BCB570 HW4

February 15, 2018

Urminder Singh

```
In [5]: '''
        This file reads .gaf files and finds out the annotation statistics
        [1]--> file path
        Urminder Singh
        '''
        from __future__ import division
        import sys
        import Bio.UniProt.GOA
        import tabletext

        class bcolors:
            HEADER = '\033[95m'
            OKBLUE = '\033[94m'
            OKGREEN = '\033[92m'
            WARNING = '\033[93m'
            FAIL = '\033[91m'
            ENDC = '\033[0m'
            BOLD = '\033[1m'
            UNDERLINE = '\033[4m'

        #read .gaf file with gafiterator

        infiles=["goa_human.gaf","goa_cow.gaf","goa_rat.gaf","goa_mouse.gaf","goa_zebrafish.ga
        tProt=["Total Proteins"]
        eProt=["Prots with exp evidence"]
        neProt=["Prots with no exp evidence"]
        perc=["%Prots with exp evidence"]
        exp_ids=(["EXP","IDA","IPI","IMP","IGI","IEP"])

        for f in infiles:
            prots_exp=[]
            evd_exp=[]
            prots_all=[]
            with open(f, 'r') as handle:
                for rec in Bio.UniProt.GOA.gafiterator(handle):
                    #make two different lists for experimentally verified and other type of pr
                    prots_all.append(rec["DB_Object_ID"])
                    if rec["Evidence"] in exp_ids:
                        prots_exp.append(rec["DB_Object_ID"])
                        evd_exp.append(rec["Evidence"])
            #remove redundant entries
            totalprots=len(set(prots_all))
            num_exp_prots=len(set(prots_exp))
            num_noexp_prots=totalprots-num_exp_prots
            print bcolors.BOLD +'Results for input file:',f,bcolors.ENDC
            print 'Total proteins:',totalprots
            print 'Total proteins with experimental evidence:', num_exp_prots,'fraction:',(num_
            print 'Total proteins without experimental evidence:', num_noexp_prots,' fraction:

            #data for table
            tProt.append(totalprots)
            eProt.append(num_exp_prots)
            neProt.append(num_noexp_prots)
            perc.append(format((num_exp_prots/totalprots*100), '.2f'))
```

```
Results for input file: goa_human.gaf
Total proteins: 19502
Total proteins with experimental evidence: 13619 fraction: 69.8338631935
Total proteins without experimental evidence: 5883  fraction: 30.1661368065
Results for input file: goa_cow.gaf
Total proteins: 20193
Total proteins with experimental evidence: 671 fraction: 3.32293368989
Total proteins without experimental evidence: 19522  fraction: 96.6770663101
Results for input file: goa_rat.gaf
Total proteins: 19710
Total proteins with experimental evidence: 5641 fraction: 28.6199898529
Total proteins without experimental evidence: 14069  fraction: 71.3800101471
Results for input file: goa_mouse.gaf
Total proteins: 21691
Total proteins with experimental evidence: 11272 fraction: 51.9662532848
Total proteins without experimental evidence: 10419  fraction: 48.0337467152
Results for input file: goa_zebrafish.gaf
Total proteins: 21842
Total proteins with experimental evidence: 4229 fraction: 19.3617800568
Total proteins without experimental evidence: 17613  fraction: 80.6382199432
```

```
In [6]: #print final table
        print bcolors.BOLD + bcolors.HEADER+'\t\t\t\n\nTable 1. Results for the input files'+bc
        tdata=[]
        header=["","human","cow","rat","mouse","zebrafish"]
        tdata=[header,tProt,eProt,neProt,perc]
        print tabletext.to_text(tdata)
```

```
                        Table 1. Results for the input files
```

```
                            human  cow    rat    mouse  zebrafish

Total Proteins              19502  20193  19710  21691      21842

Prots with exp evidence     13619    671   5641  11272       4229

Prots with no exp evidence   5883  19522  14069  10419      17613

%Prots with exp evidence    69.83  3.32   28.62  51.97  19.36
```

Table 1 summarizes all the results found by this script. If we look at raw numbers humans have most number of annotated proteins i.e. 13,619. If we compare the percent of proteins experimentally verified, we again see that human have the most i.e. 69.83% followed by mouse which has almost 52% of proteins experimentally verified.