lab09_hw

Norman Lee A18086849

1: Introduction to the RCSB Protein Data Bank (PDB) The PDB archive is the major repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. Understanding the shape of these molecules helps to understand how they work. This knowledge can be used to help deduce a structure's role in human health and disease, and in drug development. The structures in the PDB range from tiny proteins and bits of DNA or RNA to complex molecular machines like the ribosome composed of many chains of protein and RNA.

Download a CSV file from the PDB site (accessible from "Analyze" > "PDB Statistics" > "by Experimental Method and Molecular Type". Move this CSV file into your RStudio project and use it to answer the following questions:

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

sum of x ray+ EM total/ sum of all values=(199931+29978)/244730(100%)= 93.7%

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

(210566+13748+15300)/244730(100%)= 98%

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?

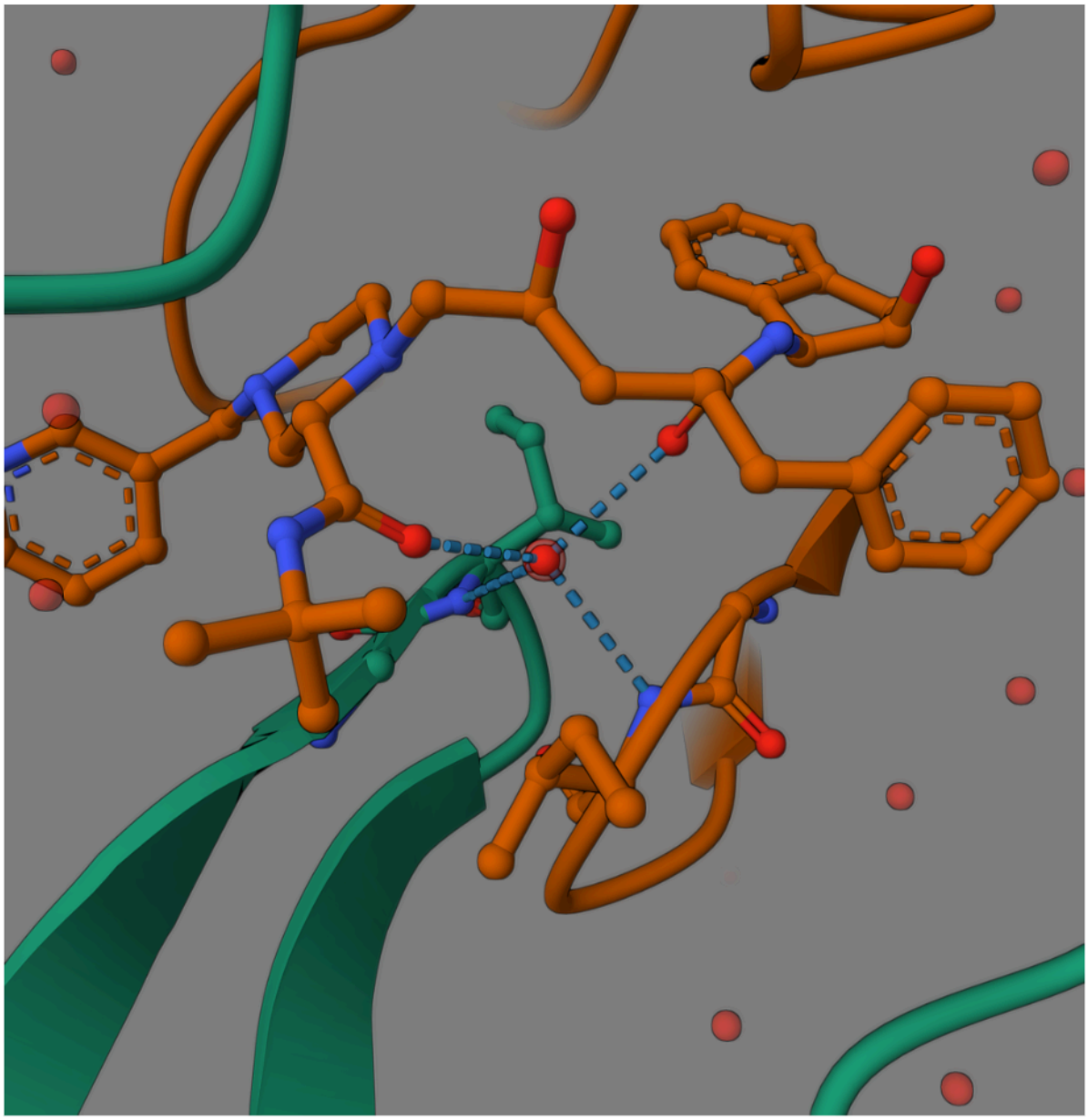
4866

2. Visualizing the HIV-1 protease structure In this section we will use the 2Å resolution X-ray crystal structure of HIV-1 protease with a bound drug molecule indinavir (PDB ID: 1HSG). We will use the Mol* molecular viewer to visually inspect the protein, the binding site and the drug molecule. After exploring features of the complex we will move on to perform bioinformatics analysis of single and multiple crystallographic stuctures to explore the conformational dynamics and flexibility of the protein - important for it's function and for considering during drug design.

Q4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure? **Each water molecule is represented by a single oxygen atom because hydrogens are invisible in X-ray structures**

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 **HOH 301**

Q6: Generate and save a figure clearly showing the two distinct chains of HIV-protease along with the ligand. You might also consider showing the catalytic residues ASP 25 in each chain and the critical water (we recommend "Ball & Stick" for these side-chains). Add this figure to your Quarto document.



3. Introduction to Bio3D in R Bio3D is an R package for structural bioinformatics. Features include the ability to read, write and analyze biomolecular structure, sequence and dynamic trajectory data.

In your existing Rmarkdown document load the Bio3D package by typing in a new code chunk:

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

> Q7: How many amino acid residues are there in this pdb object? **198** Q8: Name one of the two non-protein residues? **mk1** Q9: How many protein chains are in this structure? **2**

Quick PDB visualization in R We can use the Bio3D partner package, bio3dview, to generate quick interactive molecular visualizations. To install the development version of bio3dview from GitHub, along with the related NGLVieweR package use:

```
# install.packages("pak")
#pak::pak("bioboot/bio3dview")
#install.packages("NGLVieweR")
```

```
#library(bio3dview)
#library(NGLVieweR)

#view.pdb(pdb) |>
  #setSpin()
```

4. Comparative structure analysis of Adenylate Kinase The goal of this section is to perform principal component analysis (PCA) on the complete collection of Adenylate kinase structures in the protein data-bank (PDB). Overview: Starting from only one Adk PDB identifier (PDB ID: 1AKE) we will search the entire PDB for related structures using BLAST, fetch, align and superpose the identified structures, perform PCA and finally calculate the normal modes of each individual structure in order to probe for potential differences in structural flexibility.

set up

```
#if (!requireNamespace("BiocManager", quietly = TRUE))
    #install.packages("BiocManager", repos = "https://cran.R-project.org")
library(BiocManager)
```

```
Warning: package 'BiocManager' was built under R version 4.5.2
```

> >Q10. Which of the packages above is found only on BioConductor and not CRAN? **msa**
>
> >Q11. Which of the above packages is not found on BioConductor or CRAN?: **bioboot/bio3dview**

>Q12. True or False? Functions from the pak package can be used to install packages from GitHub and BitBucket? **TRUE**

Search and retrieve ADK structures

```
library(bio3d)
aa <- get.seq("1ake_A")
```

```
Warning in get.seq("1ake_A"): Removing existing file: seqs.fasta

Fetching... Please wait. Done.
```

```
aa
```

```
              1        .         .         .         .         .         60
pdb|1AKE|A    MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLVT
              1        .         .         .         .         .         60

             61        .         .         .         .         .         120
pdb|1AKE|A    DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDRI
             61        .         .         .         .         .         120

            121        .         .         .         .         .         180
pdb|1AKE|A    VGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPLIG
            121        .         .         .         .         .         180

            181        .         .         .   214
pdb|1AKE|A    YYSKEAEAGNTKYAKVDGTKPVAEVRADLEKILG
            181        .         .         .   214

Call:
  read.fasta(file = outfile)

Class:
  fasta

Alignment dimensions:
  1 sequence rows; 214 position columns (214 non-gap, 0 gap)

+ attr: id, ali, call
```

>Q13. How many amino acids are in this sequence, i.e. how long is this sequence? **214**

>Q14. What do you note about this plot? Are the black and colored lines similar or different? Where do you think they differ most and why? **black lines have way lower peaks than the colored lines but the red and green lines follow nearly identical peak patterns, suggesting they share similar flexibility profiles. All structures show the same general shape, but they differ most around residues 120–170**