Class 10: Structural Bioinformatics

Raquel Gonzalez (A16207442)

The PDB Database

Here we examine the size and composition of the main database of biomolecular structures - the PDB.

Get a CSV file from the PDB database and read it into R.

```
pdbstats <- read.csv("pdb_stats.csv", row.names = 1)
head(pdbstats)
```

```
                          X.ray     EM    NMR Multiple.methods Neutron Other
Protein (only)          161,663 12,592 12,337              200      74    32
Protein/Oligosaccharide   9,348  2,167     34                8       2     0
Protein/NA                8,404  3,924    286                7       0     0
Nucleic acid (only)       2,758    125  1,477               14       3     1
Other                       164      9     33                0       0     0
Oligosaccharide (only)       11      0      6                1       0     4
                          Total
Protein (only)          186,898
Protein/Oligosaccharide  11,559
Protein/NA               12,621
Nucleic acid (only)       4,378
Other                       206
Oligosaccharide (only)       22
```

> Q1: What percentage of structures in the PDB are solved by X-Ray and Electron Microscopy.

My pdbstats data frame has numbers with commas in them. This may cause us problems. Let's see:

```
pdbstats$X.ray
```

```
[1] "161,663" "9,348"   "8,404"   "2,758"   "164"     "11"
```

We need to remove the commas so the numbers are not returned as strings.

```
x <- "22,200"
as.numeric(gsub(",", "",x))
```

```
[1] 22200
```

I can turn this into a function that I can use for every column in the table.

```
commasum <- function(x) {
  sum(as.numeric(gsub(",", "",x)))
}

commasum(pdbstats$X.ray)
```

```
[1] 182348
```

Apply across all columns.

```
totals <- apply(pdbstats, 2, commasum)
```

```
round(totals/totals["Total"] * 100, 2)
```

```
           X.ray                EM               NMR Multiple.methods 
           84.54              8.72              6.57             0.11 
         Neutron             Other             Total 
            0.04              0.02            100.00 
```

84.54% of structures are solved by X-ray and 8.72% are solved by EM.

> Q2: What proportion of structures in the PDB are protein?

```
round(as.numeric(gsub(",","", pdbstats[1,7]))/totals["Total"]*100, 2)
```