

Desktop Molecular Graphics Background Essentials

1 Introduction

The visualization techniques of the structure of macromolecules are companion tools to the sequence analysis algorithms. New DNA sequences are being cloned and sequenced rapidly but the structure of the putative encoded proteins cannot be determined based only on their sequence. As the number of protein structures solved by x-ray crystallography is increasing, it will become easier to find structural homologs to fit onto newly protein sequences. Molecular graphics play a key role in understanding current structures and creating (structural) models.



Molecular graphics have evolved over the last 40 years from a simple vector display on a high performance oscilloscope to sensor-based virtual reality.

For a beautifully illustrated account of “History of Visualization of Biological Macromolecules” see

<http://www.umass.edu/microbio/rasmol/history.htm>.

Image reference: <http://www.umass.edu/molvis/francoeur/levinthal/lev-index.html>

Desktop computers are now more powerful than mainframes of the last decades and there are free and commercial desktop software to manipulate 3 dimensional structures for the creation of publication quality images to illustrate research papers, proposals and to help visualize target molecules, their structural properties or their interaction with other molecules or ligands.

To be able to manipulate 3 dimensional structures on a desktop computer with a molecular graphics software is critical for today's molecular biologist and a necessary complement to sequence analysis projects.

2 Where do 3 dimensional structures come from?

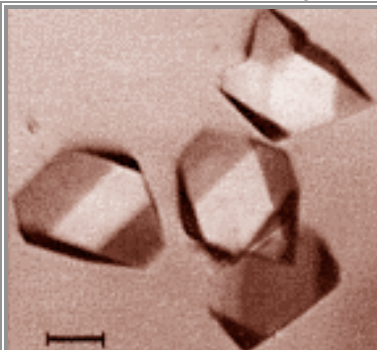
In summary there are three main methods: X-ray crystallography, NMR and 3D image reconstruction from cryo-electron microscopy.

Biochemists and crystallographers have developed techniques to crystallize macromolecules. Indeed proteins, nucleic acids or their complex can form crystals in specific biochemical conditions. The crystals are very fragile and small (often less than a millimeter) but they still can be placed inside an x-ray beam. Because of the regular

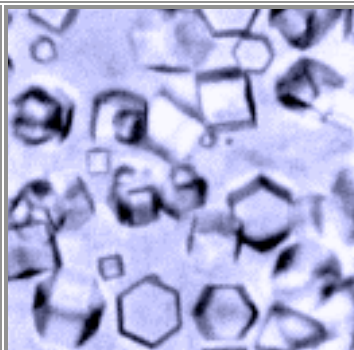
arrangement of the molecules within the crystals the x-ray will diffract in a very specific pattern that can be recorded on x-ray photographic film or an electronic array detector. With the help of powerful computer and complex software, the mathematical analysis of the diffraction pattern allows the crystallographer to calculate where the electrons (of the atoms) of the protein would be located in 3D space inside the crystal. They then fit a wireframe representation of the amino acids inside the electron density. When the position of the atoms is refined, the structure is published and usually deposited at the Protein Data Bank. These are the structures that you can fetch with a web browser and display inside on your desktop computer. A notable exception is for structures determined in the private sector these coordinates are proprietary and the authors are not obligated to their data public There used to be a lot of months or years of involved work for each solved structure, but new streamlined methods allow for faster determination in as little as one week! (High throughput structure determination, NIH, Protein Structure Initiative.)

(for more information on high throughput (not covered in class) see <http://www.uwstructuralgenomics.org/>)

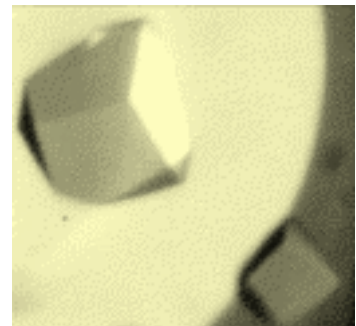
Crystals are placed into an x-ray beam. The atoms of the proteins within the crystals diffract the incident x-ray and create diffraction patterns on a film. With complex mathematical calculations crystallographers obtain an electron density map into which the amino acid sequence is fitted with help of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



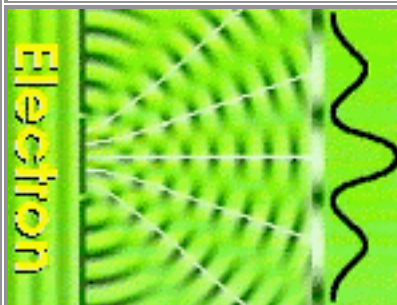
L(+)-lactate dehydrogenase crystals. Bar=100 μ m. Ostendorp et al. (1996) Protein Science 5, 862



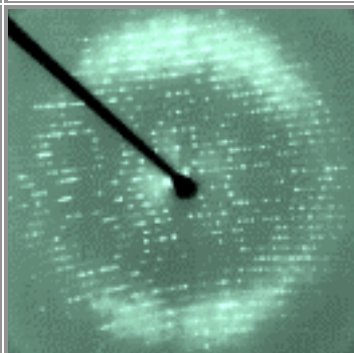
pentalenene synthase crystals. Lesburg et al. (1995) Protein Science 4, 2436



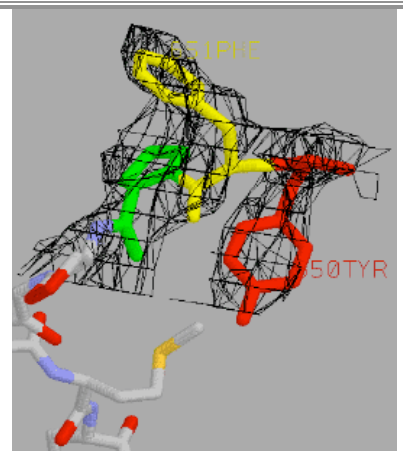
phosphoribulokinase crystals. Roberts et al. (1995) Protein Science 4, 2442



Diffraction Amplitude waves of diffracted electrons can add or subtract to each other. The result are the white dots on the diffraction image.



Diffraction image from a pentalenene synthase crystal. Lesburg et al. (1995) Protein Science 4, 2436

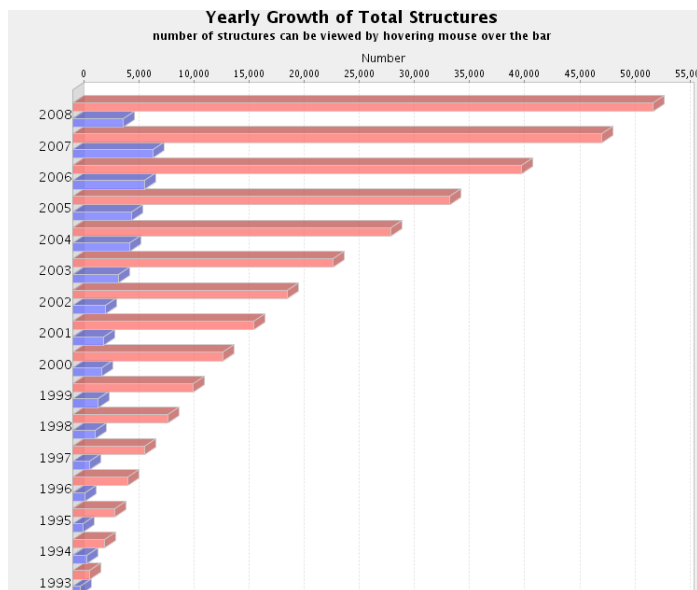


Electron density map

3 The Protein Data Bank (PDB) : web repository of published structures

3.1 The web site

"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."



On October 14, 2008 there were **53660** PDB entries.

While the majority of published structures are derived from X-ray crystallography. However, every year a larger number are derived from NMR experiments.

The number of structure has been increasing exponentially since the first structures deposited in 1972.

The PDB database home page is at <http://www.rcsb.org/pdb>

The "PDB Statistics" button on the top left of the page leads to more information about released entries.

3.2 PDB file names

Sequences are found by their "accession number." Similarly PDB files are designated by a PDB ID code, alphanumeric and only 4 characters long. Most authors publish the PDB ID within their publications.

Many proteins are represented multiple times within the database, as reported by different authors or as mutants or because they were crystallized under different conditions such as pH. For example myoglobins account for 277 entries and hemoglobins 405.

The citing of a PDB structure is done by referencing the PDB ID code, followed by the primary journal citation.

See

http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/policies_references.html

Excerpt from web site:

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 (1993)

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

PDB ID: 1CI0 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*

3.3 File Format

PDB files are plain text files and can be opened with the simplest word processors (e.g WordPad on Windows, TextEdit on Macintosh, pico, nano, vi on Linux.) The data is arranged in columns, formatted to specific width creating “fields” (70 characters wide by default).

Each line is a “record” and starts with the name of a record content type. For example, the most abundant and pertinent information within a PDB file are the ATOM lines, all starting with the ATOM keyword. Each ATOM record represents the 3D coordinates for one atom within the structure. ATOM records are used only for protein and nucleic acid structures. HETATM= records (hetero-atoms) are used for most other atomic coordinates, such as ligands (sugars, solvents, enzymatic substrate), 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 the ability to open the file with a word processor is extremely useful to understand and inspect its contents.

Each character is positioned within the file at a specific column position, and it is vital not to disrupt this arrangement when editing a PDB file. For example the following 2 records are not equivalent:

```
ATOM 1 N HIS 1 49.668 24.248 10.436 1.00 25.00 1 1GCN 50
ATOM 1 N HIS 1 49.668 24.248 10.436 1.00 25.00 1 1GCN 50
```

Because of the position-specific format in the columns, the latter record is in fact equal to:

```
ATOM 2 N HIS 1 4.966 2.424 1.043 1.00 25.00
```

The first 3 real numbers on each line the x, y and z *Cartesian* coordinates of the atom in three dimensional space. The position of this Nitrogen atom would be in a very different position with this erroneous modification.

The three dimensional XYZ coordinates are the first three real numbers on each line. For example **49.668 24.248 10.436** for the first ATOM record (Nitrogen atom of Histidine amino-acid number 1), in the previous example. In the Cartesian coordinate system the x, y and z axes are perpendicular to one another and their length is 1. The unit length is 1 Å (equal to 0.1 nm [nanometer] in the international notation).

This is a preferred system of reference for most biological users, however it is worth knowing that in some cases the frame of reference is the length of the crystallographic "unit cell". In this case the axes are labeled a, b and c. They are not necessarily perpendicular to one another and do not necessarily have the same length. If the coordinates are expressed as a function of these axes they are usually referred to as fractional coordinates. Most chemical databases give the coordinates in this fashion. In the PDB formatted file, the CRYST1 and SCALEn records are related to these axes, but for our purpose, these can be ignored.

The "ABOUT PDB" button leads to a page with a link to the PDB file format, currently the format file resides at http://www.rcsb.org/pdb/file_formats/pdb/pdbguide2.2/guide2.2_frame.html but that dates back to 1996. More information can be found under the menu "Dictionaries and File Formats" on the left panel.

A typical PDB file has a large header, title section with information about the compound, the journal into which the structure was described, technical information about the crystallization procedure and information about the crystal symmetry, the amino acid sequence of the protein(s), specific records for the secondary structure (alpha helix or beta sheet), the nucleic acid sequence etc.

The current most common record keywords are in the following table:

<i>Keyword</i>	<i>Definition</i>
TITLE SECTION: section describing experiment and biological macromolecules	
HEADER	uniquely identifies entry with the idCode field. Provides a classification for the entry. Contains the date the coordinates were deposited at the PDB.
OBSLTE	appears in entries which have been withdrawn from distribution.
TITLE	contains a title for the experiment or analysis that is represented in the entry.
CAVEAT	warns of severe errors in an entry. Use caution when using an entry containing this record.
COMPND	describes the macromolecular contents of an entry. For each macromolecular component, the molecule name, synonyms, number assigned by the Enzyme Commission (EC), and other relevant details are specified.
SOURCE	specifies the biological and/or chemical source of each biological molecule in the entry.
KEYWDS	contains a set of terms relevant to the entry.
EXPDTA	presents information about the experiment. E.g X-RAY DIFFRACTION, NMRNEUTRON DIFRACTION, THEORETICAL MODEL
AUTHOR	contains the names of the people responsible for the contents of the entry.
REVDAT	contains a history of the modifications made to an entry since its release.
SPRSDE	contains a list of the ID codes of entries that were made obsolete by the given coordinate entry and withdrawn from the PDB release set. One entry may replace many.
JRNL	contains the primary literature citation that describes the experiment which resulted in the deposited coordinate set. Other references are given in REMARK 1

REMARK	presents experimental details, annotations, comments, and information not included in other records.
REMARK 1	lists important publications related to the structure presented in the entry.
REMARK 2	states the highest resolution, in Angstroms, that was used in building the model.
REMARK 3	presents information on refinement program(s) used and the related statistics.
REMARK 4 – 999	free text annotation
PRIMARY STRUCTURE: section listing sequence of residues in each chain	
DBREF	provides cross-reference links between PDB sequences and the corresponding database entry or entries.
SEQADV	identifies conflicts between sequence information in the ATOM records of the PDB entry and the sequence database entry given on DBREF. Please note that these records were designed to identify differences and not errors. No assumption is made as to which database contains the correct data.
SEQRES	contains the amino acid or nucleic acid sequence of residues in each chain of the macromolecule that was studied. Note that sometimes the sequence is limited to the portion of the protein that is solved, other times the complete sequence is present.
MODRES	provides descriptions of modifications (e.g., chemical or post-translational) to protein and nucleic acid residues.
HETEROGEN SECTION: complete description of non-standard residues	
HET	describes non-standard residues, such as prosthetic groups, inhibitors, solvent molecules, and ions for which coordinates are supplied.
HETNAM	gives the chemical name of the compound with the given hetID i.e. a unique chemical name.
HETSYN	provides synonyms, if any, for the compound in the corresponding (i.e., same hetID) HETNAM record. This is to allow greater flexibility in searching for HET groups.
FORMUL	presents the chemical formula and charge of a non-standard group. Example for glucose: FORMUL 3 GLC C6 H12 O6
SECONDARY STRUCT.: helices, sheets, and turns in protein and polypeptides.	
HELIX	identifies the position of helices in the molecule. Helices are both named and numbered. The residues where the helix begins and ends are noted, as well as the total length.
SHEET	identifies the position of sheets in the molecule. Sheets are both named and numbered. The residues where the sheet begins and ends are noted.
TURN	identifies turns and other short loop turns which normally connect other secondary structure segments.
CONNECTIVITY: lists location of disulfide bonds and other linkages if present.	
SSBOND	identifies each disulfide bond in protein and polypeptide structures by identifying the two residues involved in the bond.
LINK	specifies connectivity between residues that is not implied by the primary structure. This record supplements information given in CONECT records: provided for convenience in searching.
HYDBND	specifies hydrogen bonds in the entry.
SLTBRG	specifies salt bridges in the entry.

CISPEP	Specifies the prolines and other peptides found to be in the cis conformation.
MISCELLANEOUS FEATURES: describes features such as the active sites.	
SITE	supplies the identification of groups comprising important sites in the macromolecule. Each listed SITE needs a corresponding REMARK 800 that details its significance.
CRYSTALLOGRAPHIC & COORDINATE TRANSFORMATION: geometry & coord system	
CRYST1	presents the unit cell parameters, space group, and Z value. If the structure was not determined by crystallographic means, CRYST1 simply defines a unit cube. The Z value is the number of polymeric chains in a unit cell. In the case of heteropolymers, Z is the number of occurrences of the most populous chain
ORIGXn	(n = 1, 2, or 3). Presents the transformation from the orthogonal coordinates contained in the entry to the submitted coordinates. ORIGX relates the coordinates in the ATOM and HETATM records to the coordinates in the submitted file.
SCALEn	(n = 1, 2, or 3). Presents the transformation from the orthogonal coordinates as contained in the entry to fractional crystallographic coordinates. The unit cell parameters are used to calculate SCALE.
MTRIXn	(n = 1, 2, or 3). Presents transformations expressing non-crystallographic symmetry.
TVECT	presents the translation vector for infinite covalently connected structures.
COORDINATES SECTION: collection of atomic coordinates	
MODEL	specifies the model serial number when multiple structures are presented in a single coordinate entry, as is often the case with structures determined by NMR or molecular dynamics. Every MODEL record has an associated ENDMDL record.
ATOM	presents the atomic coordinates for <u>standard residues</u> . They also present the occupancy and temperature factor for each atom. Heterogen coordinates use the HETATM record type.
SIGATM	presents the standard deviation of atomic parameters as they appear in ATOM and HETATM records. Each SIGATM record immediately follows the corresponding ATOM/HETATM record.
ANISOU	presents the anisotropic temperature factors.
SIGUIJ	presents the standard deviations of anisotropic temperature factors.
TER	indicates the end of a list of ATOM/HETATM records for a chain. The TER records occur in the coordinate section of the entry, and indicate the last residue presented for each polypeptide and/or nucleic acid chain for which there are coordinates. For proteins, the residue defined on the TER record is the carboxy-terminal residue; for nucleic acids it is the 3'-terminal residue. For a cyclic molecule, the choice of termini is arbitrary.
HETATM	presents the atomic coordinate records for atoms within "non-standard" groups. These records are used for water molecules and atoms presented in HET groups. 2 examples: magnesium and ion: <pre>HETATM1357 MG MG 168 4.669 34.118 19.123 1.00 3.16 MG2+ HETATM3835 FE HEM 1 17.140 3.115 15.066 1.00 14.14 FE3+</pre>
ENDMDL	paired with MODEL records to group individual structures found in a coordinate entry.
CONNECTIVITY SECTION: information on chemical connectivity.	
CONECT	specifies connectivity between atoms for which coordinates are supplied. The connectivity is described using the atom serial number as found in the

	entry.
BOOKKEEPING SECTION: some final information about the file itself.	
MASTER	is a control record for bookkeeping. It lists the number of lines in the coordinate entry or file for selected record types.
END	marks the end of the PDB file.

NOTE: The most important records for the casual user are to first verify that the correct PDB file has been retrieved, with the correct data within it: **HEADER**, **TITLE**, **COMPND** and **SOURCE**.

The recognition of the presence of non-protein and non nucleic acids moieties: **HET**

The **ATOM** and **HETATM** records, make the actual 3D coordinates set and all or part can be cut/paste in a new text document if needed.

MODEL and **ENDMDL** are not recognized by some software, such as older versions of Rasmol or VMD. The presence of these records may prevent chains beyond MODEL 1 to be shown. Manual editing, or switching visualization package is required in this case.

TER records append the end of each chain. However some (mostly older) PDB files may not have these records in place in which case it may be useful to add them (the Chimera software for example creates a line (bond) between molecules A and B if there is no TER separating their ATOM records.) On the other hand it may be useful to remove TER records in other cases.

NOTE: **ATOM** is reserved for the protein and nucleic-acid atoms.

HETATM is usually used for all other compounds (listed under HET) such as ligands (e.g. NAD, AMP, ATP), solvents (e.g. P04, S04), water molecules (HOH or WAT) and metal ions (e.g. MG, FE, CA, ZN). However some author do not always follow these conventions and it can be useful to inspect the PDB file with a word processor to know how the file is organized.

Example of ATOM records for **1BL8** (Potassium Channel (Kcsa) From Streptomyces Lividans)

#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	
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
ATOM	4	O	ALA	A	23	65.327	24.916	47.902	1.00181.62		O
ATOM	5	CB	ALA	A	23	67.503	21.981	47.702	1.00	74.09	C
ATOM	6	N	LEU	A	24	66.837	24.176	46.401	1.00163.39		N
/////											
ATOM	704	OE1	GLN	A	119	79.595	14.626	51.132	1.00193.75		O
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 710 O ALA B 23 82.400 9.868 41.107 1.00173.57 O
ATOM 711 CB ALA B 23 85.344 11.988 40.624 1.00 82.33 C
ATOM 712 N LEU B 24 82.767 11.834 40.067 1.00159.27 N
////
////
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 2827 K K 403 71.815 26.478 18.867 1.00 75.52 K
HETATM 2828 O HOH 500 69.120 26.480 12.189 1.00 66.21 O
MASTER 0 0 0 12 0 0 1 15 2824 4 0 32
END

```

The above example spans the complete ATOM and HETATM record coordinates. The text is organized in position-specific space-delimited columns (most common labeled here #1-11). The first line can be read as follows: “ATOM record serial number 1 is a Nitrogen (the N-terminus) of an amino acid ALanine. That alanine is part of chain A and is the 23rd residue on the sequence (X-ray data is not visible for residues 1-22). The next 3 columns are the XYZ Cartesian coordinates, followed by the occupancy (1.00) and the temperature factor columns.”

The file contains 4 proteins identified with chain identifiers A, B, C, and D separated by TER records. The non-protein atoms are 3 potassium ions (K) and one water (HOH).

NOTE: Most common space-delimited columns

- #1 record name
- #2 Atom serial number
- #3 Atom name
- #4 Residue name
- #5 Chain identifier
- #6 Residue sequence number
- #7-#9 Orthogonal coordinates in Angstroms for X, Y and Z respectively,
- #10 Occupancy
- #9 Temperature factor

3.4 Biological vs crystallographic sets

PDB files contain data derived from the crystal of molecules, and referred to as the “*asymmetric unit*.” As such the data may reflect the packing arrangement or the crystallographic interaction of the molecules within the crystal. Crystallographic symmetry transformations are provided under **REMARK_290** with the **SMTRYn** records as 3x4 rotation matrices to calculate the position of other molecules within the same crystal. This can be useful for crystallographers and to study the packaging within the crystal but biologists at the bench would rather prefer the coordinates for their biologically active compounds which could be a dimer, trimer or an even higher order multimer. The PDB web site now offers the option to “download the biological unit file” rather than the published PDB crystallographic dataset. However, this may not be not true for every entry, as it relies upon symmetry information that might not have been provided by the authors within the file.

A new record **BIOMTn** (biological matrix) is now commonly found within PDB files under **REMARK_350**. It consists of one or more 3x4 rotation matrices that can help calculate the biological unit. Fortunately, when this record is present, the downloaded biological unit file contains the proper biological entity of interest. Older files may still have symmetry information contained within under an old REMARK format, and may or may not be available as a biological unit.

BIOMT records of PDB file 8RUC	REMARK 350 MOLECULE CAN BE GENERATED BY APPLYING BIOMT TRANSFORMATIONS						
	REMARK 350	BIOMT1	1	-1.000000	0.000000	0.000000	0.000000
	REMARK 350	BIOMT2	1	0.000000	1.000000	0.000000	0.000000
	REMARK 350	BIOMT3	1	0.000000	0.000000	-1.000000	100.650000

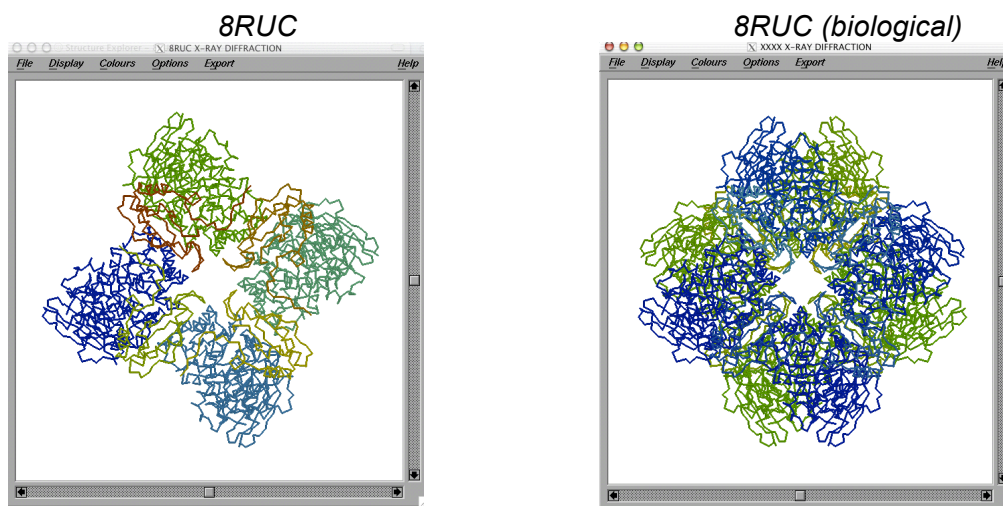
Examples:

- Properly referenced BIOMT records allow downloading of biological unit: 8RUC
- Properly referenced symmetry in older format 1985 file provides proper biological unit: 8CAT
- Non standard symmetry information within older file results in no biological unit downloadable: 1FPV
- Multimeric crystal data for monomeric biological unit: 1MME
- Multimeric crystal data for monomeric biological unit: 1P35 - (1 wrong output file has dimer instead of monomer).

Downloading biological subunits from the PDB site:

For the newer version of the PDB web site, (<http://www.rcsb.org/pdb/>) click on the “**Download Files**” menu at left, which will expand to offer download options in PDB and other formats. The biological unit is the last option on the list.

The image shows two screenshots of the PDB website interface. The left screenshot shows the main navigation menu with 'Download Files' highlighted. The right screenshot shows the expanded 'Download Files' menu, where 'Biological Unit Coordinates' is highlighted at the bottom of the list. An arrow points from the 'Download Files' menu in the left screenshot to the expanded menu in the right screenshot.



“What if I need a biological unit and it is not provided within PDB? ”

For most entries: the Protein Quaternary Structure (PQS) Web server offers multimeric coordinates generated at the (European Bioinformatics Institute) EBI : <http://pqs.ebi.ac.uk/>

Special case of viruses: Viruses are perhaps the largest crystallized structures with very large biological units, their capsid. For many viruses the PDB site provides a correct biological unit (containing ENDMDL records), but not for all. Users interested in icosahedral symmetry of viruses can generate complete or partial capsides with the “oligomer generator” on the Virus particle Explorer (VIPER) site: <http://viperdbscripps.edu/>

4 (Free) Molecular graphics software

Here is a short list of free desktop software. Only 2 or 3 of these will be studied in class.

PyMOL (user sponsored)	http://pymol.sourceforge.net
VMD (MacOSX, Windows, Linux)	http://www.ks.uiuc.edu/Research/vmd/
RasMol 2.6 (Mac, Windows, Unix). Powerful line command	http://www.umass.edu/microbio/rasmol/
Rasmol 2.7 (almost any OS)	http://www.bernstein-plus-sons.com/software/rasmol/
Protein-Explorer (depends on plug-in- Use either Chime or JMol).	http://www.proteinexplorer.org
IMol (Mac OSX only)	http://www.pirx.com/iMol/
Deep View (Swiss PDB Viewer) (Mac, Windows, Linux, SGI)	http://www.expasy.ch/spdbv/mainpage.html
MOLVIEW (Mac only)- makes QuickTime movies, beautiful renderings	http://www.danforthcenter.org/smith/molview.htm
Cn3D (Mac, Windows, Unix).	

Additional sequence alignment window.	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
QuteMol (Mac, Windows) (requires specific graphics hardware)	http://qutemol.sourceforge.net/
Chimera (UCSF)- For veteran molecular graphists!	http://www.cgl.ucsf.edu/chimera/

5 Molecular graphics basics

Computer graphics programs (either on a desktop computer or on an expensive workstation) read, interpret and display the 3D coordinates (PDB file or other formats) into graphical images that can be manipulated in real time in three dimensions and modified interactively on the computer screen.

5.1 Opening a file

Most software will provide a graphical menu interface and opening a 3D coordinates file will be part of the standard menu e.g. File > Open... Some software will infer the file format from the filename extension “.pdb” and open the file accordingly. Others may have a button switch to specify the file format. Some have a combination of both. Some software can open the file directly from a line command. In PyMol, files ending in “.pdb” are recognized by the File > Open... menu, but not files ending with .pdb1, .pdb2 etc. However, these can be loaded with the line-command option. Alternatively, the file name can be manually changed to .pdb.

5.2 Opening view, and display changes

Almost every program will display the file in a standard, graphical format on the graphical window (canvas) as soon as it is read in. Most software will display a thin-lined wireframe structure with standard CPK colors, except VMD, which displays carbons in a cyan color. Most software will have a black background by default.

Did you know?

CPK is an acronym that stands for Corey-Pauling-Koltun. CPK is used to designate both the space-filling physical models and the color scheme of these models. The original models of Corey and Pauling were made of hard wood with steel rod connectors and clamps (1 inch per angstrom), or made of plastic (0.5 inch per angstrom). In 1960 the model designs were improved by Koltun and were lighter and the conformation angles more accurate. Yankeelov and Coggins (1971) built a CPK model of myoglobin at the rate of 50 amino acid residues per week.

More historical information on physical models is available at <http://www.netsci.org/Science/Compchem/feature14b.html>

A 1951 photograph of Linus Pauling and Robert Corey with a model of a molecule can be seen at <http://osulibrary.orst.edu/specialcollections/coll/pauling/dna/pictures/paulingcorey.html> (in fact it is Photo ID 1.45-30 of The Caltech institute Archives <http://www.archives.caltech.edu/>)

The 6 essential CPK colors are:

A complete color scheme can be found online at http://www.umass.edu/microbio/rasmol/distrib/rasman.htm#cpkcolours	Element:	Color:
	Carbon	Light grey
	Oxygen	Red
	Nitrogen	Light blue
	Hydrogen	White
	Sulfur	Sulfur yellow
	Phosphorus	Orange

The complete CPK: from <http://www.umass.edu/microbio/rasmol/distrib/rasman.htm#cpkcolours> (RGB Hexadecimal are used for web design.)

ELEMENT HEXADECIMAL	COLOR NAME	RGB DECIMAL	RGB
Carbon	light grey	[200,200,200]	C8 C8 C8
Oxygen	red	[240,0,0]	F0 00 00
Hydrogen	white	[255,255,255]	FF FF FF
Nitrogen	light blue	[143,143,255]	8F 8F FF
Sulphur	sulphur yellow	[255,200,50]	FF C8 32
Chlorine, Boron	green	[0,255,0]	00 FF 00
Phosphorus, Iron, Barium	orange	[255,165,0]	FF A5 00
Sodium	blue	[0,0,255]	00 00 FF
Magnesium	forest green	[34,139,34]	22 8B 22
Zn, Cu, Ni, Br	brown	[165,42,42]	A5 2A 2A
Ca, Mn, Al, Ti, Cr, Ag	dark grey	[128,128,144]	80 80 90
F, Si, Au	goldenrod	[218,165,32]	DA A5 20
Iodine	purple	[160,32,240]	A0 20 F0
Lithium	firebrick	[178,34,34]	B2 22 22
Helium	pink	[255,192,203]	FF C0 CB
Unknown	deep pink	[255,20,147]	FF 14 93

A large portion of the user's time will be spent in creating a macroscopic view of the molecule(s) present within the 3D coordinates (e.g. changing every protein chain to different color), or create a close-up view for particular residues or set of atoms to illustrate a specific interaction (e.g. a ligand interaction). In both cases, knowing how the information is organized and arranged within the PDB data file will save invaluable time and frustration. In particular it is very useful to know which molecules have ATOM records and which are HETATM. Finding this information is as simple as opening the PDB file with a simple text processor.

There are many types of representations that can be achieved for proteins, ligands nucleic acids and other macromolecular compounds. These will be reviewed below.

5.3 Saving or creating an image file

Once a satisfying representation is achieved within the molecular graphics software 3D canvas by applying typed or menu-driven command the representation can be saved as an image within a "flat" graphical computer file from most software in popular graphics formats such as GIF, JPEG or EPS. At worst a screen-dump of the computer full or partial screen can save the day.

Practical * Sreen dump on a PC: use the "Print Screen" keyboard key and paste into
Hint another software e.g. PowerPoint.

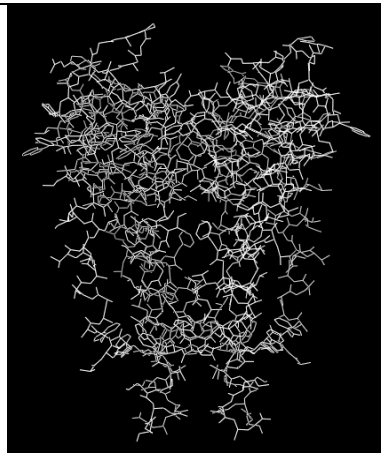
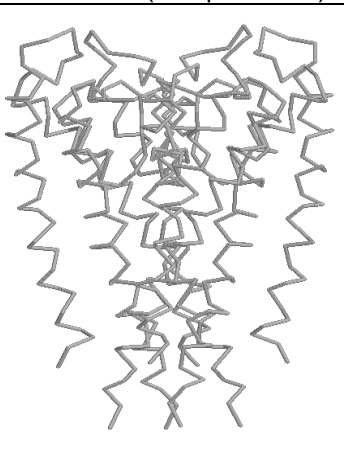
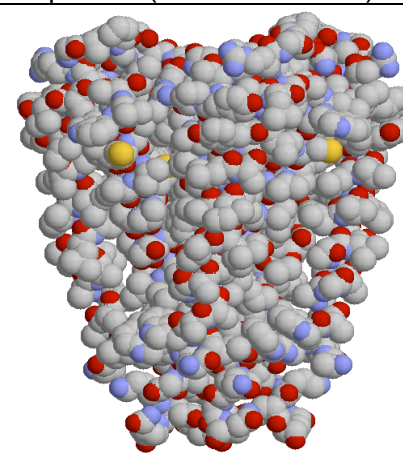
* Screen dump on a Macintosh: Apple+Shift+3 for full screen, or
Apple+Shift+4 for custom size.

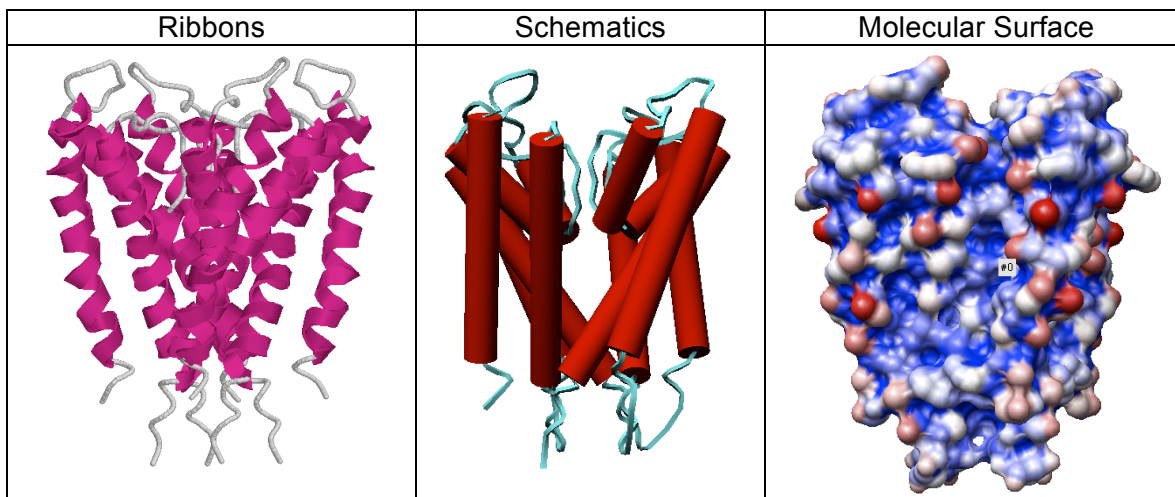
5.4 Scripting and animation

Some software can animate molecules with the help of a script that can be played back. The script can create the animation directly on the computer screen from within the molecular software (playback). The animation can be alternatively created from a series of saved image files, created manually but most likely from a script with specific commands to save graphical files. A third party software (either free or commercial) is then necessary to create the animation by assembling and compressing the saved frames.

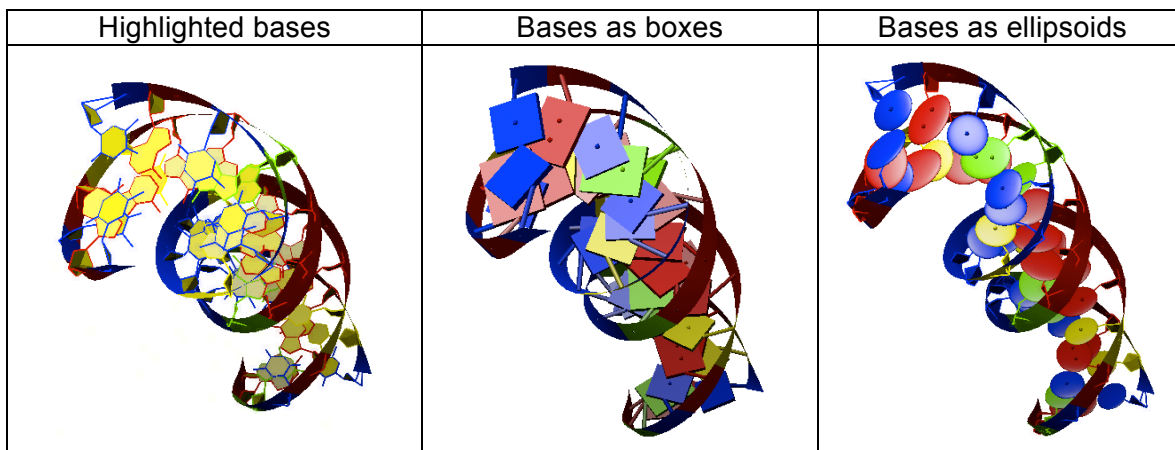
5.5 Molecular graphics representations

Most free and commercial software have many schematic representations for depicting macromolecular 3D coordinates of proteins and other biological molecules, ranging from wireframe representations to molecular surfaces. Here are the most common types for protein representations (illustrated with 1BL8; Potassium Channel (Kcsa) From *Streptomyces Lividans*). The background has been made white for most images for printing purposes.

wireframe	Backbone (C-Alpha trace)	Spacefill (also called CPK)
		



Many of the wireframe, space-fill and molecular surface versions of the protein can also be applied to nucleic acids. In addition, some software has schematic versions for representing bases and sugars as boxes or filled geometric forms. Examples created with Chimera software from the DNA portion of 2OR1 (Recognition of a DNA operator by the repressor of phage 434: a view at high resolution.)



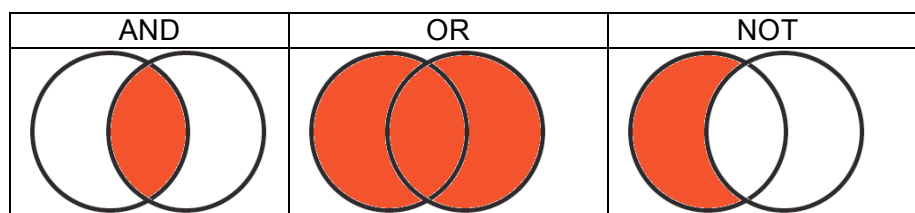
5.6 Molecular graphics software logic

Molecular graphics softwares are designed with the specific purpose of displaying biological macromolecules, primarily proteins, nucleic acids and small ligands. Therefore they all provide some level of easy selection (typed text, menu driven or both) to help select and change the appearance of specific sections of one or more molecules.

For example, most applications provide the ability to select a specific molecule chain, for example chain A. Others have more elaborate lists of predefined sets. For example within Rasmol one can select the following predefined sets on the line command: protein, hydrophobic, charged, backbone, sidechain, nucleic, purine, pyrimidine, water, etc.

Most programs will work on all the structures within the PDB file by default, and will only restrict its actions to the selected atoms if and when specific atom selections are made.

Boolean logic often applies to making a selection. Boolean operators can be illustrated by the following Venn diagrams:



For example within Rasmol the typed command to select both chains molecule A and B at the same time is: `select :A or :B`. Note that our grammatical “and” signifying that we would like both chains at the same time has to be translated to the correct Boolean “OR” operator. Since by definitions atoms can only belong to a single protein chain and have no atoms in common, “A or B” in this case would be more correctly diagrammed as follows:



6 Protein structure summary

Summary: Proteins are a linear assembly of amino-acids (a covalent polymer), the order of the amino acids is called the *sequence* and represents the *primary structure* of the protein. The chain of amino-acids folds into local *secondary structure elements* in the form of *alpha-helices* and *beta-sheets*. The overall assembly of these elements make the final protein conformation of *tertiary structure*. The assembly of multiple proteins is referred to as a *quaternary structure*.

6.1 The 20 amino-acids: Name, 3-letter and 1-letter codes, chemical structure

Name	Glycine	Alanine	Valine	Leucine	Isoleucine	Proline	Serine
Code	Gly - G	Ala - A	Val - V	Leu L	Ileu I	Pro - P	Ser - S
Chem							
Threonine	Cysteine	Methionine	Asparagine	Glutamine	Phenylalanine	Tyrosine	Tryptophan
Thr - T	Cys - C	Met - M	Asn - N	Gln - Q	Phe - F	Tyr - Y	Trp - W
Lysine	Arginine	Histidine	Aspartate	Glutamate			
Lys - K	Arg - R	His - H	Asp - D	Glu - E			

<http://www.kumc.edu/biochemistry/bioc800/aaindex.htm>

Amino acids are linked by a peptide bond resulting from a chemical reaction between the acidic end of one amino acid and the basic (amino) end of the next one in the sequence. The reaction releases a water molecule in the process.

PROTEINS

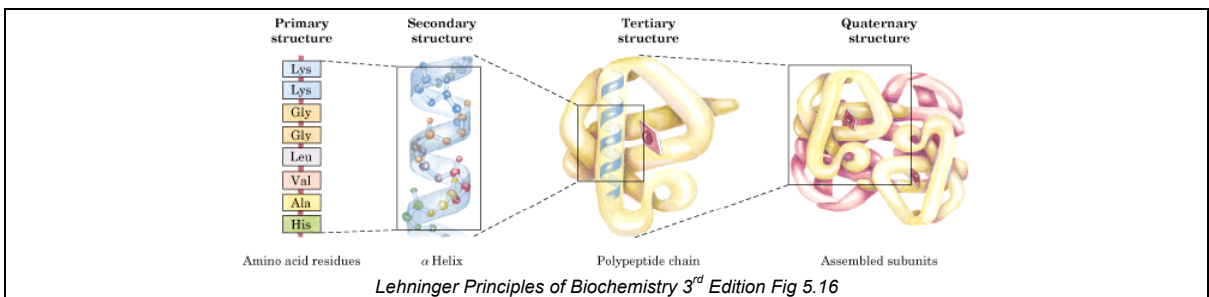
monomer amino acid

peptide bond

<http://employees.csbsju.edu/hjakubowski/classes/ch331/protstructure/olproteinmenu.html>
 Reaction between 2 amino-acids releases water and results in a dipeptide with peptide bond in the middle

The assembly of a few amino acids is an oligopeptide, when there are more the term is polypeptide. Proteins are simply large polypeptides.

1) <u>Primary</u> structure: the sequence of amino acids	2) <u>Secondary</u> structures: local alpha helices, beta-sheets and turns	3) <u>Tertiary</u> structures: the overall 3D protein structure	4) <u>Quaternary</u> structure: assembly of multiple proteins
--	--	---	---



7 Nucleic Acid Structure Summary

(Based on:

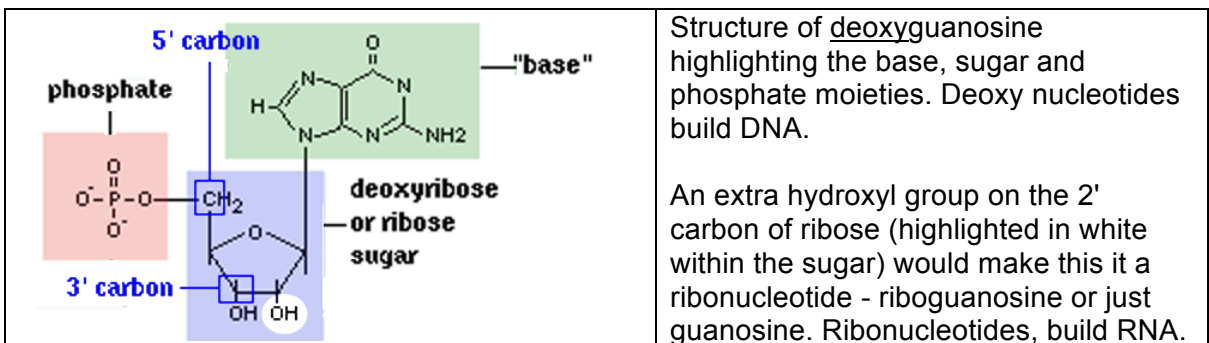
<http://www.vivo.colostate.edu/hbooks/genetics/biotech/basics/nastruct.html> & http://www.imb-jena.de/ImgLibDoc/nana/IMAGE_NANA.html)

DNA (deoxyribonucleic acid) and RNA (ribonucleic acid) 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 have a distinctive structure composed of 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 combination of a base and sugar without the phosphate is called a *nucleoside*.)



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

C^{5'} O_{3'} C^{2'}-endo B

C^{5'} O_{3'} C^{3'}-endo B

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.)

Abbr	Base	Nucleoside	Nucleic Acid
A	Adenine	deoxyadenosine	DNA
		adenosine	RNA
G	Guanine	deoxyguanosine	DNA
		guanosine	RNA
C	Cytosine	deoxycytidine	DNA
		cytidine	RNA
T	Thymine	deoxythymidine (thymidine)	DNA
U	Uracil	uridine	RNA

There are five common bases described in the table on the left. Four of the five are used to build either DNA (ACGT) or RNA (ACGU).

In double-stranded nucleic acids, base pairs are always formed between a *purine* (A or G) and a *pyrimidine* (C, T or U).

Guanine Cytosine

Adenine

One way to remember:
 purine = small word but big structure (2 rings)
 pyrimidine = big word but small structure

adenine

guanine

cytosine

thymine

uracil

Structures of the purine (R) and pyrimidine (Y) bases of nucleic acids in their dominant tautomeric forms and with the IUPAC (*International Union of Pure and Applied Chemistry*) numbering system.

All nucleic acids have two distinctive ends: 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. Polymerases add nucleotides to the 3' end of the previously incorporated base and synthesis is in a 5' to 3' direction.

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 interaction between bases are important for the local folding.

In double stranded nucleic acids, the two strands antiparallel to one another: 3'-end of one strand pairs with 5'-end of the other strand.

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.

8 Conclusion: user's adaptability

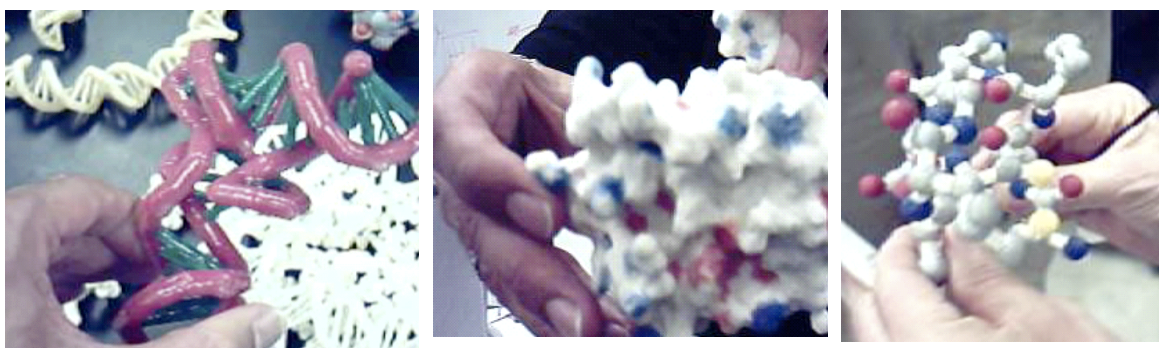
Every software, free or commercial, works in a specific way, with line-command or pull-down menus, but all have basic similar functions such as creating a ribbon diagram. However when one understands the format of the 3D data (such as the organization of the PDB files –how many protein chains, the name of the ligands, metal ions etc.-) it becomes much easier to control the software to attain the desired goals of creating still images and movies and adapt to the current software knowing what is possible based on the existing data.

8.1 The next step: from the fog of the computer memory and into the real world.

(Note: this section is not part of the normal in-class curriculum and is only informational)

The 3D coordinates are the result of carefully planned experiments and calculations under the careful watcheye of the crystallographers or the NMR scientists. The 3D data that is contained within the PDB files can be manipulated in real time in a virtual 3D world and the molecule(s) can be colored and rearranged to highlight important residues, inter- or intra-molecular interactions, or the overall folding of a protein. Some pretty images, sometimes with an artistic feel can even arise from such exploration.

However the newest way to make molecules “come alive” is to use the “rapid prototyping” modeling which can also be described as “3D printing.” The method consists in printing the 3D information of the virtual representation of the molecule(s) in a large numbers of layers, with a powdery substance such as starch or plaster, later held together by a liquid glue.



Holding a “real” tRNA molecule

....a protein surface.

...or an alpha helix.

The DMC (3rd floor Biotechnology Center) has a “3D printer” that prints with plaster layers. Some Biochemistry students may have the opportunity to create models during their career.

In Madison models are created on a Zprinter 310 (www.zcorp.com). The information for creating the 3D models is derived from PDB files and held in a special kind of computer text file as a special output from Rasmol (a proprietary version called RPRasmol -RP= rapid prototyping). PyMol is an alternative software to use to create model files.

In Wisconsin, the larger center is the “Center for BioMolecular Modeling” in Milwaukee (www.rpc.msoe.edu/cbm).

So, we have come full circle: physical models made of wood or metal used to help create the final computer data. Nowadays, the final computer data can be made into a physical model again.



IMPORTANT UPDATE ON NUCLEIC PDB FILES CONTAINING NUCLEIC ACIDS

PDB files that contain nucleic acid have a change in their ATOM nomenclature that can affect how molecular graphics software interpret the file, in particular the Rasmol program.

On the PDB site (<http://www.rcsb.org/>) files can be downloaded from 2 main expanding menus from the left side menu items: “**Download Files,**” and “**Download Original Files.**”

Files retrieved from the “Download Files” will have the new nucleic acid format. Files retrieved from the “Download original Files” will contain the old format. Note that the file names will also differ. The old format will be `pdbxxxx.ent` where `xxxx` is the PDB accession in lower case, while the new files will be named `xxxx.pdb` where `XXXX` is the PDB accession in upper case.

This is best understood with a simple example, here **3CRO**:

Old format: file `pdb3cro.ent`

TER	410		T A	20						3CRO 553
ATOM	415	O5*	T B	1	-55.231	-19.025	20.750	1.00	49.50	3CRO 554
ATOM	416	C5*	T B	1	-54.322	-19.784	19.967	1.00	56.94	3CRO 555
ATOM	417	C4*	T B	1	-53.045	-19.016	19.679	1.00	41.16	3CRO 556
ATOM	418	O4*	T B	1	-53.322	-17.632	19.347	1.00	48.36	3CRO 557

New Format: file `3CRO.pdb`

TER	406		DT A	20						
ATOM	407	O5'	DT B	1	-55.231	-19.025	20.750	1.00	49.50	O
ATOM	408	C5'	DT B	1	-54.322	-19.784	19.967	1.00	56.94	C
ATOM	409	C4'	DT B	1	-53.045	-19.016	19.679	1.00	41.16	C
ATOM	410	O4'	DT B	1	-53.322	-17.632	19.347	1.00	48.36	O

In the old format T B meant an Thymine which belongs to chainB. In the new format, T is replace by DT and one can assume that D stands for DNA. It can also be noted that the sugar atoms are marked with * in the old format and ' in the new format.

Consequences for Rasmol:

Rasmol parses the PDB file as it reads it, and remember which category each atom belongs to. In the case of nucleic acid, a T would fit in the predefined sets: *nucleic*, *pyrimidine*, *AT* and *T* and could be called upon by the command `select` such as `select nucleic`.

With the new format, Rasmol does not understand that DT stands for T (and by extension it is the same for DA, DC, and DG, and DU for RNA.) Therefore, the following commands will not work with the new formatted PDB files:

```
select nucleic
select A
select T
select AT
select C
select G
select GC
select purine
select pyrimidine
```

As a workaround, you can create your own preset with these files at the beginning of the session:

```
select DA or DC or DG or DT
define mynucleic selected
```

You can now use the defined set “mynucleic” in the same way as you would have used “nucleic.” Since “nucleic” is already predefined in Rasmol it cannot be overwritten.

-The End-

