I529 HW1-Section1

Chia-Hsuan Chou

Jan. 29, 2016

I. Goal

The goal is to generate random sequences and calculate the mean and standard deviation of every nucleotide in generated sequences.

II. Procedure

For RANSEQ1.py and RANSEQ2.py procedure, they are basically the same. The only difference is the way to generate the random sequences. First, they both read the sequence from input FASTA file. However, RANSEQ1.py calculate the frequency of every nucleotide and based on that frequency to generate random sequences. For RANSEQ2.py, it shuffles the nucleotides appearing in the input sequence and generate new sequences based on the permutation. Next, they all calculate the standard deviation and mean for every nucleotide based on the generated sequences. Finally, they all print the information above on the screen and write the information in an output file.

III. Result

```
Random sequence generation:
TTCCTTTTTT
TTCACATTAA
GTCCATGATA
CTGTATTTTA
CTCCTATTTC
TTATTTATGT
AGGCAGTTTT
GCTTCTTTTT
TCCTTATATT
TTACTATCTT
The mean of A is 0.1800
The mean of C is 0.1800
The mean of G is 0.0800
The mean of T is 0.5600
The standard deviation of A is 0.116619
The standard deviation of C is 0.097980
The standard deviation of G is 0.097980
The standard deviation of T is 0.149666
```

Figure 1. The output file from RANSEQ1.py

```
Random sequence permutation:
ATTTTCCATG
TTGCTACTAT
TCCTTGTTAA
ATCTCTATTG
TTTCTAACTG
ATGCCTATTT
TGCTTCATAT
TATGTCTATC
CGTCTATTAT
TTTCATGTAC
The mean of A is 0.2000
The mean of C is 0.2000
The mean of G is 0.1000
The mean of T is 0.5000
The standard deviation of A is 0.000000
The standard deviation of C is 0.000000
The standard deviation of G is 0.000000
The standard deviation of T is 0.000000
```

Figure 2. The output file from RANSEQ2.py

IV. Discussion

For the result of RANSEQ1.py, because the program generate the sequences based on the frequency of every input nucleotide, the mean and standard deviation will not be the same every time when we generate it. However, for the result of RANSEQ2.py, the program generate the sequences based on shuffling the input nucleotide, the mean will same as the mean of input sequence and the standard deviation will be 0 all the time.

I think that for random generate sequences, the way which RANSEQ1.py do will be better. Because for the random sequence permutation, there is no difference between the standard deviation of every nucleotide and the mean is exactly the same as input sequence and there is no any statistics analysis to do. If we think about evolution, it is not possible that every nucleotide numbers will be exactly the same via evolution. However, the frequency of every nucleotide might be the same and it can give us some clues of predicting new gene.