Python Programming & Computational Structural Biology

Computational Biology – Spring 2021

What's it about ?

Computational analysis and study of the structure and function of biological molecules (macromolecules and small molecules)

- Computational analysis and study of biological molecules (macromolecules and small molecules)
- Storage and management of structural information
- Computational algorithms for structure search and comparison
- Computational algorithms for the prediction of 2D and 3D structures of macromolecules
- Computational algorithms for the ***simulation*** of the physical and chemical behavior of biological (Biomolecular Simulation)
- Computational algorithms for protein-ligand analysis and discovery of new drugs (Computational Drug Design)

What will happen?

Learning

- Introduction of structural databases and formats of molecular representation
- ~~Introduction to methods of macromolecule structure determination~~
- Prediction of secondary and tertiary structure of proteins
- ~~Biomolecular Simulation Methods~~
- ~~Introduction to protein-ligand interactions, computational docking and virtual screening~~

Doing

- Introduction to Python Programming and Jupyter notebooks and how to use them for analysis of sequences and structures of biological macromolecules
- Learning the basics of working in the Jupyter Hub and Linux environments
- How to use the molecular visualizer PyMOL to analyze, display and build structures of molecules (and make pretty pictures!)
- Constructing a protein model by comparative (homology) modelling
- ~~Running a docking simulation with the Autodock Vina software~~

Who am I ?

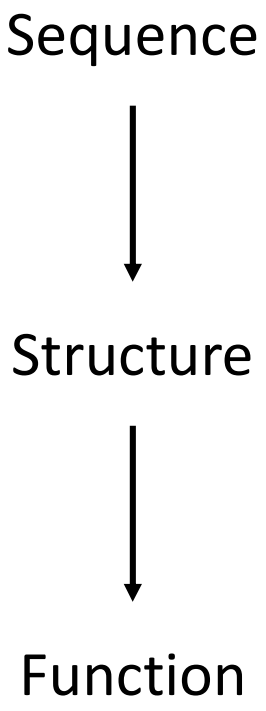
- My name is Paulo Martel
- I teach Bioinformatics, Computational Biology, Drug Design, Protein Structure and Enzymology (far too much!)
- My research focus around computational analysis of protein structure and function using biomolecular simulation techniques, namely Molecular Dynamics, Comparative Modelling and Protein Electrostatics
- My office number is 3.12 in Building 8 (but you probably won't go there…)
- My email is pmartel@ualg.pt
- Feel free to contact me or ~~drop by~~ if you have any questions regarding the course and its content
- This course will be taught in remote learning mode, but the final exam will be given in-person, in a classroom (in-person learning is back in UALG from April, 19th)

Before we start

- Most of computational activities will take place on a **remote server** with address https://compbio-2021.ddns.net
- You can login to that server from any computer you wish
- There are accounts for you on this server using, with login name same as your student number (without the leadin “a”).
- Login in with your student number as user and mas password **#123<your_student_number> …** we can change those passwords later if you prefer.
- The environment we are using is called Jupyter Hub and provides access both to a Python Notebook interface and one or more login terminals on the server
- You will be required to install the following software on your personal computer:
  - PyMOL (you can download it from the site https://www.pymol.org)
  - ~~MGLTools (the instructor will provide a link for downlad)~~

Biomolecular Structures

Structure Leads to Function



Flow of Biological Information

Gene …TTAATAAGT…

⬇ transcription

m-RNA …UUAAUAAGU…

⬇ splicing, translation

cadeia polipeptídica …LISVHDN…

⬇ post-translational modifications

proteína

Central Dogma of Molecular Biology

Exceptions: RNA viruses, prions, ribozymes (?)

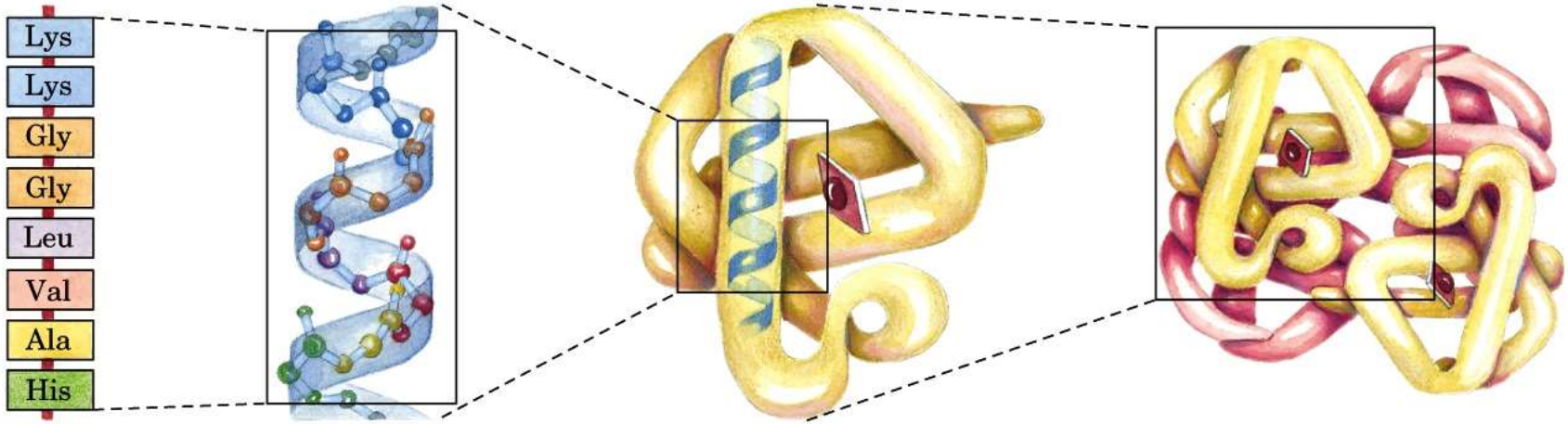
Levels of structural organization

Primary structure

Secondary structure

Tertiary structure

Quaternary structure



Aminoacid sequence

α-helix

Polypeptide chain

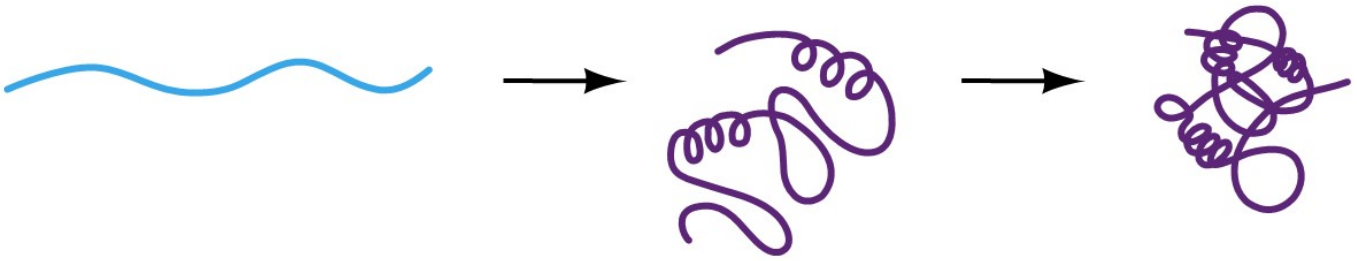
Arrangment of monomers

Sequence determines structure

The tridimensional structure of proteins arises due to the physico-chemical forces acting between atoms in the polypeptide chain and the solvente. Many proteins will spontaneously acquire their *native* structure following ribossomal synthesis. This process is called *protein folding*.

Predicting the native structure of a protein given its sequence is one of the most fundamental probblems of nodern day molecular biology.
(Folding problem)

Protein folding:

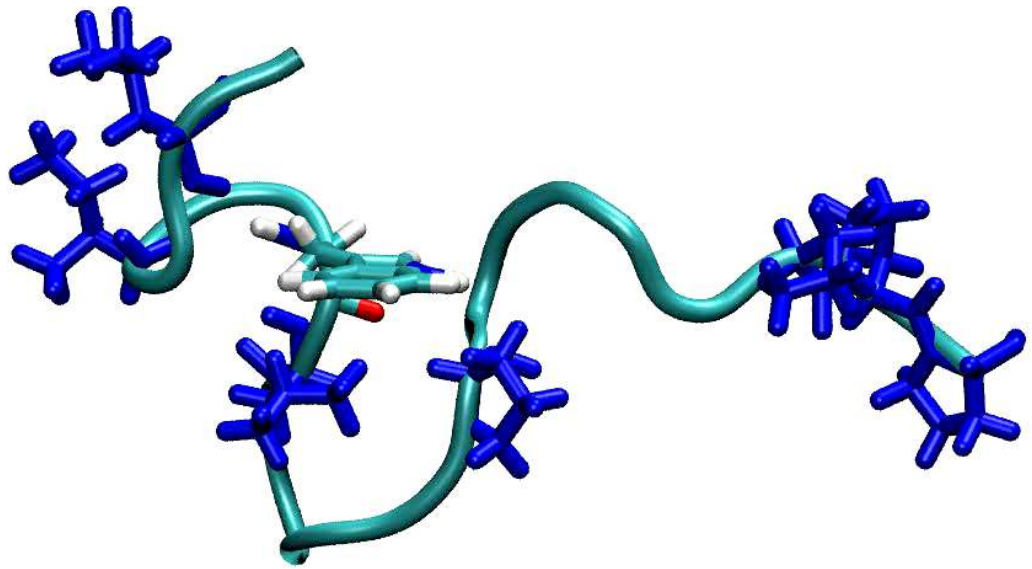


Linear protein chain

Secondary structure formation

Bending of the protein chain in a 3D shape

Computing folding



Molecular Dynamics simulation of the folding mechanism of the mini-protein ***trp*-cage**

Representing Structures

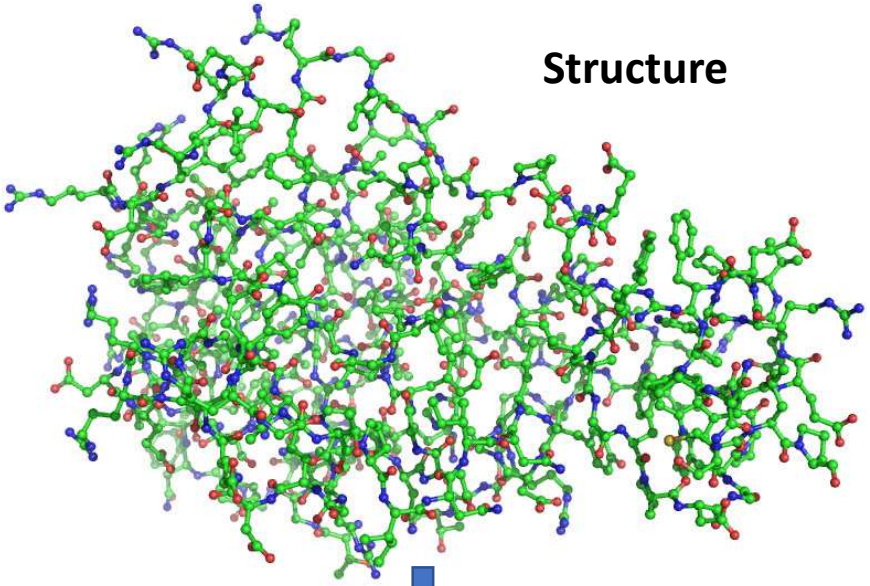
The representation of molecular structure is much more complex and storage intensive than sequence

Sequence

```
...AVAGGATILVHNQDAGEPAIVLAFG...
```

Simple sequence of one-letter symbols

Structure



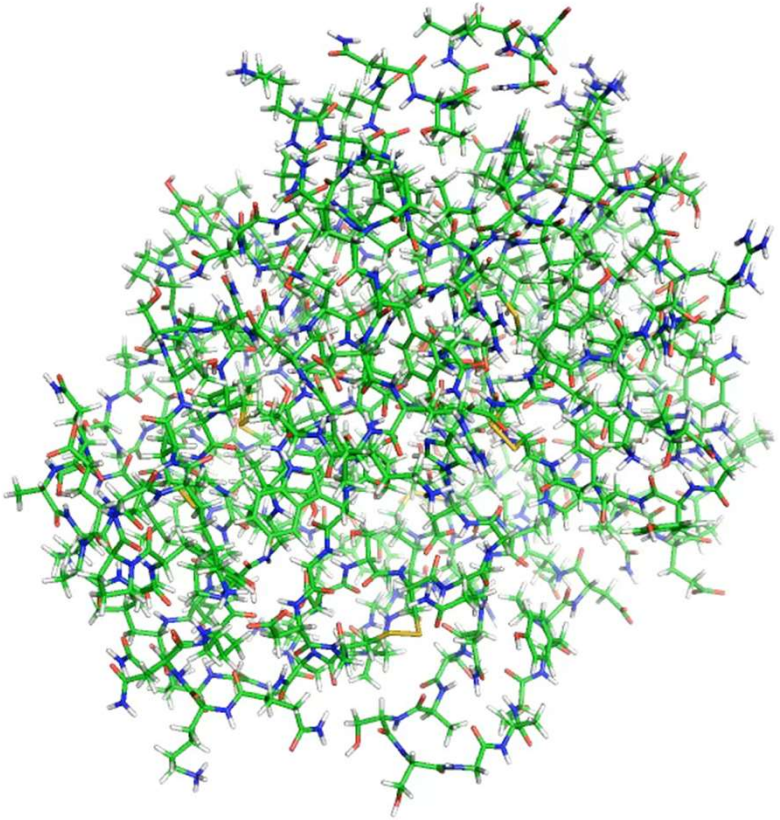
XYZ **coordinates** of each atom, their **types** and **connectivity**

Example

Human Trypsin

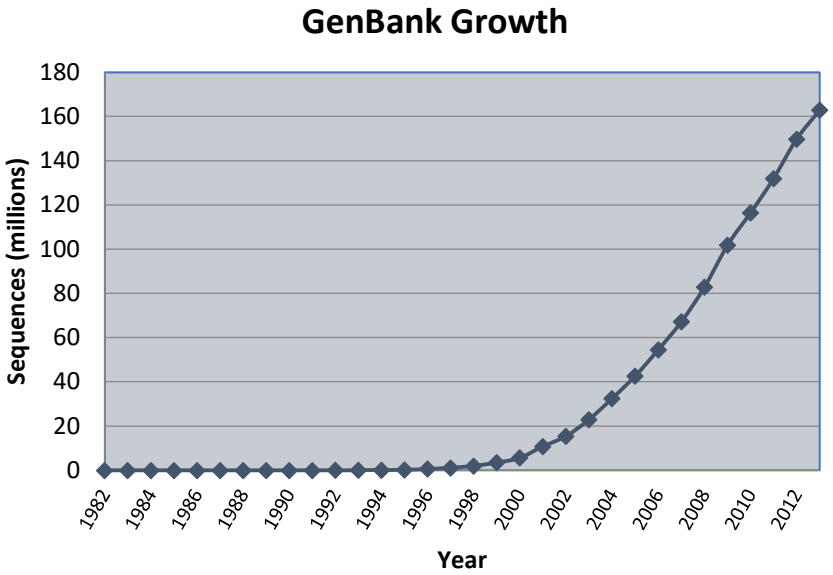
```
>sp|P07477|TRY1_HUMAN Trypsin-1 OS=Homo sapiens OX=9606 GN=PRSS1
MNPLLILTFVAAALAAPFDDDDKIVGGYNCEENSVPYQVSLNSGYHFCGGSLINEQWVVS
AGHCYKSRIQVRLGEHNIEVLEGNEQFINAAKIIRHPQYDRKTLNNDIMLIKLSSRAVIN
ARVSTISLPTAPPATGTKCLISGWGNTASSGADYPDELQCLDAPVLSQAKCEASYPGKIT
SNMFCVGFLEGGKDSCQGDSGGPVVCNGQLQGVVSWGDGCAQKNKPGVYTKVYNYVKWIK
NTIAANS
```

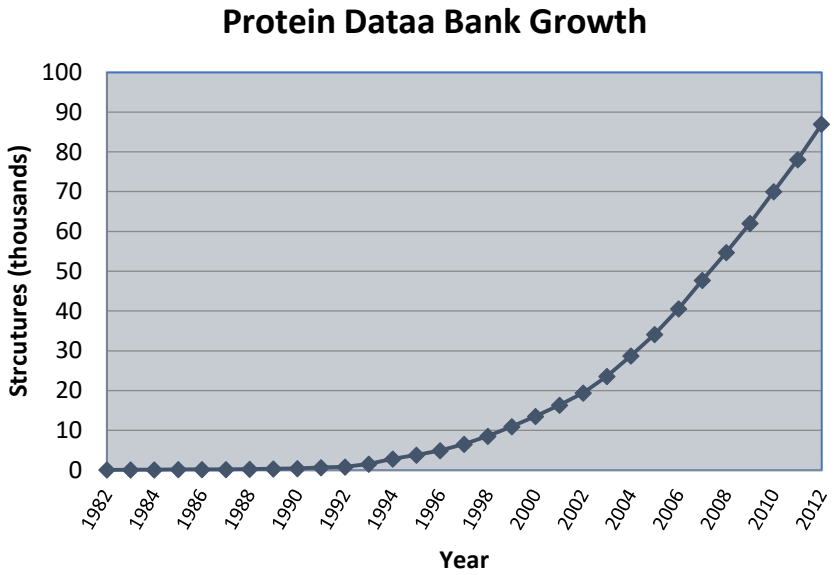
247 aminoacids



3415 atoms

Sequence *versus* structure





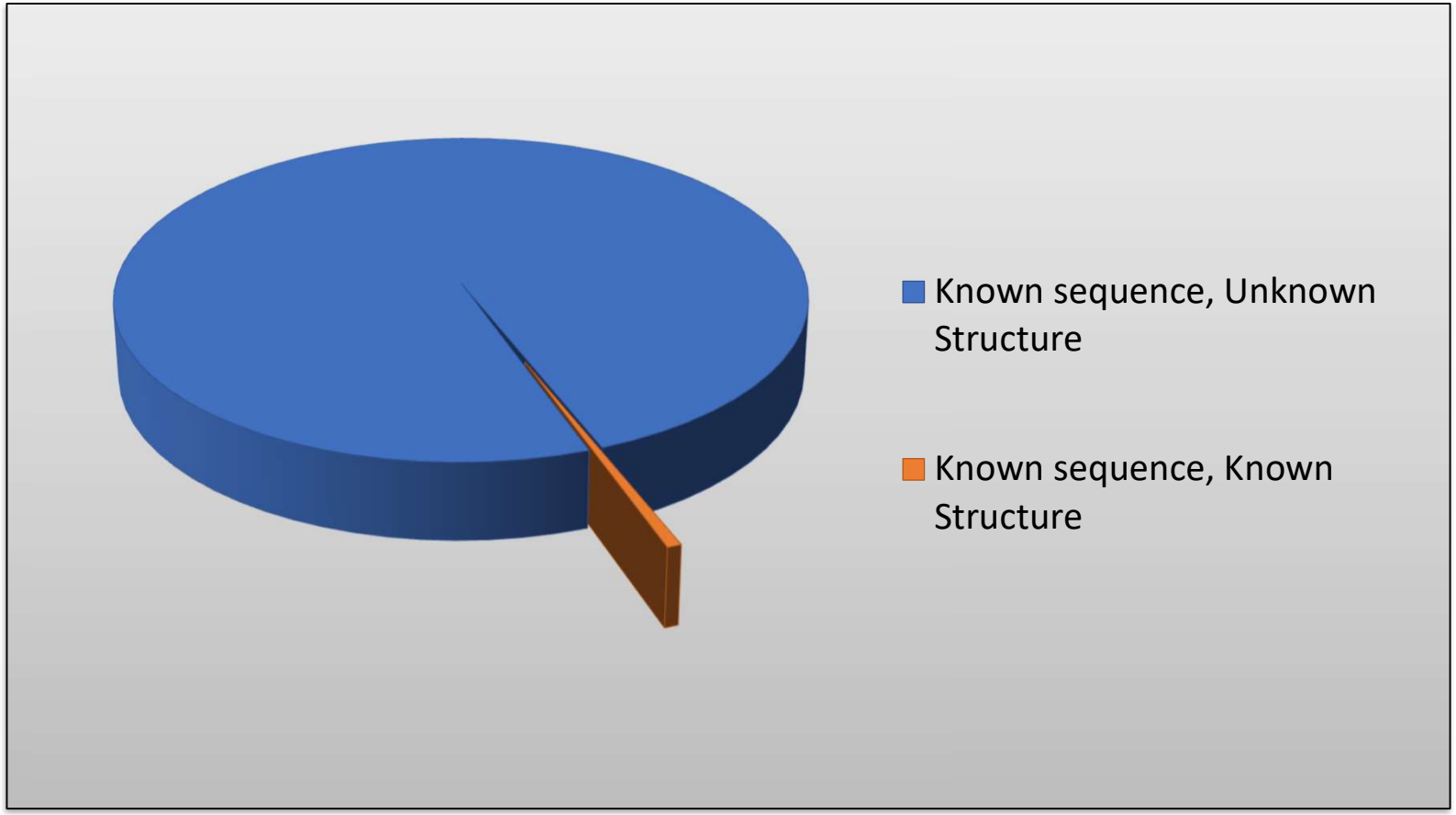
milions of sequences *versus* **thousands** of structures!

In 1982: 172 know structures and 602 sequences

Today (March 2019): 149,886 structures and 212,260,377 sequences!!

In conclusion: Sequencing is way faster than structure determination (the number of proteins of known sequence and unknown structure is growing very rapidly)!

Most knowproteins have unknow sequence !



The importance of structure prediction

The vast (and steadily growing) number of proteins of unknown structure puts a heavy demand on ever faster methods for 3D structure determination. Due to their intrincacies, such methos simply cannot cope with the fast pace of sequencing, and the gap widens more and more. This situation is not likely to change, ever.

So what can we do ?

We need to be able *predict* the 3D structure of proteins from their aminoacid sequence. In general terms this is a very hard computational problem, but there are special situations where it is very feasible. That is currently our best hope of coping up.

Prediction of the 3D structure of proteins is thus one of the most fundamental problems of bioinformatics / computational biology.

Biomolecular Databases and Formats

Macro vs. Small Molecules

- **Macromolecules**: contain the structure of biological macromolecules. The primary source is the Protein Data Bank (https://www.rcsb.org).

- **Small molecules**: databanks containing formulas, structures and other information relative to small molecules* . Some of the most important are:
  - PubChem (NCBI) :https://pubchem.ncbi.nlm.nih.gov/
  - ZINC (purchasable compounds): https://zinc.docking.org/
  - Drugbank (pharma oriented): https://www.drugbank.ca/
  - CCDC (crystallographic structures): https://www.ccdc.cam.ac.uk/

* 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 (https://www.rcsb.org).

- **Secondary Databanks**: specialized views, collections or filters on the primary databanks. Example: The PDBind database (http://www.pdbbind.org.cn)

- 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) http://www.rcsb.org

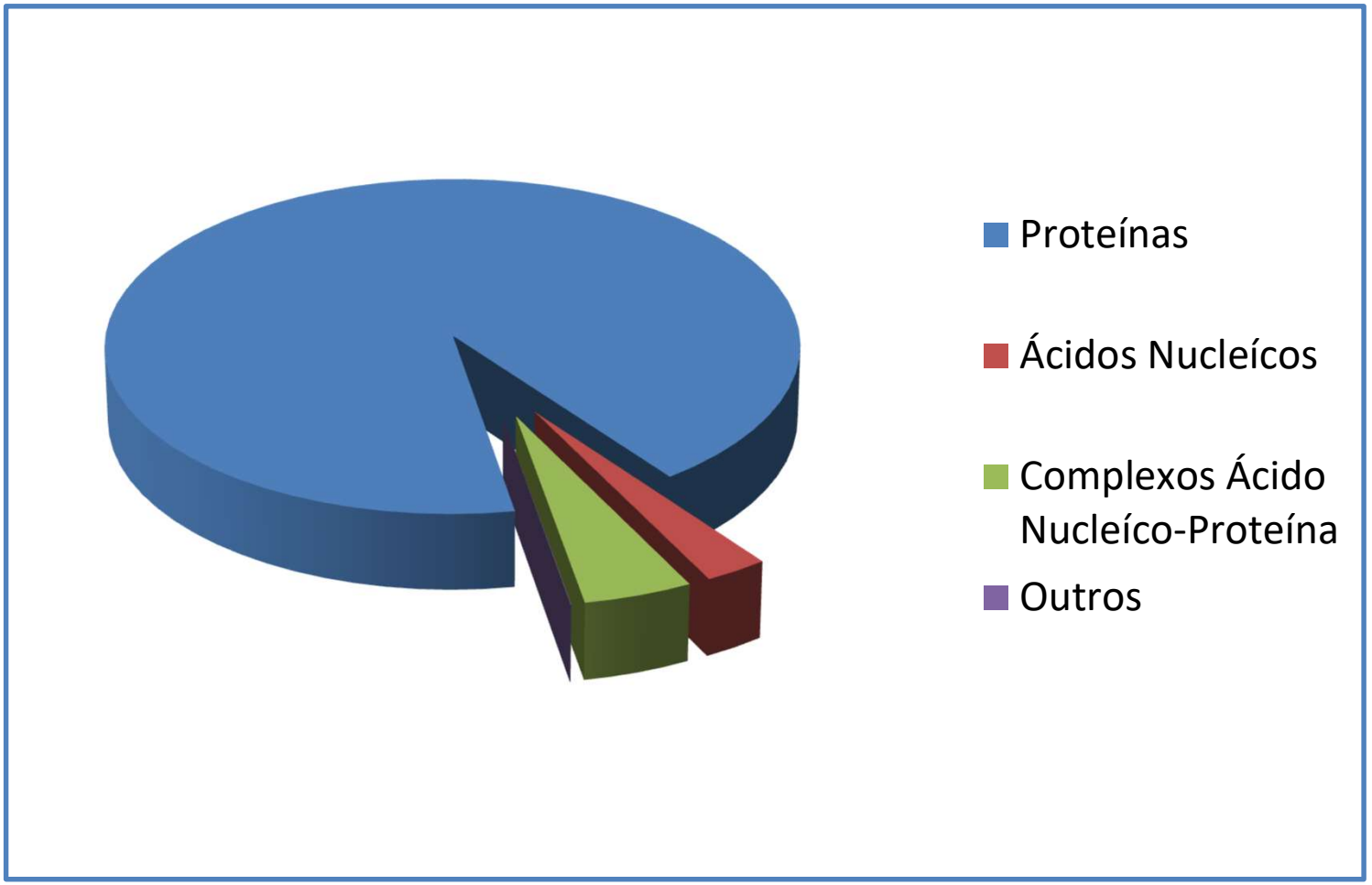
The Protein Data Bank

- 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 http://www.rcsb.org

The Protein Databank contains different types of biological macromolecules

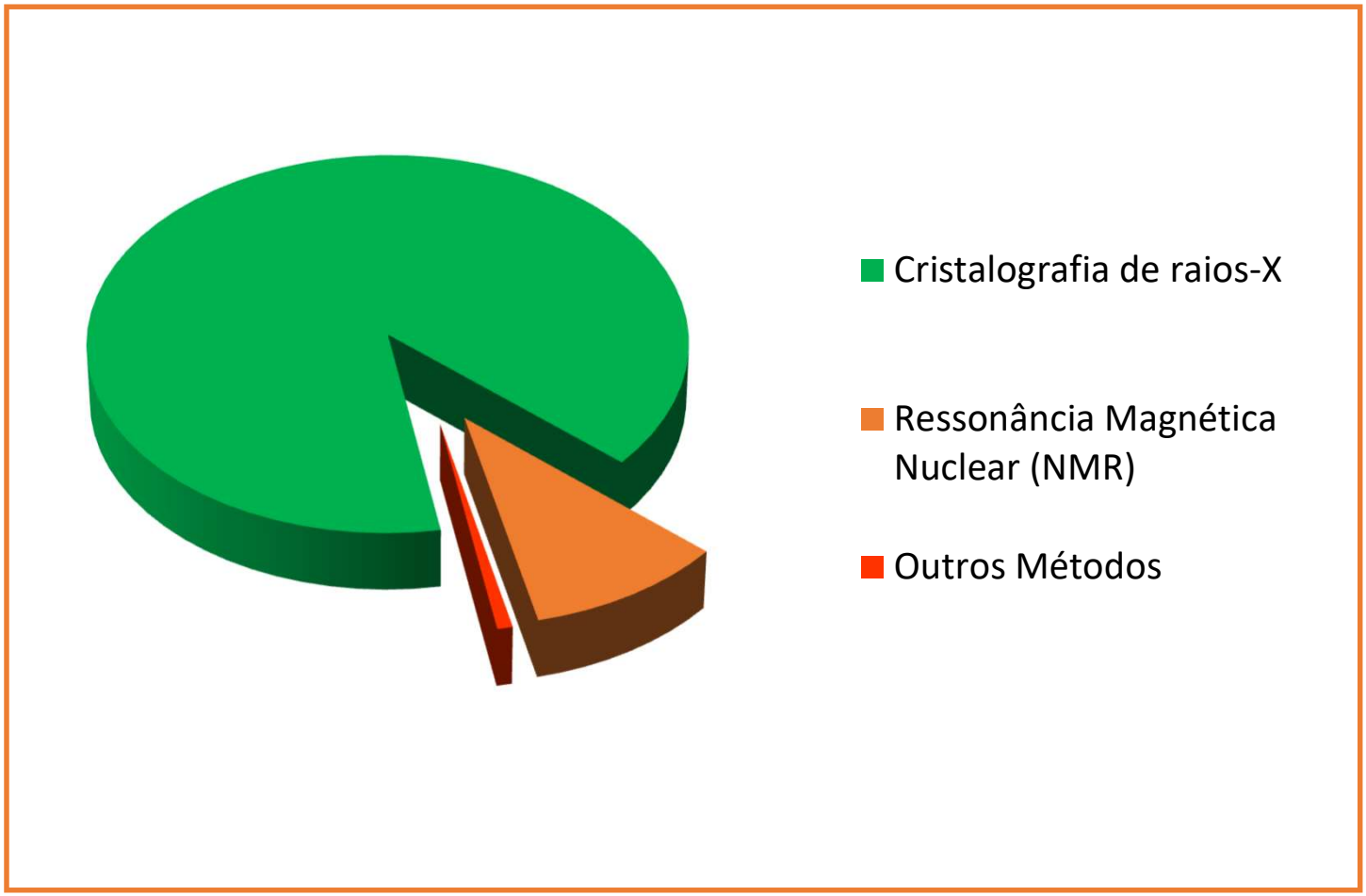


Where does all the structural information come from ?

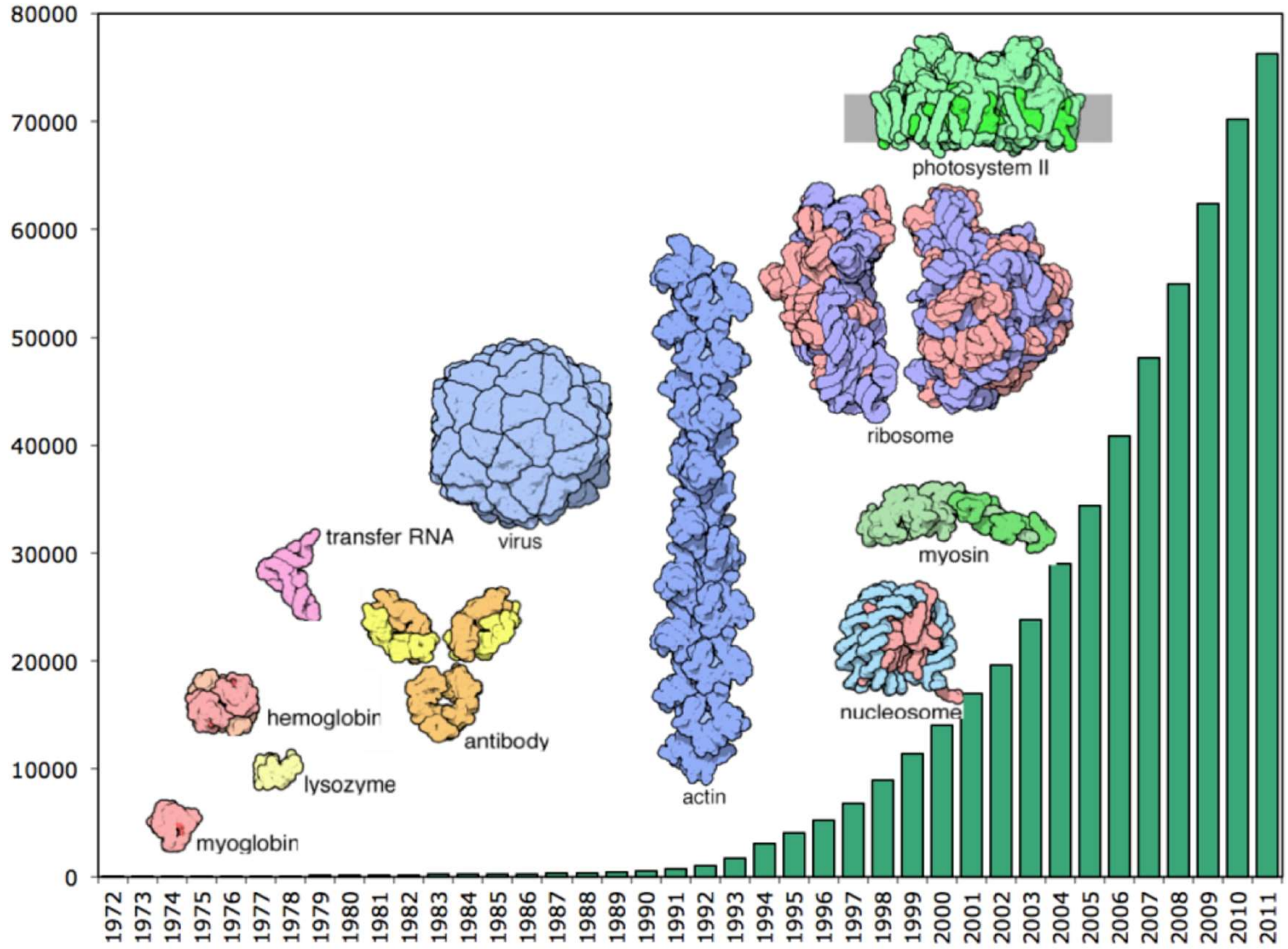
It's a combination of various types of data:

- 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)

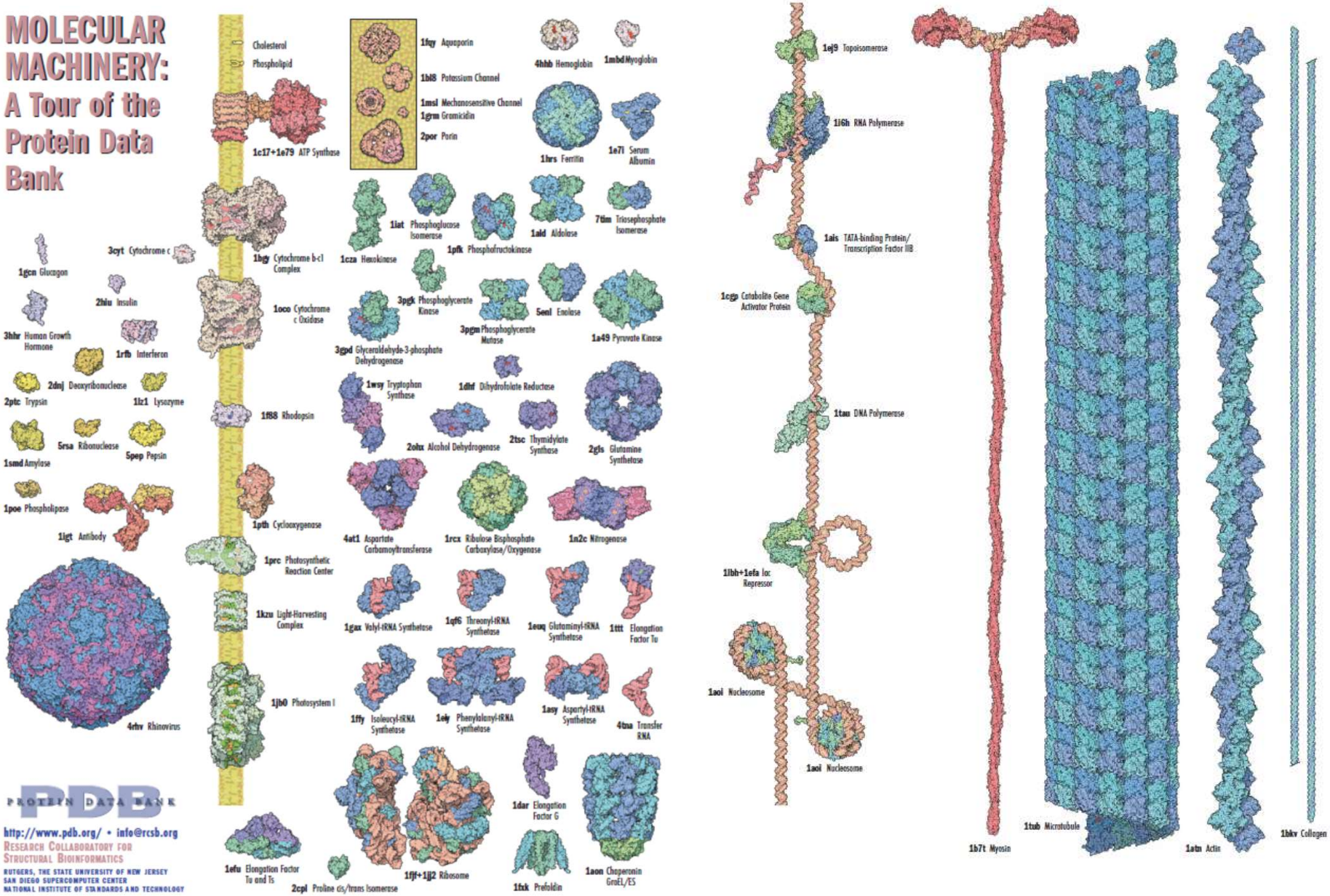
Most structures have been solved by X-ray cristallography



Progress in the determination of macromolecular structures



The PDB contains a very diverse set of structures !

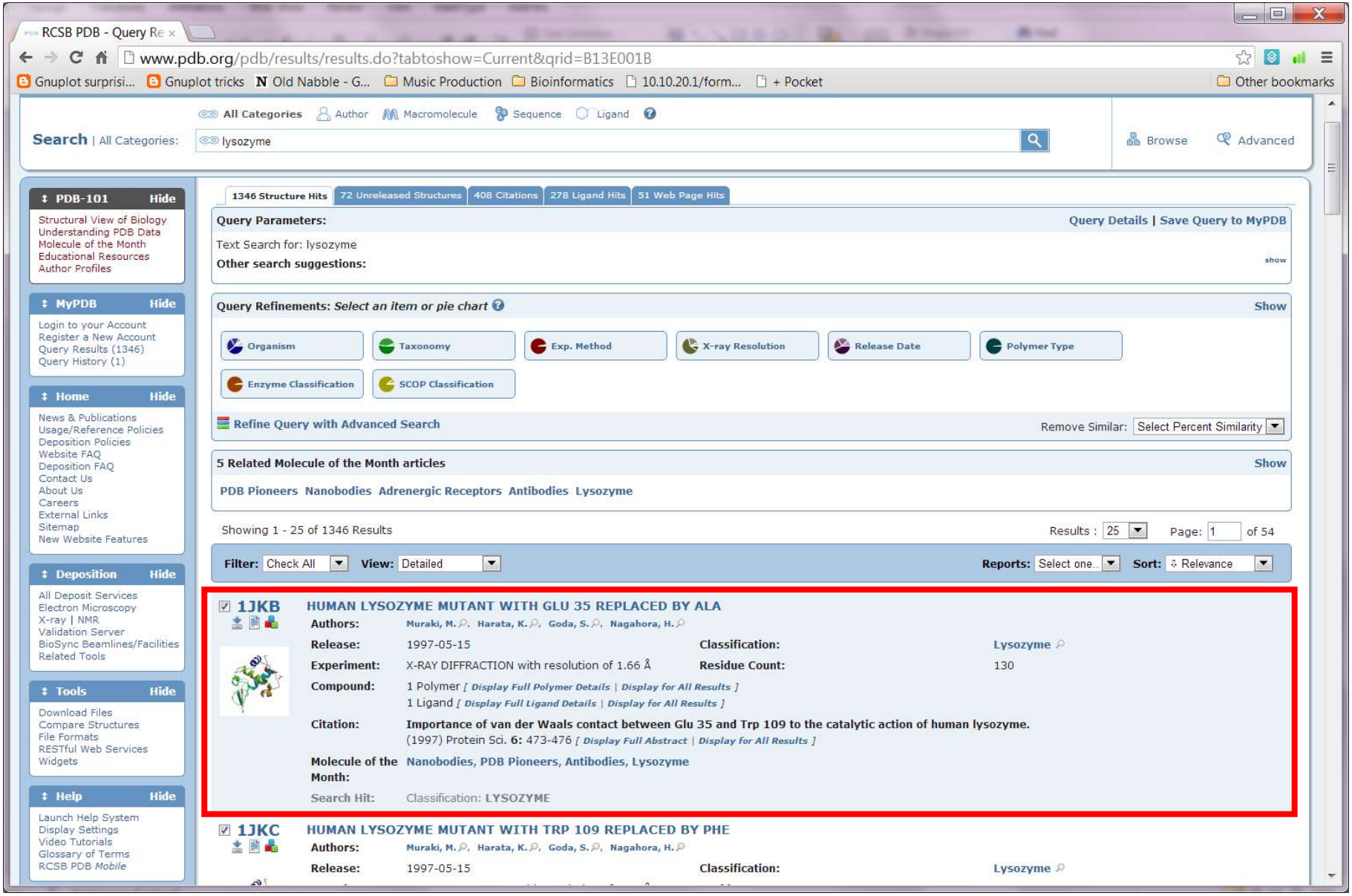


PDB Web Portal

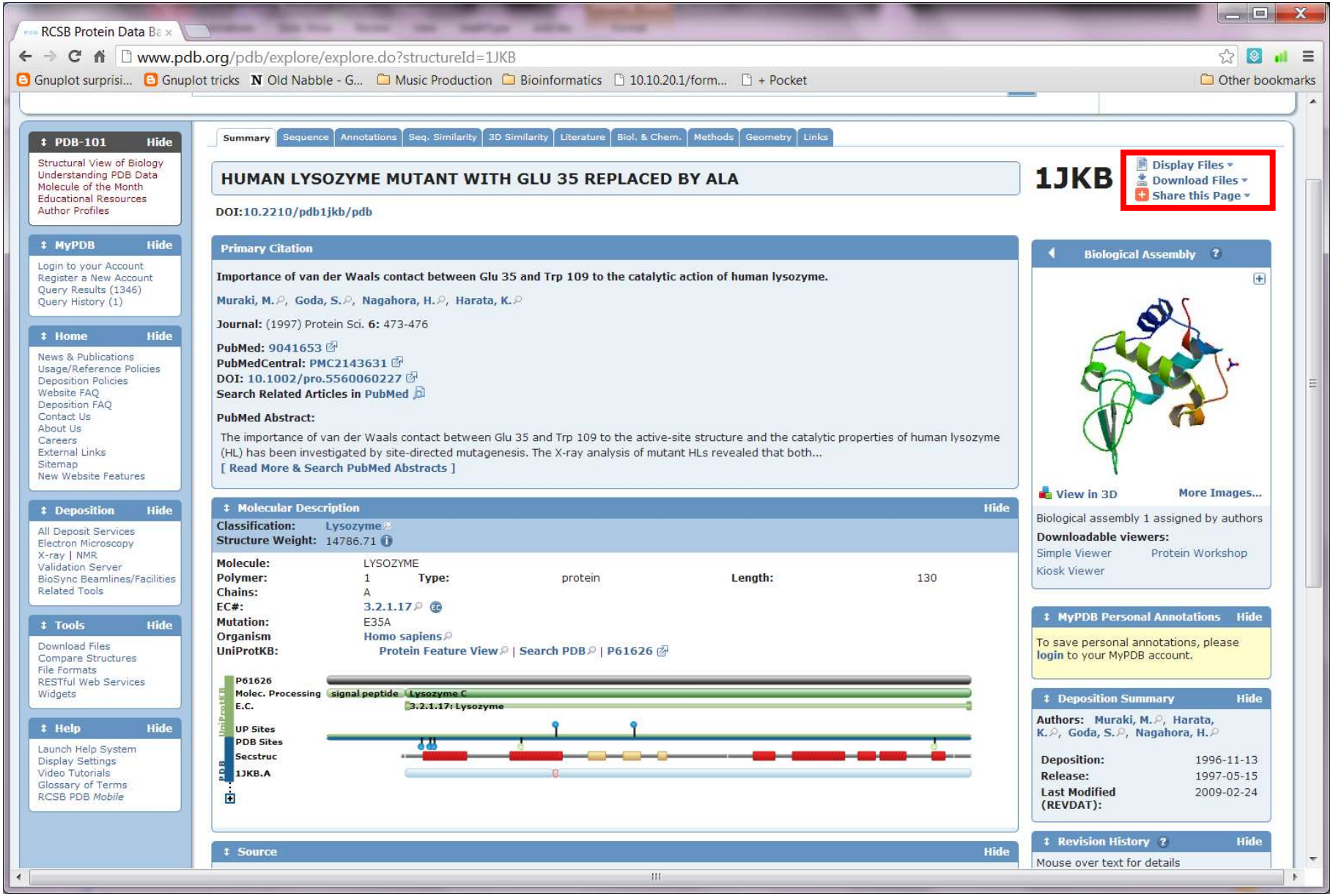
https://www.rcsb.org

PDB Web Portal

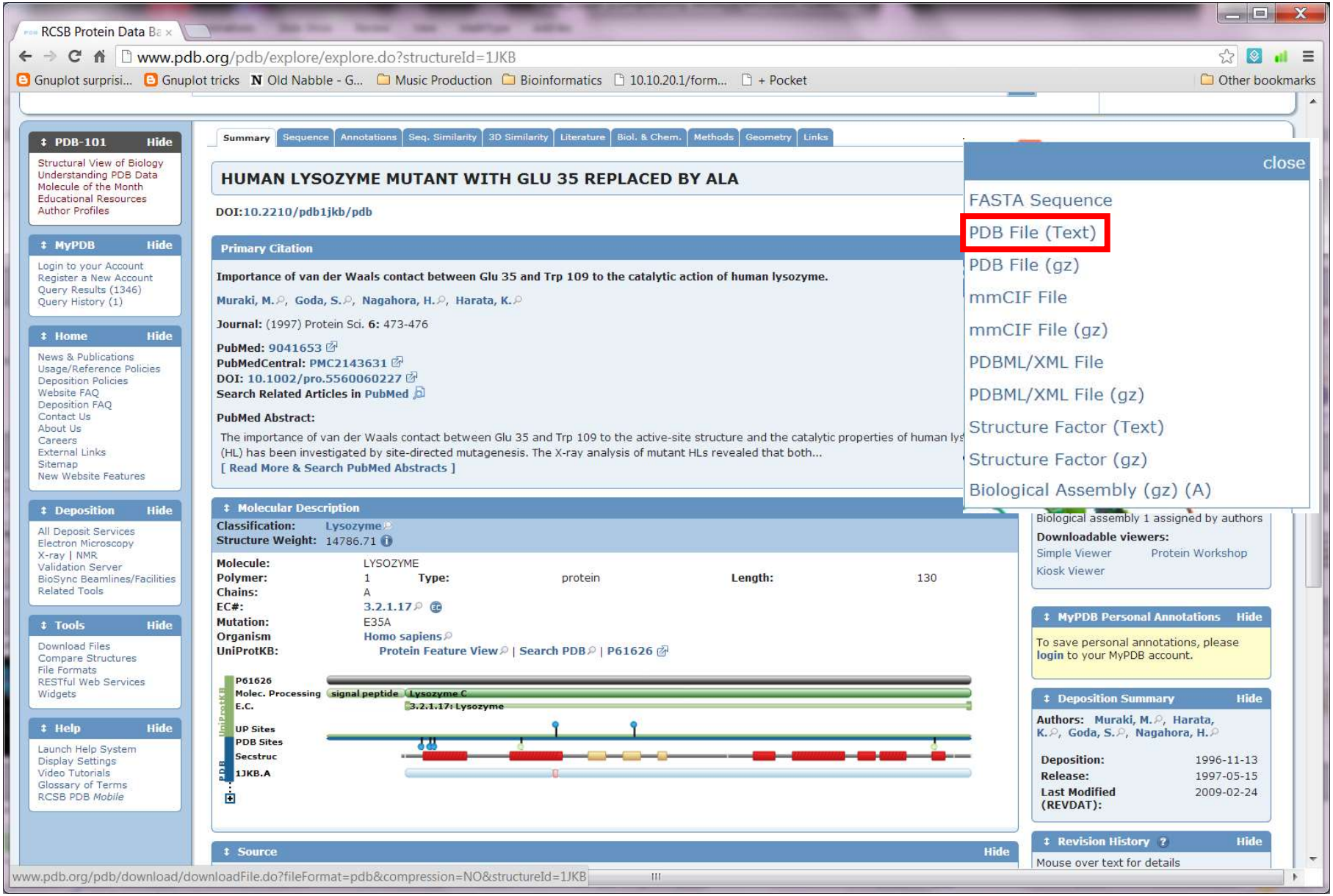
PDB Web Portal



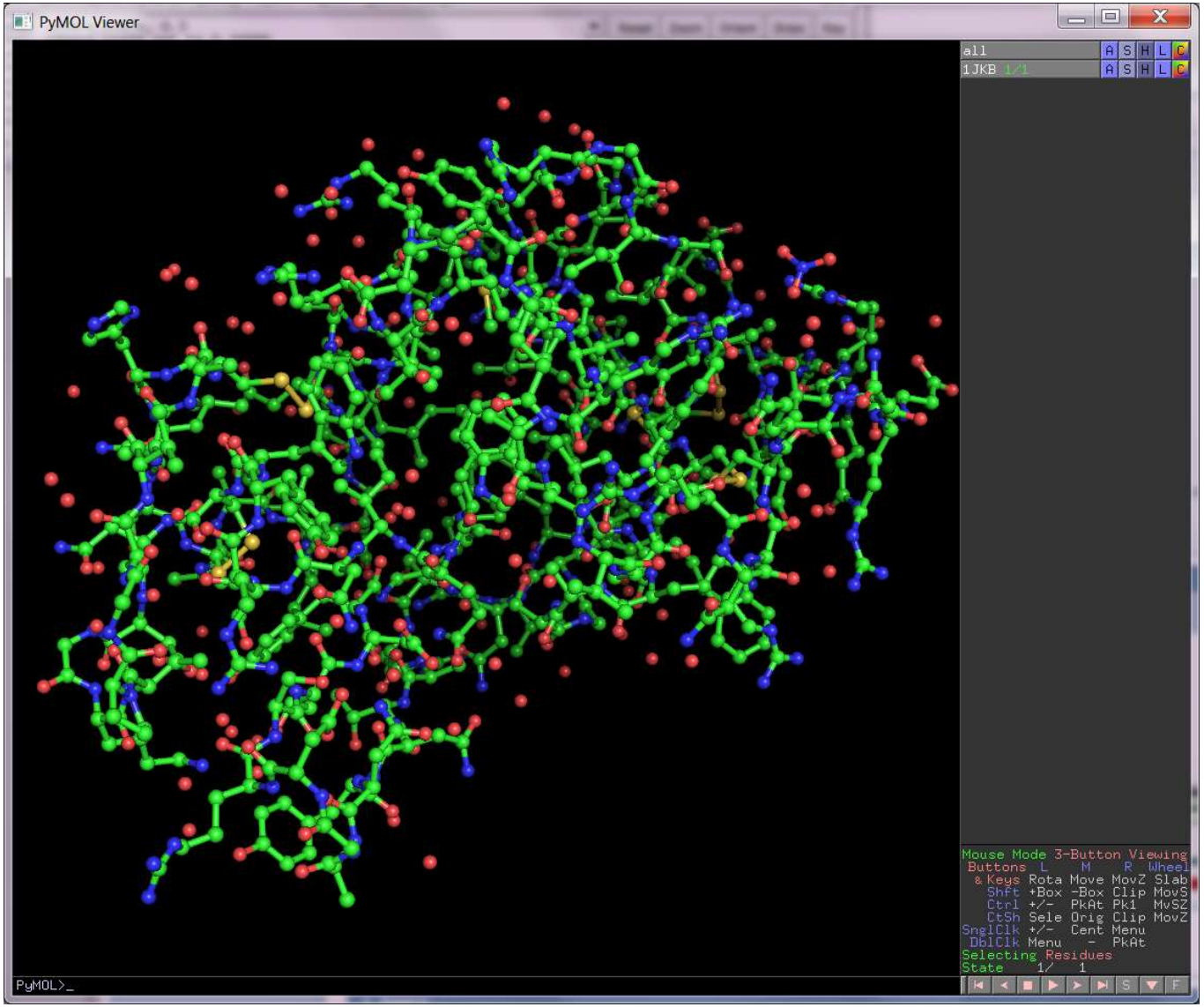
PDB Web Portal



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.

  • **mmtf:** formato binário altamente compacto, ilegível para humanos, mas ocupando muito menos espaço e oferecendo rápida transmissão, leitura e escrita.

Formato do ficheiro PDB

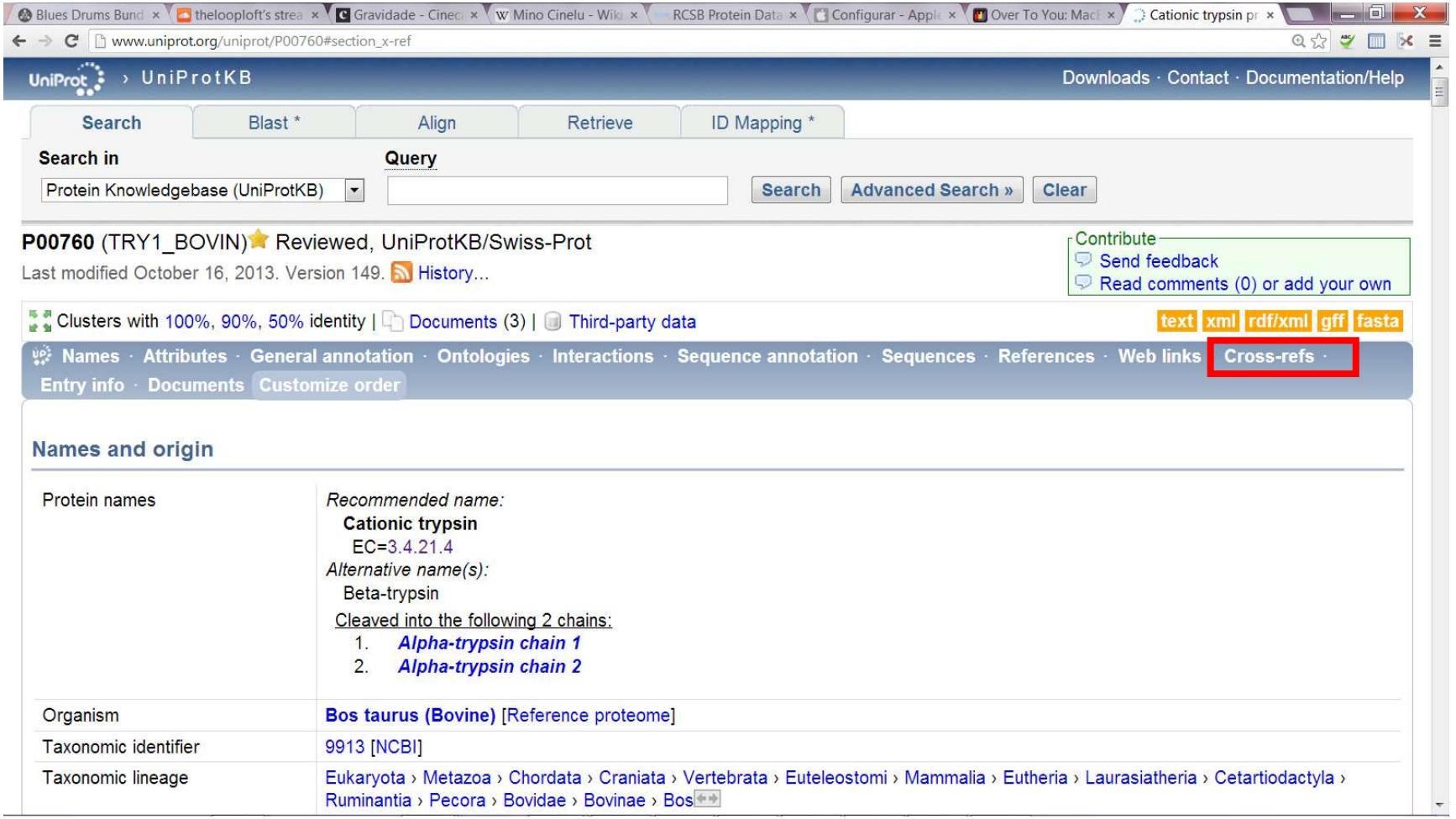
```
HEADER    METAL BINDING PROTEIN                   21-AUG-03   1Q8H
TITLE     CRYSTAL STRUCTURE OF PORCINE OSTEOCALCIN
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: OSTEOCALCIN;
COMPND   3 CHAIN: A
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: SUS SCROFA;
SOURCE   3 ORGANISM_COMMON: PIG
KEYWDS    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    0
JRNL        AUTH   Q.Q.HOANG,F.SICHERI,A.J.HOWARD,D.S.YANG
JRNL        TITL   BONE RECOGNITION MECHANISM OF PORCINE OSTEOCALCIN
JRNL        TITL 2 FROM CRYSTAL STRUCTURE.
JRNL        REF    NATURE                        V. 425   977 2003
JRNL        REFN   ASTM NATUAS  UK ISSN 0028-0836
REMARK   1
REMARK   2
REMARK   2 RESOLUTION. 2.00 ANGSTROMS.
REMARK   3
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           N
ATOM      2  CA  PRO A  13       9.718  29.013  43.919  1.00 37.33           C
ATOM      3  C   PRO A  13       9.566  29.662  42.541  1.00 37.52           C
ATOM      4  O   PRO A  13       9.275  30.855  42.444  1.00 38.00           O
ATOM      5  CB  PRO A  13       8.383  28.488  44.434  1.00 37.68           C
ATOM      6  CG  PRO A  13       7.919  29.624  45.336  1.00 36.60           C
ATOM      7  CD  PRO A  13       9.196  30.126  45.995  1.00 36.47           C
ATOM      8  N   ASP A  14       9.777  28.879  41.483  1.00 36.83           N
ATOM      9  CA  ASP A  14       9.671  29.384  40.116  1.00 36.13           C
                    ..........
MASTER      299    0    6    3    0    0    0    6  378    1   38    4
END
```

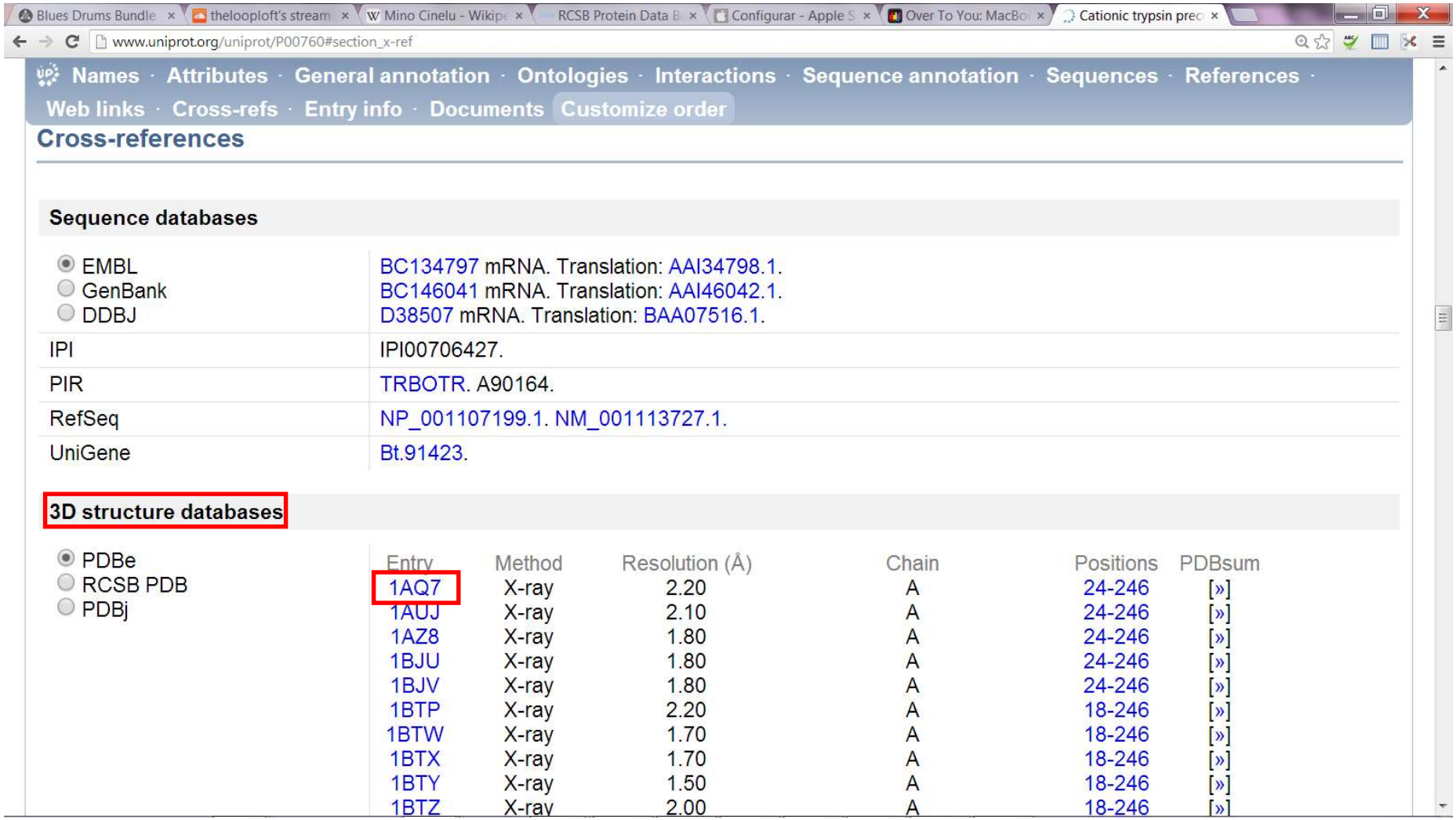
Cabeçalho

Coordenadas

Interligação entre Uniprot e PDB



Interligação entre Uniprot e PDB



Blues Drums Bundle | thelooploft's stream | Mino Cinelu - Wikipe | RCSB Protein Data B | Configurar - Apple S | Over To You: MacBo | RCSB Protein Data B

www.pdb.org/pdb/explore/explore.do?structureId=1AQ7

RCSB PDB PROTEIN DATA BANK → RCSB PDB-101

A MEMBER OF THE PDB | EMDataBa

An Information Portal to Biological Macromolecular Structu

As of Tuesday Oct 29, 2013 at 5 PM PDT there are 95113 Structures | PDB Statistics

Search
Advanced
Browse

Everything | Author | Macromolecule | Sequence | Ligand

e.g., PDB ID, molecule name, author

Search History , Previous Results

PDB-101 Hide
- Structural View of Biology
- Understanding PDB Data
- Molecule of the Month
- Educational Resources
- Author Profiles

MyPDB Hide
- Login to your Account
- Register a New Account
- MyPDB Help Page

Home Hide
- News & Publications
- Usage/Reference Policies
- Deposition Policies
- Website FAQ
- Deposition FAQ
- Contact Us
- About Us
- Careers
- External Links
- Sitemap
- New Website Features

Summary | 3D View | Sequence | Annotations | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

TRYPSIN WITH INHIBITOR AERUGINOSIN 98-B

1AQ7

Display Files
Download Files
Share this Page

DOI:10.2210/pdb1aq7/pdb

Primary Citation

Atomic Structure of the Trypsin-Aeruginosin 98-B Complex

Sandler, B., Murakami, M., Clardy, J.

Journal: (1998) J.Am.Chem.Soc. **120:** 595-596

PubMed ID is not available

Molecular Description Hide

Classification: Hydrolase/hydrolase Inhibitor
Structure Weight: 23979.18

Molecule: TRYPSIN
Polymer: 1 **Type:** protein **Length:** 223
Chains: A
EC#: 3.4.21.4

Biological Assembly