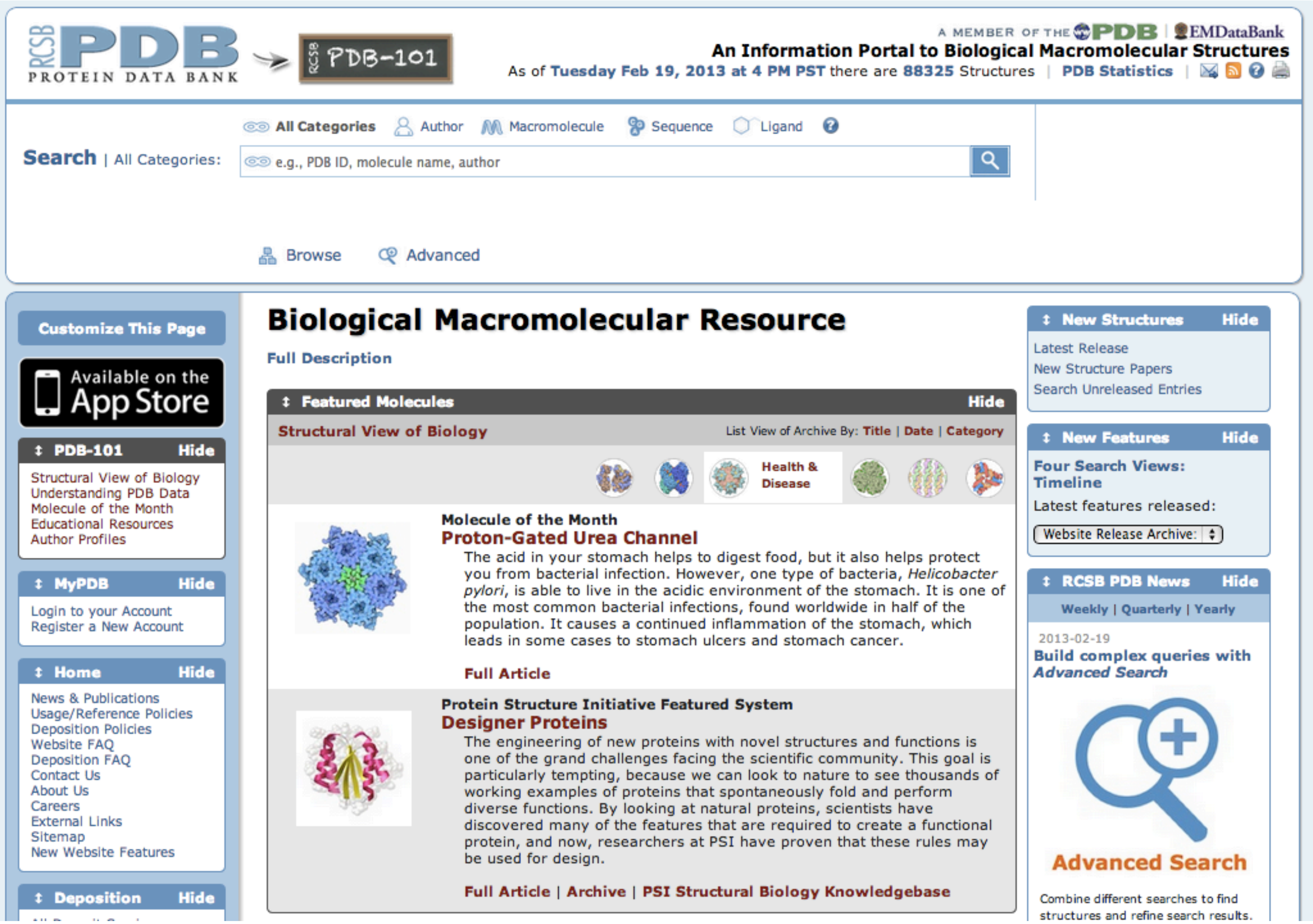
COMP364: Biopython part II

Jérôme Waldispühl, McGill University

Protein Data Bank (PDB)

http://www.rcsb.org



RCSB PDB PROTEIN DATA BANK → PDB-101

A MEMBER OF THE PDB | EMDataBank

An Information Portal to Biological Macromolecular Structures

As of **Tuesday Feb 19, 2013 at 4 PM PST** there are **88325** Structures | **PDB Statistics**

All Categories Author Macromolecule Sequence Ligand

Search | All Categories: e.g., PDB ID, molecule name, author

Browse Advanced

Customize This Page

Available on the App Store

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

Home Hide

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

Deposition Hide

Biological Macromolecular Resource

Full Description

Featured Molecules Hide

Structural View of Biology List View of Archive By: **Title** | **Date** | **Category**

Health & Disease

Molecule of the Month

Proton-Gated Urea Channel

The acid in your stomach helps to digest food, but it also helps protect you from bacterial infection. However, one type of bacteria, *Helicobacter pylori*, is able to live in the acidic environment of the stomach. It is one of the most common bacterial infections, found worldwide in half of the population. It causes a continued inflammation of the stomach, which leads in some cases to stomach ulcers and stomach cancer.

Full Article

Protein Structure Initiative Featured System

Designer Proteins

The engineering of new proteins with novel structures and functions is one of the grand challenges facing the scientific community. This goal is particularly tempting, because we can look to nature to see thousands of working examples of proteins that spontaneously fold and perform diverse functions. By looking at natural proteins, scientists have discovered many of the features that are required to create a functional protein, and now, researchers at PSI have proven that these rules may be used for design.

Full Article | **Archive** | **PSI Structural Biology Knowledgebase**

New Structures Hide

Latest Release
New Structure Papers
Search Unreleased Entries

New Features Hide

Four Search Views: Timeline

Latest features released:

Website Release Archive:

RCSB PDB News Hide

Weekly | **Quarterly** | **Yearly**

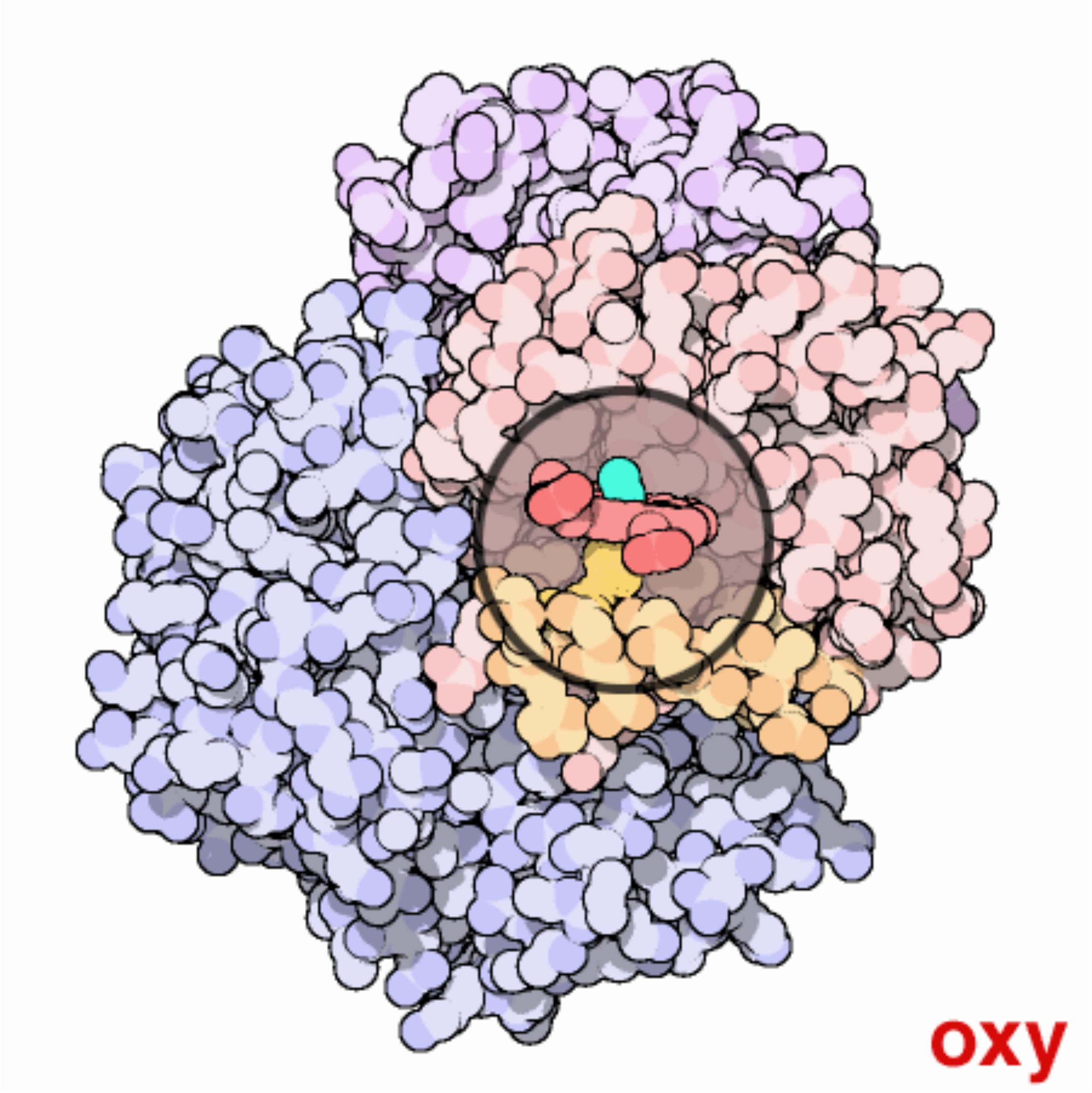
2013-02-19

Build complex queries with Advanced Search

Advanced Search

Combine different searches to find structures and refine search results.

Why Structures?



Facts about the PDB

What can I find in the PDB?

- Protein Structures determined by:
  - o Crystallography
  - o Nuclear Magnetic Resonance
  - o Theoretical Models with or without partial data
- RNA & DNA structures

How are the data stored?

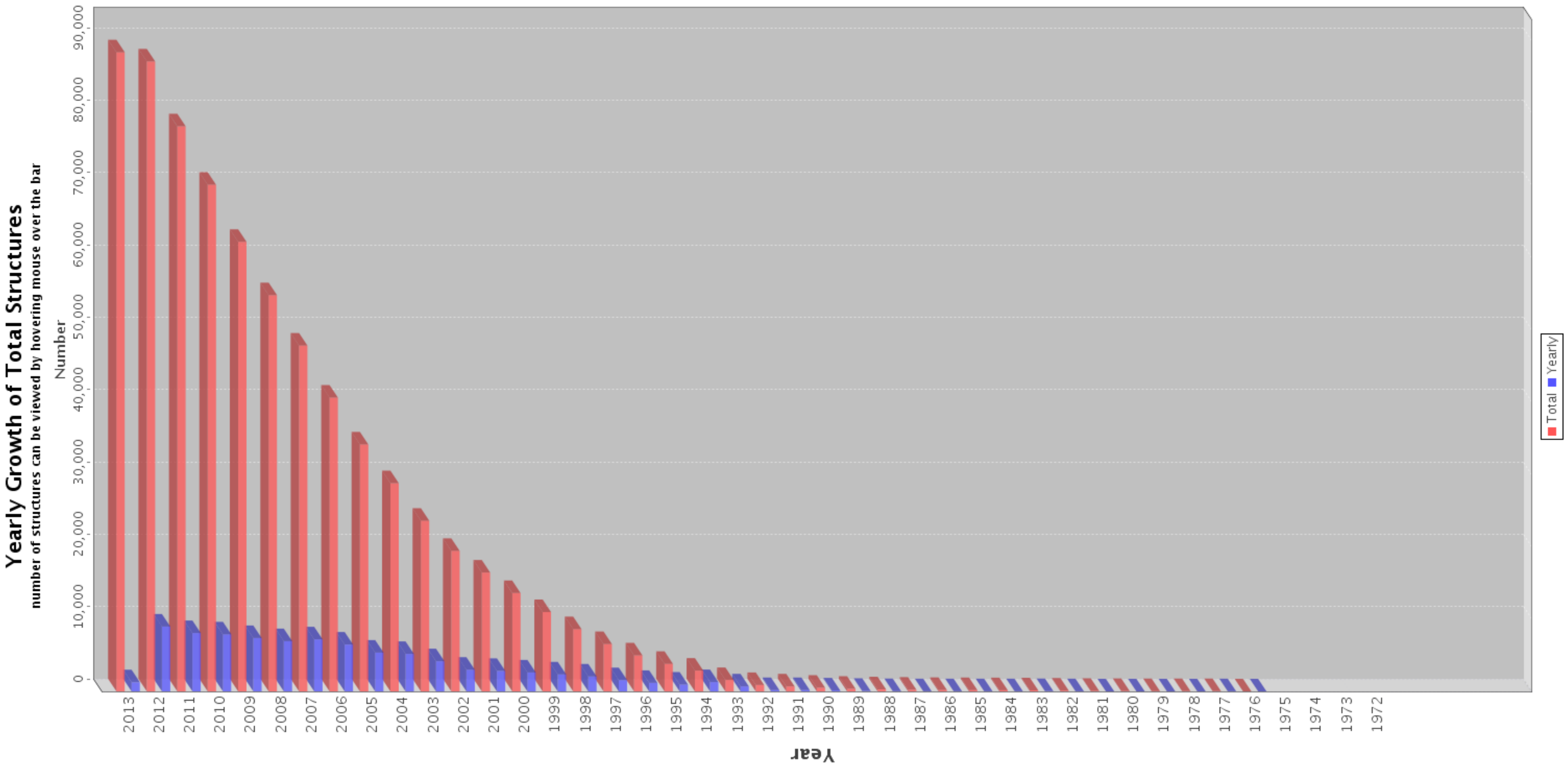
The structures are stored using a fixed-column format using the extension .pdb

What is a PDB id?

An entry number is assigned to each structure. Typically it is a number followed by 3 letters (E.g. 2POR).

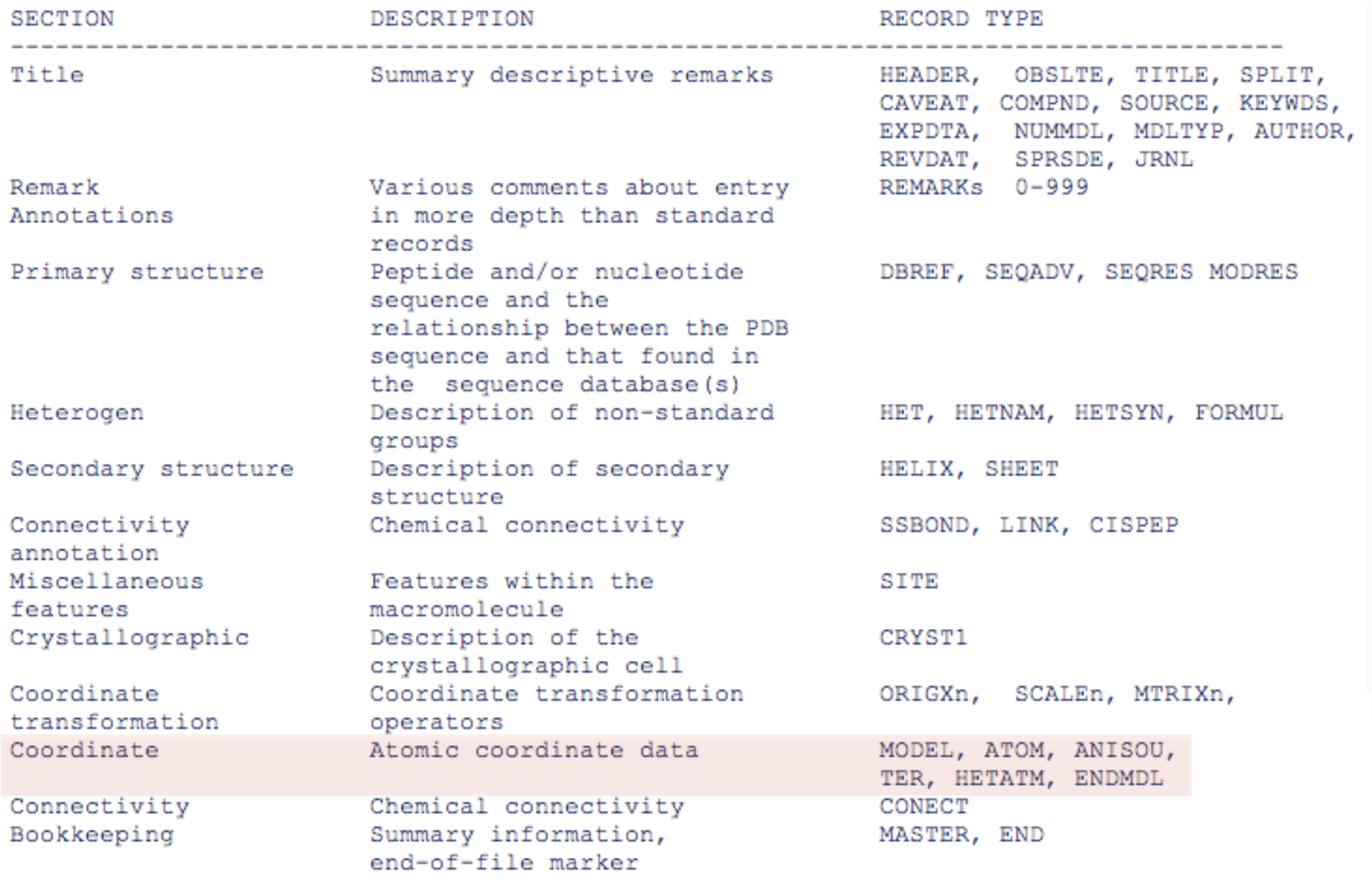
N.B.: The same molecule can have multiple entries.

PDB growth



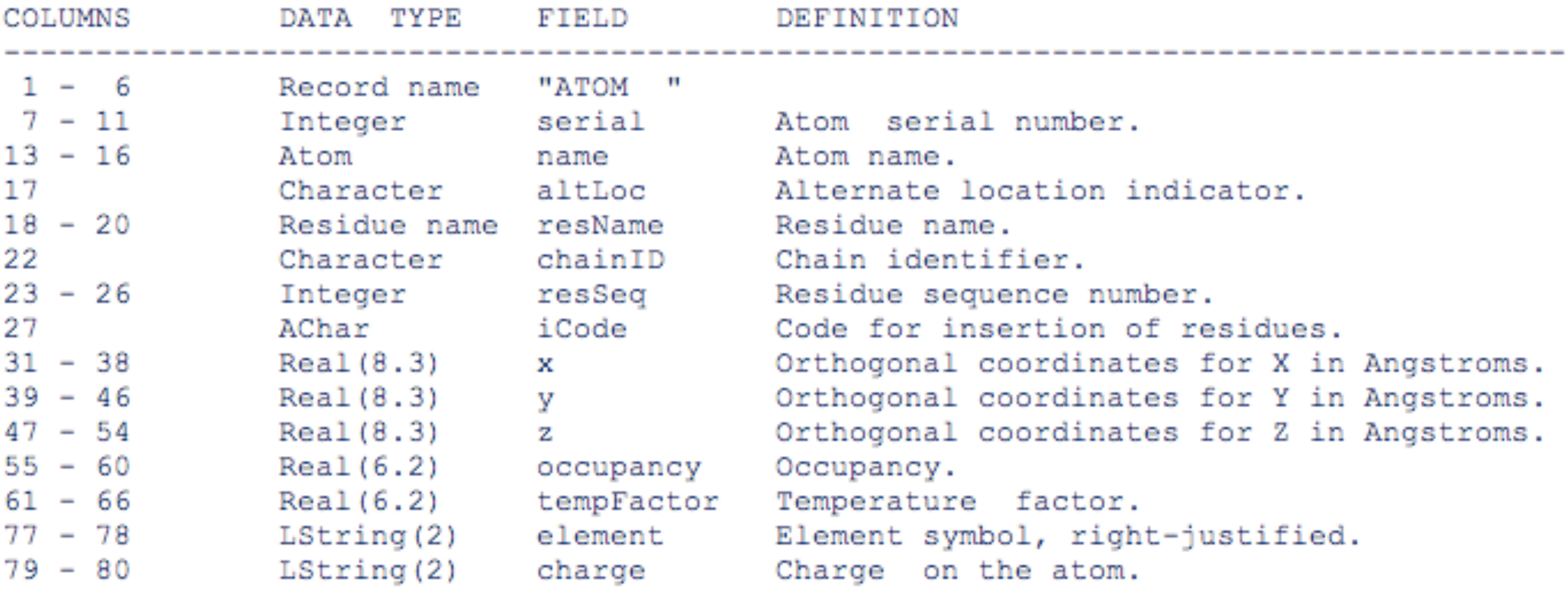
As of Tuesday Feb 19, 2013 at 4 PM PST there are 88325 Structures.

PDB file format



SECTION	DESCRIPTION	RECORD TYPE
Title	Summary descriptive remarks	HEADER, OBSLTE, TITLE, SPLIT, CAVEAT, COMPND, SOURCE, KEYWDS, EXPDTA, NUMMDL, MDLTYP, AUTHOR, REVDAT, SPRSDE, JRNL
Remark Annotations	Various comments about entry in more depth than standard records	REMARKs 0-999
Primary structure	Peptide and/or nucleotide sequence and the relationship between the PDB sequence and that found in the sequence database(s)	DBREF, SEQADV, SEQRES MODRES
Heterogen	Description of non-standard groups	HET, HETNAM, HETSYN, FORMUL
Secondary structure	Description of secondary structure	HELIX, SHEET
Connectivity annotation	Chemical connectivity	SSBOND, LINK, CISPEP
Miscellaneous features	Features within the macromolecule	SITE
Crystallographic	Description of the crystallographic cell	CRYST1
Coordinate transformation	Coordinate transformation operators	ORIGXn, SCALEn, MTRIXn,
Coordinate	Atomic coordinate data	MODEL, ATOM, ANISOU, TER, HETATM, ENDMDL
Connectivity	Chemical connectivity	CONECT
Bookkeeping	Summary information, end-of-file marker	MASTER, END

Syntax of ATOM rows



COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms.
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms.
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms.
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

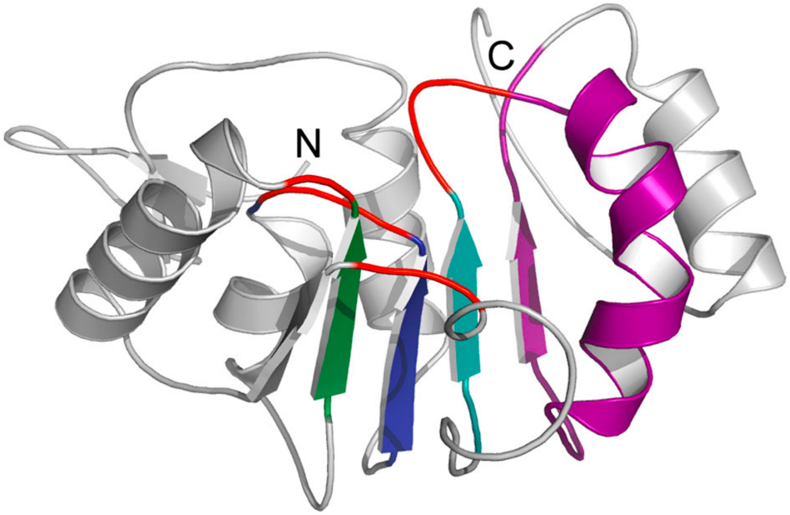
- Column-fixed format
- Derived in the 70’s from X-ray & NMR data format.

Syntax of ATOM rows

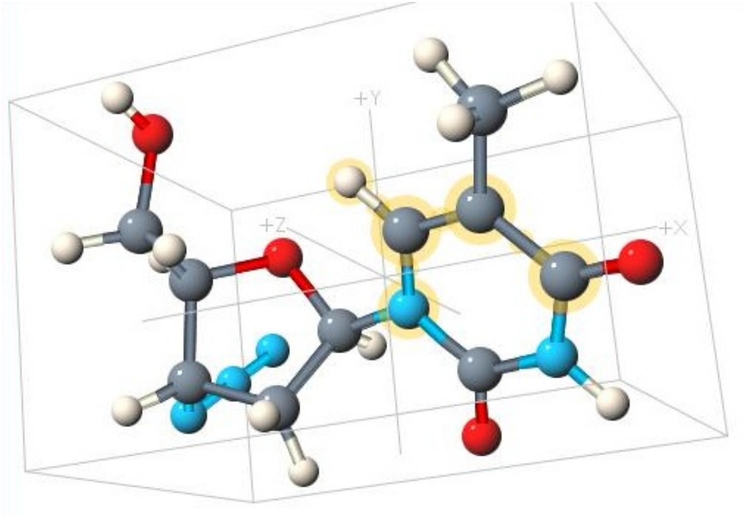
```
ATOM      1  N   MET A   1      10.263  -7.566  -4.747  1.00 47.36     N
ATOM      2  CA  MET A   1       9.077  -7.905  -5.617  1.00 47.69     C
ATOM      3  C   MET A   1       9.155  -9.333  -6.212  1.00 47.89     C
ATOM      4  O   MET A   1      10.028  -9.649  -7.048  1.00 48.03     O
ATOM      5  CB  MET A   1       8.869  -6.852  -6.731  1.00 47.38     C
ATOM      6  CG  MET A   1       7.608  -7.091  -7.622  1.00 47.57     C
ATOM      7  SD  MET A   1       5.992  -6.631  -6.851  1.00 51.09     S
ATOM      8  CE  MET A   1       6.098  -4.849  -6.823  1.00 46.57     C
ATOM      9  N   ASN A   2       8.229 -10.164  -5.758  1.00 47.66     N
ATOM     10  CA  ASN A   2       8.058 -11.566  -6.180  1.00 47.92     C
ATOM     11  C   ASN A   2       8.046 -11.829  -7.684  1.00 48.09     C
ATOM     12  O   ASN A   2       7.713 -10.959  -8.465  1.00 49.43     O
ATOM     13  CB  ASN A   2       6.732 -12.052  -5.638  1.00 48.00     C
ATOM     14  CG  ASN A   2       6.831 -13.287  -5.003  1.00 45.23     C
ATOM     15  OD1 ASN A   2       6.195 -14.238  -5.405  1.00 48.13     O
ATOM     16  ND2 ASN A   2       7.617 -13.343  -3.949  1.00 42.01     N
```

PDB Viewers

- Pymol : http://www.pymol.org
- Jmol : http://www.jmol.org/
- Many others: KiNG, QuickPDB, Webmol, Rasmol



Pymol



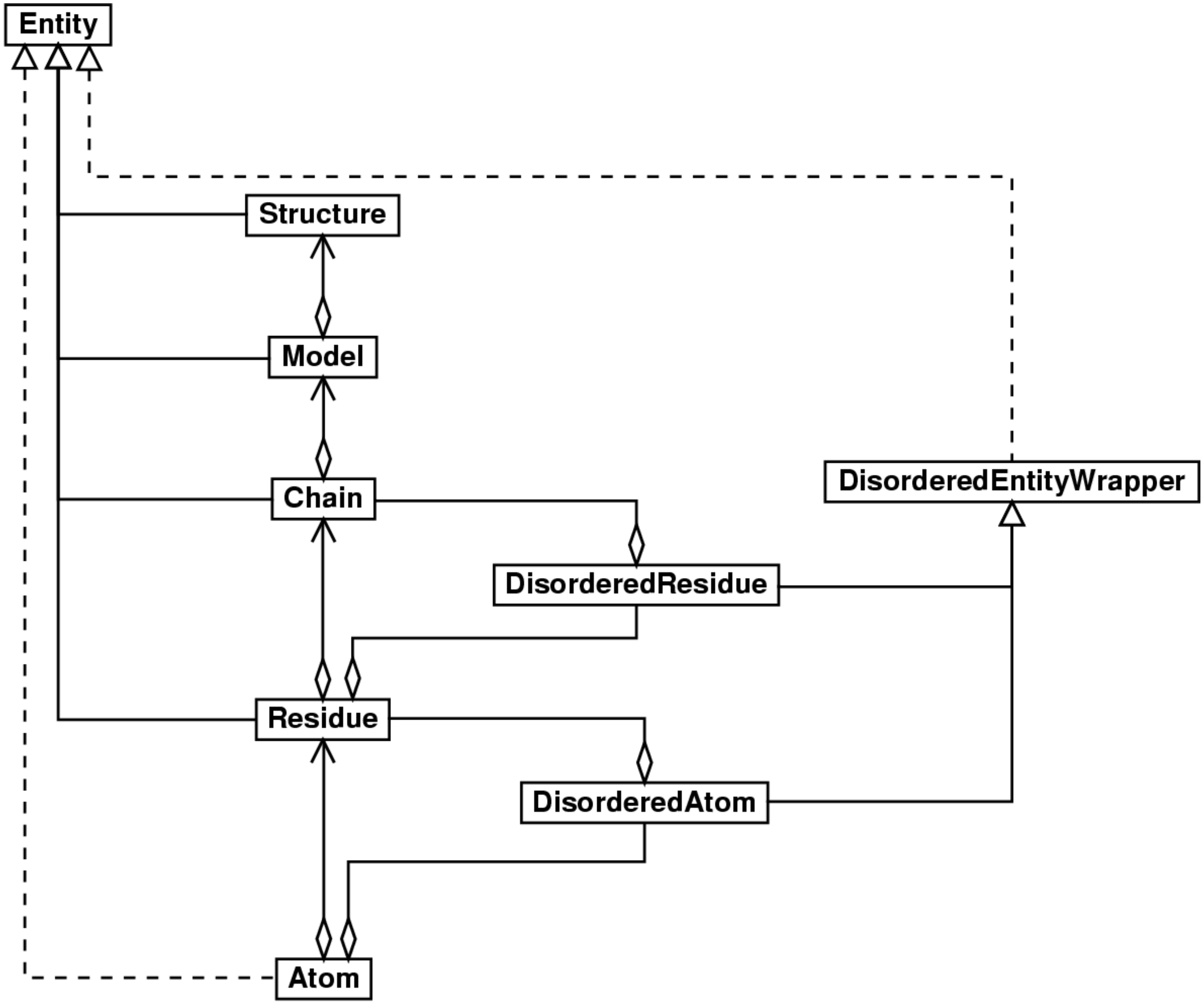
Jmol

Parsing PDB files with Biopython

```
from Bio.PDB.PDBParser import PDBParser

p=PDBParser(PERMISSIVE=1)

structure_id="1fat"

filename="pdb1fat.ent"

s=p.get_structure(structure_id, filename)
```

Create a new Structure Object

Structure representation



Working with structure objects

- Choose a model (E.g.: `first_model=structure[0]`).
- Choose a chain (E.g.: `chain_A=model["A"]`).
- Choose a residue (E.g.: `res10=chain[10]`).
- Choose a atom (E.g.: `atom=res10[”CA"]`).
- Retrieve Atom attributes:

```
a.get_name()       # atom name (spaces stripped, e.g. "CA")
a.get_id()         # id (equals atom name)
a.get_coord()      # atomic coordinates
a.get_bfactor()    # B factor
a.get_occupancy()  # occupancy
a.get_altloc()     # alternative location specifie
a.get_sigatm()     # std. dev. of atomic parameters
a.get_siguij()     # std. dev. of anisotropic B factor
a.get_anisou()     # anisotropic B factor
a.get_fullname()   # atom name (with spaces, e.g. ".CA.”)
```

Example

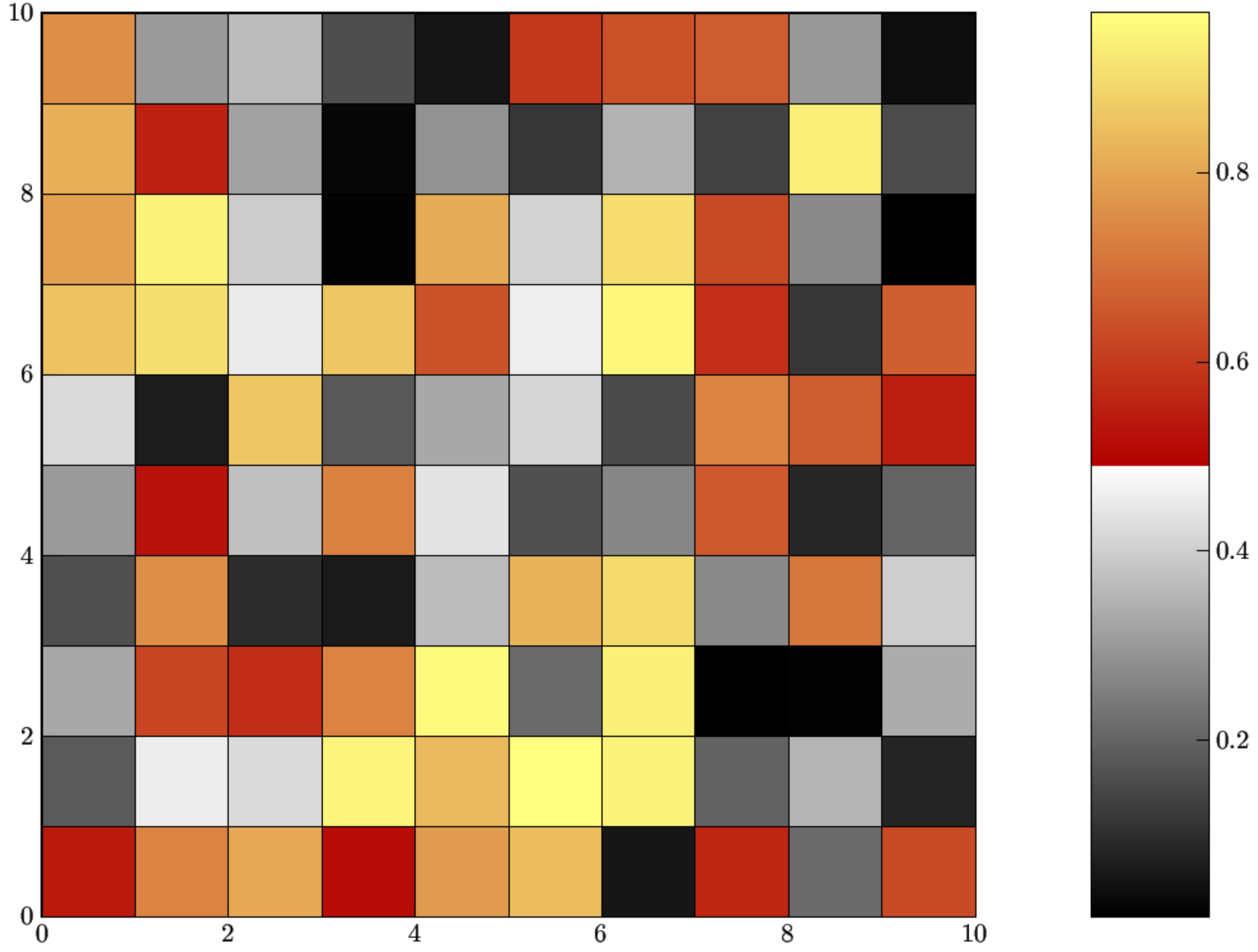
```
from Bio.PDB.PDBParser import PDBParser

parser=PDBParser()

# parse PDB file and store it in structure object
structure=parser.get_structure("test", "1fat.pdb")

# print the coordinate of CA atoms with B factor > 50
for model in structure.get_list():
  for chain in model.get_list():
    for residue in chain.get_list():
      if residue.has_id("CA"):
        ca=residue["CA"]
        if ca.get_bfactor()>50.0:
          print ca.get_coord()
```

Appendix

- User defined color maps
- GenBank record

User defined color scale

```
from pylab import *
cdict = {'red': ((0.0, 0.0, 0.0),
                 (0.5, 1.0, 0.7),
                 (1.0, 1.0, 1.0)),
         'green': ((0.0, 0.0, 0.0),
                   (0.5, 1.0, 0.0),
                   (1.0, 1.0, 1.0)),
          'blue': ((0.0, 0.0, 0.0),
                   (0.5, 1.0, 0.0),
                   (1.0, 0.5, 1.0))}
my_cmap = mpl.colors.LinearSegmentedColormap('my_cmap',cdict,256)
pcolor(rand(10,10),cmap=my_cmap)
colorbar()
```


10
8
6
4
2
0
0
2
4
6
8
10
0.8
0.6
0.4
0.2

GenBank SequenceFeatures

location : Location of the sequence.

type : This is a textual description of the type (e.g. ‘CDS’ or ‘gene’).

ref : A reference to a different sequence.

ref_db : cross sequence reference.

Strand : The strand identifier.

Qualifiers : dictionary of additional information about the features.

sub_features : additional sub_features.