

Genomics (Ecol 553) Computational Lab
Week 7: Oct 9, 2008.

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

Topics

Perl: arrays, reading files, running commands

Homework due at 11:59PM Sunday, Oct 19

All work described below should be placed in a new directory on ice.hpc.arizona.edu, named: `~/homework/homework3`. When you are ready to submit your work, you will `cd` to `~/homework`, then run the command `turnin homework3`. (In case your `PATH` is broken, the `turnin` command is: `/home/u1/twheeler/553/bin/turnin`)

1) Write a program, called `blast_smallest_pct.pl`, which does the following:

- Reads in the blast result file `/scr1/twheeler/big.blast`
- Prints out the smallest percent identity among all hits. (Not the smallest percent identity for each query sequence, the single smallest overall)

2) Write a program, called `k_smallest_pct.pl`, which does the following:

- Takes a number as an argument (I'll call that number `k`)
- Reads in the blast result file `/scr1/twheeler/big.blast`
- Prints out the `k` smallest percent identities among all hits. (Not the `k` smallest percent identity for each query sequence, the `k` smallest overall)

But be careful: your program should gracefully handle the case in which the argument (`k`) exceeds the number of records in `big.blast`.

(note: we will discuss how to accept arguments in a Perl script on Tuesday)

3) Write a program, called `get_seqs.pl`, which does the following:

- Takes a `gi` number as an argument, call it `gi`
- Reads in the blast result file `/scr1/twheeler/big.blast`
- For each hit where the query id is equal to `gi`:
 - Get the subject id from that hit line, call it `subject_gi`.
 - Build a `fastacmd` call to grab the sequence for `subject_gi`.
 - Run that command (inside the perl script), appending the results to a file called `$gi.hits.fasta`

The result should be a file with all the sequences of genes that contained matches for the query sequence `gi`.

For example, if you want the sequences for the hits that have as the first entry in the hit line `gi|15594348|ref|NP_212136.1|`, the argument should be `15594348`, and the resulting file should be `15594348.hits.fasta`

The database you should query with `fastacmd` is `/genome/nr`

Extra credit 1: instead of using `fastacmd` to get the entire sequence for the hit `gi`, get just the sequence corresponding to the range of the hit, based on `s.start`, `s.end`. Be careful to collect the correct strand. Make a new script that does this, and call it `get_seqsEC1.pl`

Extra credit 2: instead of accepting a single `gi` number as argument, accept the name of a file. That file will contain a list of `gis` numbers, and the task described above should be performed for all `gis` on that list. Make a new script that does this, and call it `get_seqsEC2.pl`

(note: we will discuss how to use Perl to create and write to a file on Tuesday)

4) First, make a copy of my `/scr1/twheeler/READS/` directory. The result should be a directory named `~/homework/homework3/READS`

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

- Creates a new directory, called `./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`