

Homework 2 (Bioinformatics majors only)

L519: Bioinformatics: Theory and Applications

School of Informatics

Indiana University – Bloomington

Feb. 6, 2003

Due date: before class on Feb. 13, 2003

Write a program that extracts and prints a set of subsequences from a input set of sequences as the ranges of the subsequences are specified in a separate input file.

```
$ extract_subseqs.pl -ranges range-file -seqs seq-file [-seqid id-format -dna -rc -protein]
```

Below is an example run of the program.

```
$ cat thisfile.faa
>gi|3323367|gb|AAC26591.1| hemolysin III (hlyIII) [Treponema pallidum]
MEKIVNADGSDAICPASAACAISIRSYQESYSLGEEIANAVTHGIGVGLSIVALVLLVVRVAVHYTPADLT
ARYVVGFSVFGSSLIVLYLCSTLYHALPRGAKYVFGVIDHCCIVVLIAGTYTASCLTTLYGAIGWTVFGV
IWGLACSGSVIYSVFGHRVRWLSLVMYIAMGWLVVVFAKPLRERLPEISFLFLVLGGVLYTVGCVFYALK
RIKWHTIWHMFVIGGSVMHFFSLYLSF
>gi|3323368|gb|AAC26592.1| bacterioferrin (TpF1) [Treponema pallidum]
MNMCTDGKKYHSTATSAAVGASAPGVPDARAIAAICEQLRQHVADLGVLYIKLHNYHWHIYGIEFKQVHE
LLEEYVSVTEAFDTIAERLLQLGAQAPASMAEYLALSGIAEETEKEITIVSALARVKRDFEYLSTRFSQ
TQVLAAESGDAVTDGIITDILRTLKAIWMLGATLKA
>gi|3323369|gb|AAC26593.1| adenine phosphoribosyltransferase (apt) [Treponema pallidum]
MRGIGRYHAPVDGHAALDRAIRKIDFPKKGILYYDITGVLNNAVFRYCLDQMVEFYRDEHVTAVAAIE
SRGFIFAAPFADRMGIPLILVRKAGKLPGDTYSCSYSLEYGKATVEVHKSDVVAGARVLLTDDLIATGGT
LNAARTMLRAGGAEVVGGFFAVVGLPFLRYHELIGDLPVRTLIEYNQETSN
```

```
$ cat thesesubseqs
3323367 15 35
3323369 1 20
3323367 10 20
```

```
$ extract_subseqs.pl -ranges thesesubseqs -seqs thisfile.faa -protein -seqid gi
>3323367:10-20
SDAICPASAAC
>3323367:15-35
PASAACAISIRSYQESYSLGE
>3323369:1-20
MRGIGRYHAPVDGHAALDRAI
```

Options :

- **-ranges range-file**
The file range-file specifies the ranges of seqs as in the example above.
- **-seqs seq-file**
- **-seqid id-format**
An optional argument. When there are several seq ids like in the .faa file, users can designate which id will be used.
- **-dna**
An optional argument. A DNA seq input file is assumed.
- **-rc**
An optional argument. For a DNA seq input file, seqs should be reverse-complemented when the first range number is greater than the second range number.
- **-protein**
An optional argument. A protein seq input file is assumed.

Data : You can use data on /bioinfo/data/L519/data on IBM SP and Solar.

Discussion Forum : hw2 on oncourse.iu.edu

If the forum is not set, please email me immediately or call me.

Directory and submission :

DIRECTORY STRUCTURE AND INSTRUCTIONS CHANGED. READ CAREFULLY.

- Create a directory /gpfs/L519/hw2/yournetwork-id
- Put the three scripts on /gpfs/L519/hw2/yournetwork-id
- Write a short **README** file about your homework. One line under the heading **SUBMIT** in your README file should clearly describe how to submit your script by simply *cut-and-paste* the line (thus, all arguments should be correctly specified and we assume that your script run on the current directory. Another line under **RESULT** should indicate which file contains the output from executing the script.

Hints :