Homework 1

Instructions. This homework is due at the beginning of lab, Wednesday, September 6th. You can edit this file and copy/paste into it by opening it in Libreoffice on Linux, PDFfiller Google App, Acrobat, and others (I can only confirm Libreoffice). Bioinformaticians are especially agile at using the internet to gather and share information, so if you do not remember or see the answer in class materials, feel free to use the web, but do not plagiarize it. Also, like anything you can obtain from the internet, information about bioinformatics may be wrong or only partially correct. Be sure to check the source of the information to help judge its quality.

1. The FASTA format is a common format for sequence data used in bioinformatics. Please answer these questions about this format. The first three are true/false questions.
   (a) ___ FASTA files are binary files.
   (b) ___ The sequence block should contain no more than 80 nucleotides per line.
   (c) ___ The header or descriptor for each sequence can contain any printable ASCII character.
   (d) The sequences in a FASTA file are encoded using IUPAC symbols. In class we mentioned the 15mer (15 base pair oligomer) YYYYYYYYYYYYYYY. What nucleotides does the IUPAC code Y represent? _____.
   (e) The newline character(s) that occur(s) at the end of each line (except perhaps the last) in a text file are encoded differently on different operating systems. Sometimes these differences can confuse a novice bioinformatician. In the `hpc-class` directory `/ptmp/bcbio444/hw1`, you will find an executable program called `seq_name` that reads FASTA files and reports the name of the nth sequence, the first by default.
      i. To run this command, you need to log onto `hpc-class`, type the full path of the program and give it two arguments, the name of a text file and an (optional) index of the sequence name sought. For example, the following command outputs the name of the third sequence in the FASTA file `permissive.fa`.
      `/ptmp/bcbio444/hw1/seq_name /ptmp/bcbio444/permissive.fa 3`
      Use this command to report the name of the 153rd sequence in the FASTA file `/ptmp/bcbio444/hw1/REF_2010_env_DNA.fasta`. Record the output below:

      It does not work because the file was created on a Windows machine. You can detect this fact by opening the file in `vim`. When you first open it, read the last line where `vim` tells you the name of the file just opened and some other information about the file. You will see `[dos]` if the file was created on the Windows OS. (You will see `[noeol]` if there is no newline after the last line, which can also confuse bioinformaticians and bioinformatics software.)

ii. To fix the file, copy it to your own directory, and learn how to run the command `dos2unix`, which converts the newlines in the file. Verify that the converted file works with the program `seq_name`. Report all the commands you ran to answer Parts i–ii (the command `history` may help you remember, but clean it up to just include the necessary commands).

iii. What one-liner UNIX command could you use to determine the name of the 153rd sequence.

2. RNA-Seq is an extremely common procedure for generating transcriptomic data. This will not be the first time we talk about it. Read the Wikipedia entry (linked) through the Methods section. (For those of you interested, notice the section on Genomic Medicine.) For engineers who have not seen biology since high school, this page describes the basic biology in a very stepwise way, which should make you comfortable, but consider asking your biological colleagues for additional help. Also, you should know the The Central Dogma of information flow in molecular biology.

   In the following question, suppose you have a pair of tissue samples, normal and tumor, from 14 patients with small cell lung cancer (SCLC). You isolate the RNA from each tissue of each patient and perform RNA-Seq on the mRNA. For the purposes of this question, you are focused on a single gene that you think plays a role in SCLC. A bioinformatician prepares the data and gives you the number of RNA-Seq *reads* (fragments sequenced) of this gene in each of the samples. Please answer the following questions.

   (a) Why is the RNA-Seq experiment an example of a random experiment?

   (b) For a single sample (tumor or normal) from a single patient, what is the sample space?

(c) Thirteen of the 14 patients show the same pattern: there are more reads for this gene in the tumor than the normal tissues.

i. If this gene has nothing to do with the cancer (this is a hypothesis, or model), what probability would you assign to the event that there are more reads in the tumor tissue for a single patient?

ii. If patients are independent, what is the probability of the event that 13 of 14 patients have more reads in the tumor samples? Show your work.

3. The file `/ptmp/bcbio444/hw1/REF_2010_env_DNA.fasta` mentioned in question 1 has parsable sequence descriptors, where the accession number is the word fragment after the final dot. Read this page on `sed` regular expressions (regexps) and this page of regexp examples.

(a) Use `sed` on the command line to extract the accession numbers from the file. One solution uses these regexp elements: `.`, `*`, `\(`*`regexp`*`\)`, `\`*`digit`*, `\`*`char`*. Record the command below.

(b) Look up the accession numbers at using Batch Entrez at NCBI and record the source species of these sequences. Are they all from the same species? Are they all from the same species? Are they all from the same species? Are they all from the same species?