Class 9: Structural Bioinformatics

Wenxi Tang

PDB statistics

```
#read the csv file
pdbstats <- read.csv("Data Export Summary.csv")
pdbstats
```

```
           Molecular.Type   X.ray    EM    NMR Multiple.methods Neutron Other
1          Protein (only) 152,914 9,495 12,121              191      72    32
2 Protein/Oligosaccharide   9,008 1,663     32                7       1     0
3              Protein/NA   8,069 2,949    282                6       0     0
4     Nucleic acid (only)   2,602    78  1,434               12       2     1
5                   Other     163     9     31                0       0     0
6  Oligosaccharide (only)      11     0      6                1       0     4
    Total
1 174,825
2  10,711
3  11,306
4   4,129
5     203
6      22
```

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

```
#use `gsub` to remove comma in the dataset and read it as numbers
xray <- as.numeric(gsub(",", "", pdbstats$X.ray))
em <- as.numeric(gsub(",", "", pdbstats$EM))
total <- as.numeric(gsub(",", "", pdbstats$Total))
```

```
#calculate the percentage
sum(xray)/sum(total)
```

```
[1] 0.8587
```

```
sum(em)/sum(total)
```

```
[1] 0.07054812
```

```
#create a function to convert characters to numbers
char2numsum <- function(x){
  sum(as.numeric(gsub(",", "", x)))
}
```

```
char2numsum(pdbstats$X.ray)/char2numsum(pdbstats$Total)
```

```
[1] 0.8587
```

```
char2numsum(pdbstats$EM)/char2numsum(pdbstats$Total)
```

```
[1] 0.07054812
```

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

```
char2numsum(pdbstats$Total[1])/char2numsum(pdbstats$Total)
```

```
[1] 0.8689288
```

> Q3: Type HIV in the PDB website search box on the home page and determine how many HIV-1 protease structures are in the current PDB?