# Biomolecular Databases and Formats

# Macro vs. Small Molecules

- Macromolecules: contain the structure of biological macromolecules. The primary source is the Protein Data Bank (<a href="https://www.rcsb.org">https://www.rcsb.org</a>).
- Small molecules: databanks containing formulas, structures and other information relative to small molecules\*. Some of the most important are:
  - PubChem (NCBI) : <a href="https://pubchem.ncbi.nlm.nih.gov/">https://pubchem.ncbi.nlm.nih.gov/</a>
  - ZINC (purchasable compounds): <a href="https://zinc.docking.org/">https://zinc.docking.org/</a>
  - Drugbank (pharma oriented): <a href="https://www.drugbank.ca/">https://www.drugbank.ca/</a>
  - CCDC (crystallographic structures): <a href="https://www.ccdc.cam.ac.uk/">https://www.ccdc.cam.ac.uk/</a>

<sup>\*</sup> The definition of small molecule may vary (~1000-atom limit)

# Macromolecules

# Macromolecular Databanks

 Primary databanks: contain the raw information, usually with a set of tools that can be accessed througha portal. Example: Protein Data Bank (<a href="https://www.rcsb.org">https://www.rcsb.org</a>).

• **Secondary Databanks**: specialized views, collections or filters on the primary databanks. Example: The PDBind database (<a href="http://www.pdbbind.org.cn">http://www.pdbbind.org.cn</a>)

- The delopment of molecular structure determination techniques lead to the acumulation of a large body of structures of proteins and nucleic acids (~160000)
- For the most part, those structures were solved using two structure determination methods, X-ray crystallography and nuclear magnetic resonance (NMR)
- The body of structures is stored in the public accessible Protein Databank (PDB) <a href="http://www.rcsb.org">http://www.rcsb.org</a>



#### The Protein Data Bank

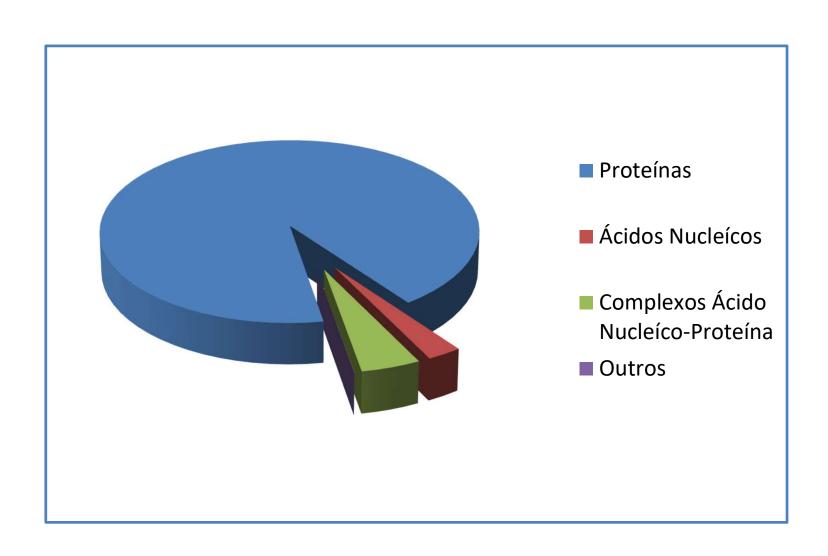
- O Protein Data Bank (PDB) foi criado em 1971 por E.Meyer e W.Hamilton, do Brookhaven National Laboratory (USA), contendo no início 7 estruturas!
- A gestão do PDB foi transferida em 1998 para os membros do RSCB (Research Collaboratory in Structural Bioinformatics) dos quais a Universidade de Rutgers é o site principal. O PDB (http://www.rcsb.org) é um banco de dados de acesso livre.
- · Contendo inicialmente estruturas de proteínas, o PDB contem hoje em dia outros tipos de moléculas, tais como ácidos nucleicos, lípidos e polissacáridos.

Número total de estruturas em 12/12/2019: 158,180

Técnica experimental	Proteínas	Ácidos nucleicos	Complexos Ac.Nuc,/Proteína	Outros	Total
Cristalografia de raios X	132004	2073	6787	8	140872
NMR	11248	1306	262	8	12824
Microscopia electrónica	2974	33	1021	0	4028
Outras	281	4	6	13	304
Combinação	144	5	2	1	152
Total	146651	3421	8078	30	158180

Dados de 12/12/2019 em <a href="http://www.rcsb.org">http://www.rcsb.org</a>

# The Protein Databank contains different types of biological macromolecules

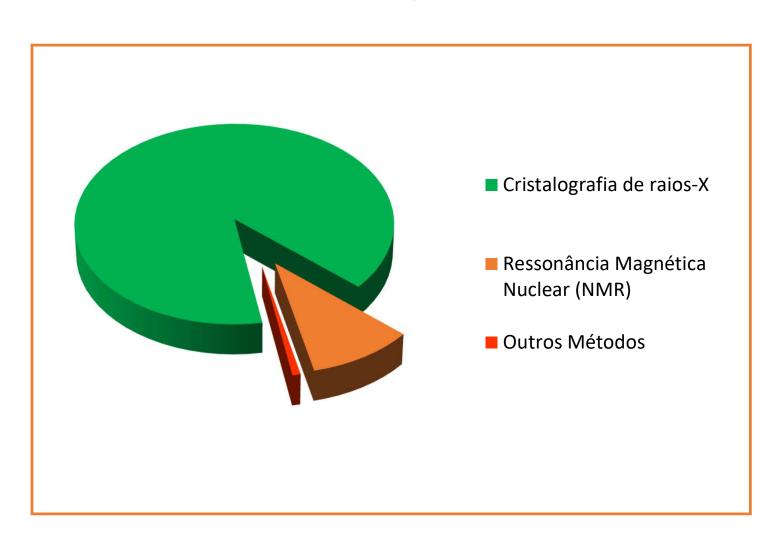


# Where does all the structural information come from?

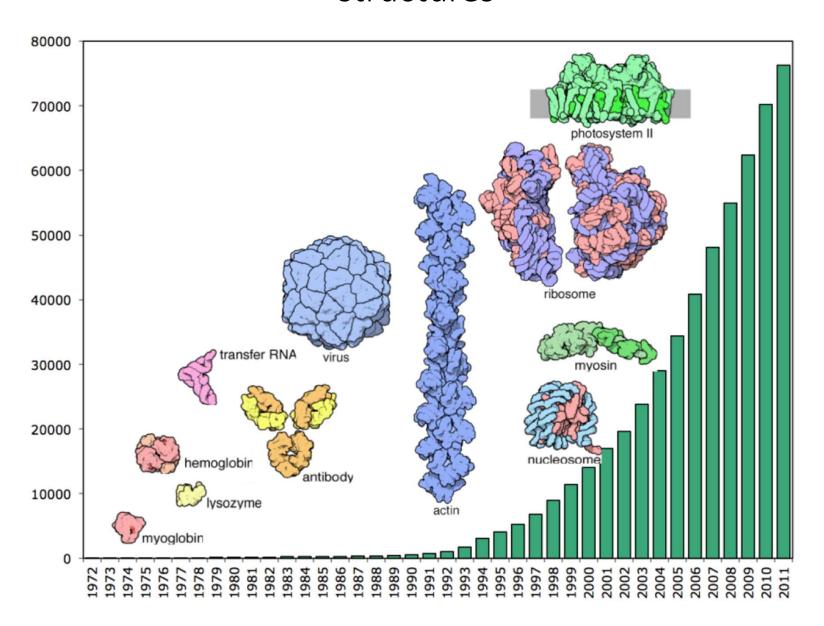
#### It's a combination of various types of data:

- Molecular geometry and bond theory
- Small molecule geometry
- Experimental Methods of Macromolecular structure determination
  - X-ray crystallography
  - ❖ Nuclear Magnetic Resonance (NMR)
  - Other methods (Cryo-EM, neutron diffraction, etc)

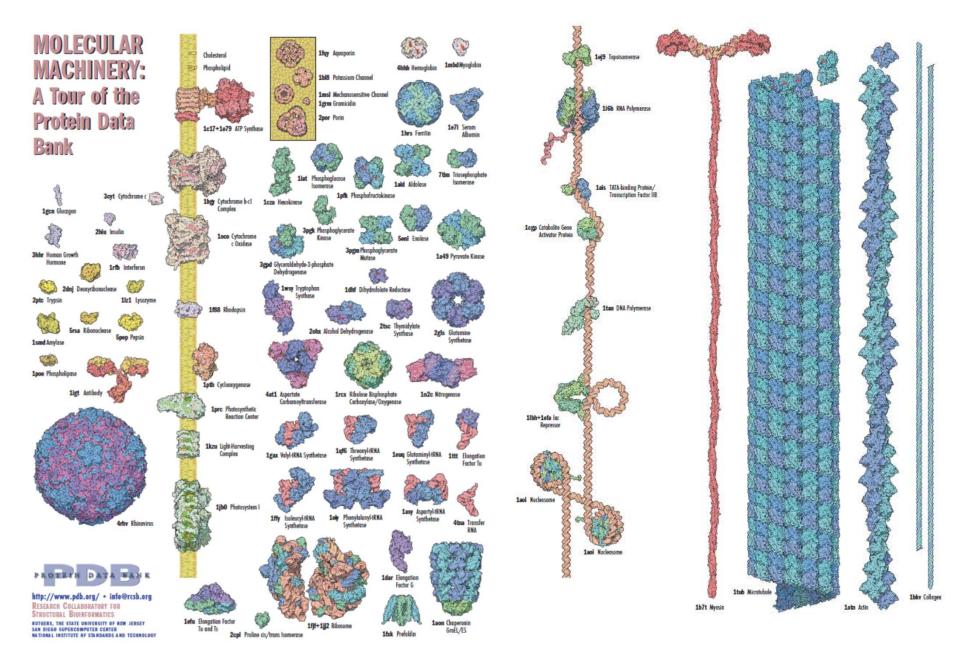
# Most structures have been solved by X-ray cristallography



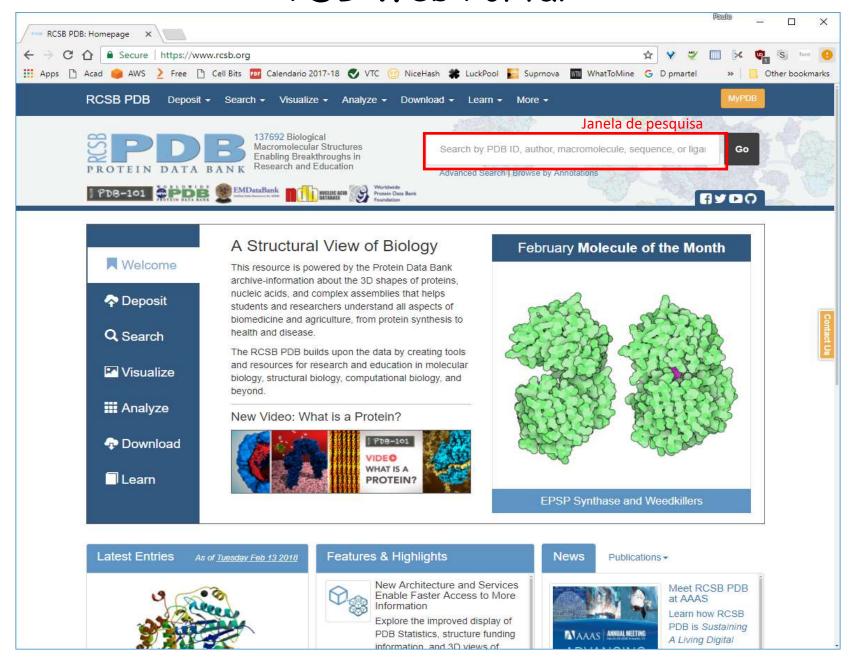
# Progress in the determination of macromolecular structures

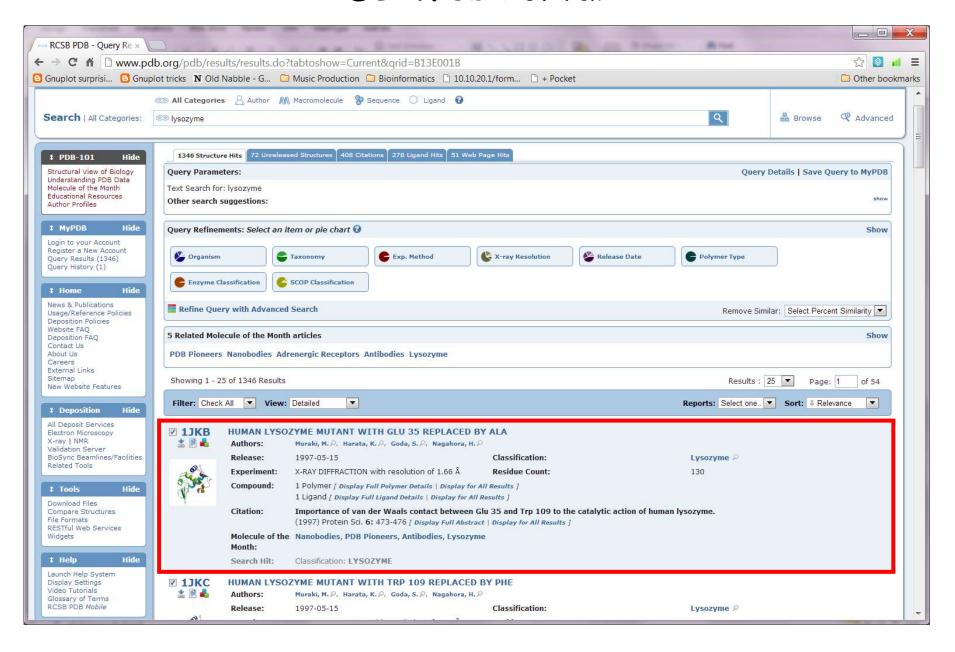


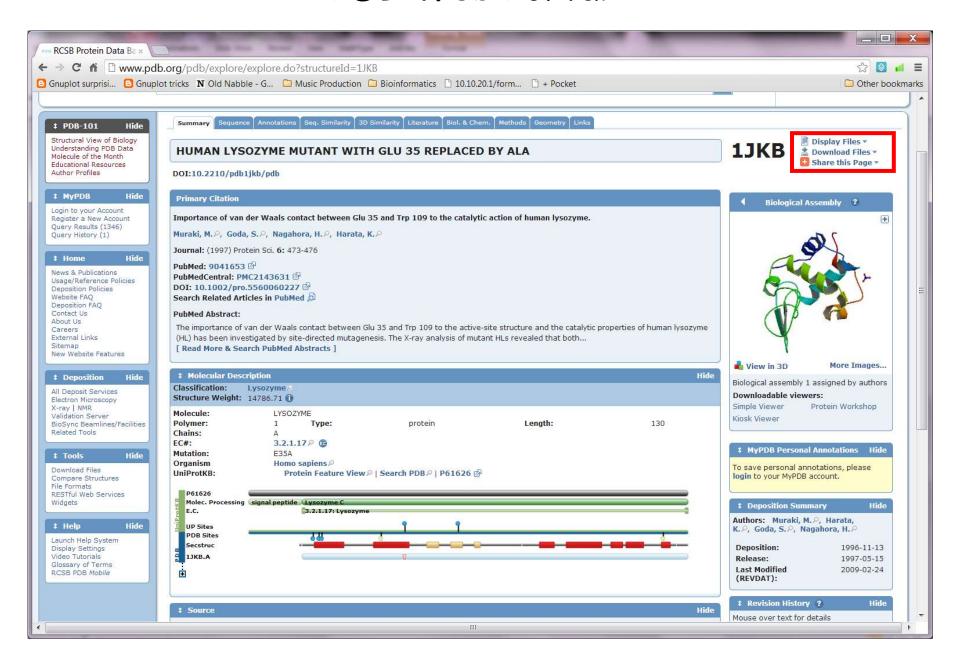
### The PDB contains a very diverse set of structures!



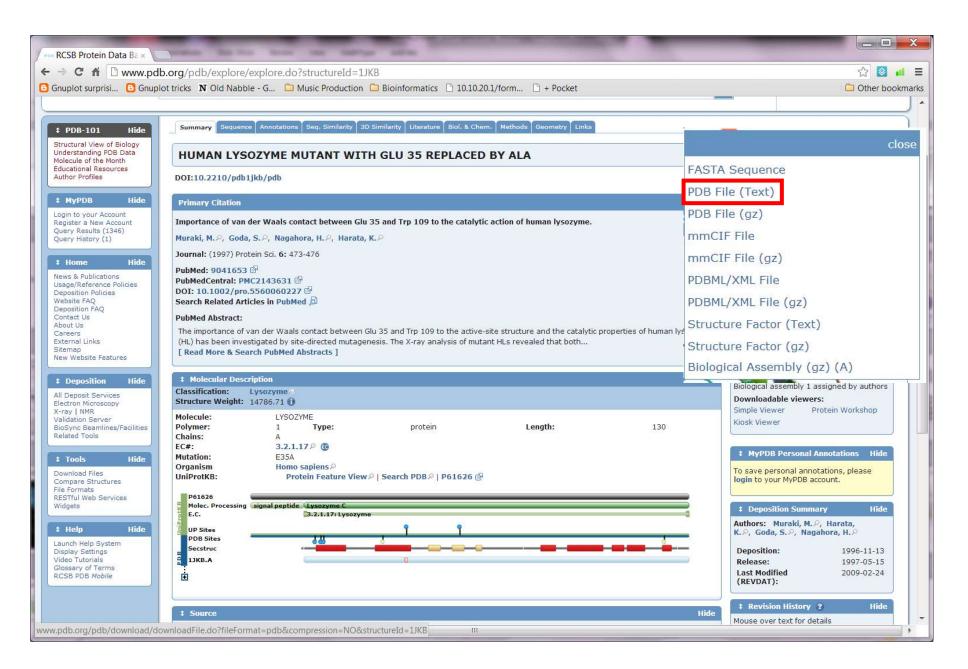




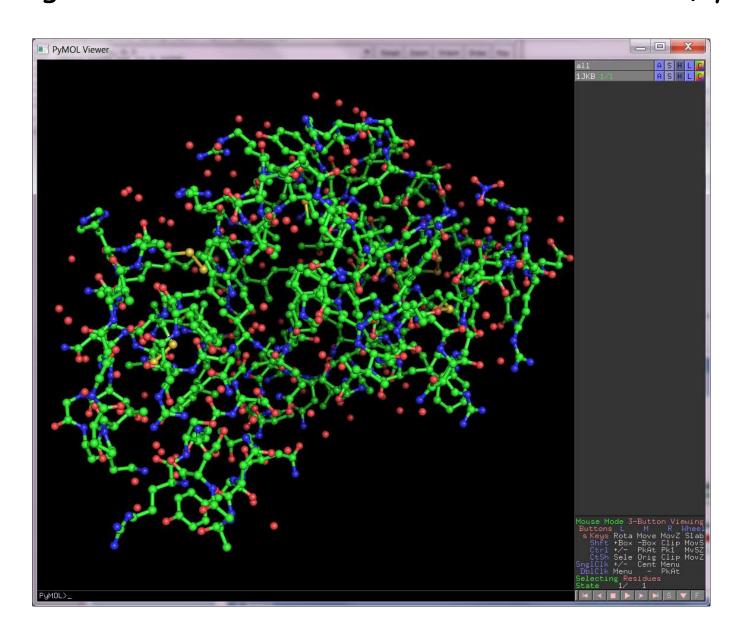




#### Download th structure in PDB format



### Viewing the structure with a molecular visualizer (PyMOL)



# Formatos de representação da estrutura

- A representação da estrutura molecular em bancos de dados passa pela descrição das coordenadas atómicas, do tipo de átomo, e das ligações químicas presentes.
- A descrição do tipo de átomos e ligações que os unem designa-se como topologia da molécula.
- No caso das proteínas, a topologia dos 20 aminoácidos standard pode ser assumida a priori, pois a estrutura dos aminoácidos é conhecida
- A topologia de outras moléculas, tais como grupos prostéticos, deverá ser especificada
- O formato "tradicional" de representação de estrutura no Protein Data Bank é o formato PDB.

# Formato da informação no Protein Data Bank

- · A informação contida no Protein Databank inclui coordenadas atómicas, topologias de ligação (descrição das ligações químicas), nomes dos átomos e grupos químicos, dados associados ao processo de determinação experimental da estruturas e outras informações sobre a função, ligandos, propriedades, etc...
- Presentemente a informação no PDB está disponível nos seguintes formatos:
  - pdb file: O formato "flat file", um tipo de ficheiro chamado "ficheiro PDB". Estes ficheiros são os mais utilizados pelos softwares de manipulação e visualização de estruturas e têm geralmente a extensão ".pdb"
  - mmCIF: um formato mais poderoso e estruturado que o ficheiro PDB, ainda não tendo sido largamente adoptado
  - XML: extended mark-up language, um formato muito geral de representação de informação, compatível com um vasto número de aplicações de software.

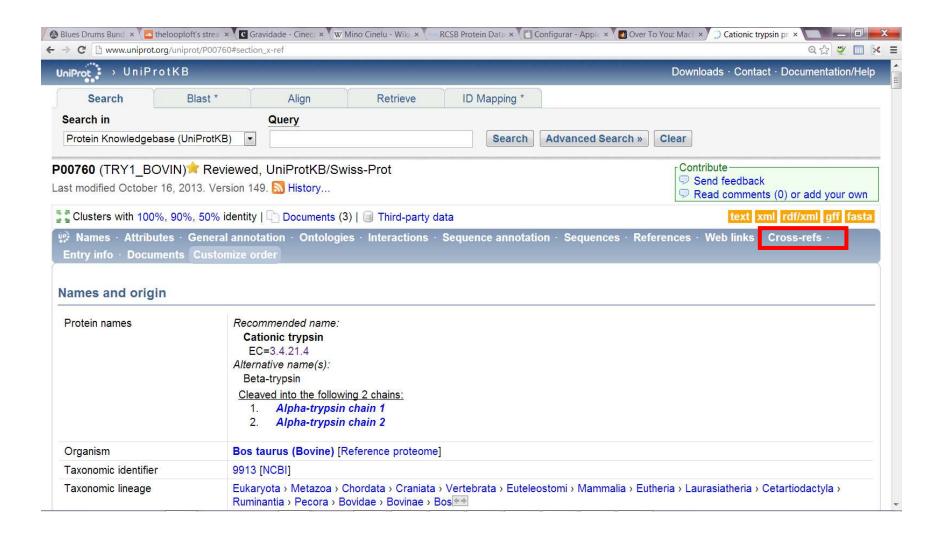
#### Formato do ficheiro PDB

```
HEADER
         METAL BINDING PROTEIN
                                                 21-AUG-03
                                                             108H
TITLE
         CRYSTAL STRUCTURE OF PORCINE OSTEOCALCIN
COMPND
         MOL ID: 1;
         2 MOLECULE: OSTEOCALCIN;
COMPND
         3 CHAIN: A
COMPND
SOURCE
         MOL ID: 1;
         2 ORGANISM SCIENTIFIC: SUS SCROFA;
SOURCE
SOURCE
         3 ORGANISM COMMON: PIG
KEYWDS
         HELIX-TURN-HELIX-TURN-HELIX, PAPER-CLIP, HYDROXYAPATITE
KEYWDS
        2 CRYSTAL SURFACE BINDING PROTEIN, CALCIUM BINDING PROTEIN,
KEYWDS
         3 BONE GLA PROTEIN
EXPDTA
         X-RAY DIFFRACTION
AUTHOR
         Q.Q.HOANG, F.SICHERI, A.J.HOWARD, D.S.YANG
REVDAT
       1 11-NOV-03 1Q8H
JRNL
           AUTH
                 Q.Q.HOANG, F.SICHERI, A.J.HOWARD, D.S.YANG
JRNL
                 BONE RECOGNITION MECHANISM OF PORCINE OSTEOCALCIN
JRNL
           TITL 2 FROM CRYSTAL STRUCTURE.
                  NATURE
                                                V. 425 977 2003
JRNL
JRNL
           REFN ASTM NATUAS UK ISSN 0028-0836
REMARK
       1
REMARK
        2
REMARK
         2 RESOLUTION. 2.00 ANGSTROMS.
REMARK
REMARK
        3 REFINEMENT.
REMARK
         3
            PROGRAM
                        : CNS 1.1
REMARK
        3 AUTHORS
                        : BRUNGER, ADAMS, CLORE, DELANO, GROS, GROSSE-
                     . . . . . . . . .
ATOM
         1 N
                PRO A 13
                               10.210 29.966 44.935 1.00 38.06
          2 CA PRO A 13
ATOM
                                9.718
                                      29.013 43.919 1.00 37.33
                                      29.662 42.541 1.00 37.52
ATOM
          3 C
                PRO A 13
                                9.566
ATOM
          4 0
                PRO A 13
                                9.275
                                      30.855
                                              42.444 1.00 38.00
                                                                           0
          5 CB
                PRO A 13
                                8.383
                                      28.488
                                              44.434 1.00 37.68
ATOM
ATOM
          6 CG
                PRO A 13
                                7.919 29.624 45.336 1.00 36.60
                                                                           С
                                      30.126
ATOM
          7 CD
                PRO A 13
                                9.196
                                              45.995 1.00 36.47
          8 N
                ASP A 14
                                9.777 28.879
                                              41.483 1.00 36.83
ATOM
                                                                           N
                                                                           С
ATOM
               ASP A 14
                                       29.384
                                               40.116 1.00 36.13
           299
                            3
                                 0
                                      0
                                           0
                                                6 378
MASTER
                                                        1 38
END
```

Cabeçalho

Coordenadas

### Interligação entre Uniprot e PDB



# Interligação entre Uniprot e PDB

