CMSC423 Homework 2
Handed out: 9/16/2008
Due: 9/23/2008

This is one of the few assignments that require you to write code in Perl.

Your assignment is to write a Perl script that can help retrieve data from the NCBI Trace Archive using the *query_tracedb* script provided by NCBI. This script is available on the glue machines in the public/ directory.

The script you write must fulfill the following specifications:

1. Must accept two parameters on the command line, using the following invocation
   my_query_tracedb <organism> <max records>

where <organism> is the name of an organism (species code) stored in the Trace Archive, and <max records> indicates the maximum number of records to be retrieved. If more than <max records> are available in the database for this organism, your script will just return the first <max records> from among them. If <max records> is not provided, your script should retrieve all records for the given query.

2. The trace archive sets an upper limit on the number of records you are allowed to retrieve at a time using *query_tracedb*. Your script should obscure this limit from the user, i.e. your script should retrieve all records requested using as many independent requests to the Trace Archive as necessary. Note: for this assignment please set the upper limit of records/chunk to 5,000.

3. Your program should retrieve just fasta and quality information in .tar.gz format

Additional details:
   1. A sample query, to get you started, is:
      SPECIES_CODE = "WOLBACHIA ENDOSYMBIONT OF DROSOPHILA MELANOGASTER"
      e.g. your script should accept the command
      my_query_tracedb "WOLBACHIA ENDOSYMBIONT OF DROSOPHILA MELANOGASTER" <nn>
      where <nn> is the number of records you want to retrieve.
   2. It is OK if your script generates multiple output files, corresponding to individual chunks of 5,000 records.
   3. Please provide us with a simple README file that indicates how to run your program. Make sure you include a sample invocation that you know for sure will work.

(see next page)

Grading! We will grade all aspects of the code, including how "pretty" it looks. Specifically pay attention to the following aspects:

1. Please make sure that your code works as advertised in the README file you provided. If your code doesn't work as indicated in the README file you will automatically lose 50% of the grade for this assignment.
2. Please provide copious comments and format your code so that it is easy to read. Part of your grade will be based on the formatting of the code.

Submission! Please submit your assignment using the 'submit' command on the glue machines, i.e.:

>   submit 2008 fall cmsc 423 0101 2 <submission_file>

For this assignment we will not accept e-mail submissions.

Please contact me as soon as possible if you have any questions regarding this assignment, or if you "get stuck" and might not be able to complete the assignment on time. Once the assignment is due I will no longer accept any excuses.

Good Luck!