Elements of Bioinformatics (autumn 2010)

Lecturers: Ilari Scheinin and Veli Mäkinen

Exercise 3

Tue 23.11, 16-18, C222, Esa Pitkänen

Choose any 5 assignments from below (each assignment gives 1 point, 5 points is maximum for each week; you can do more for better learning, of course).

1. **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.

2. **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.

3. **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.

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

   Modify the previous week's exercise 2.7 solution `http://www.cs.helsinki.fi/u/vmakinen/elements10/ex2_7solution.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.[1]

   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.

5. **PWMs and PSSMs**

   Some binding sites for hematopoietic transcription factor GATA-1 from *H. sapiens* are listed below:

[1] 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.

```
AGATAA
TGATAA
AGATAG
TGATAG
TGATCA
TTATCA
```

Compute the consensus sequence, positional weight matrix (PWM), and position-specific scoring matrix (PSSM) for the sites as described at the lecture (using pseudocounts for the latter). Compute also the sequence logo heights for the letters at each position.

6. **Searching with palindrome PSSM.**

   Modify the example given at lecture `http://www.cs.helsinki.fi/u/vmakinen/elements10/pssm.py` to work with palindrome PSSMs like `AGAACAnnnTGTTCT`.

7. **Formulating gene finding.**

   Formulate the dynamic programming algorithm for computing $S[i, j, k]$ sketched at pages 23-24 on 18.11 lecture slides `http://www.cs.helsinki.fi/u/vmakinen/elements10/elements10_lecture181110.pdf`. Include the speed-up of sliding window maxima computation.

8. **Formulating exon chaining.**

   Formulate a dynamic programming algorithm for the exon chaining problem sketched at page 26 on 18.11 lecture slides `http://www.cs.helsinki.fi/u/vmakinen/elements10/elements10_lecture181110.pdf`. To fix ideas, start from a set of weighted intervals $(\ell_i, r_i) \in I$, $\ell_i < r_i, \ell_i, r_i \in \mathbb{N}$, where $s(\ell_i, r_i)$ is the weight of the interval (ℓ_i, r_i). The the task is to find a list of non-overlapping intervals with the maximum sum of weights. Can you find an algorithm that is linear in the size of I (excluding the necessary step of sorting the endpoints)?