Class 10: Structural Bioinformatics

Chelsea A16871799

#PDB database Let's first see what is in PDB database– the main repository of protein structure

Download composition sats from: http://www.rcsb.org/stats/summary

For context: Unitprot Contain 251600,768. The PDB only contains 183,201

```
stats <- read.csv("Data Export Summary.csv",row.names=1)
stats
```

```
                         X.ray     EM    NMR Multiple.methods Neutron Other
Protein (only)         158,844 11,759 12,296              197      73    32
Protein/Oligosaccharide  9,260  2,054     34                8       1     0
Protein/NA               8,307  3,667    284                7       0     0
Nucleic acid (only)      2,730    113  1,467               13       3     1
Other                      164      9     32                0       0     0
Oligosaccharide (only)      11      0      6                1       0     4
                         Total
Protein (only)         183,201
Protein/Oligosaccharide 11,357
Protein/NA              12,265
Nucleic acid (only)      4,327
Other                      205
Oligosaccharide (only)      22
```

The R recognize 158,844 as character as it contains commas, we need to fix this.

```
x <- stats$X.ray
x
```

```
[1] "158,844" "9,260"   "8,307"   "2,730"   "164"     "11"
```

```
  as.numeric(gsub(",", "",x))
```

```
[1] 158844   9260   8307   2730    164     11
```

```
  rm.comma <- function(x){
    as.numeric(gsub(",","",x))
  }
```

```
  rm.comma(stats$EM)
```

```
[1] 11759  2054  3667   113     9     0
```

#I can use appoly() to fix the whole table

```
  pdbstats <- apply(stats,2, rm.comma)
  rownames(pdbstats) <- rownames(stats)
  head(pdbstats)
```

```
                         X.ray    EM   NMR Multiple.methods Neutron Other
Protein (only)          158844 11759 12296               197      73    32
Protein/Oligosaccharide   9260  2054    34                 8       1     0
Protein/NA                8307  3667   284                 7       0     0
Nucleic acid (only)       2730   113  1467                13       3     1
Other                      164     9    32                 0       0     0
Oligosaccharide (only)      11     0     6                 1       0     4
                         Total
Protein (only)          183201
Protein/Oligosaccharide  11357
Protein/NA               12265
Nucleic acid (only)       4327
Other                      205
Oligosaccharide (only)      22
```

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

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

```
     X.ray         EM        NMR Multiple.methods
    179316      17602      14119              226
   Neutron      Other      Total
        77         37     211377
```

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

```
     X.ray         EM        NMR Multiple.methods
     84.83       8.33       6.68             0.11
   Neutron      Other      Total
      0.04       0.02     100.00
```

84.83% is X.ray while 8.33% is EM. Q2: What proportion of structures in the PDB are protein?

```
round(pdbstats[1,"Total"]/sum(pdbstats[,"Total"])*100,2)
```

```
[1] 86.67
```

<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? SKIPPED Q4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure? #There is a 2 Angstrom structure and hydrogen is not visible at all times. water molecules are too tiny to visualize.

Q5: There is a critical "conserved" water molecule in the binding site. Can you identify this water molecule? What residue number does this water molecule have? #The water molecule HOH 308 near Mk1

Q6: Here is a lovely figure of HIP-Pr with the catalytic residues, Mk1 compound and all important water 308.

Q7: [Optional] As you have hopefully observed HIV protease is a homodimer (i.e. it is composed of two identical chains). With the aid of the graphic display can you identify secondary structure elements that are likely to only form in the dimer rather than the monomer?

The bio3d package for structural bioinformatics

```
library(bio3d)
pdb <- read.pdb("1hsg")
```

```
  Note: Accessing on-line PDB file
```

```
pdb
```

```
Call:  read.pdb(file = "1hsg")
```

```
   Total Models#: 1
     Total Atoms#: 1686,  XYZs#: 5058  Chains#: 2  (values: A B)

     Protein Atoms#: 1514  (residues/Calpha atoms#: 198)
     Nucleic acid Atoms#: 0  (residues/phosphate atoms#: 0)

     Non-protein/nucleic Atoms#: 172  (residues: 128)
     Non-protein/nucleic resid values: [ HOH (127), MK1 (1) ]

   Protein sequence:
      PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYD
      QILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFPQITLWQRPLVTIKIGGQLKE
      ALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTVLVGPTP
      VNIIGRNLLTQIGCTLNF

+ attr: atom, xyz, seqres, helix, sheet,
        calpha, remark, call
```

#predicting functional motions of a single structure

Let's finish toady with a bioinformatics calculators predict the functional motion of a PDB structure.

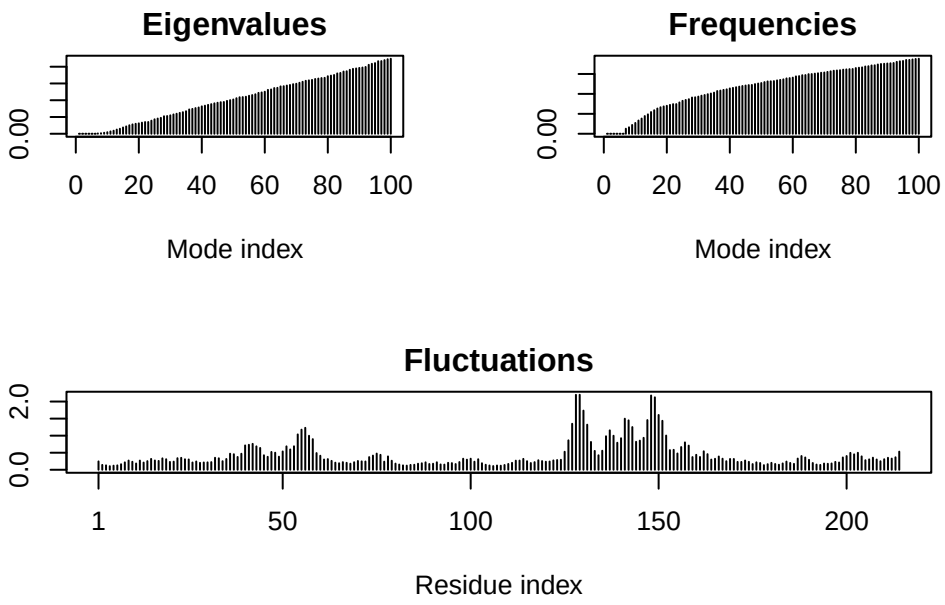
```
adk <- read.pdb("6s36")
```

```
  Note: Accessing on-line PDB file
   PDB has ALT records, taking A only, rm.alt=TRUE
```

```
m <- nma(adk)
```

```
 Building Hessian...        Done in 0.016 seconds.
 Diagonalizing Hessian...   Done in 0.278 seconds.
```

```
plot(m)
```




```
mktrj(m, file="adk_m7.pdb")
```