

Genomics (Ecol 553) Computational Lab
Week 11: Nov 7, 2008.

Course webpage: <http://genomics.arizona.edu/553/computation/>

Topics

Perl: regular expressions

Note

This is the final homework.

Homework due at 2:00PM Thursday, Nov 13

All work described below should be placed in a new directory on ice.hpc.arizona.edu, named: `~/homework/homework5`. When you are ready to submit your work, you will `cd` to `~/homework`, then run the command `turnin homework5`. (In case your `PATH` is broken, the `turnin` command is: `/home/u1/twheeler/553/bin/turnin`)

- 1) Write a program, called `get_blast_subset.pl`, which does the following:
 - Takes two arguments
 - name of a file containing a list of gi numbers (I'll call it `gi_file`)
(an example can be found at `/scr1/twheeler/gi_nums`)
 - name of a file to which all results should be printed (I'll call it `out_filename`)
 - Reads the list of gis from `gi_file`
 - Reads in the blast result file `/scr1/twheeler/big.blast`
 - For each id in `gi_file`, finds the portion of the big blast file corresponding to a query for that gi, and prints it to `out_filename`.

The result will be a file that looks just like the original blast file: a series of blocks, each of which starts with “# BLASTP”, and includes 4 comment lines followed by some number of hit lines. The difference is that the file `out_filename` will be much smaller, containing only the results for gis in `gi_file`.

- 2) Repeat problem #4 from homework 3, this time using regular expressions to do the heavy lifting. I recap the question below:

First, make a copy of my `/scr1/twheeler/READS/` directory. The result should be a directory named `~/homework/homework5/READS`

Then write a program, called `fix_filenames2.pl`, which does the following:

- Creates a new directory, called `~/homework/homework5/CHANGED_READS`
- Gets a listing of the files in `READS`, and for each one, cleans up the name:
 - The first three characters correspond to well numbers; we don't need them
 - The suffix "seq" is non-descriptive, it should be "fa" (short for fasta)
 - The character "_" splits two names, but they are in the wrong order; we'd like to reverse them.
 - For example, the name `A018_2867F.Seq` should be changed to `2867F_8.fa`
 - Preserving the original file, make a copy of the file to the `CHANGED_READS` directory, but with the new name.
 - For example the file `READS/A018_2867F.Seq` should be copied to `CHANGED_READS/2867F_8.fa`

(note: I change the new filename syntax to use a "_" instead of a "-", for clarity)

3) (I had another question planned here, but it was based on the assumption that we had class on Tuesday. No class? Goodbye question. Instead ...) What role will Perl play in your final project for this class? Please be brief - 200 or fewer words.