Exploring a Molecule

JASKIRAT SINGH BHATIA

```
knitr::opts_chunk$set(echo = TRUE,
                      warning = FALSE,
                      message = FALSE,
                      fig.align = "center",
                      fig.width = 7,
                      fig.height = 6,
                      out.width = "60%")
set.seed(12314159)
#
# Libraries you need
library(knitr)
library(loon.data)
library(loon)
#
# Directory info
imageDirectory <- "./img"
dataDirectory <- "./data"
path_concat <- function(path1, path2, sep="/") paste(path1, path2, sep = sep)
```

```
    load(path_concat(dataDirectory, "SCmolecule.Rda"))
```

```
SCitemLabels <- with(SCmolecule,
                     paste0("ID: ", id, "\n",
                            "Type: ", type, "\n",
                            "Residue Type: ", residueType, "\n",
                            "Residue: ", residue, "\n",
                            "Chain: ", chain
                            ))
```

```
head(SCmolecule)
```

```
##   group id label residue chain sequence      x      y      z displacement
## 1  ATOM  1   O5'      DC     D        1 23.081 73.401 36.511        44.77
## 2  ATOM  2   C5'      DC     D        1 24.340 73.259 35.792        46.46
## 3  ATOM  3   C4'      DC     D        1 24.267 72.789 34.262        42.04
## 4  ATOM  4   O4'      DC     D        1 25.550 72.957 33.595        41.08
## 5  ATOM  5   C3'      DC     D        1 23.957 71.289 34.142        38.19
## 6  ATOM  6   O3'      DC     D        1 23.249 71.081 32.947        33.45
##   type    mass         residueType
## 1    O 15.9994 Deoxyribonucleotide
## 2    C 12.0107 Deoxyribonucleotide
## 3    C 12.0107 Deoxyribonucleotide
## 4    O 15.9994 Deoxyribonucleotide
## 5    C 12.0107 Deoxyribonucleotide
## 6    O 15.9994 Deoxyribonucleotide
```

Basic Overview

- Units -> atom
- Variates -> group, id, label, residue, chain, sequence, x, y, z, displacement, type, mass, residueType

```
summary(SCmolecule)
```

```
##     group            id             label         residue    chain
##  ATOM  :1707   Min.   :   1.0   O      : 165   DG     :308   A:483
##  HETATM:  55   1st Qu.: 442.2   C      : 114   DC     :260   B:482
##                Median : 883.5   CA     : 114   LYS    :180   D:400
##                Mean   : 883.2   CB     : 114   ARG    :132   E:397
##                3rd Qu.:1324.8   N      : 114   LEU    :112
##                Max.   :1766.0   CG     :  78   GLU    :108
##                                 (Other):1063   (Other):662
##     sequence           x                y               z
##  Min.   :  1.0   Min.   :-2.206   Min.   : 7.05   Min.   :-1.217
##  1st Qu.: 16.0   1st Qu.:23.052   1st Qu.:25.62   1st Qu.:22.285
##  Median : 28.0   Median :29.002   Median :41.40   Median :28.593
##  Mean   : 38.4   Mean   :27.194   Mean   :41.79   Mean   :28.433
##  3rd Qu.: 41.0   3rd Qu.:33.491   3rd Qu.:57.97   3rd Qu.:34.832
##  Max.   :432.0   Max.   :43.954   Max.   :75.39   Max.   :57.128
##
##   displacement   type          mass                     residueType
##  Min.   : 2.00   C :942   Min.   : 12.01   AminoAcid          :934
##  1st Qu.:24.34   CD:  4   1st Qu.: 12.01   CadmiumIon         :  4
##  Median :31.62   N :329   Median : 12.01   Deoxyribonucleotide:773
##  Mean   :32.31   O :439   Mean   : 14.13   Water              : 51
##  3rd Qu.:38.84   P : 36   3rd Qu.: 16.00
##  Max.   :74.67   S : 12   Max.   :112.41
##
```

Looking at the Summary

- There are 2 major groups of atoms
  - ATOM
  - HEATOM
- 4 Different Chains
  - A
  - B
  - D
  - E
- 4 Different Types of Residue
  - Amino Acid
  - Cadmium Ion
  - Water
  - Deoxyribonucleotide
  - Out of these, Amino Acids is present the most and Cadmium Ion the Least
- 662 Types of Residue

- 1063 types of Labels
- The average mass of molecules is 14.13
  - But there is one type molecule with a very high mass

Exploring the Type of Atoms - ATOM and HEATOM

```
atom <- SCmolecule[which(SCmolecule$group == "ATOM"),]
heatom <- SCmolecule[which(SCmolecule$group != "ATOM"),]

summary(atom)
```

```
##     group             id              label         residue    chain
##  ATOM  :1707   Min.   :   1.0   C       : 114   DG      :308   A:467
##  HETATM:   0   1st Qu.: 428.5   CA      : 114   DC      :260   B:467
##                Median : 856.0   CB      : 114   LYS     :180   D:387
##                Mean   : 855.6   N       : 114   ARG     :132   E:386
##                3rd Qu.:1283.5   O       : 114   LEU     :112
##                Max.   :1710.0   CG      :  78   GLU     :108
##                                 (Other):1059   (Other):607
##     sequence          x               y               z
##  Min.   : 1.0   Min.   :-2.206   Min.   : 7.05   Min.   :-1.217
##  1st Qu.:15.0   1st Qu.:23.037   1st Qu.:25.71   1st Qu.:22.426
##  Median :27.0   Median :28.997   Median :41.83   Median :28.629
##  Mean   :28.8   Mean   :27.170   Mean   :41.89   Mean   :28.495
##  3rd Qu.:39.0   3rd Qu.:33.463   3rd Qu.:57.97   3rd Qu.:34.791
##  Max.   :64.0   Max.   :42.570   Max.   :75.39   Max.   :57.128
##
##   displacement   type          mass                     residueType
##  Min.   : 2.00   C :942   Min.   :12.01   AminoAcid           :934
##  1st Qu.:24.56   CD:  0   1st Qu.:12.01   CadmiumIon          :  0
##  Median :31.56   N :329   Median :12.01   Deoxyribonucleotide:773
##  Mean   :32.32   O :388   Mean   :13.84   Water               :  0
##  3rd Qu.:38.71   P : 36   3rd Qu.:16.00
##  Max.   :74.67   S : 12   Max.   :32.06
##
```

```
summary(heatom)
```

```
##     group          id           label       residue   chain
##  ATOM  : 0   Min.   :1712   O       :51   HOH     :51   A:16
##  HETATM:55   1st Qu.:1726   CD      : 4   CD      : 4   B:15
##              Median :1739   C       : 0   ALA     : 0   D:13
##              Mean   :1739   C1'     : 0   ARG     : 0   E:11
##              3rd Qu.:1752   C2      : 0   ASN     : 0
##              Max.   :1766   C2'     : 0   ASP     : 0
##                             (Other): 0   (Other): 0
##     sequence           x               y               z
##  Min.   : 67.0   Min.   : 2.158   Min.   : 8.833   Min.   : 0.267
##  1st Qu.:318.0   1st Qu.:24.024   1st Qu.:19.675   1st Qu.:19.628
##  Median :344.0   Median :29.390   Median :37.648   Median :25.427
```

```
##  Mean   :336.3   Mean   :27.932   Mean   :38.517   Mean   :26.487
##  3rd Qu.:384.5   3rd Qu.:33.755   3rd Qu.:57.801   3rd Qu.:37.856
##  Max.   :432.0   Max.   :43.954   Max.   :71.338   Max.   :56.105
##
##   displacement   type         mass                     residueType
##  Min.   : 5.77   C : 0   Min.   : 16.00   AminoAcid          : 0
##  1st Qu.:19.79   CD: 4   1st Qu.: 16.00   CadmiumIon         : 4
##  Median :33.02   N : 0   Median : 16.00   Deoxyribonucleotide: 0
##  Mean   :31.88   O :51   Mean   : 23.01   Water              :51
##  3rd Qu.:44.52   P : 0   3rd Qu.: 16.00
##  Max.   :60.85   S : 0   Max.   :112.41
##
```

Group ATOM

- There are 1707 total Atoms
- Contains all Residue
- Contains all labels
- Are present in each chain
- Do not contain type CD molecule
- Contains 2 Types of Residue
  - Amino Acids
  - Deoxyribonucleotide

Group HEATOM

- There are total of 55 Atoms
- Contains only 2 labels
  - 0
  - CD
- Contains only 2 residue
  - HOH
  - CD
- Present in All Chains
- Contains onlt 2 Types
  - C
  - O
- Contains 2 Types of residue
  - Water
  - Cadmium Ion

Exploring the Mass and Type of Molecules

```r
type.o <- as.data.frame(table(SCmolecule[which(SCmolecule$type == "O"),]$residueType))

type.c <- as.data.frame(table(SCmolecule[which(SCmolecule$type == "C"),]$residueType))

type.n <- as.data.frame(table(SCmolecule[which(SCmolecule$type == "N"),]$residueType))

type.p <- as.data.frame(table(SCmolecule[which(SCmolecule$type == "P"),]$residueType))

type.s <- as.data.frame(table(SCmolecule[which(SCmolecule$type == "S"),]$residueType))

type.cd <- as.data.frame(table(SCmolecule[which(SCmolecule$type == "CD"),]$residueType))


colnames(type.o) <- c("Residue Type", "count")
colnames(type.c) <- c("Residue Type", "count")
colnames(type.n) <- c("Residue Type", "count")
colnames(type.p) <- c("Residue Type", "count")
colnames(type.s) <- c("Residue Type", "count")
colnames(type.cd) <- c("Residue Type", "count")

print("Type O has mass 15.994 containing ")
```

```
## [1] "Type O has mass 15.994 containing "
```

```r
kable(type.o)
```

Residue Type	count
AminoAcid	164
CadmiumIon	0
Deoxyribonucleotide	224
Water	51

```r
print("Type C has mass 12.0107 containing")
```

```
## [1] "Type C has mass 12.0107 containing"
```

```r
kable(type.c)
```

Residue Type	count
AminoAcid	576
CadmiumIon	0
Deoxyribonucleotide	366
Water	0

```r
print("Type N has mass 14.0067 containing")
```

```
## [1] "Type N has mass 14.0067 containing"
```

```
kable(type.n)
```

Residue Type	count
AminoAcid	182
CadmiumIon	0
Deoxyribonucleotide	147
Water	0

```
print("Type N has mass 14.0067 containing")
```

```
## [1] "Type N has mass 14.0067 containing"
```

```
kable(type.p)
```

Residue Type	count
AminoAcid	0
CadmiumIon	0
Deoxyribonucleotide	36
Water	0

```
print("Type S has mass 32.065 containing")
```

```
## [1] "Type S has mass 32.065 containing"
```

```
kable(type.s)
```

Residue Type	count
AminoAcid	12
CadmiumIon	0
Deoxyribonucleotide	0
Water	0

```
print("Type CD has mass 112.411 containing")
```

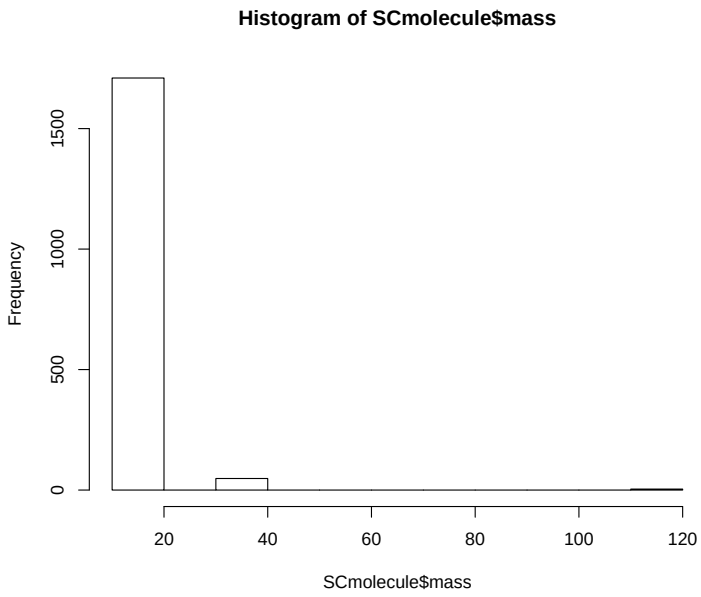
```
## [1] "Type CD has mass 112.411 containing"
```

```
kable(type.cd)
```

Residue Type	count
AminoAcid	0
CadmiumIon	4
Deoxyribonucleotide	0
Water	0

Histogram of Mass

```
hist(SCmolecule$mass)
```




```
print("Unique Masses")
```

```
## [1] "Unique Masses"
```

- From Histogram we can see that
  - Majority of the atoms have a mass of 15.994

Exploring Chains

```
type.a <- as.data.frame(table(SCmolecule[which(SCmolecule$chain == "A"),]$residueType))

type.b <- as.data.frame(table(SCmolecule[which(SCmolecule$chain == "B"),]$residueType))

type.d <- as.data.frame(table(SCmolecule[which(SCmolecule$chain == "D"),]$residueType))

type.e <- as.data.frame(table(SCmolecule[which(SCmolecule$chain == "E"),]$residueType))



colnames(type.a) <- c("Residue Type", "count")
colnames(type.b) <- c("Residue Type", "count")
colnames(type.d) <- c("Residue Type", "count")
colnames(type.e) <- c("Residue Type", "count")
```

```
print("Chain A")
```

```
## [1] "Chain A"
```

```
kable(type.a)
```

Residue Type	count
AminoAcid	467
CadmiumIon	2
Deoxyribonucleotide	0
Water	14

```
print("Chain B")
```

```
## [1] "Chain B"
```

```
kable(type.b)
```

Residue Type	count
AminoAcid	467
CadmiumIon	2
Deoxyribonucleotide	0
Water	13

```
print("Chain D")
```

```
## [1] "Chain D"
```

```
kable(type.d)
```

Residue Type	count
AminoAcid	0
CadmiumIon	0
Deoxyribonucleotide	387
Water	13

```
print("Chain E")
```

```
## [1] "Chain E"
```

```
kable(type.e)
```

Residue Type	count
AminoAcid	0
CadmiumIon	0
Deoxyribonucleotide	386
Water	11

Exploring Further with plots

```
library(loon)
# histogram
hist <- l_hist(SCmolecule$x, linkingGroup = "sc")

# Scatter plot
scatter <- l_plot(SCmolecule, linkingGroup = "sc")
```

```
# Coloring each molecule by its Chain ID
scatter["color"] <- SCmolecule$chain
```

```
# Giving a size to a molecule according to its mass
sizeByMagnitude <- (SCmolecule$mass)
sizeByMagnitude <- 2 + sizeByMagnitude - min(sizeByMagnitude)
scatter["size"] <- sizeByMagnitude
```

```
# Creating a 3-D Plot of the molecule
ng <- l_navgraph(SCmolecule[,c("x", "y", "z")],
linkingGroup = "sc", sync = "pull",
glyph = "circle")
```

```
# Showing Item Labels
scatter["itemLabel"] <- SCitemLabels
scatter["showItemLabels"] <- TRUE
```

- Chain A and B form a scisor like shape with 2 blobs in between

\begin{center}\includegraphics[width=60%]{conc_of_blobs} \end{center}

- The big blobs in the Center of the chains are 2 Cadmium Ions(2 in each chain). Which are also the largest ions present.
- Water Molecules appear to be scattered everywhere and are present in each chain
- Chain D and E form a DNA like Spiral

Figure 1: Blobs

Figure 2: Chain D and E