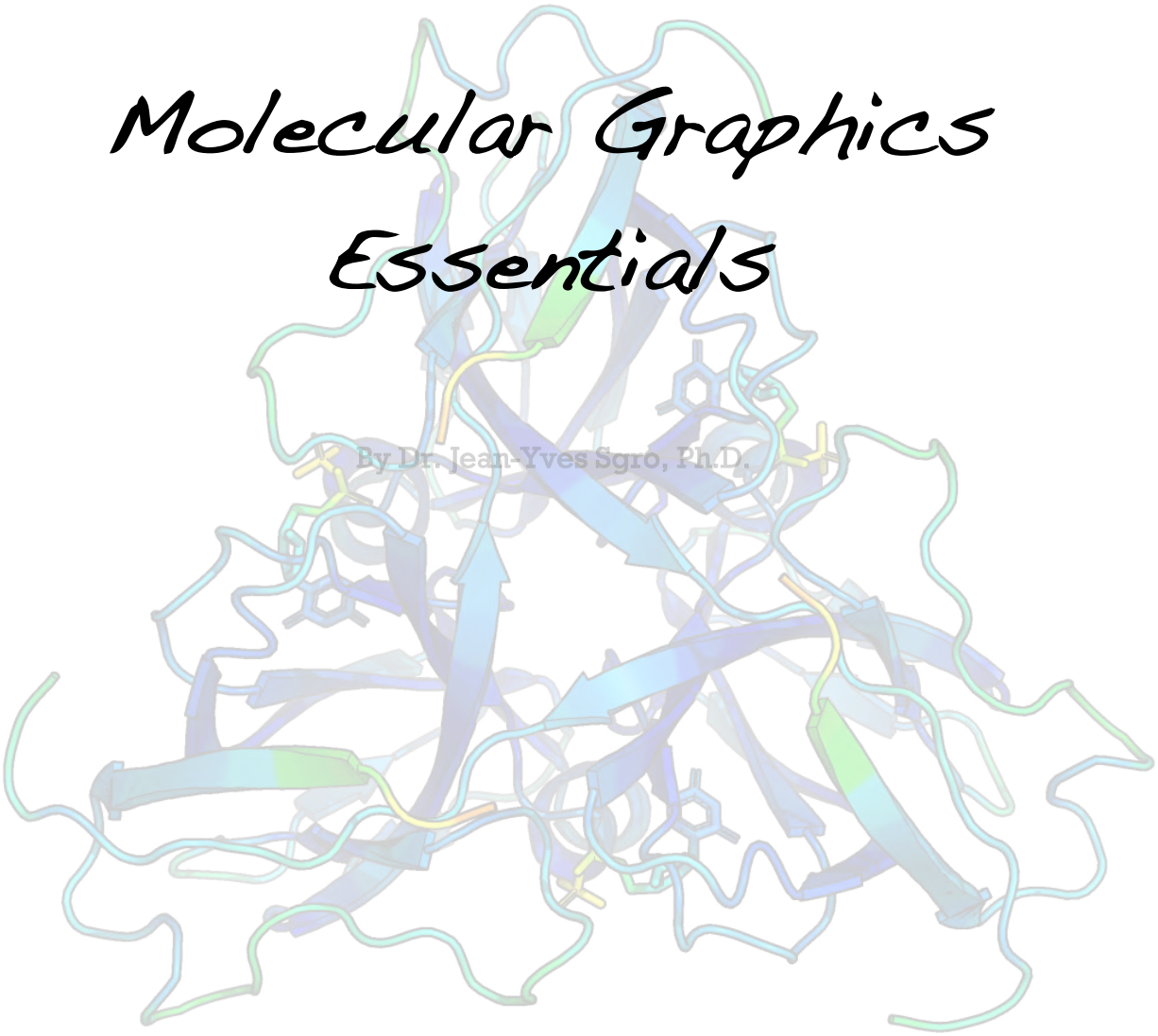
Book 1

Molecular Graphics Essentials



By Dr. Jean-Yves Sgro, Ph.D.

JYS

Cover: PyMOL rendering of PDB 1DUD (*Crystal structure of the Escherichia coli dUTPase in complex with a substrate analogue (dUDP).* Larsson, G., Svensson, L.A., Nyman, P.O. (1996) Nat.Struct.Mol.Biol. **3**: 532-538)

Preamble

For many years I participated in the teaching of Prof. Ann Palmenberg classes, in particular teaching about molecular graphics on the Desktop at a time when computer use in the classroom was not yet preeminent. This class was a useful complement to the main topic of Sequence Analysis and Evolution also using computers.

I personally used what was called "Unix Workstations" (Silicon Graphics, sgi) which had powerful graphics and rather beautiful "photorealistic" renderings. For teaching molecular graphics on the Desktop, I first used Rasmol, which was an amazing software fitting in about half of a 3.5" floppy disk or about 500 kilobytes. Rasmol included a sophisticated line command language but lacked the beautiful "photorealistic" renderings of the workstations. This was remedied by adding two other software in the course, one for the "publication quality renderings" (VMD) and the other for the modeling abilities of side-chain mutations and automated 3D superimposition of structures (Swiss PDB viewer later called DeepView.)

When PyMOL was still in preliminary development at version 0.99 I spent one intense week porting all the class material to PyMOL. Now, rather than using three different software, all was possible with only PyMOL. Over the years I extended and updated the PyMOL course material.

The UW-Madison Biochemistry students were the primary audience for these classes in courses Biochem 660 and 712, and occasionally in Biochem 511. I offered the PyMOL class in Biochem 660 for over a decade. The PyMOL tutorial and preliminary molecular graphics and file format introduction were part of a very large, made-to-order physical copy of a class book of about 500 pages that also contained tutorials on using other software. The PyMOL section was about 200 pages.

2017 was the last year that Biochem 660 was offered. I have therefore decided to release the complete PyMOL tutorial which you will find split in multiple PDF files. In this final revision, I had updated all web pages, and added links to archived pages when web site were defunct to keep the text as relevant as possible for future use.

I hope that it will be useful to you in accomplishing your molecular graphics goals.

Sincerely,

Jean-Yves Sgro, Ph.D.
Distinguished Scientist | Senior Scientist
Biotechnology Center | Biochemistry Department
University of Wisconsin-Madison - USA
Email: jsgro@wisc.edu
Formerly at the Institute for Molecular Virology and VirusWorld web site creator.

JYS

JYS

Table of Contents

Molecular Graphics Essentials

1. Introduction

MOLECULAR GRAPHICS is the discipline and philosophy of studying molecules and their properties through graphical representation[1].

PyMOL is an open-source visualization software with 3D rendering capabilities used to create high quality images of small molecules and biological macromolecules. PyMOL has become a *de facto* standard within scientific publications.

Computer representations have replaced the physical models from the early days of structural biology, though it is now possible to create a 3D physical model from a digital file with rapid prototyping manufacturing such as 3D printing; PyMOL can export the necessary files.

Suggested reading and reviews about molecular graphics:

Article Title	Reference
Review: Visualization of macromolecular structures.	O'Donoghue *et al.* Nature Methods **7**, S42 - S55 (2010) doi:10.1038/nmeth.1427
Education: An introduction to biomolecular graphics	Mura *et al.* PLoS Comput Biol 6(8): e1000918. doi:10.1371/journal.pcbi.1000918
Visualization software for molecular assemblies	Goddard TD and Ferrin TE. Curr Opin Struct Biol. 2007 **17(5)**:587-95. doi:10.1016/j.sbi.2007.06.008

[1] Dickerson, R.E.; Geis, I. (1969). *The structure and action of proteins*. Menlo Park, CA: W.A. Benjamin.

There is a large body of existing software for creating molecular graphics. Here are two authoritative sources listing software: one from the Protein Data Bank and another from the O'Donoghue Nature Methods article cited above:

Lists of molecular graphics software	Short URL
http://www.rcsb.org/pdb/static.do?p=software/software_links/molecular_graphics.html	http://bit.ly/IAQOfN
http://www.nature.com/nmeth/journal/v7/n3s/fig_tab/nmeth.1427_T1.html	http://bit.ly/Qk0Jrl

2. Brief History

THE KLUGE[2], originally known as the "Electronic Systems Laboratory Display Console" was the first computer terminal able to show three-dimensional objects and rotate them.
These were the early 1960's at the Massachusetts Institute of Technology under Project MAC (Multi-Access Computer). The user could interact with the displayed objects through a variety of interfaces, notably buttons and a light-pen. A vector-based display, it could only represent white lines on a black background.

Cyrus Levinthal used the Kludge to visualize, study and model the structure of proteins and nucleic acids. From this encounter between cutting-edge computer technology and molecular biology emerged the crucial elements for the development of a research-technology field known today as interactive molecular graphics.

[Above text adapted from Francoeur 2002, see below.]

History of molecular graphics

Title	Reference / Link
Review: Cyrus Levinthal, the Kluge and the origins of interactive molecular graphics	Eric Francoeur. Endeavour. 2002 **26(4)**:127-31. Max Planck Institute for the Hitory of Science, Berlin. doi: 10.1016/S0160-9327(02)01468-0
Web: History of Visualization of Biological Macromolecules	Eric Martz and Eric Francoeur http://www.umass.edu/microbio/rasmol/history.htm
Web: Early Interactive Molecular Graphics at MIT	Eric Francoeur. **http://www.umass.edu/molvis/francoeur/levinthal/lev-index.html**

[2] Image from http://www.umass.edu/molvis/francoeur/levinthal/lev-index.html

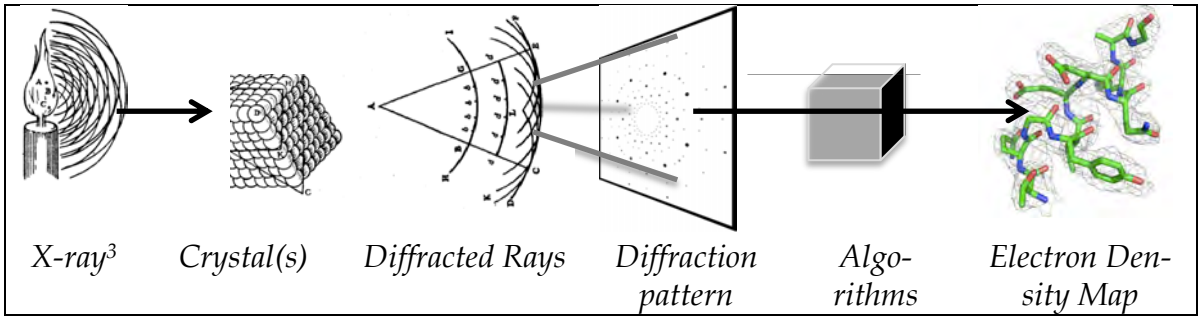
3. Methods to determine 3D structure

THREE MAIN METHODS are used to obtain 3D coordinates of biological molecules: X-ray crystallography, nuclear magnetic resonance (NMR), and 3D image reconstruction from cryo-electron microscopy.

3.1 X-ray crystallography

SINGLE CRYSTAL X-ray diffraction is a method that determines the arrangement of atoms within a crystal by analyzing the *diffraction pattern* and *intensities* of spots, which are the *reflections* created by the *interference* of X-ray waves elastically scattered by the electron clouds of the atoms from one set of evenly spaced planes within the crystal.

As a summary here is a simplified schematic of the key steps to calculate a correct atomic structural model:



Crystals are placed into an X-ray beam. The atoms of the proteins within the crystals diffract the incident X-ray and create diffraction patterns. With complex mathematical calculations (algorithms) crystallographers obtain an electron density map into which the amino acid sequence is fitted with help of computer graphics.

In the context of a short description, the algorithms used to calculate and refine structure are perhaps best presented in a poetic form, from the blog of a scientist[4]:

[3] Images from Christiaan Huygens "*Treatise on light*" http://www.gutenberg.org/files/14725/14725-h/

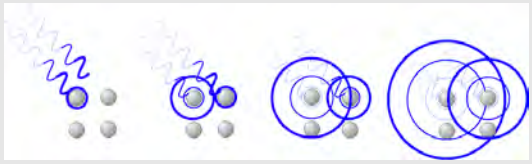
[4] http://twentyfirstfloormirror.wordpress.com/2010/09/09/the-crystal-clear-cell/ [http://bit.ly/1oAx8r2] Archived at: http://bit.ly/1MRQXvh

The script runs on an ancient Linux server and spits arcane numbers and symbols to the screen. To get to the biological, we must delve through the mathematical. Mercifully hidden from our caffeine-addled brains, the software sings the songs of past masters: Euler, Bragg, Fourier. Numeric incantations that, we hope, will sculpt the shape of our protein. […] The computer program has finished its song and we now stare triumphant at a ghostly blue-on-black image of our protein's structure. I slump back in the chair in quiet joy. The room sits too silently for me to scream.

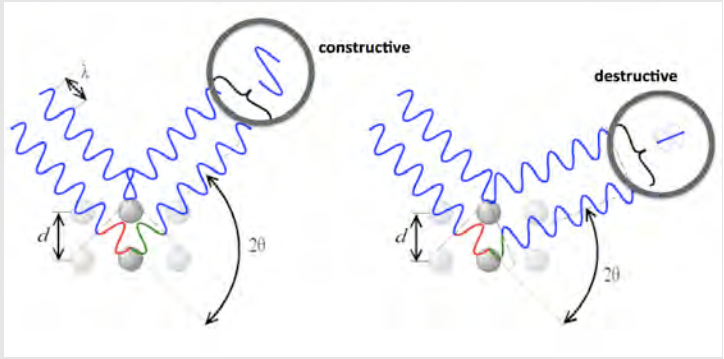
For the Physics enthusiasts: (Excerpt from: http://en.wikipedia.org/wiki/Bragg%27s_law, or http://bit.ly/MDiMIN)

In physics, **Bragg's law** gives the angles for coherent and incoherent scattering from a crystal lattice. When X-rays are incident on an atom, they make the electronic cloud move as does any electromagnetic wave. The movement of these charges re-radiates waves with the same frequency (blurred slightly due to a variety of effects); this phenomenon is known as Rayleigh scattering (or elastic scattering). The scattered waves can themselves be scattered but this secondary scattering is assumed to be negligible. These re-emitted wave fields interfere with each other either constructively or destructively (overlapping waves either add together to produce stronger peaks or subtract from each other to some degree), producing a diffraction pattern on a detector or film. The resulting wave interference pattern is the basis of diffraction analysis.

Rayleigh diffusion and diffraction: X-rays interact with the atoms in a crystal. The diffraction is due to the interference of waves that are elastically scattered:



*According to the 2**θ** deviation, the phase shift causes constructive (left figure) or destructive (right figure) interferences.*



The interference is constructive when the phase shift is a multiple of 2π; this condition can be expressed by Bragg's law[1] : $n\lambda = 2d \sin\theta$ where n is an integer, λ is the wavelength of incident wave, d is the spacing between the planes in the atomic lattice, and θ is the angle between the incident ray and the scattering planes.

[1] W.L. Bragg, "The Diffraction of Short Electromagnetic Waves by a Crystal", *Proceedings of the Cambridge Philosophical Society*, **17** (1913), 43–57.

Methods using femtosecond X-ray pulses (1 fs = 10^{-15} s) with X-Ray Free-Electron Lasers (XFEL) have been recently demonstrated by two teams of scientists as they solved structures from nano-crystals (photosystem I protein)[5] or non crystallized materials (mimi virus.)[6]

In Madison the Center for Eukaryotic Structural Genomics (CESG, uwstructuralgenomics.org) uses high throughput methods to help the Protein Structure Initiative (PSI) achieve the goal to "solve 10,000 protein structures in 10 years and to make the three-dimensional atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences."

Go further with the web

Crystallography 101 (with 17 min video)
http://www.ruppweb.org/Xray/101index.html
[http://bit.ly/Y91Mm5]
Archived version: **http://bit.ly/1powpdJ**
video alone: http://vimeo.com/7643687

Explanation of X-ray pattern: Franklin's X-ray diffraction
http://www.dnalc.org/view/15014-Franklin-s-X-ray-diffraction-explanation-of-X-ray-pattern-.html
[http://bit.ly/MXG8I9]

Bragg's Law and Diffraction: How waves reveal the atomic structure of crystals
http://www.eserc.stonybrook.edu/ProjectJava/Bragg/
[http://bit.ly/KW8EFD]

3.2 NMR spectroscopy

NUCLEAR MAGNETIC RESONANCE (NMR) SPECTROSCOPY uses radiofrequency radiation to induce transitions between different nuclear spin states of samples in a magnetic field. Different atoms in a molecule experience slightly different magnetic fields (chemical shift) and therefore transitions at slightly different resonance frequencies in an NMR spectrum.

[5] Chapman et al. *Nature* **470**, 73–77 (2011); *Femtosecond X-ray protein nanocrystallography*

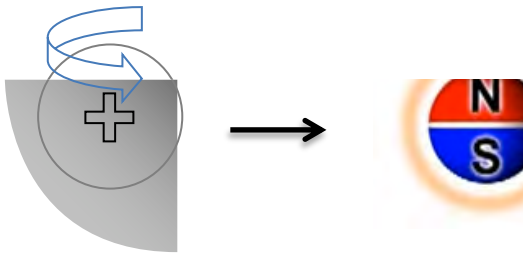
[6] Seibert et al. *Nature* **470**, 78–81 (2011); *Single mimivirus particles intercepted and imaged with an X-ray laser*

Magnetic resonance (MR) in all its forms (spectroscopy, imaging and relaxometry) is based on four simple facts[7]:

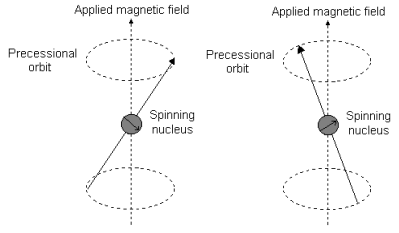
(1) Many isotopes possess *permanent magnetic moments* and, when placed into an *external magnetic field*, tend to align themselves along the field. The magnetic moments of all nuclides present in a sample sum up to a macroscopic vector quantity called ***nuclear magnetization***.
(2) By applying a suitable *radiofrequency pulse*, the *nuclear magnetization* can be rotated by any desired angle, thus brought into a non-equilibrium state and no longer aligned with the field.
(3) After a powerful but short *excitation pulse* the component of the excited *nuclear magnetization vector* (which is transversal to the magnetic field) rotates around the field direction with a frequency proportional to the field strength with a proportionality constant characteristic of the particular nuclide.
(4) After excitation, *nuclear magnetization* returns back to its equilibrium state. This process is called *relaxation* and the return paths (relaxation curves) can be quite complex.

Some of this can be illustrated with the simplest isotope of hydrogen ^{1}H. The unique proton of the nucleus is charged positively, and as it spins this creates a very small magnetic field, making the proton a "mini magnet" of sort. Importantly, the direction of the magnetic field (North and South poles) depends on the direction of spin.

The spinning of a charged object induces a magnetic field.



Schematic visualization of the spining around the magnetic field.



[7] http://www.ebyte.it/library/educards/nmr/OnePageMrPrimer.html

When placed within a magnetic field the nuclei align themselves either *with* (low energy state) or *against* (high energy state) the magnetic field (see (1) above.)

Then a pulse of radio waves with a specific energy is sent, some of the low energy absorb it and switch to a high energy state (see (2) above.)

When the magnetic field stops, they emit the energy they had absorbed which is intercepted and drawn as a peak by a receiver ((3) and (4) above.)

The most important method used for structure determination of proteins utilizes NOE (The **N**uclear **O**verhauser **E**ffect) experiments to measure distances between pairs of atoms within the molecule. Subsequently, the obtained distances are used as constraints to generate a 3D structure of the molecule by solving a distance geometry problem.

Go further with the web

The Basics of NMR: http://www.cis.rit.edu/htbooks/nmr/inside.htm
[http://bit.ly/1ONfj]

RADPAGE: Magnetized Nuclear Spin Systems
http://www.umkcradres.org/education/neuro/Spec/RADPAGE/Magnetized%20nuclear%20spin%20systems.htm
[http://bit.ly/L44XTa] Archived: http://bit.ly/1rBepjD

PDF: Fundamentals of NMR:
http://www.ias.ac.in/initiat/sci_ed/resources/chemistry/James.T.pdf
[http://bit.ly/Pg4Ntn]

PDF: Protein NMR spectroscopy in a nutshell:
http://www.helsinki.fi/~aannila/molbiophys/nmr3.pdf
[http://bit.ly/MCFfGw]

Video: NMR spectroscopy - Introduction to proton nuclear magnetic resonance 2:56 min
http://youtu.be/Oj-yxbPz5Ao

Video: Nuclear magnetic resonance NMR spectroscopy - 14:52 min
http://youtu.be/BirHLLz3aXc

Video: How NMR Work? – 8:30 min
http://youtu.be/SQkcu_qUF7U

Video: Simple demonstration of magnetic resonance as used in NMR and MRI - 5:09 min
http://youtu.be/1OrPCNVSA4o
http://www.drcmr.dk/MR (software)

Simulation: Analytical Nuclear Magnetic Resonance (NMR) Principles
http://vam.anest.ufl.edu/simulations/nuclearmagneticresonance.php
[http://bit.ly/1vuvayH]

3.3 Cryo-electron microscopy with 3D reconstruction

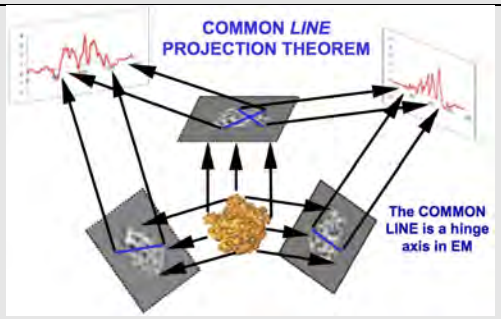
CRYO-ELECTRON MICROSCOPY (cryo-EM), or electron cryomicroscopy, is a form of transmission electron microscopy (EM) where the sample is studied at cryogenic temperatures. The biological material is spread on an electron microscopy grid and is preserved in a frozen-hydrated state by rapid freezing, usually in liquid ethane near liquid nitrogen temperature, protecting the sample in vitreous ice.

The 1984 paper[8] of the group of Jacques Dubochet at the European Molecular Biology Laboratory (EMBL) in Heidleberg, Germany, showing images of adenovirus embedded in a vitrified layer of water is generally considered to mark the birth of cryoelectron microscopy, and the technique has been developed to the point of becoming routine at several laboratories throughout the world.

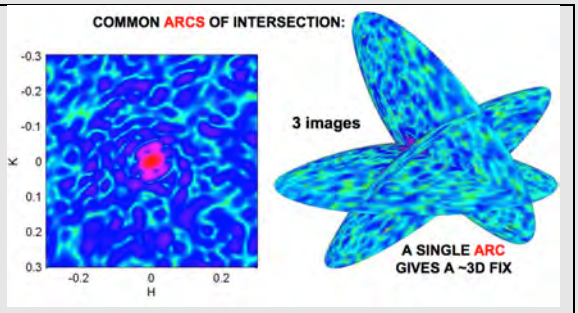
Each 2D image (micrograph) can be considered a projection of a 3D object.

Common line theorem:

- A 3D dataset can be assembled from individual images based on common lines of intersections.
- A 3D dataset can be assembled from individual images based on common arcs of intersections



EM Tomograms



Diffraction images:

Reference: Huldt, Szöke, Hajdu: Diffraction imaging of single particles and biomolecules. *J. Struct. Biol.***144**, 219 –227 (2003).
and
http://xfel.desy.de/localfsExplorer_read?currentPath=/afs/desy.de/group/xfel/wof/PM/Meetings/STI_June04/Hajdu.pdf
or http://bit.ly/R4IV5m *Archived:* http://bit.ly/1rBk890

[8] Adrian M, Dubochet J, Lepault J, McDowall AW. *Cryo-electron microscopy of viruses.* **Nature**. 1984 Mar 1-7;**308**(5954):32-6.

CryoEM maps and models are archived at the Unified Data Resource for Cryo Electron Microscopy: "**http://EMDataBank.org** is a joint effort of the Protein Data-bank in Europe (PDBe), the Research Collaboratory for Structural Bioinformatics (RCSB), and the National Center for Macromolecular Imaging (NCMI) to create a global deposition and retrieval network for cryoEM map, model and associated metadata, as well as a portal for software tools for standardized map format con-version, map, segmentation and model assessment, visualization, and data integra-tion." (Note: *see below for discussion on PDB*.)

3.4 All methods

THE COMMON GOAL OF ALL THESE METHODS is to create 3D coordinates representing the molecules studied.

The coordinate files are deposited to public repositories as explained in the next section.

The Protein Data Bank (PDB)

IN 1971 THERE WAS a total of **2** structures within the database.
In 1974 that number had grown to **12** protein structures.
On July 6, 2012 there were **82,809** structures, on September 23, 2014 there were **103,557** and on August 23, 2015 there were **111,385**[9].

1. Web repository of published structures

"The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of biological macromolecules, serving a global community of researchers, educators, and students. The archives contain atomic coordinates, bibliographic citations, primary and secondary structure information, as well as crystallographic structure factors and NMR experimental data."

Experimental Method (September 21, 2017)	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	112121	1880	5725	4	119730
NMR	10503	1224	245	8	11980
ELECTRON MICROSCOPY	1250	30	438	0	1718
HYBRID	103	3	2	1	109
Other	199	4	6	13	222
Total	124176	3141	6416	26	**133759**

[9] http://www.rcsb.org/pdb/statistics/holdings.do

109519 structures in the PDB have a structure factor file.
9320 structures in the PDB have an NMR restraint file.
3072 structures in the PDB have a chemical shifts file.
1721 structures in the PDB have a 3DEM map file.

Name	Link
The Worldwide Protein Data Bank (wwPDB) [parent site to regional hosts (below)]	`www.wwpdb.org`
RCSB Protein Data Bank (USA)	`www.rcsb.org`
PDBe (Europe)	`www.pdbe.org`
PDBj (Japan)	`www.pdbj.org`
Biological Magnetic Resonance Data Bank	`www.bmrb.wisc.edu`
wwPDB Documentation — documentation on both the PDB and PDBML file formats	`www.wwpdb.org/docs.html`
Introductory PDB tutorial (now paid access)	`www.openhelix.com/pdb`

1.1 PDB File names

PDB FILES AND STRUCTURES are designated by a PDB ID code. The code is always 4 characters long. The first character is a digit between 1 and 9, the other three characters are alphanumeric.

PDB codes are *not* case sensitive.

PDB codes can become obsolete and retired; the first digit can occasionally indicate subsequent updates, but not necessarily.

For example the 1971 structure of insulin (deposited in 1980) was **1INS**, but the current version of this structure is now **4INS**. Similarly the original PDB code for rhinovirus 14 was **1RHV** and is now **4RHV**. But the first digit does not always mean something: the PDB code **1PLV** is not in use, but **2PLV** is that of a poliovirus type 3, and **3PLV** is that of a small protein complex, and **4PLV** is not in use either.

Typically PDB codes are presented in the paper(s) describing published structures. In turn the codes can also be used to reference both the structure and the paper (see box below.)

Citing PDB structures:

Structures should be cited with the PDB ID and the primary reference.
For example, structure `102L` should be referenced as:

> PDB ID: **`102L`**
> D.W. Heinz, W.A. Baase, F.W. Dahlquist, B.W. Matthews *How Amino-Acid Insertions are Allowed in an Alpha-Helix of T4 Lysozyme*. *Nature* 361 pp. 561

Structures without a published reference can be cited with the PDB ID, author names, and title:

> PDB ID: **`1CIO`**
> Shi, W., Ostrov, D.A., Gerchman, S.E., Graziano, V., Kycia, H., Studier, B., Almo, S.C., Burley, S.K., New York Structural GenomiX Research Consortium (NYSGXRC) The Structure of PNP Oxidase from *S. Cerevisiae*

1.2 PDB file format

PDB FILES ARE **plain text files** spanning **80 columns of printable characters**.

Each line within the file is a "**record**" and starts with a keyword representing the "record type." For example lines starting with the keyword SEQRES provide information on sequence for a protein or nucleic acids; lines starting with ATOM contain the 3D coordinates of standard atoms.

Keywords have only up to 7 characters. A few records are mandatory and the various record types are expected in a specific order along the file.

PDB files are organized in three main sections:

- **Header**
- **3D data itself**
- **Miscellaneous connectivity and End records**

Minute details of the format can be found at `http://www.wwpdb.org/docs.html` and `http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/coordinates.html` [or `http://bit.ly/Pyhgdb`]

1.2.1 Header

The header is a record keeping space for archiving information about the structure contained within the PDB file. Except for optional secondary structure data that might be useful, header information is not critical for molecular graphics.

Early PDB files contained just a few lines mostly pertinent to crystallographers. Headers in current files easily have a few hundred lines made or record from 35 different keywords. For example the 29 amino acids hormone glucagon (1GCN) exhibits a header of 384 lines.

The header can be subdivided in multiple subsections (title, primary structure, secondary structure, connectivity, non-standard residues). However, sections are not specifically noted but can be easily recognized by the keywords starting each of their record lines.

1.2.2 Three-D data

The most abundant and pertinent information within a PDB file are the record lines starting with the keyword **ATOM**, used exclusively for protein and nucleic acid atom records. Each record represents the 3D coordinates for one atom within the structure.

HETATM records (hetero-atoms) are used for most other atomic coordinates, such as ligands (sugars, solvents, enzymatic substrates), metallic ions (iron, zinc, calcium etc.), and also modified amino acids. However not all authors choose to label their structures in the same way, and inspecting the file contents with a simple word processor can be crucial and save a lot of frustration!

Proteins can comprise multiple polypeptide chains; a simple double-stranded nucleic acids would contain 2 chains. The keyword **TER** is used to mark the end of a chain.

More recently, multiple versions of the structure data can be included in a single PDB file, typically from NMR data, multimeric repeating units or molecular dynamics calculations. Each set of the structure data is recognized by one **MODEL** record followed by a number. The end of a model is provided by one **ENDMDL** record.

The text structure is most important for the coordinate entries and appears visually as columns of text and numbers of various but fixed lengths.

Example of ATOM records for **1BL8** (Potassium Channel (Kcsa, Streptomyces Lividans). (The gray areas are not part of the actual file.)

```
        8         1         2         3         4         5         6         7         8
        0123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
columns
Col #   #1      #2   #3  #4  #5 #6      #7       #8       #9      #10 #11              #12
        ATOM      1  N   ALA A  23      65.191   22.037   48.576  1.00181.62           N
        ATOM      2  CA  ALA A  23      66.434   22.838   48.377  1.00181.62           C
        ATOM      3  C   ALA A  23      66.148   24.075   47.534  1.00181.62           C
1BL8    ////
atomic  ATOM    704  OE1 GLN A 119      79.595   14.626   51.132  1.00193.75           O
coords. ATOM    705  NE2 GLN A 119      79.635   16.611   50.129  1.00193.75           N
        TER     706      GLN A 119
        ATOM    707  N   ALA B  23      85.298    9.520   40.592  1.00173.57           N
        ATOM    708  CA  ALA B  23      84.639   10.739   41.145  1.00173.57           C
        ATOM    709  C   ALA B  23      83.162   10.775   40.768  1.00173.57           C
        ////
        ATOM   2822  OE1 GLN D 119      72.820   31.038   53.833  1.00171.94           O
        ATOM   2823  NE2 GLN D 119      74.344   30.617   52.270  1.00171.94           N
        TER    2824      GLN D 119
        HETATM 2825  K     K   401      67.868   26.595    9.017  1.00 57.73           K
        HETATM 2826  K     K   402      70.574   26.590   15.816  1.00 74.76           K
        HETATM 2828  O   HOH   500      69.120   26.480   12.189  1.00 66.21           O
        ////
```

Columns legend:

#1 Record keyword
#2 Atom serial number
#3 Atom name
#4 Residue name
#5 Chain identifier
#6 Residue sequence number
#7- 9 x, y, z orthogonal coordinates in Angstroms
#10 Occupancy
#11 B-value (or temperature factor)
#12 Optional column repeating atom names (shown here), the PDB code or other data.

Notes:

- An occupancy value of 1.00 indicates that the atom is found at the same place in all of the molecules in the crystal. Partial values are assigned in special cases such as side chains in multiple conformations or metal ions binding half the molecules in the crystal etc.

- B-values are an indication of the movement away from the coordinates. Atoms in much the same position in all the molecules have smaller B-values.

- Nucleic acids: older files did not distinguish between RNA and DNA except for U and T nucleotides. Newer conventions call for DNA nucleotides to be named DA, DC, DG and DT and A, C, G, U for RNA.

1.2.3 End records

After all the coordinates have been given, two sections finish the PDB file: a connectivity section and one or two lines signifying the end of the file.

The connectivity section is optional and the records start with the keyword **CONECT** specifying connectivity between atoms for which coordinates are supplied. The connectivity is described using the atom serial number as found in the entry.

Finally a "bookkeeping" section ends the file with two mandatory lines in published files. The first record starts with the keyword **MASTER** giving checksums of the number of records in the entry, for selected record types, and only for the first model.

The **END** record marks the end of the PDB file. The END keyword is the single entry on that last line finishing the PDB text file.

1.2.4 Summary

Schematic flow for a PDB files with 2 chains and 3 models
HEADER
MODEL 1 ATOM LINES FOR CHAIN A TER ATOM LINES FOR CHAIN B TER ENDMDL
MODEL 2 ATOM LINES FOR CHAIN A TER ATOM LINES FOR CHAIN B TER ENDMDL
MODEL 3 ATOM LINES FOR CHAIN A TER ATOM LINES FOR CHAIN B TER ENDMDL
END

Note: in PyMOL models are called states.

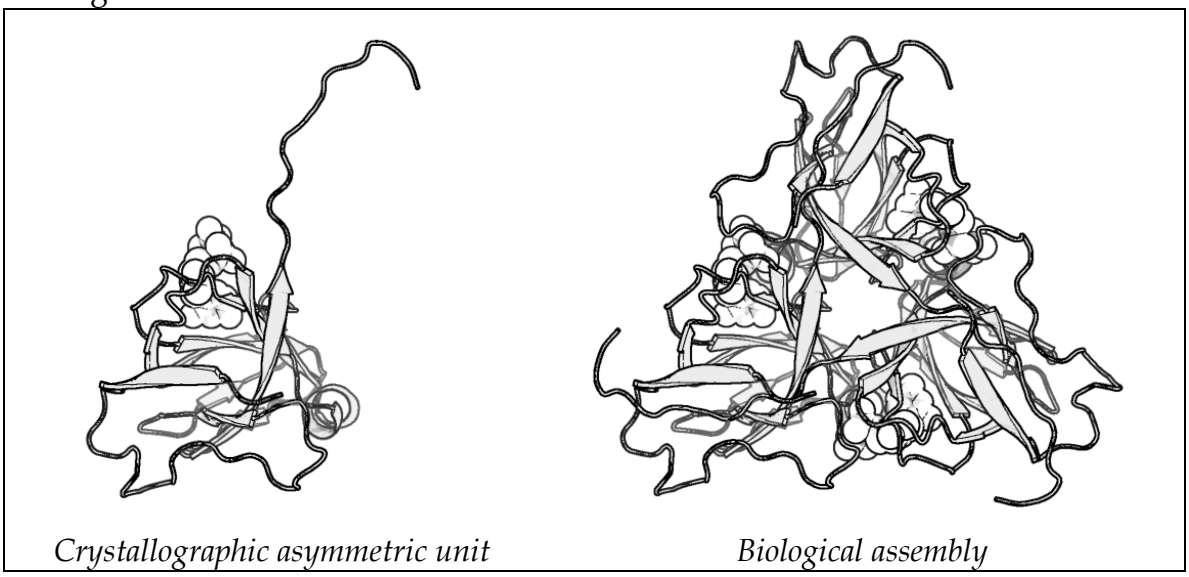
1.3 Biological *vs* crystalographic sets[10]

FOR DATA DERIVED BY X-RAY CRYSTALLOGRAPHY, the published coordinates are those of the "asymmetric unit" of the crystal. The asymmetric unit is the smallest portion of a crystal structure to which symmetry operations such as rotations and translations can be applied in order to generate the complete unit cell (the crystal repeating unit.)

The biological assembly (also sometimes referred to as the biological unit) is the macromolecular assembly that has either been shown to be or is believed to be the functional form of the molecule. For example, the functional form of hemoglobin has four chains. The records **REMARK 350** contain the rotation/translation matrices to build the biological assembly from the published coordinates of the asymmetric unit.

The protein databank web site offers the option to download the biological assembly. The PDB text file format will be similar, but the PDB code on the first line, the HEADER record will be masked to XXXX to note that the file no longer represents the deposited, original data set.

Example: the PDB entry **1DUD** (a dUTP diphosphatase, an enzyme that catalyzes the chemical reaction dUTP + H2O ⇌ dUMP + diphosphate) is biologically active as a trimer but the original PDB entry only contains one monomer. The remaining 2 monomers are calculated and delivered by the web site when downloading the biological unit.



Crystallographic asymmetric unit *Biological assembly*

[10] See
`http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/bioassembly_tutorial.html` [or `http://bit.ly/N842xQ`]

In the same way virus structures can be downloaded from the protein data bank.

However, for viruses with icosahedral symmetry an alternate site offers coordinates that are all in the same, standard orientation:
`http://viperdb.scripps.edu`

In addition, Viperdb may offer atomic models based on Cryo-EM data that might not have a standard PDB entry. That is particularly true for any published 3D "model."

1.4 CIF file format

This is now the default format of files fetched by PyMOL.

1.4.1 Limitations of PDB format

The PDB plain text format (see above) is a tabular format that caused problems when large structures such as the ribosome were solved as the number of characters per column is limited.

For example, here is a line from PDB ID code **4v49** for a ribosome:

```
HETATM66784  CA  MSE g 209     144.749 -58.573  68.549  1.00940.29           C
```

The firs column is limited to 6 characters and therefore HETATM fits in it. The next column is limited to 5 digits, and after 99999 it would break and not able to take 100000.

The easy "fix" for these large files was to simply split them. Therefore structure **4v49** is available in PDB –formatted file as 2 separate files.

To address this issue it was decided to start implementing a different format that has been in progress since the early 1990's called CIF: *Crystallographic Information File*. CIF is now the default format that PyMOL uses to download structures with the command `fetch`.

1.4.2 CIF format

The CIF: *Crystallographic Information File* format started to germinate in the mind of Ian David Brown (I. David Brown) already in 1983 as a new "Standard Crystallographic File Structure." This was published in preliminary form in Acta Crystallographica in 1991[11].

The PDB format was defined by each line being a *record* of a specific type, implying a given number of columns for the tabular data it contained.

The main difference is that the CIF format requires the definition of each column (field) being given before each "chunk" of data based on providing a "dictionary" of column names.

A 1995 online guide[12] from the San Diego Supercomputer Center (SDSC) offers a insights on the format: here is what becomes of the one-line header of a PDB file when converted to a CIF file:

PDB line:

```
HEADER    PLANT SEED PROTEIN                      11-OCT-91   1CBN
```

becomes in CIF format:

```
_struct.entry_id                  '1CBN'
_struct.title                     'PLANT SEED PROTEIN'

_struct_keywords.entry_id         '1CBN'
_struct_keywords.text             'plant seed protein'

_database_2.database_id            PDB
_database_2.database_code          1CBN

_database_PDB_rev.num                     1
_database_PDB_rev.date_original 1991-10-11
```

Here each line was provided with a definition. Below we'll see how to deal with multiple columns.

---

[11] Hall SR, Allen FH, Brown ID (1991). "The Crystallographic Information File (CIF): a new standard archive file for crystallography". *Acta Crystallographica*. **A47** (6): 655-685. doi:10.1107/S010876739101067X – also as an online reprint at **http://www.iucr.org/__data/iucr/cif/standard/cifstd1.html#reprints** [archived at **http://bit.ly/2wujXYJ**] - See also **http://journals.iucr.org/a/issues/1991/06/00/es0164/es0164.pdf** [archived: **http://bit.ly/2w5EYou**]

[12] **Macromolecular Crystallographic Information (mmCIF) Tutorial**: **http://www.sdsc.edu/pb/cif/tutorial_mm.html** [Archived http://bit.ly/2hdZNLV]

The amount of definitions for the dictionaries is vast, and to make the example more relevant here is a file that only contains information for the XYZ coordinates with all the necessary elements for the amino acid histidine (HIS.)

```
data_HIS
#
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM 1   N N   . HIS A 1 1  ? 49.668 24.248 10.436 1.00 25.00 ? 1  HIS A N   1
ATOM 2   C CA  . HIS A 1 1  ? 50.197 25.578 10.784 1.00 16.00 ? 1  HIS A CA  1
ATOM 3   C C   . HIS A 1 1  ? 49.169 26.701 10.917 1.00 16.00 ? 1  HIS A C   1
ATOM 4   O O   . HIS A 1 1  ? 48.241 26.524 11.749 1.00 16.00 ? 1  HIS A O   1
ATOM 5   C CB  . HIS A 1 1  ? 51.312 26.048 9.843  1.00 16.00 ? 1  HIS A CB  1
ATOM 6   C CG  . HIS A 1 1  ? 50.958 26.068 8.340  1.00 16.00 ? 1  HIS A CG  1
ATOM 7   N ND1 . HIS A 1 1  ? 49.636 26.144 7.860  1.00 16.00 ? 1  HIS A ND1 1
ATOM 8   C CD2 . HIS A 1 1  ? 51.797 26.043 7.286  1.00 16.00 ? 1  HIS A CD2 1
ATOM 9   C CE1 . HIS A 1 1  ? 49.691 26.152 6.454  1.00 17.00 ? 1  HIS A CE1 1
ATOM 10  N NE2 . HIS A 1 1  ? 51.046 26.090 6.098  1.00 17.00 ? 1  HIS A NE2 1
```

The first line needs to be present to provide a title. For coordinate files it would normally be the PDB ID code.

The second line, #, is a separator that could be omitted but provides clarity in spacing the content.

The word `loop_` on the third line indicates that the following dictionary of definitions for column should be used for all the lines that follow until such time where the file ends (as here) or continues with a new dictionary for other types of data and a new dictionary.

It is to be noted that fields that do not have a current value will be reported as a `.` or a `?` within the data set.

1.4.3 Further information on CIF

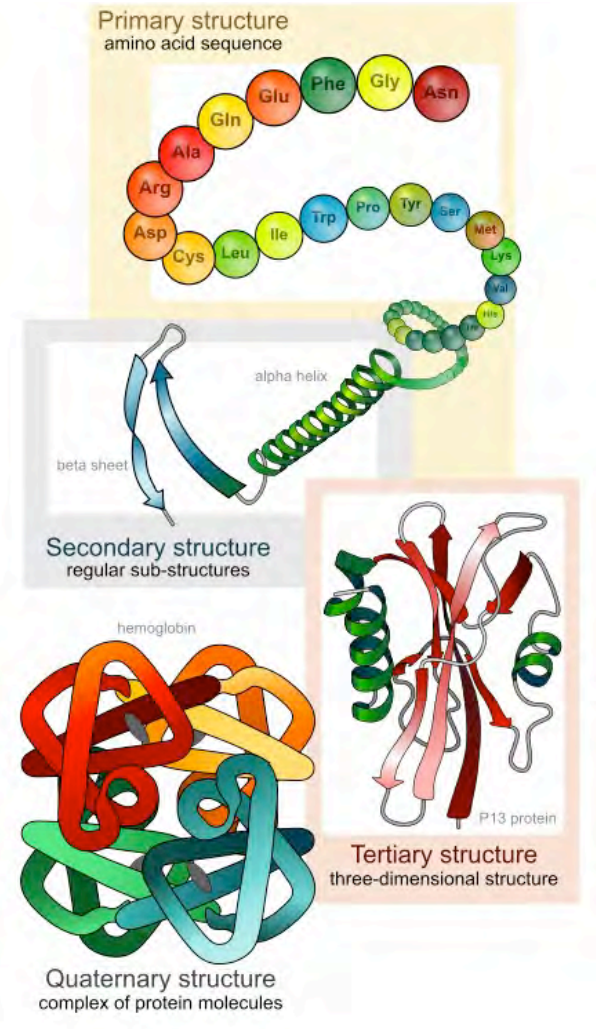
Further information on CIF can be found in these documents:

- **A Guide to CIF for Authors**:
  http://www.iucr.org/__data/assets/pdf_file/0019/22618/cifguide.pdf
  Archived: **http://bit.ly/2jIJAzn**

However, there is now a CIF 2.0 that has emerged:

- **Specification of the Crystallographic Information File format, version 2.0**:
  J. Appl. Cryst. (2016). **49**, 277-284
  https://doi.org/10.1107/S1600576715021871
  http://journals.iucr.org/j/issues/2016/01/00/aj5269/index.html
  [archived: http://bit.ly/2wF8lNW]

Protein Structure Short Summary[13]



The **primary structure** refers to amino acid linear sequence of the polypeptide chain. The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation.

The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity.

Secondary structure refers to highly regular local sub-structures. Linus Pauling and coworkers[14] suggested the **alpha helix** and the **beta strand** or **beta sheets**. These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups.

The **supersecondary structure** refers to a

[13] https://en.wikipedia.org/wiki/Protein_structure - *image from* https://en.wikipedia.org/wiki/File:Main_protein_structure_levels_en.svg

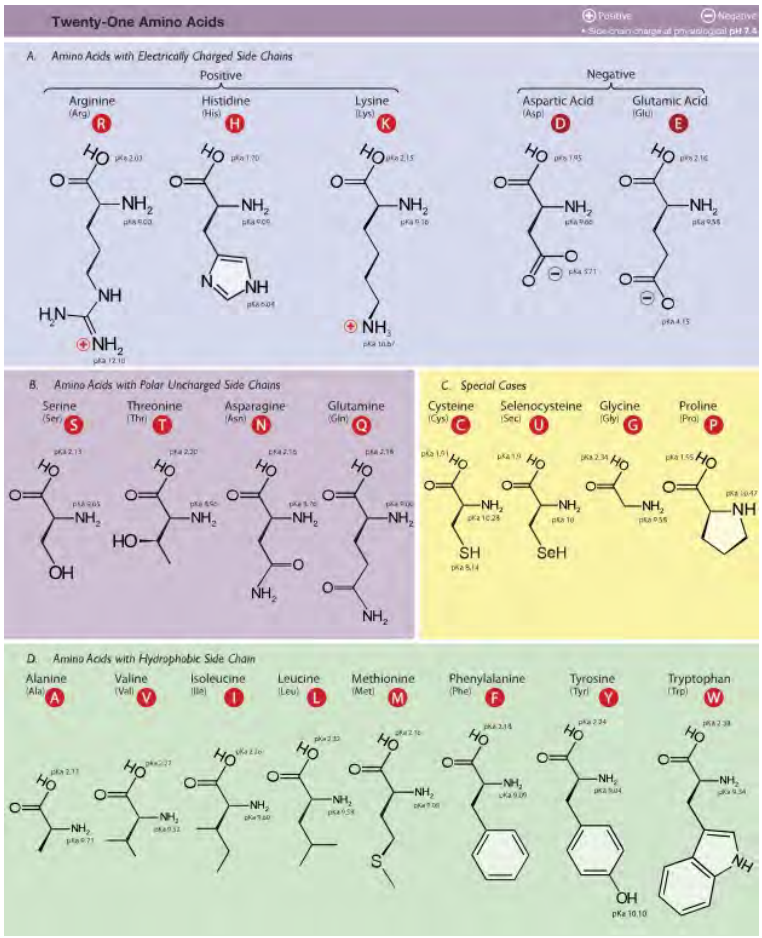
[14] Pauling L, Corey RB, Branson HR (1951). "The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain". *Proc Natl Acad Sci USA* **37** (4): 205–211.

specific combination of secondary structure elements, such as beta-alpha-beta units or helix-turn-helix motif. Some of them may be also referred to as structural motifs.

Tertiary structure refers to three-dimensional structure of a single protein molecule. The alpha-helices and beta-sheets are folded into a compact globule.

Quaternary structure is the three-dimensional structure of a multi-subunit protein and how the subunits fit together in this context. The quaternary structure is stabilized by the same non-covalent interactions and disulfide bonds as the tertiary structure.

21 amino acids (includes selenocysteine)[15]



20 natural amino acid notation[16]

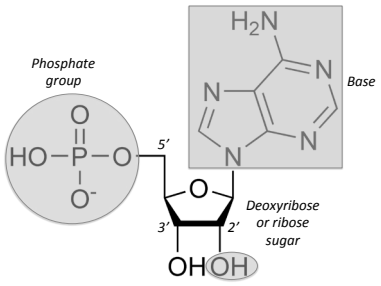
Amino Acid	3-Letter	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

[15] https://upload.wikimedia.org/wikipedia/commons/a/a9/Amino_Acids.svg

[16] https://en.wikipedia.org/wiki/Protein_primary_structure

Nucleic Acids, DNA and RNA, Summary

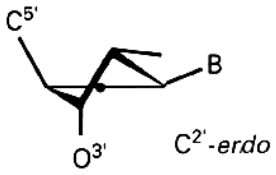
DNA (**d**eoxyribo**n**ucleic **a**cid) and RNA (**r**ibo**n**ucleic **a**cid) are polymers of *nucleotides* linked in a chain through phosphodiester bonds between the 3' carbon of one nucleotide and the 5' carbon of another nucleotide.



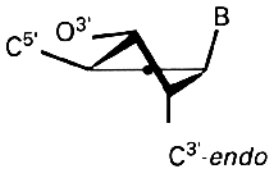
Nucleotides are three components covalently bound together:

- a nitrogen-containing "base" - either a pyrimidine (one ring) or purine (two rings)
- a 5-carbon sugar - ribose (RNA) or deoxyribose (DNA)
- a phosphate group

Note the 5' and 3' carbons on the sugars. This is critical to understanding polarity of nucleic acids (see below.) The 5' carbon has an attached phosphate group, while the 3' carbon has a hydroxyl group.



DNA – B form helix

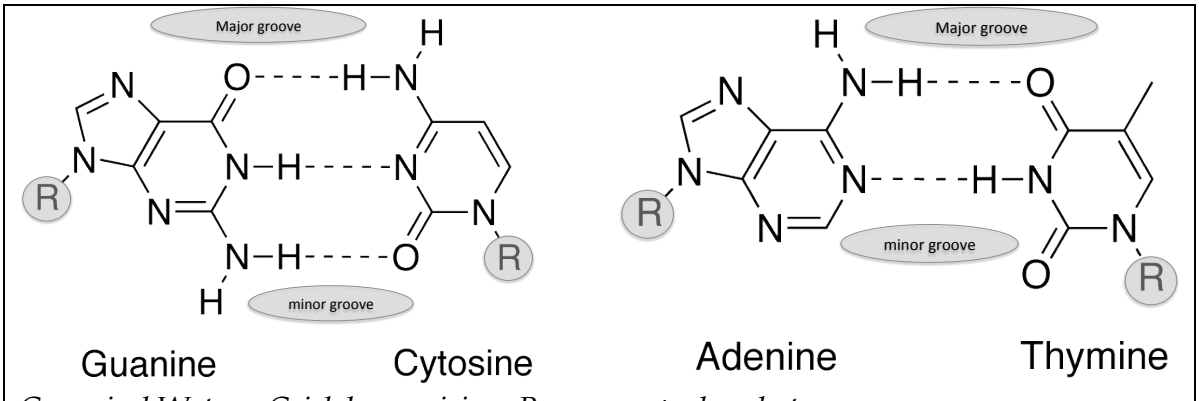


RNA – A form helix

The presence, or absence of the OH in position 2 ' of the sugar influences its 3D conformation (pucker.) The most common form of DNA (B-form) adopts the Sugar pucker C2'-endo

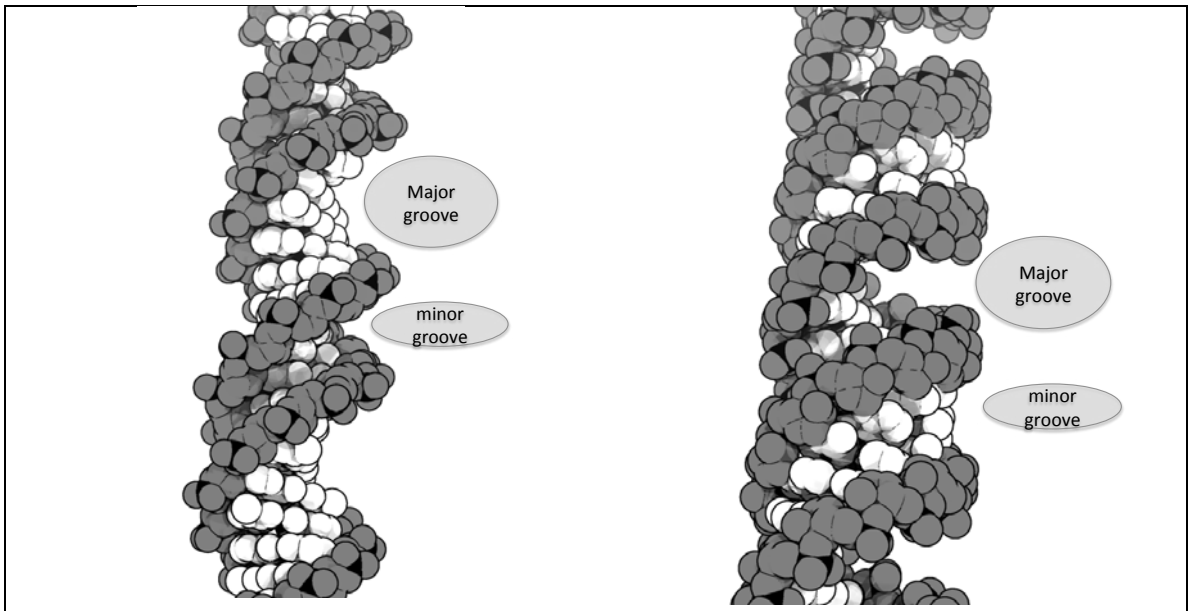
In RNA 2'-OH inhibits C2'-endo conformation and adopts the sugar pucker C3'-endo (A-form of the helix.)

G-C base pairs have 3 hydrogen bonds, whereas A-T base pairs have 2 hydrogen bonds: one consequence of this disparity is that it takes more energy (*e.g.* a higher temperature) to disrupt GC-rich DNA than AT-rich DNA.



Canonical Watson-Crick base pairing; R = sugar + phosphate

All nucleic acids have two distinctive ends (polarity): the 5' and 3' ends named after the carbons on the sugar. For both DNA and RNA, the 5' end bears a phosphate, and the 3' end a hydroxyl group. In double stranded nucleic acids, the two strands are anti-parallel to one another: 3'-end of one strand pairs with 5'-end of the other strand.



Most DNA exists in the double stranded B-form. The major force promoting formation of this helix is complementary base pairing.

For RNA A-form, stacking interactions between bases are important for the local folding.

Go further with the web

A few Wikipedia definitions:

Nucleic Acid en.wikipedia.org/wiki/Nucleic_acid	**DNA** en.wikipedia.org/wiki/DNA	**RNA** en.wikipedia.org/wiki/RNA
Nucleic acid double helix http://en.wikipedia.org/wiki/Nucleic_acid_double_helix		**Nucleic acid design** http://en.wikipedia.org/wiki/Nucleic_acid_design

Build your own 3D version of DNA or RNA sequence:
http://structure.usc.edu/make-na/server.html
[http://bit.ly/LKhTKX] [archived-non functional: http://bit.ly/2w9Jy5n]

History: double helix: 50 years of DNA: http://www.nature.com/nature/dna50/
[archived: http://bit.ly/2xqUJtj]

History: Watson and Crick describe structure of DNA 1953:
http://www.pbs.org/wgbh/aso/databank/entries/do53dn.html
[http://to.pbs.org/cTYOSz] [archived: http://bit.ly/2wDfza0]

Software: 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures:
http://nar.oxfordjournals.org/content/31/17/5108.full
[http://bit.ly/NkBDDl] [archived: http://bit.ly/2hiLVjm]