

Genome 373 14 April 2009 Python practice session

Sequence data (DNA or protein sequences) are often stored in a specific format known as FASTA. Here is an example of a short file containing 2 sequences entries:

```
>MSRA Homo sapiens methionine sulfoxide reductase A (MSRA)
CAGCCGGTACGGCCCCGGGTTTGGGCAACCTCGATTACGGGCGGCCTCCAGCCCCGCCAGCAGCGCCCCG
CGCCCCCGCCCGCCCGCCCCCTGCCGCCCCCGGTTCCGGCCGCGGACCCCACTCTCTGCCGTTCCGGCTG
CGGCTCCGCTGCCGGTAGCGCCGTCCCCCGGGACCACCCTTCGGCTGGCGCCCTCCCATGCTCTCGGCCA
>APOBEC3B Homo sapiens apolipoprotein B mRNA editing enzyme
ACAGAGCTTCAAAAAAGAGCGGGACAGGGACAAGCGTATCTAAGAGGCTGAACATGAATCCACAGATCA
GAAATCCGATGGAGCGGATGTATCGAGACACATTCTACGACAACCTTTGAAAACGAACCCATCCTCTATGG
TCGGAGCTACACTTGGCTGTGCTATGAAGTGAATAAAGAGGGGCCGCTCAAATCTCCTTTGGGACACA
GGGTCTTTTCGAGGCCAGGTGTATTTCAAGCCTCAGTACCACGCAGAAATGTGCTTCCTCTCTTGGTTCT
GTGGCAACCAGCTGCCTGCTTACAAGTGTTCAGATCACCTGGTTTGTATCCTGGACCCCTGCCCGGA
CTGTGTGGCGAAGCTGGCCGAATTCCTGTCTGAGCACCCCAATGTACCCTGACCATCTCTGCCGCCCGC
```

The key things to note about the FASTA format are that each sequence record begins with a header line. Header lines begin with a '>' character which is immediately followed by the name of the sequence. Following the name is a description of the sequence and other relevant data separated by spaces.

As a review of loops and file processing, write two programs to compute simple tasks on a file of sequences in the FASTA format. First, write a program `count-fasta-seqs.py` that takes as a command line argument the name of a file, and prints out the name of each sequence and the total number of sequences in the file. For example, if run on the sequence fragments above:

```
python count-fasta-seqs.py sample.fa
MSRA
APOBEC3B
There are 2 records in the file sample.fa
```

Now, write a program that calculates the total number of sequences in a FASTA file and reports the average sequence length:

```
python average-length.py sample.fa
There are 2 records in the file sample.fa
The average sequence length is 315.0
```

Make sure your program works on a small file that you generate, and then run it on a list of the genes on *S. cerevisiae* chromosome I. A fasta file of the sequences of these genes can be downloaded here:

ftp://ftp.ncbi.nih.gov/genomes/Saccharomyces_cerevisiae/CHR_I/NC_001133.ffn

```
python average-length.py NC_001133.ffn
There are 94 records in the file NC_001133.ffn
The average sequence length is 1503.0
```

count-fasta-seqs.py

```
import sys
fname=sys.argv[1]
myFile=open(fname,'r')
numSeqs=0;
for line in myFile:
    if line[0] == '>': #records start with '>'
        numSeqs+=1
        words=line.split() #split based on white space
        name=words[0] #get the first one
        name=name[1:] #remove the '>' char
        print name
myFile.close()
print "There are",numSeqs,"records in the file",fname
```

average-length.py

```
import sys
fname=sys.argv[1]
myFile=open(fname,'r')
numSeqs=0;
totLen=0
for line in myFile:
    if line[0] == '>': #records start with '>'
        numSeqs+=1
    else : #must be sequence
        line=line.rstrip() #remove '\n'
        totLen+=len(line) #add to the total length
myFile.close()
print "There are",numSeqs,"records in the file",fname
aveLen=float(totLen)/numSeqs
print "The average sequence length is %.1f" % aveLen
```