class09

Vince (PID: A15422556)

2/15/2022

## The PDB Database

The PDB is the main repository for 3d structure data of biomolecules. Here we explore its composition.

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

## X.ray NMR EM Multiple.methods Neutron Other Total  
## Protein (only) 144301 11877 6676 182 70 32 163138  
## Protein/Oligosaccharide 8528 31 1116 5 0 0 9680  
## Protein/NA 7617 274 2153 3 0 0 10047  
## Nucleic acid (only) 2393 1398 61 8 2 1 3863  
## Other 150 31 3 0 0 0 184  
## Oligosaccharide (only) 11 6 0 1 0 4 22

Q1: What percentage of structures in the PDB are solved by X-Ray and Electron Microscopy. 87.197% of structures are solved by X-ray and 5.354% of structures are solved by EM.

tot.method <- colSums(pdbData)  
round(tot.method/tot.method["Total"] \* 100, 3)

## X.ray NMR EM Multiple.methods   
## 87.197 7.284 5.354 0.106   
## Neutron Other Total   
## 0.039 0.020 100.000

Q2: What proportion of structures in the PDB are protein? 87.27 proportion of structures in the PDB are protein.

ans <- pdbData$Total[1] / sum(pdbData$Total) \* 100  
round(ans, 3)

## [1] 87.27

The answer to this question is 87.27 % of total structures.

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? 4483 structures.

## Visualizing the HIV-1 Protease Structure

Q4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure? The hydrogen atom is very small in comparison to the rest of the molecule.

Q5: There is a conserved water molecule in the binding site. Can you identify this water molecule? What residue number does this water molecule have (see note below)? Residue number 308

VMD generated image of HIV-protease, PDB code: 1hsg



## Sequence Viewer Extension

Q6: 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 and the sequence viewer extension can you identify secondary structure elements that are likely to only form in the dimer rather than the monomer? Beta-pleated sheets and alpha helices are likely to form in the dimer.

## Bio3D in R

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

Q7: How many amino acid residues are there in this pdb object? 198 residues

Q8: Name one of the two non-protein residues? MK1

Q9: How many protein chains are in this structure? 2 protein chains

## Comparative structure analysis of Adenylate Kinase

Extract the sequence for ADK

aa <- get.seq("1ake\_A")

## Warning in get.seq("1ake\_A"): Removing existing file: seqs.fasta

## Fetching... Please wait. Done.

aa

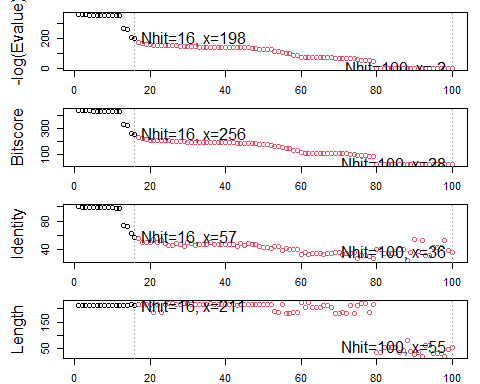
## 1 . . . . . 60   
## pdb|1AKE|A MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLVT  
## 1 . . . . . 60   
##   
## 61 . . . . . 120   
## pdb|1AKE|A DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDRI  
## 61 . . . . . 120   
##   
## 121 . . . . . 180   
## pdb|1AKE|A VGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPLIG  
## 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

blast <- blast.pdb(aa)

## Searching ... please wait (updates every 5 seconds) RID = 0SVEC8J3013   
## .................................  
## Reporting 100 hits

hits <- plot(blast)

## \* Possible cutoff values: 197 -3   
## Yielding Nhits: 16 100   
##   
## \* Chosen cutoff value of: 197   
## Yielding Nhits: 16



hits$pdb.id

## [1] "1AKE\_A" "4X8M\_A" "6S36\_A" "6RZE\_A" "4X8H\_A" "3HPR\_A" "1E4V\_A" "5EJE\_A"  
## [9] "1E4Y\_A" "3X2S\_A" "6HAP\_A" "6HAM\_A" "4K46\_A" "4NP6\_A" "3GMT\_A" "4PZL\_A"

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? bio3d-view

Q12. True or False? Functions from the devtools package can be used to install packages from GitHub and BitBucket? True

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

## Normal Mode Analysis (NMA)

pdb <- read.pdb("1ake")

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

chainA <- trim.pdb(pdb, chain="A")  
modes <- nma(chainA)

## Building Hessian... Done in 0.04 seconds.  
## Diagonalizing Hessian... Done in 0.39 seconds.

m7 <- mktrj.nma(modes, mode=7, file="mode\_7.pdb")

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? There are regions that are similar as well as regions that are dissimilar. The black and colored lines are sometimes so similar that they overlap, but other times they vary as the colored lines spike upwards compared to the black lines that stay lower on the graph. I think they differ most at regions that bind nucleotides because of the flexibility that is required for this process to take place.

## Find-A-Gene Project predicted structure using AlphaFold.

