012.

Assignments

Study the SQL & Ensembl biodatabase lectures by Ilari Scheinin at http://www.cs.helsinki.fi/u/ischeini/eob/. Copy the example python programs to users.cs.helsinki.fi, and run them there.

1. **Biodatabases and SQL I.**

   Write a python program that takes a gene name as user input, finds the corresponding Ensembl Gene ID using SQL, and outputs this ID back to the user.

2. **Biodatabases and SQL II.**

   Write a python program that takes an Ensembl Gene ID as input, and prints out all the corresponding transcript and translation IDs.

3. **Biodatabases and SQL III.**

   Write a python program that takes a gene name as user input, and prints out the length of the gene.

4. **Combining SQL and sequence analysis.**

   Complete the template python program at http://www.cs.helsinki.fi/u/vmakinen/biodbproject/assignment4.py to do the following task: Given input gene names, use SQL to query for the IDs of corresponding transcripts that also have a protein translation, read the corresponding entries from the FASTA file containing all transcripts as cDNA, convert the cDNAs into protein sequences, do pair-wise local alignment for each pair of proteins encoded by a different gene, and output the score of the best local alignment for each pair of genes.

   The motivation for the pipeline is for example a *gene expression study* that indicates a set of genes is active in a certain experimental setting, and one would be interested in finding out if they may share a common function based on sequence similarity.

5. **From gene name to content of its transcripts.**

   Write a python program that asks the user for a name of a gene, and prints out the names of transcripts from this gene, along with the number of exons for each transcript.

6. **From marker to genes.**

   Write a python program that asks the user for a name of a marker, and prints out the list of genes that overlap with the position of this marker.

[1] You can only take this project if you have *not* studied similar content as part of your earlier studies, e.g. in *Elements of Bioinformatics (Autumn 2010 version)*, *Practical Course in Biodatabases*, or *Practical Bioinformatics, Modul 2, Biodatabases*.

7. **From gene name to GO terms.**

   Write a python program that asks the user to provide a gene name, retrieves all GO terms associated with protein translations from this gene, and prints out the list of GO terms to the user.

8. **From gene names to motif finding.**

   Modify your solution to assignment 4. `http://www.cs.helsinki.fi/u/vmakinen/biodbproject/assignment4.py` to read the upstream regions (say, 1000 nucleotides preceding the gene) given a set of gene names as the input. Print the results into a FASTA file with gene names as the headers and upstream sequences as the content. Feed the output to a motif finder program (e.g. `Weeder`: `http://159.149.109.9/modtools/`) and describe the results.[2]

   The chromosomes of human genome are located at `users` server in directory `/home/tkt_mbie/fasta/genome/`. *Hint.* Query gene location inside the chromosome using SQL and then read the corresponding block from the chromosome file. Take care of gene orientation.

[2]For some interesting inputs to the program, browse the JASPAR database for PWMs and follow the description of some PWM to find the genes whose upstream regions contain the transcription factor binding locations constituting that PWM.