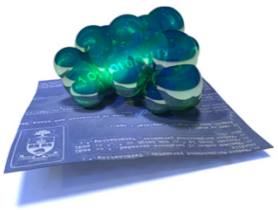
A
Bioinformatics
Course

PDB

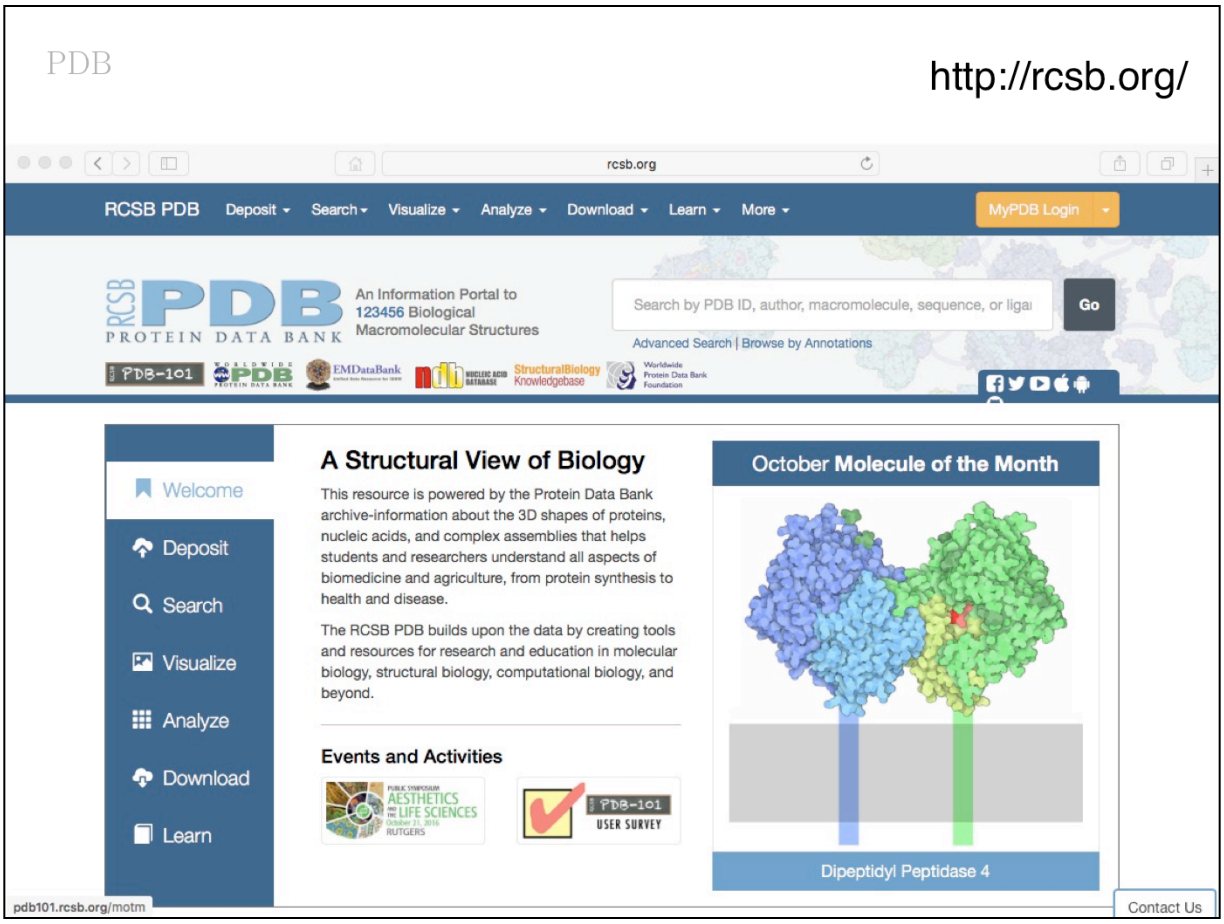


Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

PDB

http://rcsb.org/

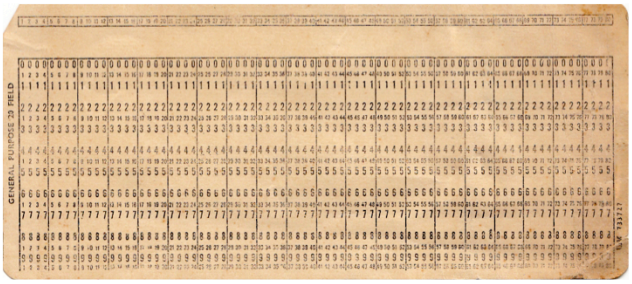


...the **PDB** (Protein structure DataBase) is **the** central repository for 3D structural data of proteins and nucleic acids. RCSB is the Research Collaboratory for Structural Bioinformatics.

If you look closely at the screenshot, you will see that the PDB held 123,456 coordinate files (as of October 16. 2016). That's the actual number, not a placeholder. How cool is that! (Aplogies for this completely useless bit of trivia. It's 134,091 in October 2017.)

PDB FILE FORMAT

- Flat file, strictly column oriented, 80 character records
- Human readable
- **Human editable**



80-character punch card. (From Wikipedia)

Flat File: A datafile without indexing structure or hierarchy. In contrast, to *relational database*, or *data grammar*.

The "PDB" file format is the most widely used format to exchange structure data. It is a keyword controlled, column-oriented, fixed-width flat file format, derived from the information processing constraints of computer punch cards.

For decades, some researchers have been advocating for a more rigorous and flexible XML based data gramar, but the existing solutions have never been widely adopted.

PDB FILE FORMAT COMPONENTS: HEADER

```
HEADER    IMMUNOGLOBULIN                          01-MAR-93   2IMM      2IMM   2
COMPND    IMMUNOGLOBULIN VL DOMAIN (VARIABLE DOMAIN OF KAPPA LIGHT      2IMM   3
COMPND   2 CHAIN) OF MCPC603                                            2IMM   4
SOURCE    HUMAN (HOMO $SAPIENS) RECOMBINANT SYNTHETIC M603 GENE         2IMM   5
AUTHOR    B.STEIPE,R.HUBER                                              2IMM   6
REVDAT   1   15-JUL-93 2IMM    0                                        2IMM   7
REMARK   1                                                              2IMM   8
REMARK   1 REFERENCE 1                                                  2IMM   9
REMARK   1  AUTH   B.STEIPE,A.PLUCKTHUN,R.HUBER                         2IMM  10
REMARK   1  TITL   REFINED CRYSTAL STRUCTURE OF A RECOMBINANT           2IMM  11
REMARK   1  TITL 2 IMMUNOGLOBULIN DOMAIN AND A                          2IMM  12
REMARK   1  TITL 3 COMPLEMENTARITY-DETERMINING REGION 1-GRAFTED MUTANT  2IMM  13
REMARK   1  REF    J.MOL.BIOL.                   V. 225   739 1992      2IMM  14
REMARK   1  REFN   ASTM JMOBAK  UK ISSN 0022-2836                 070   2IMM  15

    [...]

REMARK   2                                                              2IMM  23
REMARK   2 RESOLUTION. 2.00  ANGSTROMS.                                 2IMM  24
REMARK   3                                                              2IMM  25

   [...]
```

The header section of a PDB file.

PDB FILE FORMAT COMPONENTS: SEQUENCE

```
[...]
SEQRES   1     114  ASP ILE VAL MET THR GLN SER PRO SER SER LEU SER VAL  2IMM  35
SEQRES   2     114  SER ALA GLY GLU ARG VAL THR MET SER CYS LYS SER SER  2IMM  36
SEQRES   3     114  GLN SER LEU LEU ASN SER GLY ASN GLN LYS ASN PHE LEU  2IMM  37
SEQRES   4     114  ALA TRP TYR GLN GLN LYS PRO GLY GLN PRO PRO LYS LEU  2IMM  38
SEQRES   5     114  LEU ILE TYR GLY ALA SER THR ARG GLU SER GLY VAL PRO  2IMM  39
SEQRES   6     114  ASP ARG PHE THR GLY SER GLY SER GLY THR ASP PHE THR  2IMM  40
SEQRES   7     114  LEU THR ILE SER SER VAL GLN ALA GLU ASP LEU ALA VAL  2IMM  41
SEQRES   8     114  TYR TYR CYS GLN ASN ASP HIS SER TYR PRO LEU THR PHE  2IMM  42
SEQRES   9     114  GLY ALA GLY THR LYS LEU GLU LEU LYS ARG              2IMM  43
```

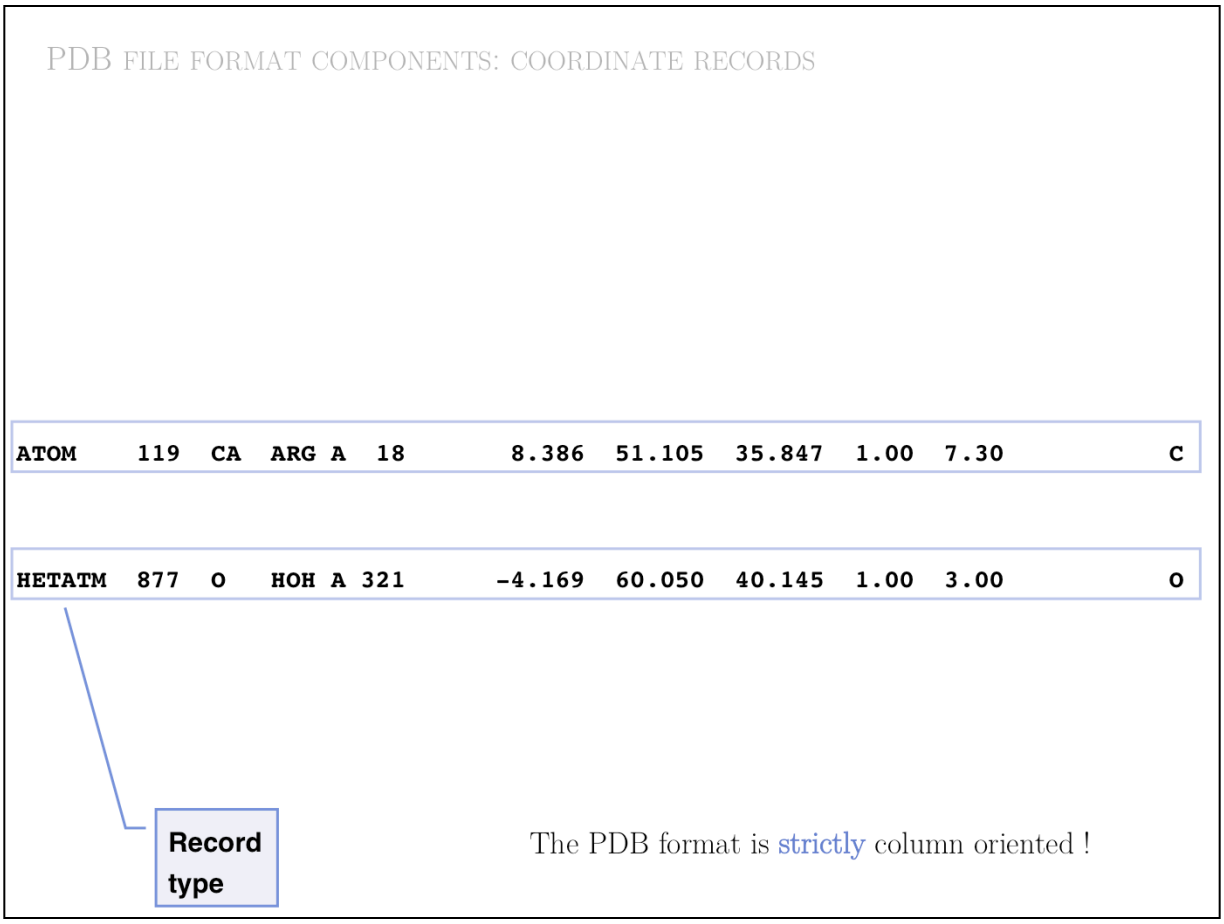
Explicit (*above*) and implicit (*from coordinates*) sequence may differ !

Always be extremely cautious with sequence numbers! There is no well defined starting point for numbering sequence!! The numbers you count in a FASTA file and the numbers you read in a PDB record may be totally different!!!

The ***explicit*** sequence is found in the SEQRES section of the PDB file. This corresponds to what the crystallographer put into the experiment and it is in general the sequence that is recorded as the PDB-sequence for the purpose of database searches.

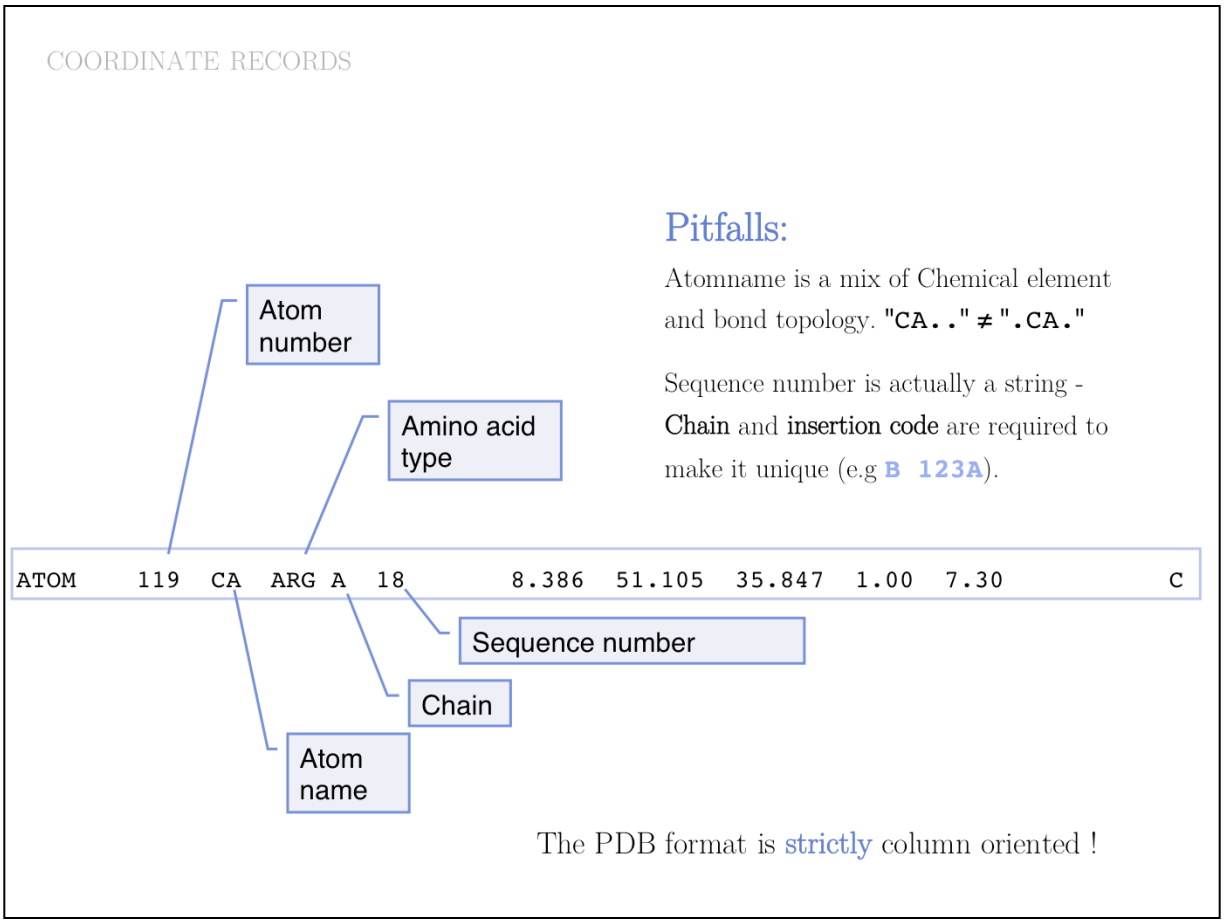
The ***implicit*** sequence can be derived from the actual coordinates. It is frequently not the same because flexible, disordered termini and loops are not recorded.

Additional complications arise from "insertion codes". These are letters that allow the insertion of residues for a common numbering scheme for families of homologous sequences. In principle this is a good idea, since this makes comparison of residues much easier. But strings such as "23A" can no longer be treated as sequence *numbers* - they are sequence *labels* and using them correctly can be a challenge. The existence of insertion codes breaks common assumptions that novice programmers make about the structure of PDB files.



The coordinate records contain the actual 3D-data. Their structured is the same for ATOM (protein and DNA components) and HETATM records (ligands, cofactors, solvent etc.).

Read the Coordinate section of the PDB format specification.
(http://www.wwpdb.org/documentation/file-format)



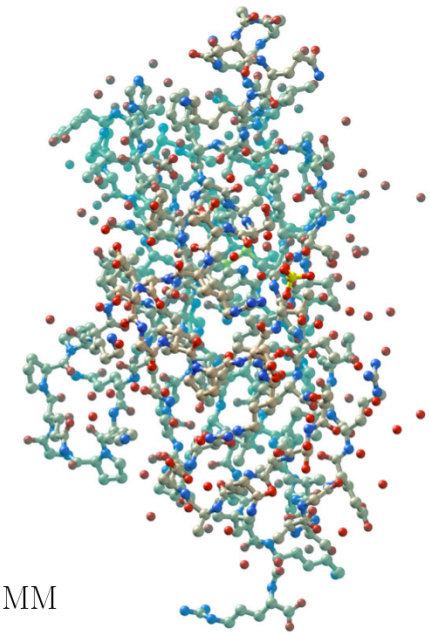
Potential pitfalls (very incomplete list):

- Record type: changes not consistently applied for modifications
- Atom number: rarely used and a nuisance to update when changing.
- Atom name: be careful about columns " CA " is C^{α}, but "CA " is calcium. Usually this is obvious from context though.
- Amino acid type: careful about non-natural amino-acids e.g. selenocysteine, selenomethionine etc. Some very old files use TRY for TRP.
- Chain: may be blank (" ") in older files. Recently was changed to ("A") even in files that contain only a single chain.
- Alternate location: only sometimes given in very high resolution structures.
- Sequence number might not be unique if insertion codes are used.
- X,Y, and Z are given in Å; (10^{-10}m = 0.1 nm) values in a cartesian (i.e. orthogonal) coordinate system; but origin and orientation is arbitrary!
- Occupancy can describe: special locations, partially bound ligands, unobserved fragments of structure...
- B-values, (also called temperature factors) are a measure of the volume of space around into which a given electron density is distributed. Usually considered to be isotropic!

COORDINATE RECORDS

X Y Z

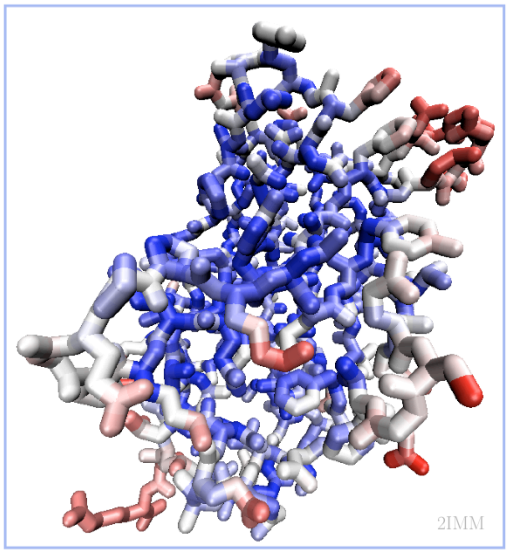
```
ATOM    119  CA  ARG A  18       8.386  51.105  35.847  1.00  7.30           C
```



2IMM

XYZ coordinates are vectors in an orthogonal coordinate system, in Å (0.1nm) – with arbitrary origin and orientation

COORDINATE RECORDS



B

```
ATOM    119  CA  ARG A  18       8.386  51.105  35.847  1.00  7.30           C
```

Occupancy

$$B = 8\pi^2 \left\langle \vec{x}^2 \right\rangle \qquad B = 80; \vec{x} \approx 1.0 \text{Å}$$

Occupancy describes the fraction of sites that are occupied by that atom in the crystal. Occupancy can be less than 1.0 for ligands of partially occupied sites, or for atoms in special crystallographic positions – e.g. on a two-fold axis of symmetry. Modelled coordinates that are not supported by electron density from the experiement and have only been added for cosmetic reasons are usually given an occupancy of 0.0

B factors describe the thermal fluctation of atoms and can be interpreted as a mean coordinate displacement IF the thermal fluctuation were isotropic (the same in each direction). A B factor of 80 is approximately what you would expect of bulk, disordered solvent, it corresponds to a mean displacement of approximately one Å. B factors are commonly very low in the well-ordered and conformationally constrained core of a protein, they are high at the surface, esepcially for side-chains and flexible loops.

The interpretation of B factors is not entirely straightforward, because poor quality crystals can give rise to higher values, actual movement in proteins is anything but isotropic, and static disorder / alternate conformations are modelled through increased B factors as well.

A PDB File Contains ....

Contents of a PDB file:

- Polypeptide Chain(s)
- Nucleic Acid(s)
- Heteroatoms e.g. prosthetic groups, cofactors, ligands ...
- Typically, H-Atoms are invisible ...

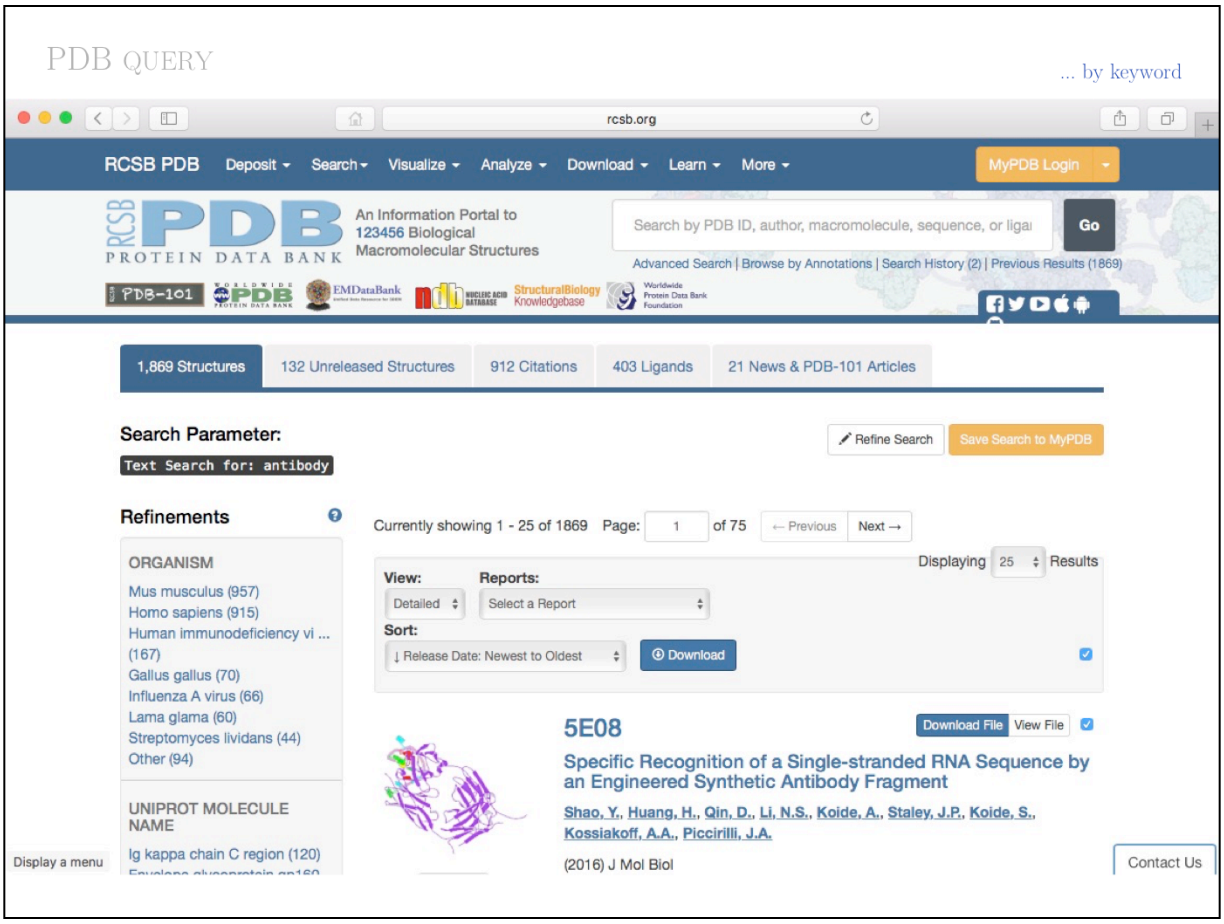
Structures are in a crystal lattice, have crystal contacts.

Asymmetric units are not guaranteed to be biological units.

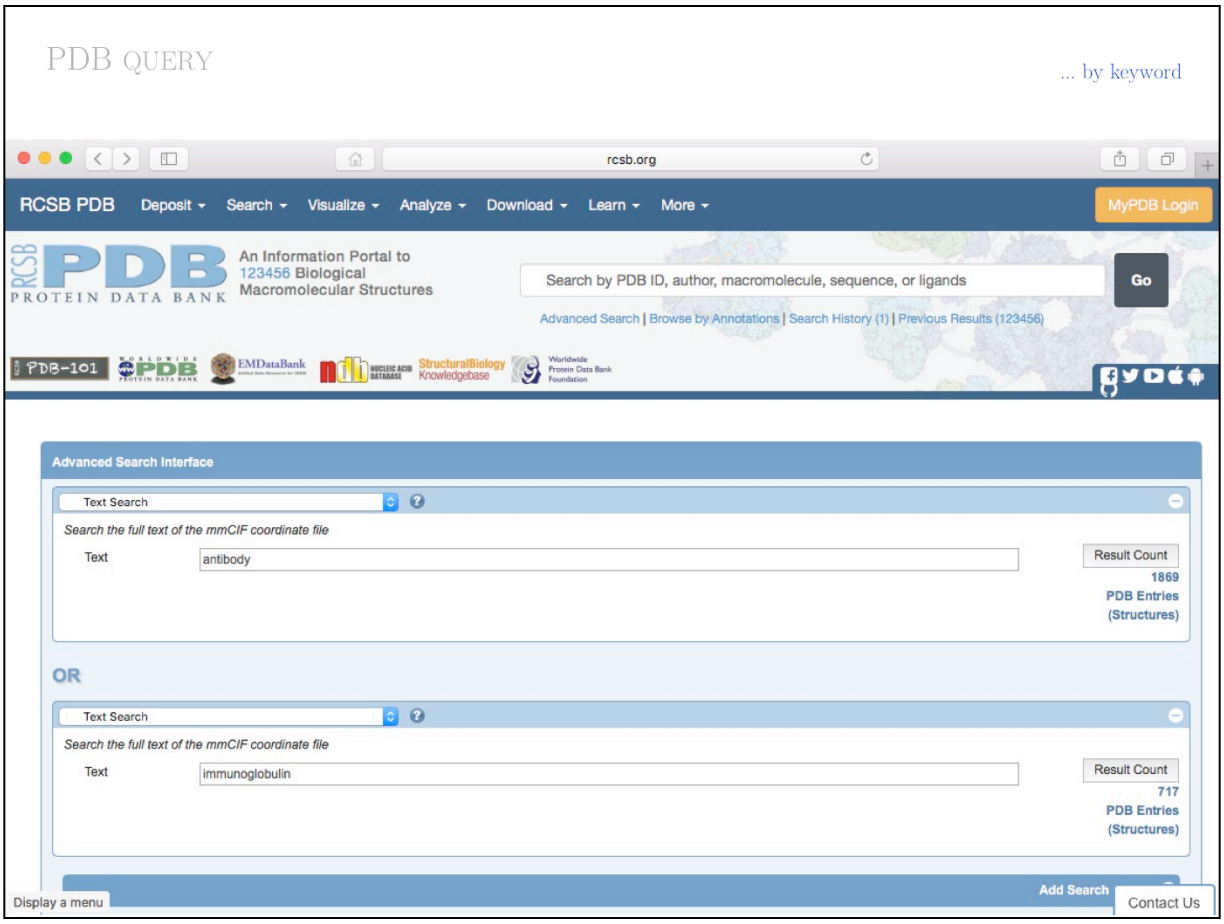
Origin of the coordinate system and rotation is arbitrary.

Coordinates are time- and population- averaged experiments.

The most straightforward access to a coordinate set in the PDB is to enter the four-character PDB ID into the search field. PDB IDS have a digit as the first character, then three (letter/digit)s. Thus the keyspace is 10x36x36x36 = 466,560. At the time the PDB was created with seven structures in 1971, each of which took years to complete in a herculean effort, no one could foresee that we would run out of IDs! Lesson learned.



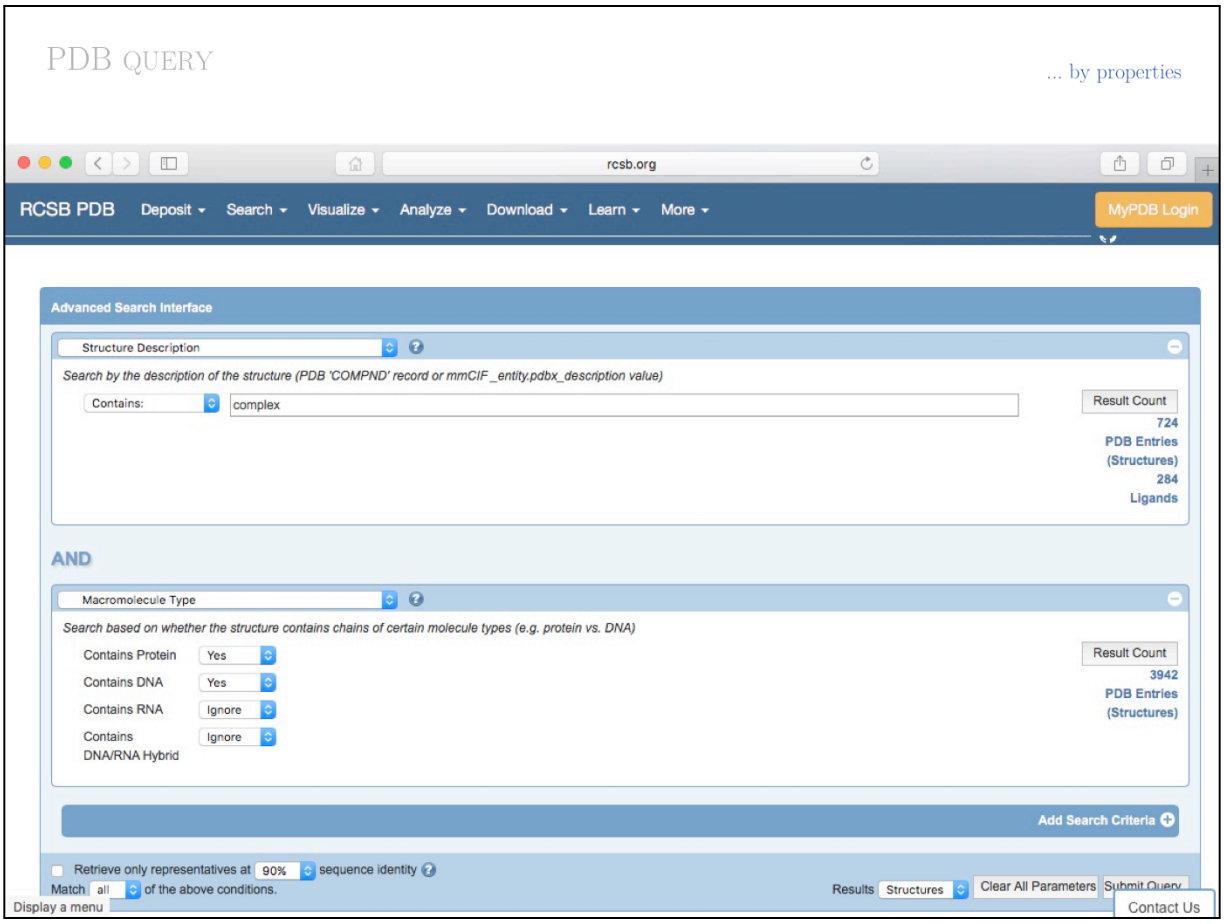
Keyword searches are useful, but unreliable. Antibody? Immunoglobulin? Fab fragment? Bence-Jones protein?



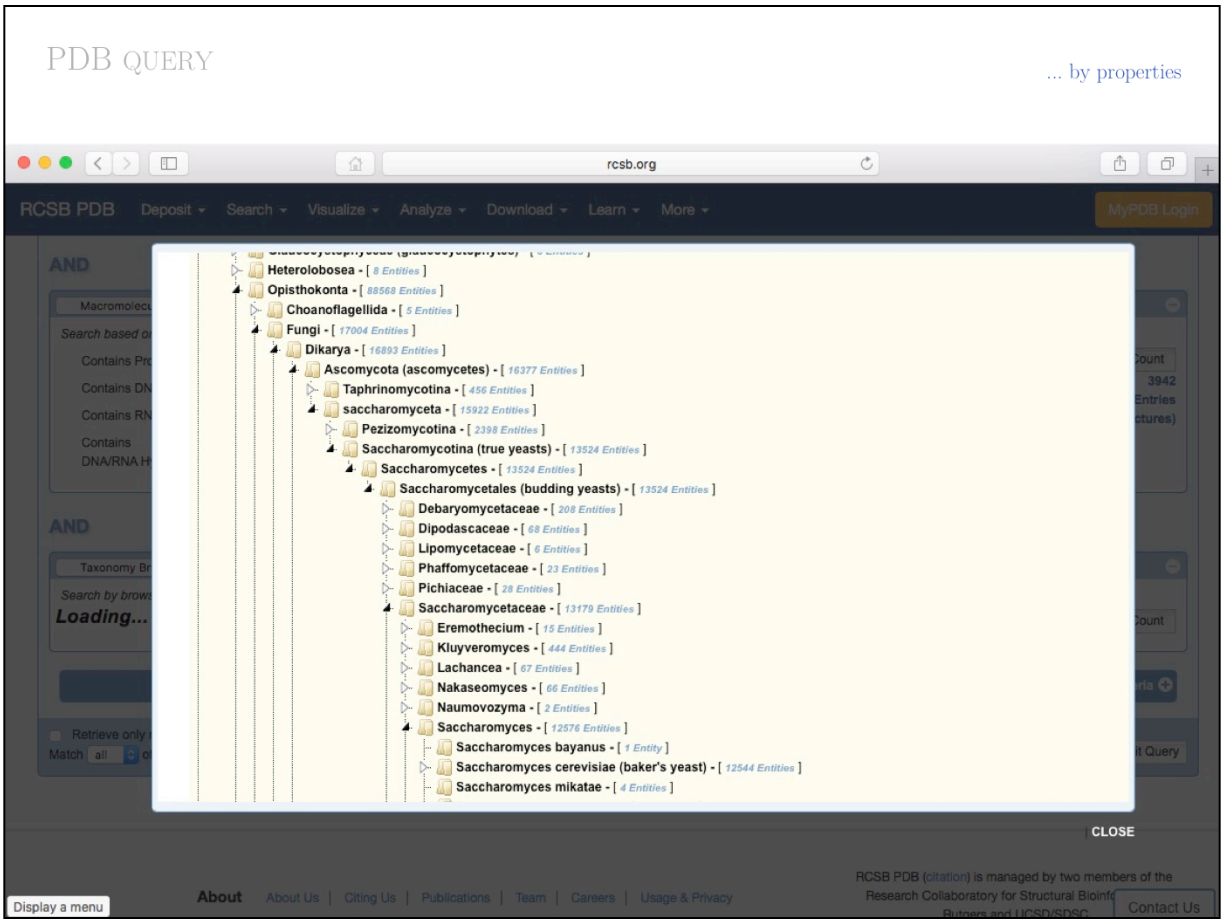
There are 1,869 structures annotated with they keyword “antibody” and 717 structures annotated with “immunoglobulin”.

Note that the number of results – 2,399 structures – is not the same as the sum of the two individual keyword searches. 1,869 "antibody" structures plus 717 "immunoglobulin" structures would have given 2,586 structures.

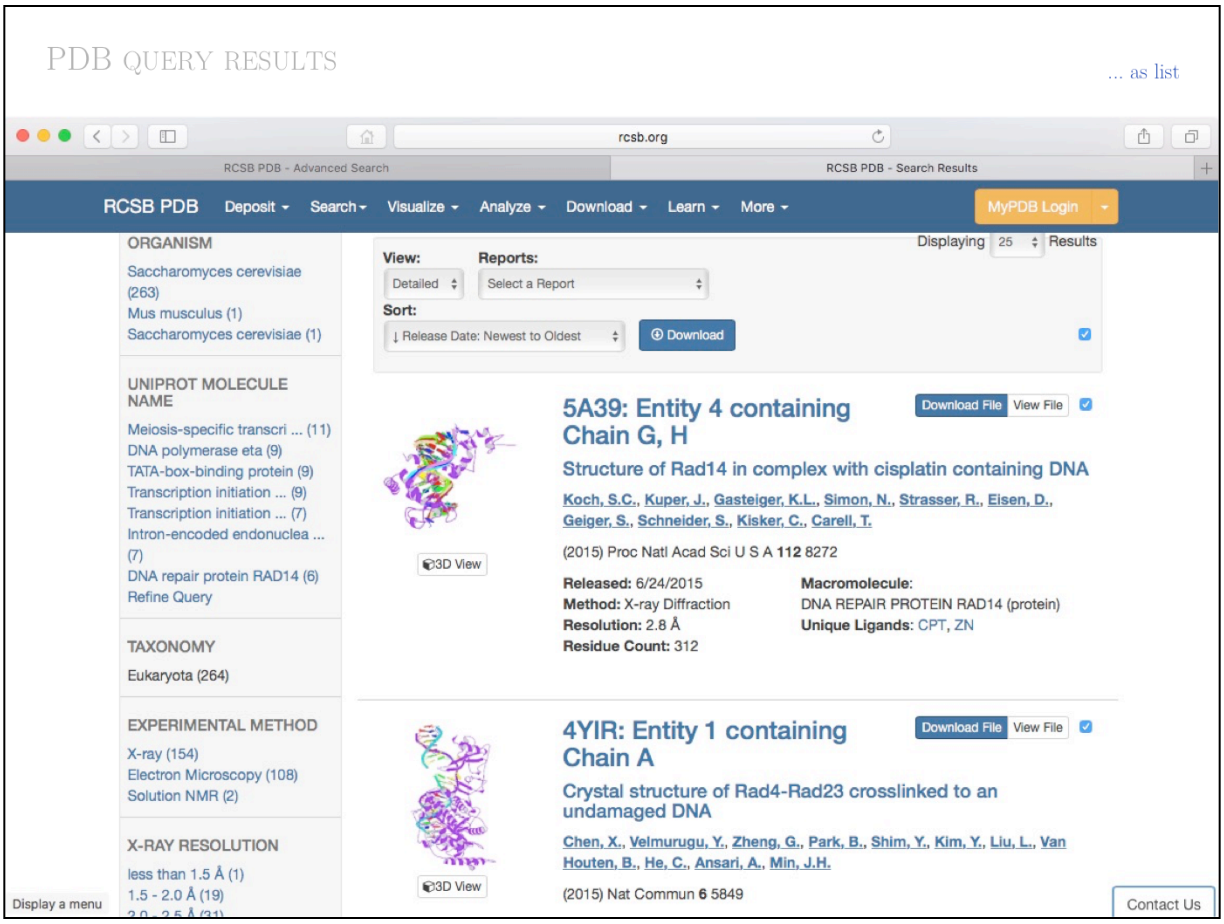
Keyword searches are good to find *some* structures, but they will virtually never find *all* structures of a desired type.



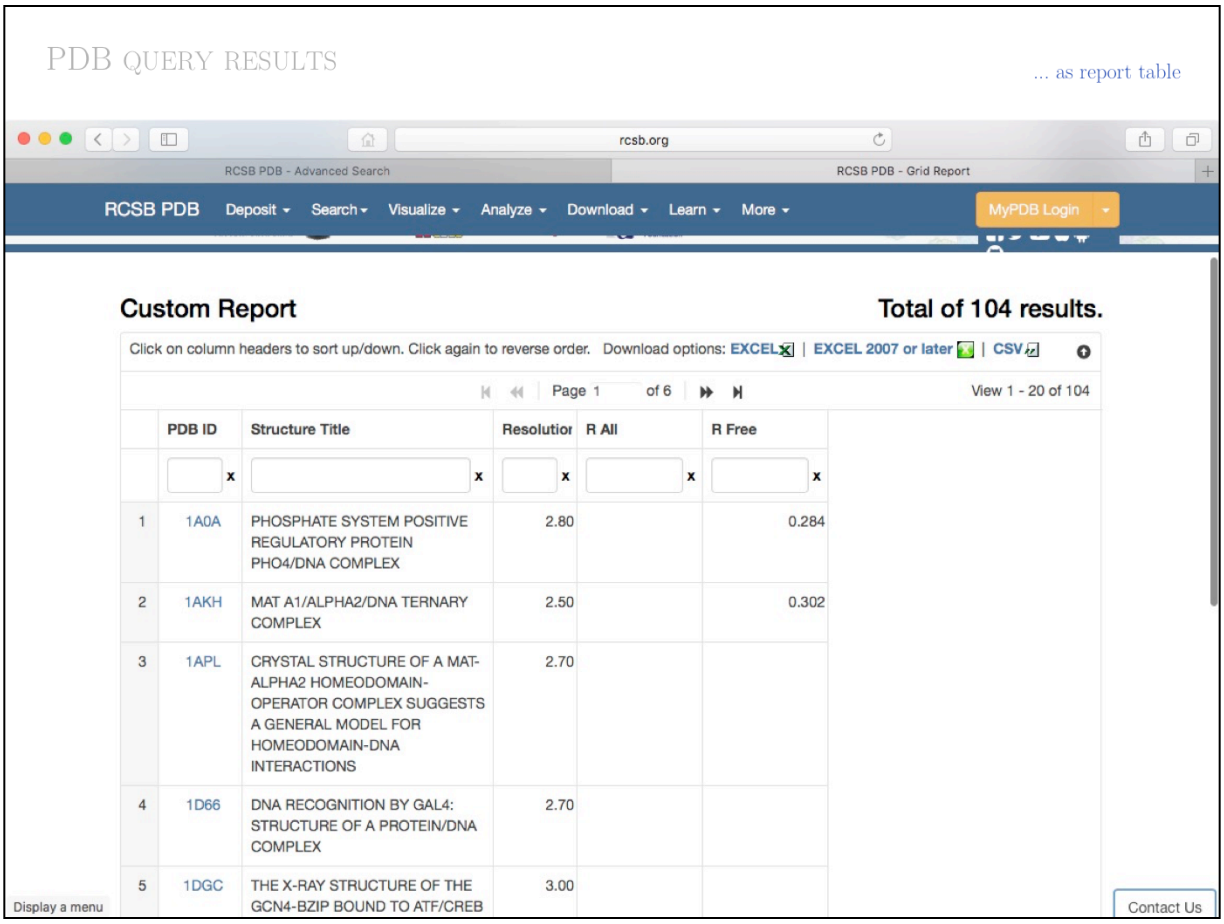
The Advanced Query interface is a flexible and powerful search tool for more sophisticated queries. Here a search for potential transcription factor/DNA complexes.



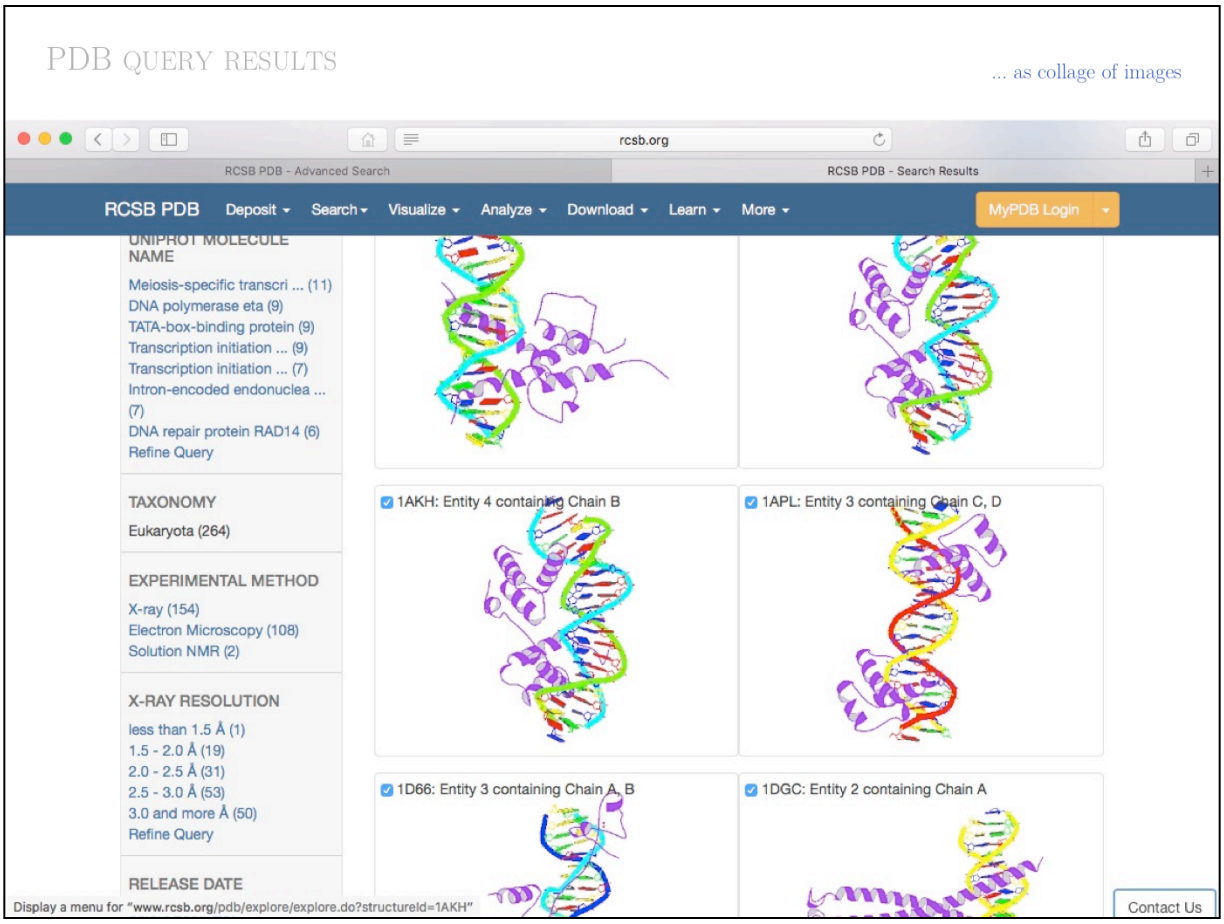
You can search for structures from specific species via the taxonomy browser.



Output can be in list format ...



... customizable tables ...



... or a gallery overview that allows you to quickly scan for the type of structure you want to consider.

LIMITATIONS

In principle, structures can be obtained at **atomic resolution**.
This means we can identify the location of individual water molecules!

In practice, structures are **time-averaged and population averaged**.
We see atomic resolution only for well-ordered atoms.

Experimental artefacts are possible, in particular crystal packing artefacts.

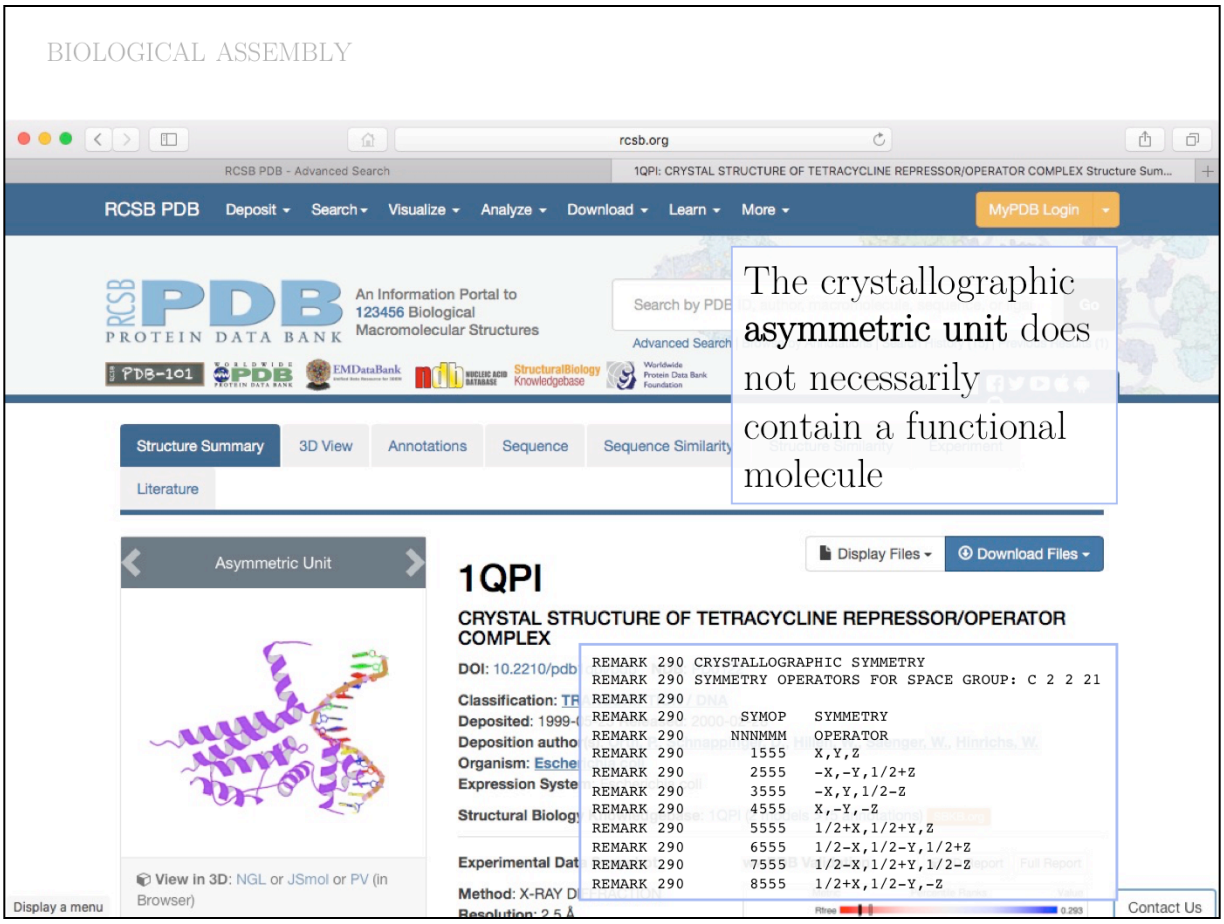
Experimental errors are possible, in particular when modeling disordered structure (B-factors!); *fraud is uncommon but has occurred at times.*

All (refined) structures combine experimental information with idealized stereochemistry.

Fortunately, experience shows that structures obtained with different experimental methods are very similar.

Quality metrics exist, but are not trivial to interpret.

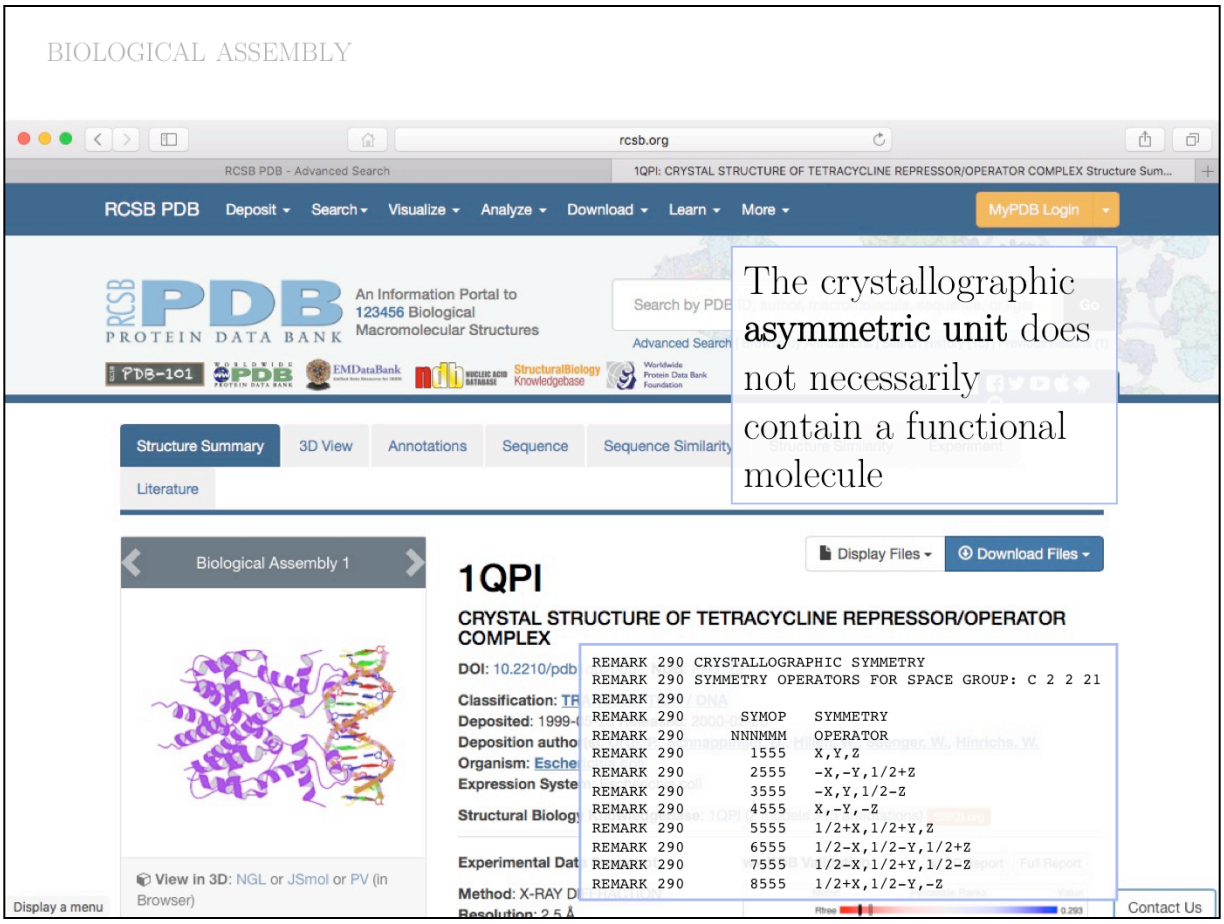
Biological relevance has to be taken into account.



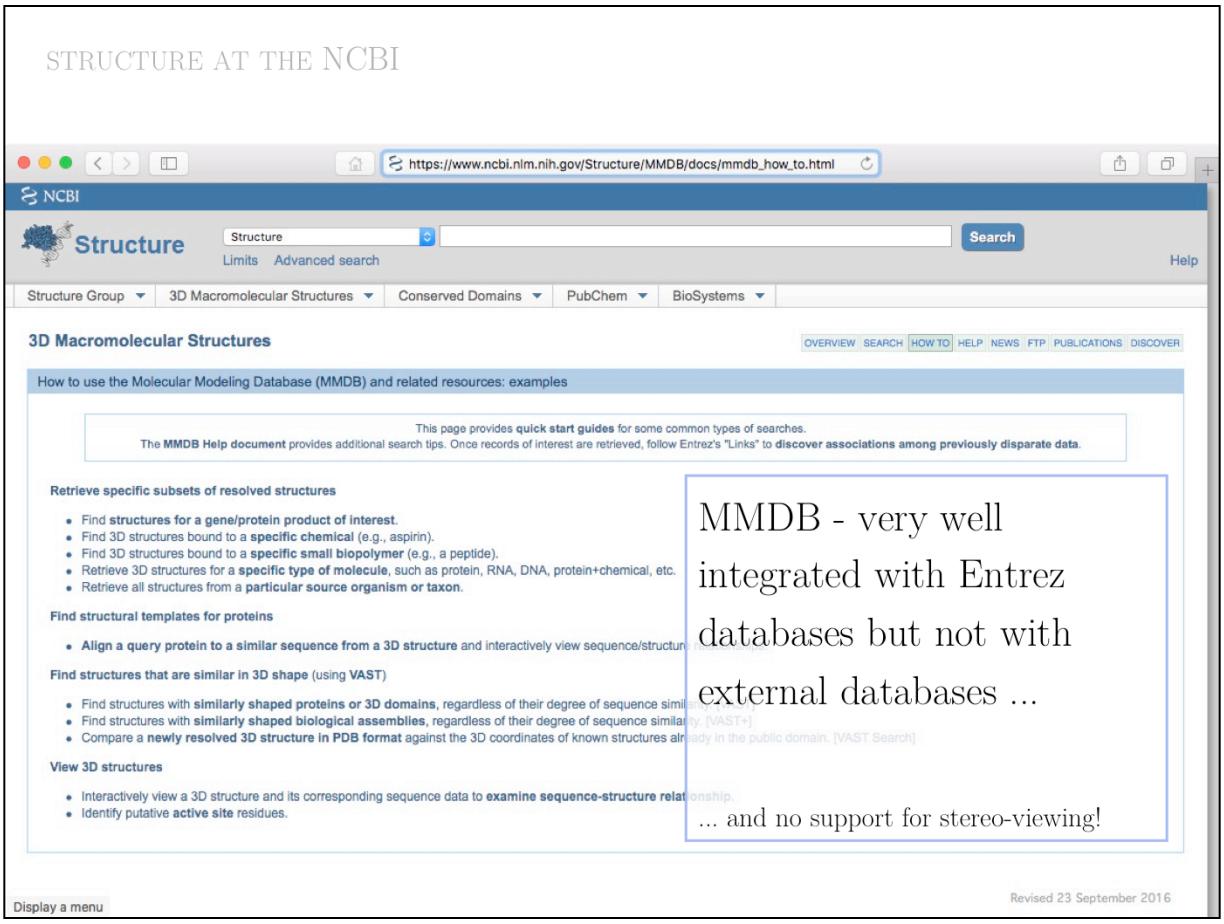
In the example above, only one chain of the tet-repressor dimer is seen bound to only one strand of B-DNA. The second chain and strand can be generated through a symmetry operation (180° rotation and translation), thus it contains the same coordinate information and does not need to be separately stored.

However, in order to study the functional molecule, the redundant coordinates have to be combined to a homodimer. The term **biological unit** describes a coordinate set that (presumably) depicts a homooligomer in its *functional* state.

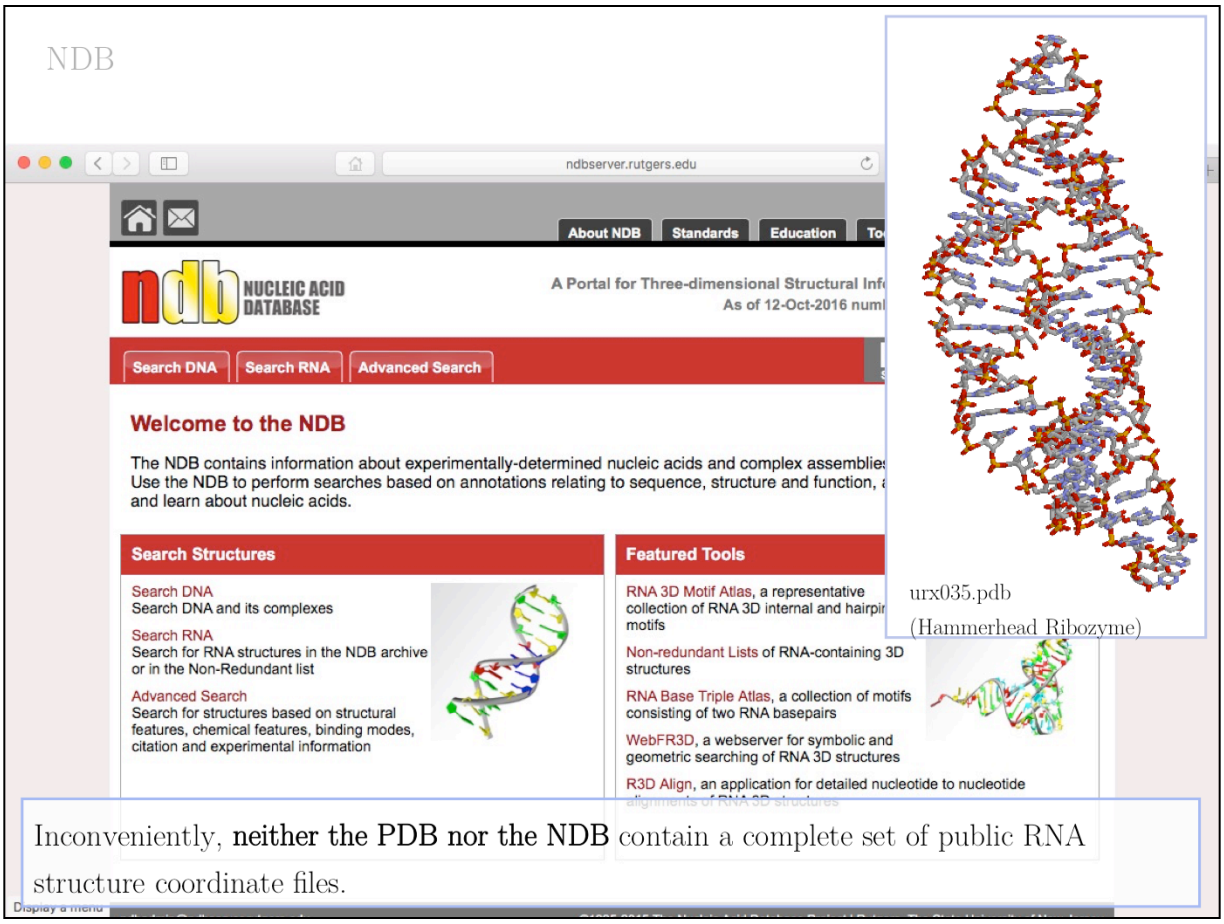
All molecules in the crystal lattice can be generated from the crystallographic symmetry operations specified in a PDB file, for the space group of the crystal. But it may not be obvious which of the symmetry replicates might actually be involved in a physiological interaction and which ones have only been induced by the crystallization process. In the tet-repressor example, the crystallographic space group has eight(!) symmetry related monomers in the unit cell of the crystal lattice.



The "Biological Assembly" complements the single chain to a complete repressor homodimer/double-stranded DNA complex.



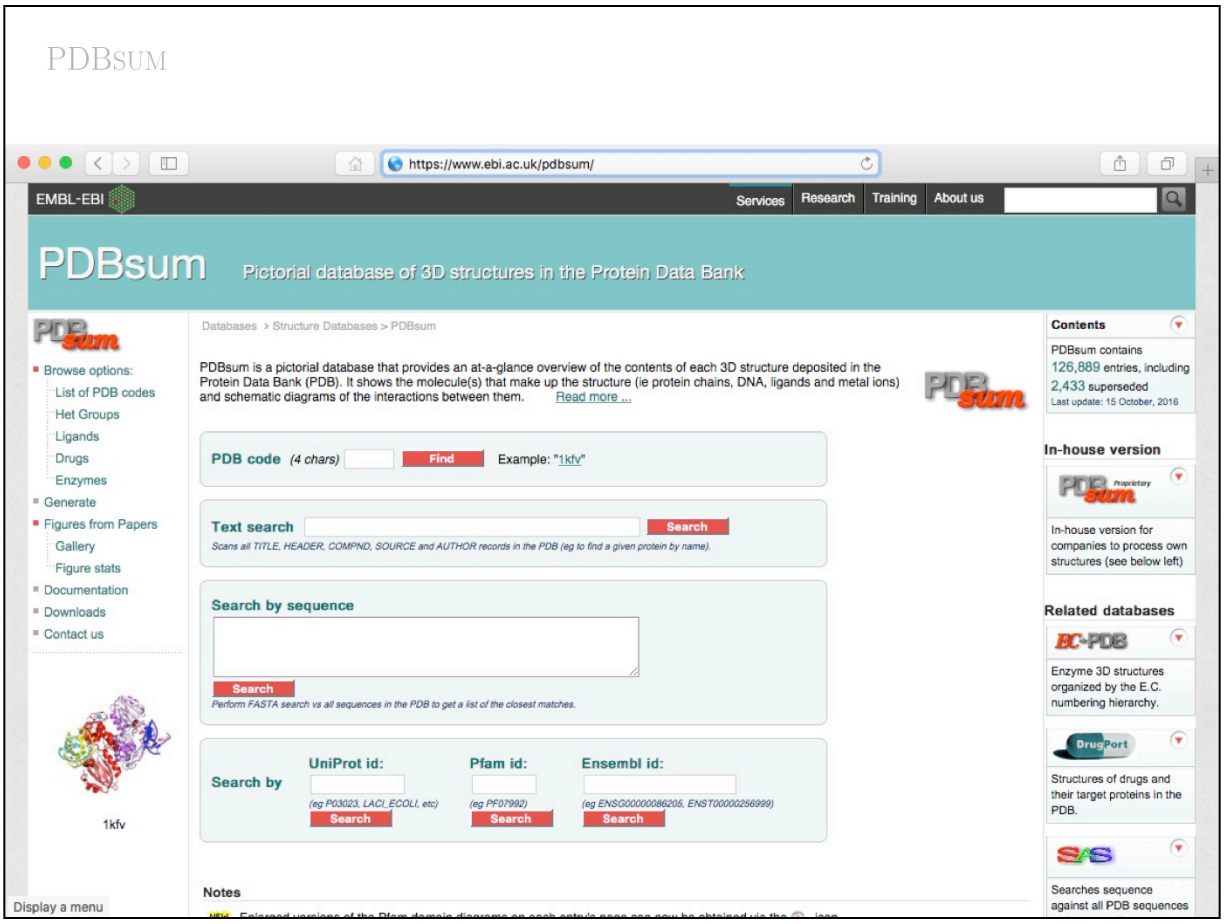
Structure services exist at the NCBI and the EBI with distinct services. Explore.



The Nucleic Acid Structure Database

http://ndbserver.rutgers.edu/

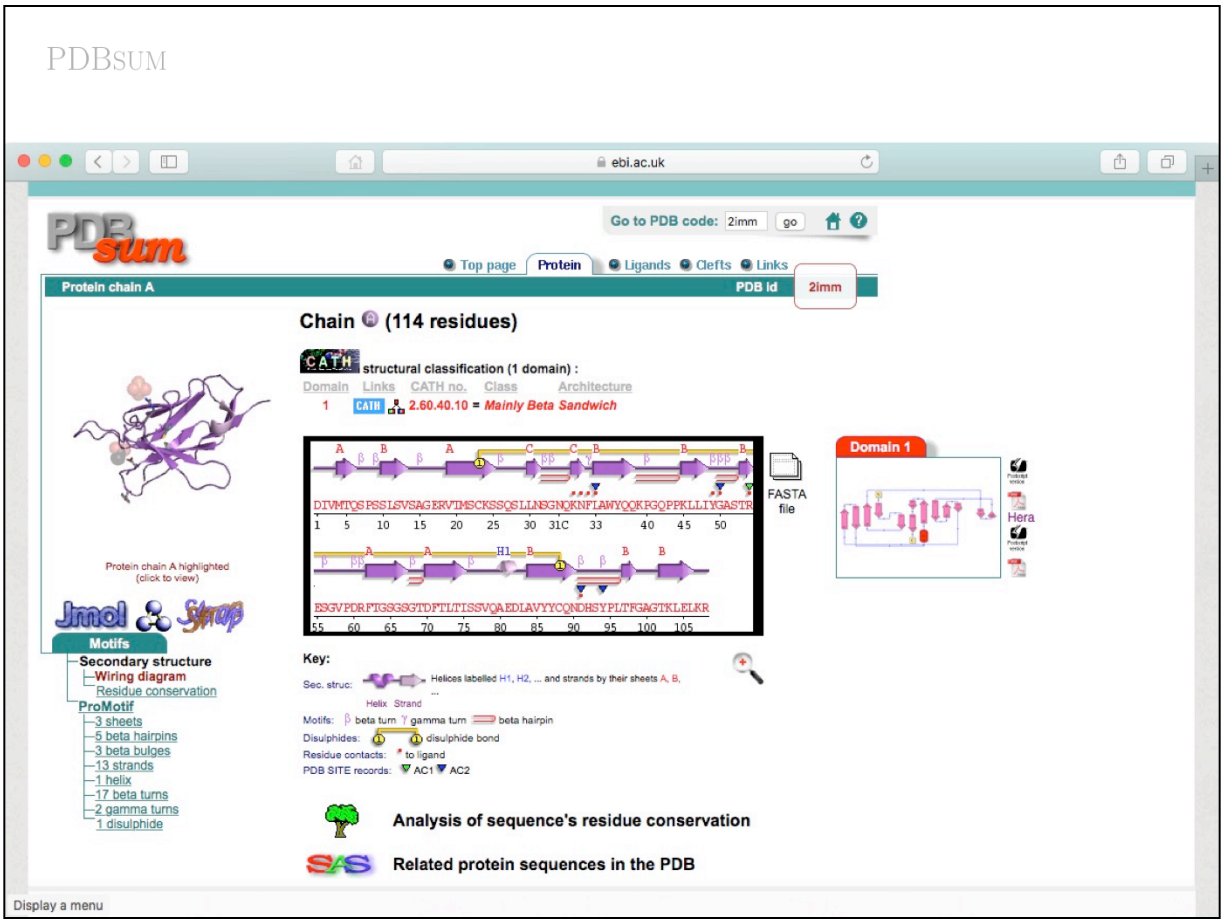
PDBsum



PDBsum is a secondary database that stores analysis and interpretation information for PDB coordinate sets.

http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/

PDBsum



Schematics like this "wiring diagram" help to get an overview of protein topology and sequence organization. I find the high-level descriptions in PDBsum very useful to begin the study of a particular protein.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA