Class 9: Structural Bioinformatics pt 1

Achyuta (PID: A16956100)

The main database for structural data is called the PDB (Protein Data Bank). Lets see what it contains.

I need to remove the comma and convert to numeric to do math:

```
stats <- read.csv("pdb_stats.csv")
as.numeric(sub(",", "", stats$Total))
```

```
[1] 186898  11559  12621   4378    206     22
```

```
stats
```

```
           Molecular.Type   X.ray     EM    NMR Multiple.methods Neutron Other
1          Protein (only) 161,663 12,592 12,337              200      74    32
2 Protein/Oligosaccharide   9,348  2,167     34                8       2     0
3              Protein/NA   8,404  3,924    286                7       0     0
4     Nucleic acid (only)   2,758    125  1,477               14       3     1
5                   Other     164      9     33                0       0     0
6  Oligosaccharide (only)      11      0      6                1       0     4
    Total
1 186,898
2  11,559
3  12,621
4   4,378
5     206
6      22
```

I could turn this into a function to fix the whole table or any future table I read like this:

```
comma2numeric <- function(x) {
  as.numeric(sub(",", "", x))
}
```

```
apply(stats, 2, comma2numeric)
```

```
Warning in FUN(newX[, i], ...): NAs introduced by coercion
```

```
     Molecular.Type  X.ray    EM   NMR Multiple.methods Neutron Other  Total
[1,]             NA 161663 12592 12337              200      74    32 186898
[2,]             NA   9348  2167    34                8       2     0  11559
[3,]             NA   8404  3924   286                7       0     0  12621
[4,]             NA   2758   125  1477               14       3     1   4378
[5,]             NA    164     9    33                0       0     0    206
[6,]             NA     11     0     6                1       0     4     22
```

```
library(readr)
pdbdb <- read_csv("pdb_stats.csv")
```

```
Rows: 6 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): Molecular Type
dbl (3): Multiple methods, Neutron, Other
num (4): X-ray, EM, NMR, Total

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sum(pdbdb$Total)
```

```
[1] 215684
```

```
sum(pdbdb$`X-ray`)/sum(pdbdb$Total) * 100
```

```
[1] 84.54406
```

```
sum(pdbdb$EM)/sum(pdbdb$Total) * 100
```

```
[1] 8.724337
```

Q1: What percentage of structures in the PDB are solved by X-Ray and Electron Microscopy. 84.5% for X-ray and 8.7% for electron microscopy.

```
sum(pdbdb$Total)
```

```
[1] 215684
```

```
pdbdb$Total[1]/sum(pdbdb$Total) * 100
```

```
[1] 86.65362
```

Q2: What proportion of structures in the PDB are protein? 86.65%%

##Mol*

PDB code: 1hsg

Figure 1: A first image from molstar

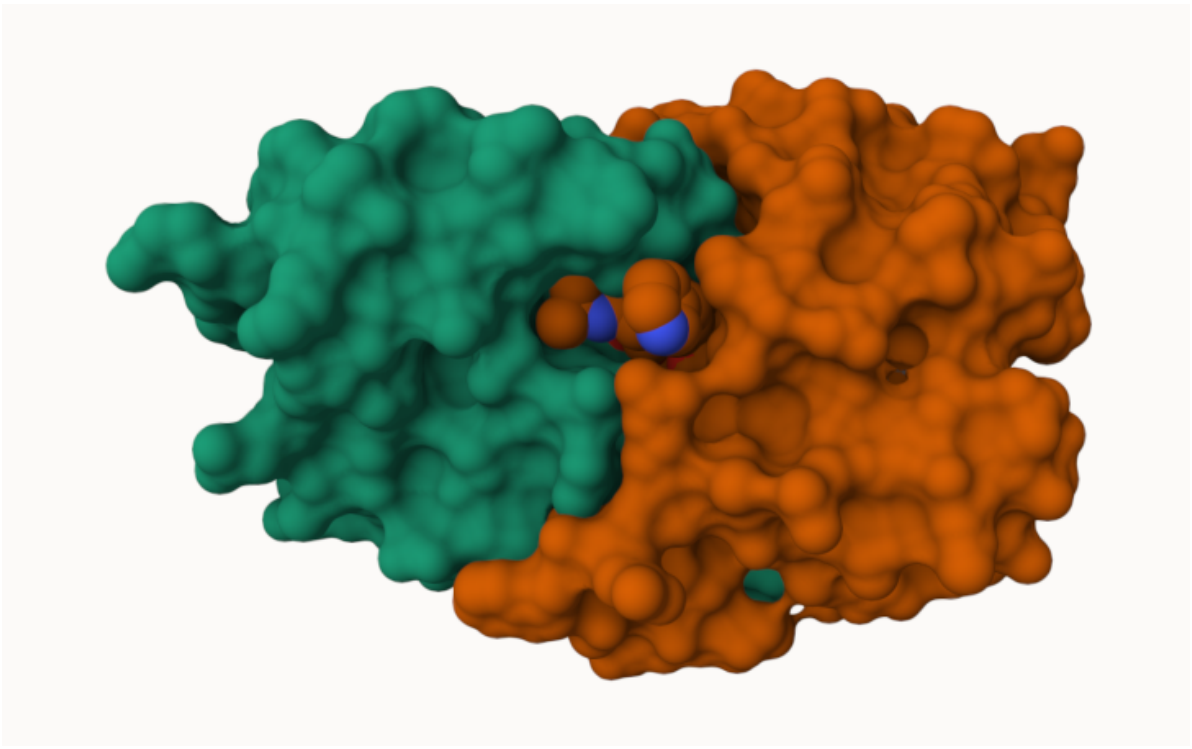

Figure 2: Another image from molstar