Functions, Scope & Arguments

HORT 530

Lab 12

Instructor: Kranthi Varala

Create TF family FASTA files

- Write a program that does the following:
  - Read in a protein FASTA file
  - Add each sequence longer than 100 aa to a dictionary.
  - Read in a tab-delimited file of protein IDs and their TF annotation.
    - 3 columns: TF_ID Gene_ID Family
  - Find all transcription factors in the list of proteins.
  - Create a set of new FASTA files such that protein sequences from each TF family are in separate files.

Create TF family FASTA files

- Protein fasta file : Arapep.fasta

- Tab-delimited TF file : Ath_TF_list

- Output file (example) :

```
>AT4G29930.3 | bHLH | basic helix-loop-helix (bHLH) DNA-binding superfamily protein
MEDLDHEYKNYWETTMFFQNQELEFDSWPMEEAFSGSGESSSPDGAATSPASSKNVVSER
NRRQKLNQRLFALRSVVPNISKLDKASVIKDSIDYMQELIDQEKTLEAEIRELESRSTLL
ENPVRDYDCNFAETHLQDFSDNNDMRSKKFKQMDYSTRVQHYPIEVLEMKVTWMGEKTVV
VCITCSKKRETMVQLCKVLESLNLNILTTNFSSFTSRLSTTLFLQADEEESSAVEAKIQM
AIAAYNDPNCLINF
>AT3G55370.3 | Dof | OBF-binding protein 3
MVFSSLPVNQFDSQNWQQMISILVFFSTSRLFKKLFLVDKNLFSCLLQGLMYNVFLTGLI
FSLQGNQHQLECVTTDQNPNNYLRQLSSPPTSQVAGSSQARVNSMVERARIAKVPLPEAA
LNCPRCDSTNTKFCYFNNYSLTQPRHFCKTCRRYWTRGGSLRNVPVGGGFRRNKRSKSRS
KSTVVVSTDNTTSTSSLTSRPSYSNPSKFHSYGQIPEFNSNLPILPPLQSLGDYNSSNTG
LDFGGTQISNMISGMSSSGGILDAWRIPPSQQAQQFPFLINTTGLVQSSNALYPLLEGGV
SATQTRNVKAEENDQDRGRDGDGVNNLSRNFLGNININSGRNEEYTSWGGNSSWTGFTSN
NSTGHLSF
```

PseudoCode: Create TF family FASTA files

```
def readFASTA(<fasta>,<minLength>):
    Open fasta file
    Read each sequence and check length
    If length>=minLength: Add 'id' : 'sequence' to dictionary

get names of FASTA and TF files

if input FASTA file exists:
    call readFASTA(<fasta>,<minLength>)

else: quit program with error message

if input TF annotation file exists:
    Read tab-delimited file
    Build dictionary of Family => [List of TFs]

else: quit with error message

Iterate over TFfamily dictionary:
    .
    .
    .
```

Getting user input and checking files

```python
import sys, getopt, os

if len(sys.argv) == 3:
        fastaFile = sys.argv[1]
        TFlist = sys.argv[2]
else:
        print "Usage: python " + sys.argv[0] + " <FASTA file> <TF list>"
        sys.exit(1)

fastaExists = os.path.isfile(fastaFile)
TFExists = os.path.isfile(TFlist)

if fastaExists:
        print "Found fasta File %s" % fastaFile
else:
        print "Cannot find fasta File %s" % fastaFile
        sys.exit(1)

if TFExists:
        print "Found TF list %s " % TFlist
else:
        print "Cannot find TF File %s" % TFlist
        sys.exit(1)
```