Final Problem Set

iCB2 – Introduction to Systems Biotechnology
13 November, 2015

Problem 1

Write a python script that takes in input a fasta file of only one protein sequence and calculate the frequency of each amino acid.
Check the output of for the files:

http://www.uniprot.org/uniprot/P53_HUMAN.fasta
http://www.uniprot.org/uniprot/BRCA1_HUMAN.fasta

Write a shell script that automatically downloads the UniProt file (*wget*) and run the previous python script on the downloaded files.

```
> ./testprog.sh
> P53_HUMAN.txt
A: 0.061 C: 0.025 E: 0.076 D: 0.051 G: 0.059 F: 0.028 I: 0.020 H: 0.031 K: 0.051 M:
0.031 L: 0.081 N: 0.036 Q: 0.038 P: 0.115 S: 0.097 R: 0.066 T: 0.056 W: 0.010 V:
0.046 Y: 0.023
> BRCA1_HUMAN.txt
A: 0.045 C: 0.024 E: 0.106 D: 0.046 G: 0.047 F: 0.026 I: 0.041 H: 0.026 K: 0.074 M:
0.016 L: 0.084 N: 0.065 Q: 0.052 P: 0.052 S: 0.120 R: 0.041 T: 0.060 W: 0.005 V:
0.054 Y: 0.017
```

Suggestion: Use the syntax – print aa+”:”,“%7.3f” %frequency – to show only the first 3 decimal digits.

Problem 2

The PFAM domain distribution for human proteome can be found at

ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/proteomes/9606.tsv.gz

Unzip the file. The first column of this file is the protein accession number. The location of the domain hits is given by the columns 2-5. Columns 2-3 are alignment start and end. Columns 4-5 are envelope start and end. Envelopes are generally considered the location of a domain on a protein. Write a python program that takes 9606.tsv.gz file as a first argument, a protein accession number as a second argument, and a location (integer) as a third argument. The program should print the domain name (hmm_name), the envelope starting and end, if the location falls within a domain for a given protein accession. The program should return nothing if the position is outside the boundaries of domains. We should be able to run the program from terminal like this

```
> python nameprog.py ../data/9606.tsv O95931 20
> Chromo 11 60
```

Suggestion: A *if* statement with three conditions needs to be used.