Annotation frequencies in the UniProt database

Laurent Gatto lg390@cam.ac.uk

May 8, 2014

1 Introduction

We will use the Bioconductor[1] annotation infrastructure to assess the number of cellular compartment GO terms for UniProt entries.

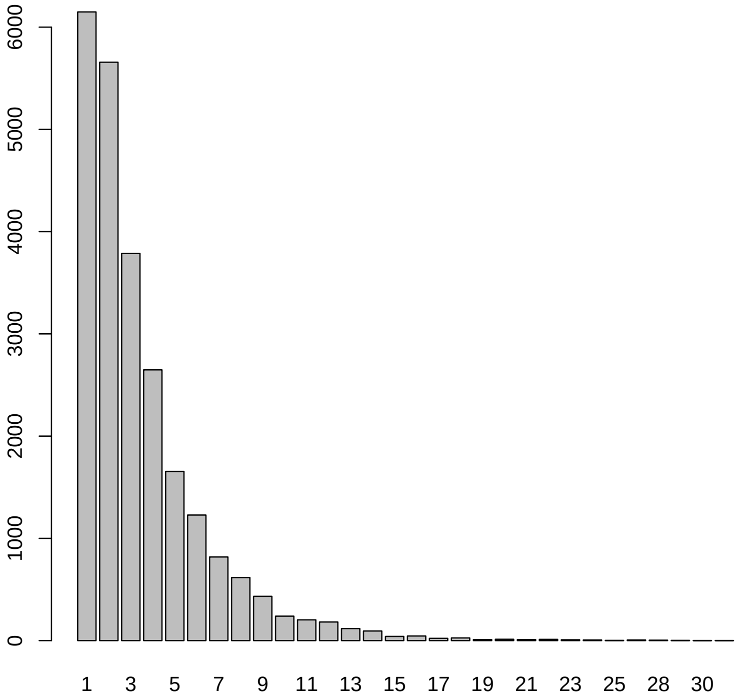
2 Homo sapiens

The following code first extracts all UniProt identifiers (allup variable) and then selects all the GO terms (allgo) for each UniProt entry. All cellular compartment (CC) terms (allcc) are then extracted and the number of different GO terms per UniProt entry is calculated (xx) and tallied as a table (txx).

```
library("Homo.sapiens")
keytypes(Homo.sapiens)
allup <- keys(Homo.sapiens, keytype = "UNIPROT")
allgo <- select(Homo.sapiens,
keys = allup, columns = "GOID",
keytype = "UNIPROT")
allcc <- allgo[allgo$ONTOLOGY == "CC", ]
xx <- tapply(allcc$GOID, allcc$UNIPROT, length)
txx <- table(xx)
```

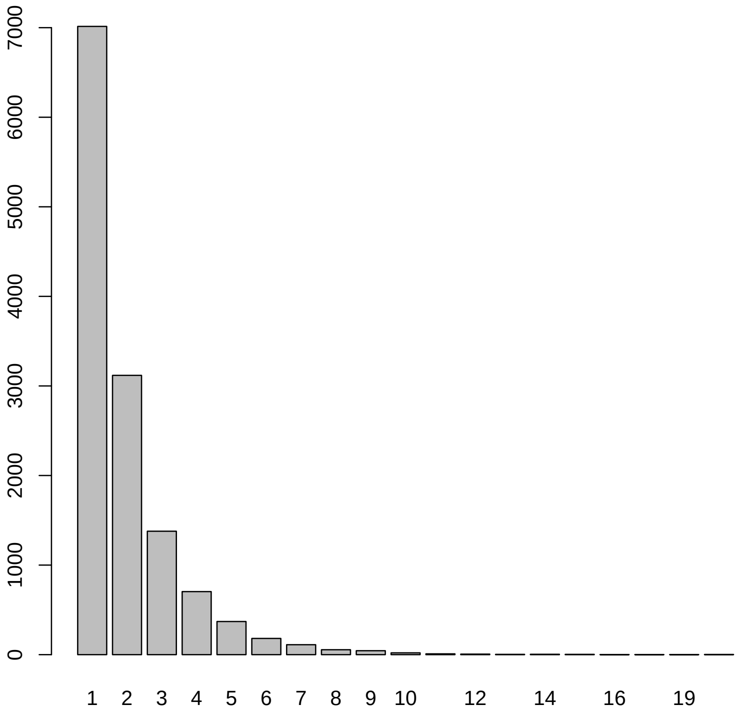
As can be seen on the resulting barplot, 6149 proteins have unique GO terms while 17888 have two and up to 36 CC annotations.

[1] http://www.bioconductor.org



Below, we limit the GO CC entries that have been inferred by direct assays, as opposed to computational methods. 7014 proteins have unique GO terms while 6011 have two and up to 22 different annotations.

```
allccida <- allcc[allgo$EVIDENCE %in%
  c("IDA", "IPI", "IMP", "IGI","IEP" ) , ]
xx2 <- tapply(allccida$GOID, allccida$UNIPROT, length)
txx2 <- table(xx2)
```


7000
6000
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8
9
10
12
14
16
19