I519: Introduction to Bioinformatics, Fall, 2015

September 11th, 2015

Homework 2 (Due via canvas by **11:59pm on Thursday, Sep 24th, 2015**)

Background: From Genome to Proteins

Genes in prokaryotic genomes are continuous segments in the genome that encode for proteins or RNA molecules (eukaryotic genomes have more complex gene structure). The conversion of the information in DNA (through the intermediate RNA molecules) into protein represents a translation of the genetic information (encoded in 4 nucleotides) into a different language that uses different symbols (for 20 amino acids). And we now know that this translation of genetic information follows the rules that are known as the genetic code, which specifies the meaning of each of the codons of 3 nucleotides, either encoding for an amino acid or none (then it is a stop codon). The genetic code is used universally in all species, but variations in the genetic code do exist. For example, *Candida albicans* (the most prevalent human fungal pathogen) translates CUG as Ser (usually it is Leu). But for this assignment, you only deal with the universal genetic code.

1. Six frame translator (40pts)

Given a genome, it is not a trivial task to predict all the genes. We will talk about gene prediction later. At this moment we want to try out something that is easier. Given a DNA sequence, we will translate it in 6-frames as demonstrated below,

```
>seq1
DNA: CTCGCCATTAACCGTTTCAGCCCCAGGTGCCTTTCTTGAGGC
 +3:   R  H  *  P  F  Q  P  Q  V  P  F  L  R
 +2:  S  P  L  T  V  S  A  P  G  A  F  L  E
 +1: L  A  I  N  R  F  S  P  R  C  L  S  *  G
   : GCCTCAAGAAAGGCACCTGGGGCTGAAACGGTTAATGGCGAG
 -1: A  S  R  K  A  P  G  A  E  T  V  N  G  E
 -2:  P  Q  E  R  H  L  G  L  K  R  L  M  A
 -3:   L  K  K  G  T  W  G  *  N  G  *  W  R
```

Why we need to consider the reverse complement of a sequence? We know DNA forms double helix. But we only need to keep the nucleotide sequence of one chain in the genome sequence file, since the other chain has to be complementary to this one. And genes can be found on either chain.

Write a program **BioTranslator.py** that gets a FASTA file of nucleotide sequences, and translates each of the sequences in all six possible frames: frame +1, +2, and +3, starting from position 1, 2, 3 on the forward chain; frame -1, -2, -3, starting from position 1, 2, 3 on the reverse complement. The input of your program is a file with a DNA sequence, and your program outputs the peptides translated from the DNA sequence as shown above. You will use the standard codon table (codon.txt). And you may use sampleseq.txt as an example input.

2. A random genome generator (40 pts)

Write a program SimuGenome.py to generate a "random" genome that has the same length and the same nucleotide composition (i.e., the same percentages of A, T, C and G) as an input genome (e.g., NC_010698.fna), but otherwise the sequence is more or less random. As a sanity check, your program needs to compute the compositions of the genome that is simulated, and compare them with the compositions of the real genome.

Note: you may need to use module random for this assignment.

```
import random  #import module before you use its functions
random.randrange(4) #this call returns a random integer n ∈ [0, 3]
```

3. Paper reading (20 points)

Paper: Po-Ru Loh et al, Compressive genomics. Nature Biotechnology 30, 627-630, 2012, http://www.nature.com/nbt/journal/v30/n7/full/nbt.2241.html

Write a one page summary for this paper. The summary should be a condensed version of the paper, written in your own words. Don't copy and paste the abstract. Dig deep into the algorithms presented in the paper, and include in your summary important details of the algorithms (e.g., how they work and their complexity). Although it is not required (and not recommended in general for summary writing), you are encouraged to write about your opinions—just like a reviewer—about the presented work in your summary!